Centre for Governance and Sustainability
NUS Business School

Proceedings of
Tencent – CGS Academic Conference

Technology for Good: Driving Social Impact

Volume I
Research Papers and Commentaries

December 2025

Supported by
CHINA ASSOCIATION FOR NGO COOPERATION
中国国际民间组织合作促进会

Proceedings of
Tencent – CGS Academic Conference

Technology for Good: Driving Social Impact

Volume I

Research Papers and Commentaries

Publishers
Tencent & Centre for Governance and Sustainability

Editors
Lawrence LOH
Bima Satria

Date of Conference
19 November 2025

Date of Publication
December 2025

About the Conference

The Tencent and CGS Academic Conference was held on 19 November 2025 at the NUS Business School under the theme *Technology for Good, Driving Social Impact*. The conference brought together about 100 academics, industry leaders, and policymakers to discuss how digital tools can support fairness, strengthen public institutions, and improve daily life across Asia and beyond. The call for papers drew strong interest, with 134 submissions from five continents and 60 proposals carried forward as full papers, commentaries, or videos.

The keynote speakers opened the conference with wide-ranging reflections on AI and digital transformation. Professor Andrew Rose, Distinguished Professor and Dean of the NUS Business School at the National University of Singapore, delivered the welcome address and highlighted the growing influence of AI on society, emphasising the need for ethical and inclusive approaches. Mr Xiao Liming, Vice President for Sustainable Social Value at Tencent, reaffirmed the company's commitment to placing social value at the centre of its work and noted the continued cooperation between China and Singapore. Ms. Xu Xiaoxiao, Deputy Secretary-General of the China Association for NGO Cooperation (CANGO), highlighted how cross-sector partnerships under the "Tech for Good" framework empower social organisations, connect to global agendas, and leverage digital innovation to advance inclusive development. Dr Ming Tan, Senior Fellow and Founding Executive Director of the Tech for Good Institute and Senior Fellow at the Centre for Governance and Sustainability, NUS Business School, called for people-centred innovation and demonstrated how thoughtful digital design can widen access to opportunity.

The presentation session featured 20 selected works reflecting a wide range of research interests. Several papers focused on responsible and inclusive AI, addressing issues such as psychological safety in chatbots, governance frameworks, cross-cultural alignment, information pollution, and safer agentic systems. Other studies examined the societal effects of digital platforms, including philanthropic models, community technologies, migrant entrepreneurship, and public-sector cooperation.

The programme continued with a panel discussion on AI for global equity, adding further perspective. The speakers considered whether rapid AI adoption in Southeast Asia will widen opportunity or widen divides, and discussed investment trends, local digital readiness, and the need for systems that are both safe and practical. The session underlined the importance of sound governance and public education as AI becomes more embedded in everyday life. The conference concluded with awards recognising outstanding research papers, commentaries, and posters.

This conference proceedings present the research and posters presented during the conference. The publication is organised into two volumes. Volume I contains research papers and commentaries presented at the conference. Volume II contains posters submitted by the invited participants.

Event Agenda

9.00 am **Registration**

9.30 am **Welcome Address**

Prof Andrew ROSE
Distinguished Professor and Dean
NUS Business School, National University of Singapore

9.35 am **Keynote Address**

Mr XIAO Liming
Vice President, Sustainable Social Value
Tencent

9.45 am **Keynote Address**

Ms XU Xiaoxiao
Deputy Secretary-General
China Association for NGO Cooperation (CANGO)

9.55 am **Guest Keynote**

Dr Ming TAN
Senior Fellow and Founding Executive Director
Tech for Good Institute
Senior Fellow, Centre for Governance and Sustainability
NUS Business School, National University of Singapore

10.05 am **MoU Signing Ceremony**

10.15 am **Research Presentations**
Presentations of 20 selected papers and commentaries, including Q&A sessions with key expert judges

3.30 pm **Panel Discussion**

AI for Global Equity: Bridging the Digital Divide and Unlocking Potential

Moderator
Ms Miro LU
Founder & Managing Director at Perspective Media
Editor-in-Chief at Asia Tech Lens

Panellists
Mr Benjamin GOH
Senior Assistant Director, National AI Group
Ministry of Digital Development and Information, Singapore
Dr Jingyang HUANG
Assistant Professor
School of Public Policy
The Chinese University of Hong Kong (Shenzhen)

Mr Kenneth SIOW
Regional Director, Southeast Asia and General Manager
Tencent Cloud International
Mr Tim ZHANG
Founder & CEO, Edge Research Pte Ltd
Senior Fellow, Centre for Governance and Sustainability
NUS Business School, National University of Singapore

4.15 pm **Award Ceremony**

4.25 pm **Closing Remarks**

Prof Lawrence LOH
Director, Centre for Governance and Sustainability
NUS Business School, National University of Singapore

List of Research Presenters

(Sorted according to research category and title)

Research paper category

1. AI for Global Health Equity: A Legal Preparedness Index, China, EU, USA, Australia Case Studies
María Tresierra LAGUNAS[1], HAN Jin[1], Grégoire LAMBRECHT[2]
[1] Center for Global Health Equity, New York University, Shanghai
[2] Computer Science, Data Science, Engineering Division, NYU Shanghai

2. An Assemblage-Centred Lens on Cloud for Good: Tencent Cloud's Diffusion and Localisation in Emerging Markets
HUANG Jingyang[1], XING Linzhou[2], KANG Siqin[3]
[1] School of Public Policy, The Chinese University of Hong Kong (Shenzhen), China
[2] Fairbank Center for Chinese Studies, Harvard University, USA
[3] School of Humanities and Social Sciences, The Chinese University of Hong Kong (Shenzhen), China

3. Beyond Addiction: Reconceptualising and Measuring Mobile Digital Overuse Among Generation Z
GONG Hongcun[1], ZHANG Yiqin[2], SHEN Yutian[3], DENG Sanhong[2]
[1] School of Economics and Management, Shanxi University, China
[2] School of Information Management, Nanjing University, China
[3] School of Public Administration, Zhejiang University, China

4. Cross-Cultural Value Alignment Frameworks for Responsible AI Governance: Evidence from China-West Comparative Analysis
LIU Haijiang[1], GU Jinguang[1], WU Xun[2], Daniel HERSHCOVICH[3], XIAO Qiaoling[4]
[1] Wuhan University of Science and Technology, China
[2] The Hong Kong University of Science and Technology (Guangzhou), China
[3] University of Copenhagen, Denmark
[4] WUST-Madrid Complutense Institute, China

5. Enhancing AI Chatbots with Psychological Safety: A Middleware Framework for Mental Health-Sensitive and Risk-Aware Conversations
ZHOU Chuanbei[1], HUANG Yi[2], Zainal Nur Hani[3]
[1] Information Technology, National University of Singapore, Singapore
[2] Centre for Research and Development, Pan-European University, Slovakia
[3] National University of Singapore, Singapore

6. Generative Governance Paradigm: Cultivating Responsible UGC through Explainable AI
NAN Haonan[1], WANG Zixiao[2], SONG Ruizhen[1], CHEN Hongquan[3], QIU Zhipeng[1]
[1]Antai College of Economics and Management, Shanghai Jiao Tong University, China
[2]Department of Human Centered Design & Engineering, University of Washington, USA
[3]Sino-US Global Logistics Institute, Shanghai Jiao Tong University, China

7. Inclusive AI-Driven Collaborative Innovation with Consumer and New Product Success: Evidence from WeGame
LIN Fuxin[1], QIAO Qianyue[2], WANG Dongqi[1], CHEN Zhi[1], FU Xiaolan[2], WU Xiaobo1
[1] School of Management, Zhejiang University, China
[2] Technology and Management Centre for Development, Department of International Development, University of Oxford, UK

8. Manipulated Emotions: How Algorithms Shape Negativity in Public Opinion on Chinese Social Media
WANG Qing[1], LI Jinhui[1]
[1] School of Journalism and Communication, Jinan University

9. Market Design Interventions for Safer Agentic AI
Swapneel MEHTA[1], Nina MAŽAR[1], Aaron NICHOLS[1], Marshall VAN ALSTYNE[1]
[1] Boston University, USA

10. Open Source as a Channel for AI Learning and Adaptation? An Analysis of Open Source Approaches in Tencent Hunyuan AI
Christopher FOSTER
[1] Global Development Institute, University of Manchester, UK

11. Struggling in the Middle: Understanding Adolescents' Self-Regulation Amid Family, School, and Peers in Singapore
ZHU Aaron Pengyu[1], CUI Shenglan, CHO Janghee[1]
[1] Division of Industrial Design, National University of Singapore, Singapore
[2] School of Design, Hunan University, China

12. Technology for Good? How Digital Platforms Reshape Collaboration in China's Civic Sector
ZHENG Wenjuan[1], ZHANG Delin[1]
[1] Hong Kong University of Science and Technology, China

13. The Economics of Information Pollution in the Age of AI: General Equilibrium, Welfare, and Policy Design
ZHANG Yukun[1], ZHANG Tianyang[2]
[1] The Chinese University of Hong Kong, Hong Kong, China
[2] University of Macau, Macau, China

14. WeChat without Borders: Grassroots Innovation and the Unintended 'Technology for Good' in a Kenya Slum
HU Lingrui[1], Fred Oluoch JIENDA[2], ZHAO Hanyu[3], SUN Luoyan[1]
[1] School of Government, Sun Yat-Sen University, China
[2] Office of the Spouse of the Deputy President of Kenya, Kenya
[3] Stephen M. Ross School of Business, University of Michigan, USA

Commentary category

Acknowledgements

We gratefully acknowledge Tencent for its funding and essential support, without which the Tencent–CGS Academic Conference on *Technology for Good: Driving Social Impact*, held on 19 November 2025, would not have been possible. Tencent's commitment to advancing responsible and socially beneficial uses of technology provided a strong foundation for the conference and its outcomes. We also extend our sincere thanks to the China Association for NGO Cooperation (CANGO) for its engagement.

Our appreciation goes to the keynote speakers for their time, insight, and thoughtful contributions.

1) Welcoming address by Professor Andrew ROSE, Distinguished Professor and Dean at NUS Business School;
2) Keynote address by Mr XIAO Liming, Vice President of Sustainable Social Value at Tencent;
3) Keynote address by Ms XU Xiaoxiao, Deputy Secretary-General at China Association for NGO Cooperation (CANGO); and
4) Guest Keynote by Dr Ming TAN, Senior Fellow and Founding Executive at Tech for Good Institute, and Senior Fellow at Centre for Governance and Sustainability NUS Business School.

We extend particular thanks to the expert judges for their constructive feedback and critical assessment were central to upholding the research quality of the conference.

1) Dr Derrick KON, CEO of CEO Solutions Pte Ltd and Adjunct Associate Professor at NUS Business School;
2) Mr Alvin LIM, CEO and Co-founder of Climate Bridge International and Senior Fellow at the Centre for Governance and Sustainability, NUS Business School;
3) Professor Lawrence LOH, Director of the Centre for Governance and Sustainability and Professor of Strategy and Policy at NUS Business School;
4) Dr Ming TAN, Senior Fellow and Founding Executive Director of the Tech For Good Institute and Senior Fellow at the Centre for Governance and Sustainability, NUS Business School;
5) Professor WONG Poh Kam, Emeritus Professor at NUS Business School and Senior Fellow at the Centre for Governance and Sustainability, NUS Business School; and
6) Mr Tim ZHANG, Founder and CEO of Edge Research Pte Ltd and Senior Fellow at the Centre for Governance and Sustainability, NUS Business School.

We also appreciate the moderator and panellists of the panel discussion on AI for global equity for their perspectives, which helped frame key debates and enriched the overall discussion.

1) Mr Benjamin GOH, Senior Assistant Director of National AI Group at the Singapore Ministry of Digital Development and Information;

2) Dr Jingyang HUANG, Assistant Professor at School of Public Policy The Chinese University of Hong Kong (Shenzhen);

3) Ms Miro LU (Moderator), Founder & Managing Director at Perspective Media and Editor-in-Chief at Asia Tech Lens;

4) Mr Kenneth SIOW, Regional Director, Southeast Asia and General Manager at Tencent Cloud International; and

5) Mr Tim ZHANG, Founder and CEO of Edge Research Pte Ltd and Senior Fellow at the Centre for Governance and Sustainability, NUS Business School.

We are thankful to all research presenters and participating researchers for their strong interest and high-quality contributions, addressing key issues related to technology and emerging artificial intelligence for social good. Their research, discussion, and exchange of ideas significantly enriched the conference and form the core of these proceedings.

Finally, we thank all conference participants for their active engagement and contributions to the exchanges and discussions. Their presence and participation were central to the success of the conference.

Centre for Governance and Sustainability (CGS)
NUS Business School
National University of Singapore

Project Structure

CENTRE FOR GOVERNANCE AND SUSTAINABILITY, NUS BUSINESS SCHOOL

Project Lead
Lawrence LOH

Project Organisation
Bima Satria, ANG Hui Min, Verity THOI, KONG Zhen Wei, HUANG Minjun, Trang NGUYEN, Annette SINGH, and Phoebe OW

Project Support
CHEN Fan, CUI Xiaoshan, Agatha Celia SANTOSA, SHI Haiqing, Harshali Ravindra SHINDE, WANG Bo, and ZHANG Bingqian

Editors

Lawrence LOH

Bima Satria

Tencent – CGS Academic Conference
Technology for Good: Driving Social Impact
Singapore, November 19th, 2025

Organised by: Tencent & Centre for Governance and Sustainability (CGS)
Supported by: China Association for NGO Cooperation (CANGO)

About Editors

Prof. Lawrence LOH is the Director of the Centre for Governance and Sustainability (CGS) at NUS Business School and serves as Professor in Practice of Strategy and Policy. He holds a PhD in Management from the Massachusetts Institute of Technology. He leads ESG market studies across Asia and the Pacific, ASEAN, and Singapore. He works as a consultant to Fortune 500 companies and international organisations and delivers executive education programmes for senior corporate leaders. He is a recipient of the NUS Annual Teaching Excellence Award and the NUS Business School Teaching Excellence Award. At CGS, he leads corporate governance and sustainability initiatives across Singapore and ASEAN. He teaches governance and sustainable business, regularly advises businesses and policymakers, and contributes commentary to major media outlets.

Bima Satria is a Research Associate at the Centre for Governance and Sustainability (CGS) at NUS Business School. He holds a master’s degree in Public Policy from National University of Singapore. He has experience working as a consultant for international development organisations and national governments, contributing to policy design, programme preparation, and stakeholder coordination across issues on MSMEs development, local economic development, urban transformation, and sustainability agendas in Indonesia and Southeast Asia. At CGS, he is involved in multiple projects on corporate sustainability and governance.

Content

(Sorted according to title)

Research Paper

Beyond Addiction: Reconceptualising and Measuring Mobile Digital Overuse among Generation Z

GONG Hongcun[1]
ZHANG Yiqin[2]
SHEN Yutian[3]
DENG Sanhong[4]

[1] School of Economics and Management, Shanxi University, China
[2,4] School of Information Management, Nanjing University, China
[3] School of Public Administration, Zhejiang University, China

Abstract

Purpose/Significance

Digital overuse has become increasingly prevalent among Generation Z, posing substantial risks to individual well-being and generating new challenges for social governance. Clarifying its behavioural manifestations, constructing a reliable and valid measurement framework, and examining its association with Internet addiction are of both theoretical and practical significance. Building upon the global discourse of Technology for Good, this study reconceptualises digital overuse not as a pathological disorder but as a manifestation of social imbalance arising from mobile digital overuse. Understanding this phenomenon is essential for advancing digital well-being, informing responsible technology design, and fostering inclusive and ethical digital governance in both Chinese and international contexts.

Method/Process

Following the standard paradigm of scale development, this study adopted a three-stage progressive design. In the first stage, semi-structured interviews combined with grounded theory were used to extract key dimensions and develop an initial 28-item scale. In the second stage, exploratory and confirmatory factor analyses were conducted to verify the scale's structural validity and reliability, resulting in its refinement and validation. In the third stage, competing structural models were tested to examine the relationship between digital overuse and Internet addiction. Additionally, normative data were established based on the sample distribution, and K-means clustering was applied to identify distinct subgroup typologies.

Results/Conclusion

The findings reveal that digital overuse among Generation Z comprises three interrelated dimensions – social manifestation, psychological state, and physiological response – forming a three-order hierarchical structure. The final Digital Overuse Scale demonstrated satisfactory reliability and construct validity. Although structurally distinct, digital overuse and Internet addiction were significantly correlated, confirming the nomological validity of the construct. Normative and cluster analyses further classified participants into non-overuse, latent overuse, and severe overuse groups. Beyond measurement, these findings offer actionable insights for digital well-being design, educational intervention, and corporate social responsibility initiatives. Overall, this study refines the conceptual and empirical understanding of digital overuse and contributes to the development of inclusive, ethical, and human-centred digital ecosystems.

Keywords: Generation Z; mobile digital overuse behaviour; measurement system; Internet addiction

Introduction

The rapid evolution of mobile digital technologies in terms of speed, capacity, and application scope has profoundly reshaped modern social life. The proliferation of instant messaging, social media, online education, telemedicine, and e-commerce has significantly enhanced social efficiency and individual convenience. However, a lifestyle marked by persistent connectivity and technological dependence has also generated notable adverse consequences. Empirical studies have demonstrated that digital overuse of mobile digital technologies is strongly associated with psychological issues such as anxiety and depression (He et al., 2025), and negatively affects academic performance (Han et al., 2025), social relationships (Xiao et al., 2025), and overall life satisfaction. As this phenomenon becomes increasingly pervasive, its implications have extended beyond the individual level, emerging as a pressing social concern.

In response, researchers have developed various measurement instruments based on pathological frameworks, such as those used for gambling and substance addiction. Yet, these instruments are inherently shaped by pathological assumptions. They focus predominantly on clinical symptoms and maladaptive behaviours in small populations (Song et al., 2025), while neglecting the more subtle and prevalent forms of "overuse" embedded in everyday life. Furthermore, most existing scales are rooted in Western contexts, raising questions about their sociocultural validity and applicability in China. The country's highly digitalised environment offers a distinctive context for examining this issue. With over 1.3 billion mobile internet users, China's Generation Z (born between 1995 and 2009) constitutes the world's largest cohort of "digital natives," numbering more than 2 billion globally and accounting for 18.6% of China's population (Fastdata, 2024). This generation is both deeply familiar with and heavily reliant on digital technologies. Nevertheless, existing research on this group remains largely descriptive and fragmented, lacking systematic and validated tools to capture the nuances of digital overuse (Sharma et al., 2023).

Accordingly, this study seeks to address three interrelated questions. First, how can researchers move beyond the limitations of pathological frameworks to reconceptualise mobile digital overuse as a social issue? Second, what methodological processes can ensure a standardised approach to scale development and validation, thereby guaranteeing conceptual clarity and psychometric robustness? Third, how can the applicability of such instruments be empirically tested among Generation Z to support their practical use in education, public health, and social governance?

To answer these questions, this study focuses on Generation Z as the primary research population. It aims to systematically conceptualise mobile digital overuse within a social problem framework and to develop and validate a rigorous, operational measurement instrument through a standardised scale development process.

Theoretical Foundation and Related Research

Despite extensive scholarly attention, there remains no unified consensus on the terminology or conceptualisation of mobile digital overuse. Commonly employed terms such as Internet addiction and problematic Internet use stem largely from psychiatric and psychological traditions that emphasise pathological characteristics. While these concepts hold explanatory value in clinical or severe behavioural contexts, their pathological orientation constrains understanding and measurement of mobile digital overuse in broader, non-clinical settings.

This study adopts the term mobile digital overuse and situates it within a social framework. First, overuse is not simply a matter of usage frequency, nor is it synonymous with clinical addiction; rather, it represents a progressively accumulated state of imbalance. As individuals deepen their engagement within digital ecosystems, the time and attention allocated to offline activities are increasingly eroded, blurring the boundary between digital and real life. Second, digital overuse constitutes a behavioural continuum ranging from mild to severe manifestations, which may, in extreme cases, escalate into addiction. Third, the defining feature of overuse lies not in using more, but in using out of balance.

Symptoms such as physical fatigue, psychological strain, social withdrawal, diminished productivity, and withdrawal-like emotional disturbances are external indicators of this imbalance. In essence, mobile digital overuse reflects an erosion of digital well-being – a socialised manifestation of disordered human-technology relations.

Existing measurement instruments can be broadly categorized into two types. The first category derives from the clinical addiction framework, typically informed by theories of gambling and substance dependence. Common dimensions include salience, loss of control, withdrawal, impulsivity, mood regulation, relapse, and negative consequences. Representative instruments include Young's (1998a) Internet Addiction Test (IAT), the Internet Addiction Diagnostic Questionnaire (IADQ; Young, 1998b), the Bergen Social Media Addiction Scale (BSMAS; Andreassen et al., 2012), the Smartphone Addiction Scale (SAS; Kwon et al., 2013), the Smartphone Addiction Inventory (SPAI; Lin et al., 2014), and the Information Addiction Scale (Wang et al., 2023). Although widely used in pathological addiction studies, these scales inherently equate digital overuse with addiction, rendering them insufficient for assessing non-pathological forms of overuse.

The second category attempts to transcend the addiction paradigm, approaching digital overuse from broader social and behavioural perspectives. Notable examples include the Generalized Problematic Internet Use Scale 2 (GPIUS2; Caplan, 2010), based on cognitive-behavioural theory, and the Perceived Digital Overuse scale developed within digital inequality research (Gui & Büchi, 2021). Other instruments include the Problematic and Risky Internet Use Screening Scale (PRIUSS; Jelenchick et al., 2014) and the 16-item Mobile Phone Addiction Tendency Scale (MPATS; Xiong et al., 2012). Although these instruments exhibit greater general applicability, traces of pathological framing persist. Moreover, certain studies adopt time or frequency of use as proxy indicators of overuse – such as in online gaming addiction or smartphone use frequency measures – yet substantial empirical evidence indicates that high-frequency use does not inherently produce negative outcomes (Smahel et al., 2012). Whether use is deemed overuse depends on individual contexts and motivations; for instance, extended engagement with social media may benefit job seekers but become burdensome after entering the workforce (O'Reilly & Mohan, 2023).

In summary, existing research displays pronounced conceptual and terminological fragmentation. Many studies remain confined within clinical addiction frameworks, limiting their ability to capture non-pathological manifestations of digital overuse. Others, though attempting to depart from addiction-based models, still exhibit conceptual and methodological constraints. This fragmentation undermines comparability across studies and impedes a nuanced understanding of the social dimensions of digital overuse. Consequently, there is an urgent need for a theoretically grounded and psychometrically robust measurement instrument that systematically captures the multidimensional characteristics of mobile digital overuse among Generation Z.

Research Design

Drawing on the established paradigm of scale development and incorporating insights from relevant literature and empirical evidence (MacKenzie et al., 2011; Zhang et al., 2022; Thatcher et al., 2018), this study employs a mixed-methods approach that integrates data from multiple sources. The research was conducted in three sequential stages, with specific procedures summarised in Table 1.

Table 1. Steps in the Development of the Measurement System

Research Steps	Details
1. Initial Scale Generation	
Domain Definition	Define the conceptualisation and definition of mobile digital overuse behaviour in this paper
Framework and Item Generation	Identify the dimensions of digital overuse among Gen Z to form an initial scale pool
Surface Validity, Content Validity Test, Sensitivity Test	A focus group of field experts (n = 5) rated the importance of the items Conduct pilot testing on the population with digital overuse to assess readability (n = 10) Examine the rating distribution of items and delete items where ratings significantly exceed 70% on one side of the rating system Retain 28 items for validation in the next stage
2. Scale Purification and Validation (n = 600)	
Representative Sample	Collect data using an online questionnaire survey platform Compare the demographic data of the sample with the baseline in China's industry reports

Research Steps	Details
	Randomly divide the sample into two sub-samples for exploratory and confirmatory factor analysis
Exploratory Factor Analysis (n = 300)	Perform principal component analysis (PCA) to explore the factor structure and item reliability of the proposed scale Further refine the proposed scale based on the results of PCA and parallel analysis Retain 28 items at the end of this stage
Confirmatory Factor Analysis (n = 300)	Conduct confirmatory factor analysis to test the reliability, convergent validity, and discriminant validity (including Cronbach's alpha, composite reliability, average variance extracted, etc.) of the factor structure of the proposed scale Verify the structure of the second-order model
3. Further Scale Validation and Application (n = 317)	
Prepare Representative Sample	Collect data using the China Representative Panel Service Compare the demographic data of the sample with the baseline in China's industry reports
Analyse the Overlap and Distinction with Internet Addiction Scale	A competing CFA model examines the overlap and distinction between mobile digital overuse behaviour and internet addiction
Naming Validity	Examine the relationship between digital overuse and internet addiction
Scale Application	Establish a norm for digital overuse among Gen Z Classify the degree of digital overuse among Gen Z

Analysis and Results

Scale Generation

Domain Definition

A precise conceptual definition and a robust theoretical framework are essential prerequisites for effective scale development. In this study, mobile digital overuse among Generation Z is defined *as an individual's overuse of mobile digital devices, technologies, or online platforms. This construct encompasses both absolute overuse – in terms of time and intensity – and relative overload across psychological, physiological, and social dimensions, representing an imbalanced form of human-technology interaction*. Such imbalance adversely affects multiple aspects of daily life, including physical and mental health, interpersonal relationships, and overall productivity, thereby diminishing users' digital well-being. Importantly, mobile digital overuse is conceptualised here as a non-pathological behavioural continuum, ranging from mild to severe manifestations rather than as a discrete or clinical condition.

Framework and Item Generation

This study adopted a grounded theory approach to identify and conceptualise the underlying dimensions of mobile digital overuse. Semi-structured interviews with Generation Z digital users were conducted to explore the concrete manifestations of digital overuse in depth. The qualitative insights obtained from these interviews were subsequently integrated with a review of relevant literature to generate hierarchical dimensions and corresponding measurement items within a grounded theoretical framework.

Prior to data collection, the initial interview guide was reviewed by both subject-matter experts and target users. Using convenience sampling, five experts in information systems and five Generation Z users were consulted, and their feedback was used to refine the interview guide. The final version consisted of five sections and seven core questions. Given the study's focus, participants were recruited from individuals aged 15–29. A total of 34 participants who self-reported meeting the criteria for mobile digital overuse were recruited. Each semi-structured interview lasted approximately 25 minutes, was audio recorded with informed consent, and subsequently transcribed verbatim for analysis.

Following the standardised analytical procedures of classical grounded theory, the interview transcripts were analysed through open coding, axial coding, and selective coding (Corbin & Strauss, 2014). During open coding, 28 initial concepts were identified. In axial coding, these were refined into seven core categories: high-intensity interaction, functional imbalance, redundant consumption, emotional fluctuation, psychological dysregulation of use, and somatic – cognitive fatigue. During selective coding, manifestations of digital overuse among Generation Z was established as the central category, forming the foundation for a multi-level dimensional structure aligned with the social-physiological-psychological model.

Ultimately, the seven categories were integrated into three second-order constructs: social overuse, psychological overuse, and physiological overuse. Specifically, social overuse encompasses high-intensity interaction, functional imbalance, and redundant consumption, reflecting disruptions in social functioning such as reduced offline engagement and diminished efficiency due to frequent, multi-device, and purposeless digital activity. Psychological overuse includes psychological dysregulation and emotional fluctuation, characterised by over-dependence on digital information, immersive use, and negative affective responses such as anxiety, regret, and fatigue following online interaction.

Physiological overuse refers primarily to somatic and cognitive fatigue, manifesting as eye strain, neck and shoulder discomfort, reduced attention, and memory decline. Detailed results are presented in Table 2.

Table 2. Main Categories, Sub-Categories, Definitions, and Theoretical Sources

Main Category	Sub-Category	Definition	Theoretical Source
Social Manifestations	High-Intensity Interaction	"Excessive interaction" refers to users' overuse of digital devices in terms of time or content, including frequent interactions, exceeding expected usage time, prolonged actual usage, and simultaneous interactions across multiple devices or content.	Gui M et al. argued that excessive online time and multitasking are two key dimensions of digital overuse (Gui & Büchi, 2021). Summary from this study's interviews (P2, P10, P12, P16, P18).
	Functional Imbalance	"Functional imbalance" describes a user's unbalanced functional performance when using digital systems, or inappropriate digital device use that disrupts normal life functionality, thereby affecting	Bianchi A et al. argued that decreased work efficiency, disrupted plans, neglected offline activities, and irregular routines are key aspects of smartphone addiction (Bianchi & Phillips, 2005).

Main Category	Sub-Category	Definition	Theoretical Source
		overall well-being and efficacy. Includes reduced work efficiency, neglect of offline activities, disruption of other planned schedules, and irregular daily routines.	Summary from this study's interviews (P1, P7, P13, P22).
	Redundant Consumption	"Redundant consumption" refers to unnecessary and non-value-added digital usage behaviours. Specifically, it includes aimless use, use without substantial benefit, deviation from original needs, and use conflicting with high-priority tasks or needs.	Summary from this study's interviews (P8, P14, P19, P20).
Psychological Manifestations	Emotional Fluctuations	Emotional fluctuations refer to negative emotional changes experienced by users after digital device usage, including contradiction, anxiety, regret, guilt, boredom, and stress.	Summary from this study's interviews (P4, P12, P17, P24).
	Usage Psychological Dysfunction	Usage psychological dysfunction refers to the psychological states and behavioural patterns exhibited when individuals use digital devices (e.g., smartphones, tablets, computers), including unease about missing information, excessive dependence on the device, and deep immersion during usage.	Kwon M et al. suggested that internet-oriented relationships are an important dimension of smartphone addiction (Kwon et al., 2013). Significance is one of the seven classic dimensions of behavioural addiction. Summary from this study's interviews (P1, P3, P5, P12).
	Usage Control	Usage control refers to users' conscious efforts to regulate their digital usage after recognising overuse, ensuring it aligns with their intended level. Includes two aspects: conscious usage control and intermittent overuse.	Recurrence is one of the seven classic dimensions of behavioural addiction. (ii) Summary from this study's interviews (P6, P18).
	Somatic-Cognitive Fatigue	Somatic-cognitive fatigue refers to the decline in physical and cognitive functions after prolonged digital device usage. This includes physical discomfort (e.g., exhaustion, vision deterioration, neck pain) and cognitive issues (e.g., forgetfulness, attention deficits).	Chen et al. argued that health problems are a critical component of problematic internet use (Chen et al., 2003). Summary from this study's interviews (P2, P5, P6, P11).

Building on the dimensional structure derived from the grounded analysis, items were generated by tracing back to the corresponding interview data within each dimension. Established measurement frameworks from prior studies were also selectively referenced to enhance theoretical robustness. The resulting items were carefully revised to ensure conceptual clarity and contextual relevance to mobile digital overuse among Generation Z. At the conclusion of this stage, an initial item pool comprising 29 items was developed, as presented in Table 3.

Surface Validity, Content Validity, and Sensitivity Testing

Face validity assesses the apparent appropriateness of items, while content validity evaluates the extent to which items adequately represent the construct of interest. Both experts and participants can assess items in terms of relevance, comprehensiveness, and clarity (Mokkink et al., 2024). To enhance the content validity and applicability of the developed scale, two complementary evaluation procedures were conducted.

First, five experts in the field of information behaviour were invited to review the initial scale. Each item was rated according to its representational accuracy using three categories: clearly expresses, partially expresses, and does not express. Experts were also asked to provide comments and recommendations for items rated as partially expresses or does not express.

Second, ten Generation Z users who exhibited characteristics of digital overuse were invited to evaluate the same version of the scale, applying identical rating criteria.

Items that received a does not express rating from two or more evaluators (either experts or users) were eliminated. Items rated as partially expresses were revised based on qualitative feedback. Following this process, one item (DF4) was removed, resulting in a refined scale comprising 28 items.

Table 3. Retained Items in the Initial Scale

First-order Dimension	Second-order Dimension	Measurement Items
High-intensity Interaction (HI)	Frequent Interaction	HI1 I always check or search for interesting content on my mobile device from time to time.
	Exceeding Expected Usage Time	HI2 I find that the time I spend using my mobile device exceeds my expectation.
	Long Actual Usage Duration	HI3 My friends or colleagues say that I spend a lot of time on my mobile device.
	Multi-device or Multi-content Interaction	HI4 I often use multiple digital devices simultaneously or do multiple things on the same device at the same time.
Functional Imbalance (DF)	Decreased Work Efficiency	DF1 Spending too much time on my mobile device leads to a decline in my study/work efficiency.
	Ignoring Offline Activities or Relationships	DF2 I often neglect offline activities and interactions (such as not going out, affecting relationships with family and friends, etc.) due to digital overuse of my mobile device.
	Affecting Other Arrangements and Schedules	DF3 I find that I procrastinate or miss work/study plans because of using my mobile device.
	Irregular Rest	DF4 Due to digital overuse of my mobile device, I have

First-order Dimension	Second-order Dimension	Measurement Items
		been disrupted in my rest schedule (staying up late and unable to get up early, etc.) more than once.
Redundant Consumption (RC)	Using Aimlessly	RC1 I often use my mobile device aimlessly when I'm bored.
	Using Without Real Help	RC2 Even when the activities I engage in on my mobile device are meaningless or unhelpful to me, I still continue to use it.
	Using Deviating from Original Needs	RC3 When using my mobile device for work or study, I'm often distracted by other irrelevant content, causing me to deviate from my original task objectives.
	Using in Conflict with High-priority Tasks	RC4 Even when I have high-priority tasks such as studying or working, I still can't help but use my mobile device to do other unrelated things.
Maladaptive Use Psychology (MD)	Fear of Missing Out	MD1 I constantly check my mobile device so as not to miss messages from others or updates on social media.
	Flow	MD2 When using my mobile device, I always feel that time passes very quickly.
	Preference for Mobile Search	MD3 I prefer to search on my mobile device rather than through books or asking others.
	Checking upon Notification	MD4 When my mobile device has any message prompts, I can't help but check immediately.
Emotional Fluctuation (ED)	Regret	ED1 After using my mobile device, I feel regretful.
	Guilt	ED2 After using my mobile device, I feel guilty.
	Pressure	ED3 After using my mobile device, I feel weary or have inexplicable pressure.
	Anxiety	ED4 After using my mobile device, I feel anxious.
Usage Control (UC)	Conscious Control	UC1 When I think I'm overusing my mobile device, I will consciously control its use.
	Recurrence	UC2 I always think that I should shorten the usage time of my mobile device.
	Intentional but Short-lived Control	UC3 I try to shorten the usage time of my mobile device again and again, but I can only maintain it for a short period and cannot achieve long-term control.
	Multiple Attempts to Control	UC4 I have tried various measures to control my mobile device usage, such as downloading focus apps and putting my phone in places hard to reach.
Corporeal Cognitive Fatigue (CF)	Physical Exhaustion	CF1 After using my mobile device, I have experienced symptoms like dizziness or blurred vision.

First-order Dimension	Second-order Dimension	Measurement Items
	Vision Decline	CF2 After using my mobile device, I have felt pain in my wrists or back of my neck.
	Tiredness	CF3 After using my mobile device, I have experienced tiredness and lack of sleep, along with irregular sleep patterns.
	Forgetfulness	CF4 After using my mobile device, I feel that my memory has decreased somewhat.
	Lack of Concentration	CF5 After using my mobile device, I find it difficult to concentrate during study or work.

Scale Refinement and Validation

Sampling and Data Collection

The finalised scale was formatted as a 7-point Likert-type questionnaire. Participants were recruited according to the following criteria: (i) individuals born between 1995 and 2009, corresponding to Generation Z, and (ii) successful completion of attention-check items to ensure data integrity. Questionnaire items were randomly ordered and distributed via the Credamo sampling service.

In accordance with COSMIN guidelines, the recommended sample size for scale validation is five to ten times the number of items to ensure reliability and validity. A total of 600 valid responses were obtained and randomly divided into two subsamples for exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The gender distribution was approximately equal (1:1), and most participants possessed at least a high school education. Overall, the demographic structure closely aligned with the general characteristics of China's Generation Z population, ensuring adequate representativeness (Deloitte, 2024).

Exploratory Factor Analysis

Exploratory factor analysis (EFA) was conducted using SPSS 26.0. The Kaiser-Meyer-Olkin (KMO) value was 0.931 (> 0.80), and Bartlett's test of sphericity was significant ($p < 0.001$), indicating that the data were suitable for factor analysis. A principal component analysis (PCA) was then performed on a randomly selected subsample ($n = 300$) to identify the latent factor structure.

Seven factors with eigenvalues greater than 1 were extracted, collectively explaining 73.158% of the total variance. All item communalities exceeded 0.60, confirming strong representation of the underlying constructs; thus, all items were retained.

To further evaluate item discrimination, Howard's "40–30–20" criterion was applied to assess factor loadings and cross-loadings (Howard, 2016). According to this standard, an item is deemed appropriate for retention if:

(a) its primary factor loading exceeds 0.40, (b) its loading on alternative factors is below 0.30, and (c) the difference between primary and secondary loadings is at least 0.20. As presented in Table 4, all 28 items met these thresholds, confirming their suitability for inclusion in the final factor structure.

Table 4. Rotated Component Matrix

	1	2	3	4	5	6	7
HII1	0.188	0.171	0.811	0.118	0.150	0.099	0.162
HII2	0.173	0.086	0.762	0.133	0.232	0.114	0.208
HII3	0.236	0.185	0.712	0.158	0.163	0.125	0.238
HII4	0.142	0.126	0.787	0.138	0.146	0.182	0.118
DF1	0.183	0.032	0.169	0.149	0.153	0.182	0.798
DF2	0.154	0.180	0.218	0.110	0.074	0.109	0.788
DF3	0.166	0.175	0.264	0.185	0.197	0.093	0.739
RC1	0.131	0.112	0.182	0.291	0.734	0.192	0.124
RC2	0.146	0.166	0.178	0.177	0.744	0.205	0.146
RC3	0.226	0.155	0.167	0.128	0.752	0.072	0.121
RC4	0.104	0.085	0.154	0.238	0.747	0.160	0.076
MD1	0.191	0.239	0.114	0.727	0.236	0.017	0.141
MD2	0.132	0.122	0.188	0.745	0.242	0.182	0.059
MD3	0.190	0.184	0.160	0.756	0.202	0.146	0.107
MD4	0.193	0.198	0.087	0.747	0.158	0.129	0.198
ED1	0.162	0.771	0.131	0.200	0.114	0.210	0.136
ED2	0.178	0.802	0.132	0.131	0.177	0.174	0.093
ED3	0.204	0.757	0.154	0.176	0.131	0.197	0.057
ED4	0.175	0.727	0.152	0.241	0.095	0.252	0.157
UC1	0.113	0.213	0.183	0.185	0.130	0.783	0.128
UC2	0.240	0.173	0.093	0.155	0.204	0.738	0.095
UC3	0.219	0.280	0.122	0.092	0.153	0.673	0.216
UC4	0.246	0.181	0.129	0.050	0.142	0.757	0.041
CF1	0.777	0.145	0.137	0.139	0.099	0.189	0.103
CF2	0.759	0.172	0.179	0.181	0.111	0.170	0.153
CF3	0.753	0.160	0.171	0.097	0.162	0.211	0.040
CF4	0.780	0.097	0.186	0.183	0.089	0.146	0.152
CF5	0.717	0.197	0.099	0.147	0.215	0.124	0.180

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) was conducted using AMOS 26.0 on the remaining subsample (n = 300) to validate the structural integrity and dimensional composition of the Generation Z Excessive Digital Use Scale. As shown in Table 6, the Cronbach's α values for all seven first-order factors exceeded 0.70, and the composite reliability (CR) values were all above 0.80, indicating strong internal consistency and reliability (Yadav & Rahman, 2017).

For convergent validity, the average variance extracted (AVE) and factor loadings were examined.

All AVE values exceeded 0.50, and all factor loadings were greater than 0.70, confirming satisfactory convergent validity for the first-order constructs. Discriminant validity was further assessed by comparing the square roots of the AVE values with the inter-factor correlations. All AVE values were greater than 0.50, and the square root of each factor's AVE exceeded its correlations with other factors, indicating good discriminant validity (Fornell & Larcker, 1981).

The model's overall fit indices demonstrated an excellent fit to the data: χ^2 = 422.380, df = 329, χ^2/df = 1.284, RMSEA = 0.031, CFI = 0.981, TLI = 0.978, and IFI = 0.981. All indices met or exceeded the commonly accepted thresholds, providing strong evidence for the construct validity and structural soundness of the scale.

Table 5. Reliability and Discriminant Validity Analysis

	HI	DF	RC	MD	ED	UC	CF
HI	0.791						
DF	0.604	0.773					
RC	0.525	0.436	0.790				
MD	0.460	0.441	0.614	0.809			
ED	0.502	0.395	0.502	0.524	0.797		
UC	0.445	0.447	0.474	0.525	0.688	0.783	
CF	0.575	0.437	0.436	0.479	0.491	0.495	0.807
C'α	0.867	0.814	0.869	0.883	0.874	0.863	0.902
AVE	0.625	0.597	0.624	0.655	0.636	0.613	0.651
CR	0.869	0.816	0.874	0.888	0.868	0.863	0.943

Note: The bolded numbers on the diagonal are the square roots of AVE.

Third-Order Factor Measurement Model

The developed Generation Z Mobile digital overuse Scale adopts a third-order factor structure. The third-order factor comprises three second-order constructs – social manifestations, psychological responses, and physiological manifestations – each represented by seven first-order constructs.

Following Burke JC's decision rules, first-order factors were modelled as reflective, whereas the second- and third-order factors were modelled as formative (Jarvis et al., 2003). Hierarchical latent variable modelling was performed using SmartPLS 3.

As shown in Figure 1, the second-order factor physiological manifestations completely overlap with its first-order constructs, exhibiting a path coefficient of 1. The path coefficients for the remaining second-order constructs ranged from 0.374 to 0.448, while those for the third-order construct ranged from 0.239 to 0.444. All coefficients exceeded the threshold of 0.2 and were statistically significant ($p < 0.001$), meeting the established criteria for formative measurement models (Ranjan & Read, 2016).

These results confirm the robustness of the third-order factor structure and substantiate the hierarchical and theoretically coherent organisation of dimensions within the Generation Z Mobile Digital Overuse Scale.

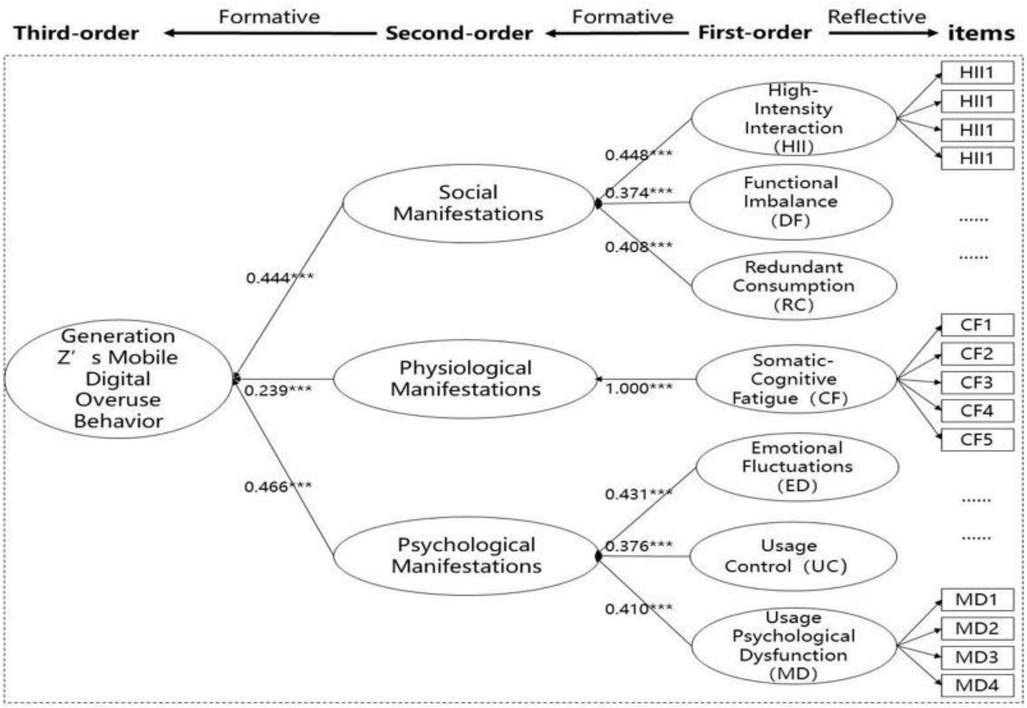

Third-order
Formative
Second-order
Formative
First-order
Reflective
items
Generation Z' s Mobile Digital Overuse Behavior
0.444***
0.239***
0.466***
Social Manifestations
Physiological Manifestations
Psychological Manifestations
0.448***
0.374***
0.408***
1.000***
0.431***
0.376***
0.410***
High-Intensity Interaction (HII)
Functional Imbalance (DF)
Redundant Consumption (RC)
Somatic-Cognitive Fatigue (CF)
Emotional Fluctuations (ED)
Usage Control (UC)
Usage Psychological Dysfunction (MD)
HII1
HII1
HII1
HII1
......
......
CF1
CF2
CF3
CF4
CF5
......
......
MD1
MD2
MD3
MD4

Figure 1. Three-Factor Structure

Secondary Validation of the Scale

Questionnaire Design and Sample Collection

The proposed scale was developed to assess mobile digital overuse among Generation Z. The Internet addiction measure was adapted from Young's Internet Addiction Diagnostic Questionnaire[8], originally derived from the pathological gambling framework and consisting of eight items. These items were slightly revised to align with the seven-point Likert scale employed in this study, ranging from "strongly disagree" to "strongly agree."

Study 3 utilised the same data collection platform and procedures as Studies 1 and 2 (Questionnaire Star) to recruit participants. A total of 321 responses were obtained; after excluding cases that failed attention-check items, 317 valid responses were retained for analysis.

Overlap and Distinction Between Mobile digital overuse and Internet Addiction

This section investigated the latent relationship between mobile digital overuse and Internet addiction through competing confirmatory factor analysis (CFA) models to determine whether the two constructs represent distinct or overlapping theoretical dimensions (Mathes et al., 2018).

Two alternative models were specified:

(i) a correlated-factors model, in which the first-order factors of mobile digital overuse were freely correlated with those of Internet addiction, treating them as related but conceptually distinct constructs; and

(ii) a single-factor model, in which the first-order factors of both constructs were combined into a unified general factor.

Results indicated that the correlated-factors model provided a significantly better fit to the data than the single-factor model ($\Delta\chi^2$ = 2137.697, df = 6, p < .001), demonstrating a substantial improvement in model fit (Mathes et al., 2018). As shown in Table 6, the two constructs were significantly correlated (r = 0.560, p < .001), suggesting that while mobile digital overuse and Internet addiction are closely related, they remain empirically and theoretically distinguishable.

Table 6. Model Fit Indices for Competing Models of Generation Z Mobile Digital Overuse and Internet Addiction

Mobile Digital Overuse and Internet Addiction	χ2	df	CFI	TLI	RMESEA (90%CI)
Single-Factor Model	5335.324	594	0.418	0.383	0.159(0.155,0.163)
Correlated Factors Model	3197.627	588	0.680	0.657	0.118(0.114,0.122)

Validation of Nomological Validity

The mobile digital overuse model was conceptualised as a second-order, three-factor structure. Using this structural equation model, the effects of the overall construct of mobile digital overuse and its three specific dimensions – social manifestations, psychological responses, and physiological responses – on Internet addiction were examined. It was hypothesised that the general factor of mobile digital overuse would exert a stronger influence on individual levels of Internet addiction than the specific second-order factors, as it captures the comprehensive behavioural state of digital overuse.

It should be noted that within the three-factor structural equation model, the factor loadings of scale items are shared between subfactors and the general factor; therefore, the average variance extracted (AVE) of subscales is not required to meet the conventional threshold of 0.50 (Pian et al., 2024). Additionally, to address potential common method bias (CMB), Harman's single-factor test was conducted using SPSS 26.0. The results indicated that the total cumulative variance exceeded 60%, while the variance explained by the first factor was 33.13%, below the critical threshold of 40%, suggesting the absence of substantial common method bias (Podsakoff et al., 2003).

Structural equation modelling results confirmed the hypothesised relationships. Mobile digital overuse showed a significant positive association with Internet addiction (β = 0.562, p < 0.001), whereas the three specific second-order factors demonstrated no statistically significant effects (β = 0.192, p = 0.797; β = 0.148, p = 0.853; β = 0.137, p = 0.689). These findings indicate that the third-order model of mobile digital overuse is both statistically robust and conceptually stable, thereby confirming the nomological validity of the developed scale.

Norm Establishment for Mobile Digital Overuse Among Generation Z

Norms represent the score distribution characteristics of a standardised sample, with percentile norms being the most commonly applied approach. In this study, scores from a representative subset of participants were statistically analysed to establish a reliable distribution, upon which corresponding percentile norms were determined.

Given that the Mobile Digital Overuse Scale adopts a third-order model – with first-order factors modelled as reflective and second- and third-order factors as formative – score computation proceeded in two steps. First, the mean scores of the seven first-order factors were calculated across all 28 items, each rated on a 7-point scale. Second, total scores for the three second-order factors were computed, yielding values of 21, 21, and 7 points, respectively. The overall score for the third-order construct was thus 49 points.

Based on the total score distribution, five percentile norms were established to categorise the varying degrees of mobile digital overuse. Table 7 presents the percentile norms ranging from the 5th to the 95th percentile, providing a standardised reference framework for subsequent assessment and comparative analysis.

Table 7. Norms for Mobile Digital Overuse

Percentage / Secondary Indicators	Social Manifestations	Psychological Manifestations	Physiological Manifestations	Total Score
5%	8.1	8.3	2.0	19.2
25%	12.8	13.5	5.0	32.5
50%	15.6	16.0	5.8	36.5
75%	17.9	17.8	6.2	40.1
95%	19.7	19.5	6.6	45.8

Classification of Mobile Digital Overuse Levels Among Generation Z

To more accurately delineate the severity of mobile digital overuse among Generation Z, a K-means cluster analysis was conducted based on the finalised scale data. Prior to clustering, the optimal number of clusters was determined using the elbow method: as the number of clusters (K) increased, the sum of squared errors (SSE) gradually declined, and a distinct inflection point was observed at K = 3, where the rate of SSE reduction noticeably plateaued. Consequently, the sample was divided into three relatively distinct clusters.

According to the K-means clustering results (see Table 8), three groups exhibited significant differences across the dimensions of social manifestations (SP), psychological states (MS), and physiological responses (PR). The first cluster showed the highest scores, the second cluster intermediate scores, and the third cluster the lowest. Accordingly, these groups were labelled as “severe digital overusers,” “potential digital overusers”, and “non-digital overusers.” Among the 317 valid responses, 233 participants were classified as severe digital overusers, 44 demonstrated a tendency toward overuse, and 40 were categorised as non-digital overusers.

By analysing the mean scores of the key indicators – SP, MS, and PR – researchers can determine whether participants exhibit characteristics of mobile digital overuse and further classify the severity of overuse. This classification provides an empirical foundation for designing targeted intervention strategies aimed at mitigating mobile digital overuse and promoting healthier digital behaviour among Generation Z.

Table 8. Cluster Results of Mobile Digital Overuse Levels Among Generation Z

Indicators	Cluster Categories (Mean ±SD)			F_value	P_value
	Category 1 (n=233)	Category 2 （44）	Category 3 （40）		
Social Manifestations	15.804±2.99	15.46±2.71	8.73±2.45	112.113	0.000***
Psychological Manifestations	16.24±2.58	15.37±1.96	8.45±1.53	197.701	0.000***
Physiological Manifestations	16.24±0.41	8.45±1.53	3.27±1.4	723.321	0.000***

Summary and Future Directions

This study developed and validated a comprehensive measurement system for mobile digital overuse among Generation Z, addressing the current absence of systematic quantitative tools within a non-clinical framework. The findings hold both theoretical and practical significance.

Theoretical Contributions. The study proposed and empirically tested a multi-level structural model comprising three second-order dimensions – social, psychological, and physiological – and seven first-order factors, thereby deepening the conceptualisation and operationalisation of mobile digital overuse. This model provides a novel theoretical perspective for advancing research in information behaviour and digital addiction studies. Moreover, the study delineated the theoretical and structural distinctions between mobile digital overuse and Internet addiction, contributing to ongoing debates on their conceptual boundaries. The establishment of normative data and hierarchical classifications further offers baseline references for future comparative research.

Practical Implications. The validated Scale provides a reliable tool for identifying different levels of mobile digital overuse among Generation Z. It can be applied in educational, managerial, and policy contexts to design personalised interventions and promote healthier digital habits. The study also highlights the importance of responsible and human-centred technology design. By integrating digital well-being principles into education, governance, and corporate practice, these efforts collectively contribute to building an inclusive and sustainable digital society.

Limitations. The study sample was limited to Chinese Generation Z, and thus the cross-cultural generalizability of the findings requires further verification. In addition, the reliance on self-reported data may introduce response bias, and the cross-sectional design limits the ability to capture dynamic behavioural changes over time. Future research should pursue cross-cultural validation, incorporate objective behavioural data to improve measurement precision, and explore the social, academic, occupational, and psychological consequences of mobile digital overuse within diverse contexts.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical Approval

The study protocol was approved by the Ethic Committee of the School of Information Management, Nanjing University.

References

Andreassen, C. S., Torsheim, T., & Brunborg, G. S. (2012). Development of a Facebook addiction scale. *Psychological Reports*, *110*(2), 501–517.

Bianchi, A., & Phillips, J. G. (2005). Psychological predictors of problem mobile phone use. *CyberPsychology & Behavior*, *8*(1), 39–51. https://doi.org/10.1089/cpb.2005.8.39

Caplan, S. E. (2010). Theory and measurement of generalised problematic internet use: A two-step approach. *Computers in Human Behavior*, *26*(5), 1089–1097.

Chen, S.-H., Weng, L.-J., Su, Y.-J., Wu, H.-M., & Yang, P.-F. (2003). Development of Chinese Internet Addiction Scale and its psychometric study. *Chinese Journal of Psychology*, *45*(3), 251–266.

Corbin, J., & Strauss, A. L. (2014). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (4th ed.). Sage Publications.

Deloitte. (2024). *Deloitte global 2024 Gen Z and millennial survey*. https://www2.deloitte.com/content/dam/Deloitte/cn/Documents/about-deloitte/deloitte-cn-2024-genz-millennial-survey-zh-240529.pdf

Fastdata. (2024). *Global Generation Z consumer insight report 2024*. Fastdata Publishing.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.1177/002224378101800104

Gui, M., & Büchi, M. (2021). From use to overuse: Digital inequality in the age of communication abundance. *Social Science Computer Review*, *39*(1), 3–19.

Han, M., Jing, Y., & Zhu, X. (2025). How does social networking site addiction induce academic burnout? An explanation based on self-concordance and core self-evaluation. *Chinese Journal of Clinical Psychology*, *33*(2), 272–276, 422.

He, C., Yan, H., & Yang, Z. (2025). The association between problematic or addictive use of the internet, mobile phones, and games and non-suicidal self-injury: A meta-analysis. *Chinese Journal of Clinical Psychology*, *33*(2), 223–233.

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human–Computer Interaction*, *32*(1), 51–62. https://doi.org/10.1080/10447318.2015.1087664

Livingstone, S., Kirwil, L., Ponte, C., & Staksrud, E. (2012). *Excessive internet use among*

European children. LSE Research Online. https://eprints.lse.ac.uk/47344/1/Excessive%20internet%20use

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, *30*(2), 199–218. https://doi.org/10.1086/376806

Jelenchick, L. A., Eickhoff, J., & Christakis, D. A. (2014). The problematic and risky internet use screening scale (PRIUSS) for adolescents and young adults: Scale development and refinement. *Computers in Human Behavior*, *35*, 171–178.

Kwon, M., Kim, D. J., & Cho, H. (2013). The smartphone addiction scale: Development and validation of a short version for adolescents. *PLOS ONE*, *8*(12), Article e83558.

Kwon, M., Lee, J. Y., Won, W. Y., Park, J. W., Min, J. A., Hahn, C., Gu, X., Choi, J. H., & Kim, D. J. (2013). Development and validation of a smartphone addiction scale (SAS). *PLOS ONE*, *8*(2), Article e56936. https://doi.org/10.1371/journal.pone.0056936

Lin, Y. H., Chang, L. R., & Lee, Y. H. (2014). Development and validation of the smartphone addiction inventory (SPAI). *PLOS ONE*, *9*(6), Article e98312.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, *35*(2), 293–334.

Mathes, B. M., Norr, A. M., Allan, N. P., Schmidt, N. B., & Zvolensky, M. J. (2018). Cyberchondria: Overlap with health anxiety and unique relations with impairment, quality of life, and service utilization. *Psychiatry Research*, *261*, 204–211. https://doi.org/10.1016/j.psychres.2017.12.064

Mokkink, L. B., Prinsen, C. A., Patrick, D. L., & Terwee, C. B. (2024). *COSMIN study design checklist for patient-reported outcome measurement instruments*. COSMIN. https://www.cosmin.nl

O'Reilly, C., & Mohan, G. (2023). Parental influences on excessive internet use among adolescents. *Internet Research*, *33*(7), 86–110.

Pian, W., Zheng, R., Potenza, M. N., Chen, L., & Liu, Y. (2024). Health information craving: Conceptualization, scale development and validation. *Information Processing & Management*, *61*(4), Article 103717. https://doi.org/10.1016/j.ipm.2024.103717

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Ranjan, K. R., & Read, S. (2016). Value co-creation: Concept and measurement. *Journal of the Academy of Marketing Science*, *44*(3), 290–315. https://doi.org/10.1007/s11747-014-0397-2

Sharma, M., Kaushal, D., & Joshi, S. (2023). Adverse effect of social media on Generation Z users' behavior: Government information support as a moderating variable. *Journal of Retailing and Consumer Services*, *72*, Article 103256.

Smahel, D., Helsper, E., & Green, L. (2012). *Excessive internet use among European children*. London School of Economics.

Song, Q., Gan, H., & Liu, Z. (2025). Why are short videos “addictive”? Development of a short video addiction behavior scale and investigation of addictive effects based on cognitive pretests. *Journalism and Communication Review*, 1–13. Advance online publication. https://doi.org/10.15897/j.cnki.cn51-1046/g2.20250407.001

Thatcher, J. B., Wright, R. T., Sun, H., Zagenczyk, T. J., & Klein, R. (2018). Mindfulness in information technology use: Definitions, distinctions, and a new measure. *MIS Quarterly*, *42*(3), 831–848.

Wang, W., Tang, S., & Qian, P. (2023). Construction and empirical study of an information addiction measurement scale. *Library and Information Service*, *67*(23), 99–110.

Xiao, L., Li, H., & Feng, F. (2025). The impact of social exclusion trajectories among college freshmen on online game addiction: Parallel mediating effects of sense of control and face consciousness. *Chinese Journal of Clinical Psychology*, *33*(3), 535–540.

Xiong, J., Zhou, Z., & Chen, W. (2012). Development of the mobile phone addiction tendency scale for college students. *Chinese Journal of Mental Health*, *26*(3), 222–225.

Yadav, M., & Rahman, Z. (2017). Measuring consumer perception of social media marketing activities in e-commerce industry: Scale development & validation. *Telematics and Informatics*, *34*(7), 1294–1307. https://doi.org/10.1016/j.tele.2017.06.001

Young, K. S. (1998a). *Caught in the net: How to recognise the signs of internet addiction and a winning strategy for recovery*. John Wiley & Sons.

Young, K. S. (1998b). Internet addiction: The emergence of a new clinical disorder. *CyberPsychology & Behavior*, *1*(3), 237–244.

Zhang, N., Wang, C., & Karahanna, E. (2022). Peer privacy concerns: Conceptualization and measurement. *MIS Quarterly*, *46*(1), 491–530.

Commentary

Beyond Geopolitics: Immigrant Entrepreneurs' Digital Bridges Facilitate Trade Breakthroughs

ZHANG Jiefan[1]
Arul CHIB[2]

[1]School of Journalism and Communication, Beijing Institute of Graphic Communication, China
[2]International Institute of Social Studies, Erasmus University Rotterdam, The Netherlands

Abstract

This commentary investigates how Chinese immigrant entrepreneurs (xinyimin) in the Netherlands leverage specific digital practices to navigate geopolitical challenges and manage trade. Drawing on interviews and ethnographic engagement, we propose the concept of "digital bridges," a form of digital guanxi resilience rooted in complex co-ethnic social relational networks.

Our findings reveal that entrepreneurial strategies are dynamically balanced across three symbiotic relationship types: Jiaoyi (transactional/familial), Jiaowang (communal/reputation-based), and Jiaoqing (emotionally-grounded partnerships). This framework highlights how entrepreneurs make selective digital choices, particularly in their cautious adoption of communication technologies, to protect their cultural identity and build social capital. Ultimately, these digital bridges ensure continuous social adaptation and empower entrepreneurs to build economic resilience. We argue for a governance system that supports this selective digital engagement, balancing technological adoption with cultural preservation for sustained transnational entrepreneurship.

Introduction and Concepts

Extant research has examined the role of digital platforms in immigration as both facilitator of integration and as a contentious political topic. For immigrants, high levels of well-being have been associated with adaptation to the host society (Paloma et al., 2021). However, their use of digital ecosystems is intricately linked to socio-economic inclusion (Clayton & Macdonald, 2013; Selwyn, 2004), influencing economic advancement and social integration (AbuJarour et al., 2019; Alam & Imran, 2015). For example, immigrants in the Netherlands use specific digital platforms to navigate and negotiate beyond uncertainty, thereby achieving collective digital action (Miellet, 2021). Digital platforms also create better opportunities to challenge dominant discourses surrounding bias and stereotypes (Colliver, Coyle, & Silvestri, 2019). Consequently, our research broadly aims to understand how migrant groups engage in specific digital behaviours to enhance economic engagement and improve well-being.

We note that recent research reveals the complexity and vulnerability of minority and non-Western groups. Serving as a representative case, Asian immigrants aim to enhance their economic status through mobility (McAuliffe & Jayasuriya, 2016; Missbach, 2017). As a major source country, Chinese migrants demonstrate notable strategic awareness and initiative in transnational flows (Ramsay & Pang, 2017). The emerging category of "*xinyimin*" (新移民) refers to individuals, now distributed worldwide, who emigrated from mainland China after the opening-up reformist policy in 1978 (Teresita & Carmelea, 2019). While *xinyimin* exhibit significant diversity in transnational practices, diaspora formation patterns, and integration outcomes across host societies, a common thread is that trade relations and digital platforms provide opportunities for entrepreneurship.

Within digital ecosystems, immigrant entrepreneurs develop media routines and create a sense of place (Alencar et al., 2019; Christensen et al., 2012), leveraging the convergence of global outreach and local embeddedness through digital place-making (Chen & Wellman, 2009). This allows for enhanced socio-spatial relationships, access to key information, building networks of support, and connecting identity with physical place (Lassalle and Shaw, 2021; Webster and Kontkanen, 2021). Digital place(-making) can assist immigrant entrepreneurs to address uncertainties and enhance their online well-being (Ozduzen et al., 2021).

Further, migrant communities relying on Chinese social media platforms (e.g., Douyin/TikTok, REDnote, Kuaishou) for small-scale online trade play a subtle yet significant role in bridging cultural differences and mitigating trade conflicts accordingly. They leverage their co-ethnic networks on, and the entrepreneurial capital promised by, digital ecosystems to diffuse cultural values, expanding digital interactions originally limited to family and friend networks to broader local audiences in Europe. Furthermore, foreign audiences engage in transnational economic commerce (e.g., Temu, Ochama, Shein), with more extensive and multidimensional participation, thereby potentially fostering digital inclusion.

Trade is increasingly becoming a decisive factor in the transnational livelihoods of xinyimin. Not only does contemporary trade influence migration flows under favourable bilateral conditions, but the opportunities created by trade relations also largely fulfil needs traditionally met by work or skilled migration visas. Conversely, controversies surrounding Chinese high-tech firms such as Huawei, BYD, and DJI illustrate how disruptions in trade, capital, and technology can impact migration. For example, President Trump recently signed an executive order imposing annual fee on visa applications, leading Silicon Valley internet companies, including TikTok, to suspend plans for recruiting foreign talent. This demonstrates how trade wars, to some extent, amplify challenges for Chinese migrants, motivating them to pursue social adaptation through digital practices, thereby achieving unconventional trade breakthroughs.

Building on this foundation, we adopt a relational relationship to emphasise the importance of familial networks, community resources, and co-ethnic transnational enclaves in supporting immigrant entrepreneurship by providing access to information, capital, mentorship, and customer bases. Specifically, we study the role of communication technologies in facilitating or impeding immigrant entrepreneurship by allowing for (dis)-embeddedness with host, co-ethnic, and home networks. We take a culture-oriented approach by using Chinese concepts, which can enhance the sociological understanding of social networks on which (dis)embeddedness depends.

Methodology and Findings

To understand the liminality of digital practices, we interviewed 20 Chinese entrepreneurs in Den Haag, Netherlands along with ethnographic engagement by volunteering at their business establishments. Axial coding of the transcripts revealed three categories of relationship types, digital non-use, and the core theme of guanxi resilience.

Our findings indicate that the central role of physical business spaces – and by extension, the relationships they sustain – in shaping professional and personal identities influences the extent to which entrepreneurs embrace digital place-making. We argue that these strategic choices are shaped and complicated by the need to embed simultaneously into the host society and co-ethnic entrepreneurial networks, while also maintaining relationships with family and home country contexts. In digital communication, we identify three dominant forms of social relations structured around co-ethnicity: Jiaoyi (交易) in family business, Jiaowang (交往) in community interaction, and Jiaoqing（交情）in business partnership. These of social relations regulate immigrant entrepreneurship, creating varied dependencies and tensions in digital place-making.

We conceptualise the digital practices developed by immigrant entrepreneurs to bridge cultural identities as “digital bridges,” which are essentially a form of digital *guanxi* rooted in social relational networks. While existing research often simplifies such digital bridges as technical instruments, our study reveals that they are grounded in the contested identities and complex social networks of immigrant entrepreneurs, manifested through the dynamic balance among three symbiotic relationships: Jiaoyi, Jiaowang, and Jiaoqing. These elements do not constitute a linear, frictionless channel but rather form an organic system characterised by internal dependencies, tensions, and negotiation. It is precisely this inherent complexity and capacity for dynamic adaptation that enable immigrant entrepreneurs to demonstrate resilience and recuperation under external political pressures, providing continuous navigation for both social adaptation and trade practice.

Jiaoyi

Jiaoyi (交易) within the family context presents an initial tension as an intergenerational compulsion in business operations, characterised by a weakening of home-country identity and a strengthening of host-country identity. Jiaoyi refers to a relational model operating within familial or close kinship networks, grounded in explicit economic reciprocity and familial obligations. Such relationships, built on blood ties and high trust, enable rapid and reliable resource mobilisation during crises, thereby forming a secure buffer against risks in digital place-making. With financial factors being central, Jiaoyi functions as an instrumental guarantee in economic relations, becoming a critical factor for budgeting and securing start-up capital in immigrant self-employment, particularly within family-based labour. As Sanders and Nee (1996) noted early on, cash registers in such business operations are often controlled exclusively by family members, reflecting highly motivated and cooperative group trust. Drawing on the digital communication

practices of Chinese immigrant entrepreneurs, we identify three strategies through which they advance Jiaoyi and sustain stability within kinship groups amid tense trade relations:

(1) *Raising Capital*: Immigrant families utilise WeChat groups to swiftly organise crowdfunding for opening new stores or digital marketing. While these transactions are trust-based and procedurally simple, they reinforce relational bonds and emotional labour. In this process, digital technologies are gradually integrated into family business networks, promoting transactional automation while simultaneously challenging traditional seniority-based authority. This shift empowers younger generations to lead the adoption of efficient technological tools.

(2) *Labor Allocation*: Family members serve as default administrators of digital platforms, such as food delivery apps and restaurant management systems. Typically, immigrant youth manage social media and handle online orders, while parents oversee daily restaurant operations and expense tracking. Such arrangements are often taken for granted, with wages frequently set below market rates or left unpaid informally. Consequently, many businesses remain small in scale, limiting upward mobility for the family as a whole.

(3) *Supply Chain Management*: Through trusted kinship networks in China, immigrants conduct real-time product inspections via WeChat video calls, source low-cost equipment through platforms like Douyin (Tiktok) and Kuaishou, and facilitate small-scale cross-border payments backed by relative guarantees. This approach enhances decision-making efficiency and transforms transnational cultural identity into a stabilising mechanism for preserving enterprise interests.

Jiaowang

Jiaowang (交往) within the ethnic community emerges as the second tension, rooted in communal hierarchies inherited from the home country identity. It refers to a form of ritualised and public engagement conducted within broader co-ethnic business communities or hometown associations, aimed at constructing collective identity and social reputation. Jiaowang constitutes a central component of social relations in business interactions, emphasising the exchange of business information within immigrant entrepreneurial networks (Stephens, 2013), which originates from network ties or membership in social groups. Through Jiaowang, information becomes a critical support resource, flowing easily and at low cost within immigrant networks. Co-ethnic networks help immigrant entrepreneurs access vital intelligence regarding the business climate, feasible models, and regulatory systems (Benson-Rea & Rawlinson, 2003). For instance, Lin and Zhou (2022) show that Chinese immigrant boutique salon owners in New York City primarily rely on social networks—such as hometown associations and other co-ethnic salon owners and workers – for social support. On the positive side, Jiaowang creates opportunities to seek assistance and identify new market niches across both physical and digital spaces (Sequeira et al., 2009; Zhou, 2004; Zolin & Schlosser, 2013).

However, less-educated or lower-skilled entrepreneurs tend to depend more heavily on family and co-ethnic networks rather than professional bodies or trade associations (Benson-Rea & Rawlinson, 2003). In contrast, highly educated or skilled entrepreneurs are often drawn to opportunities emerging from co-ethnic communities and actively expand the scale of their Jiaowang networks (Wang & Warn, 2018). Within this dynamic, we observe that a good reputation is necessary for meaningful participation in Jiaowang.

(1) *Engagement in Virtual Communities*: Immigrants' personal reputation is often tied to their social standing, leading many to actively participate in local Chinese chambers of commerce or hometown associations, and meanwhile, maintaining visibility in WeChat groups by joining discussions and sharing industry news to build a public profile. In contrast, these individuals show little interest in localised Facebook groups, LinkedIn communities, or WhatsApp chats, as non-co-ethnic platforms are perceived as lacking trustworthy public discourse and may increase communication costs through informational redundancy.

(2) *Investment in Reputation*: Reputation among immigrants is closely linked to business competence. To demonstrate trade capability and brand influence within co-ethnic circles, some immigrant entrepreneurs actively manage their corporate presence on transnational social media – such as Google Maps, TikTok, and Xiaohongshu (REDnote) – by publishing content, responding attentively to all reviews, and projecting professionalism and accountability. They perceive such image-building as enhancing their prestige within Jiaowang, thereby securing discursive authority within the co-ethnic community as a means of self-adaptation in the host country.

(3) *Assertion of Cultural Identity*: Immigrant entrepreneurs value the openness afforded by cultural interaction and thus actively join co-ethnic community events, applying information and resources gained through Jiaowang to subsequent business operations. This engagement provides a sense of security in the host society. To capture overseas Chinese markets, entrepreneurs create digital places aligned with their brands on platforms like Douyin (Tiktok), Instagram, and Facebook, asserting cultural identity and even collaborating with mainland Chinese enterprises online.

Jiaoqing

Unlike the two aforementioned forms of social relations, Jiaoqing (交 情) is more emotionally oriented, emphasising affective support in non-kinship relationships. Some entrepreneurs explicitly resist mixing family and business (Smans et al., 2014), turning instead to emotionally-grounded partnerships with friends, colleagues, or former superiors (Karadal et al., 2021). Jiaoqing thus describes a hybrid relationship – combining emotional bonds and instrumental reciprocity – formed among co-ethnic business partners or close peers. As Guercini et al. (2017) observe, business ties may supplant social ties in consolidating insider-ship in new business environments and evaluating opportunities, reflecting a nuanced understanding of Jiaoqing. The evolution of ties within Jiaoqing signals the subdivision and selective narrowing of social support within cooperative groups, moving beyond undifferentiated co-ethnic networks. This enables immigrant entrepreneurs to link multiple business endeavours (Håkansson et al., 2009; Johanson & Mattsson, 2013) while establishing a framework for emotional exchange that facilitates value creation (Guercini et al., 2017). Some entrepreneurs cultivate closeness and unity (Bird & Wennberg, 2016), integrating resources in innovative ways to attract partners and employees (Salaff, Greve, & Wong, 2006) beyond the scope of ordinary Jiaowang.

Our findings highlight how Jiaoqing elevates the specificity and importance of *guanxi* (关系，relationship), *renqing* (人情，reciprocal favour), and *mianzi* (面子，face) over formal laws and rules, demonstrating the dynamic strategies involved in mobilising personal connections (Hong et al., 2018).

(1) *Avoidance of Online Information Sharing*: Immigrant entrepreneurs frequently share operational experiences and trade opportunities in person, often choosing one partner's business premises as the setting for such exchanges. Digital platforms that emphasise rules, transparency, and data-driven interaction may conflict with the *renqing* and *mianzi* foundations of Jiaoqing. Consequently, collective offline activities that nurture both emotional and instrumental gains remain common. Through tangible, immediate information exchange – such as drawing fellow entrepreneurs into one's industry or sharing price-comparison data for logistics providers in China – Jiaoqing deepens in-group solidarity and reinforces circle cohesion.

(2) *Risk Sharing and Emergency Support*: Grounded in both emotion and mutual benefit, Jiaoqing creates a resource pool that is broader and more diverse than familial networks, enhancing resilience to market volatility. For example, when a restaurant faces shortages of specific ingredients, peers connected by Jiaoqing will prioritise allocating supplies at a "friendship price", enabling mutual support during difficult periods.

(3) *Collaborative Digital Marketing*: Immigrant entrepreneurs adopt collaborative digital marketing to align interests without blurring the lines between personal relations and commercial gain. By co-hosting online campaigns or sharing customer traffic, they sustain a delicate balance between partnership and competition, leveraging Jiaoqing in ways that maintain business relations. However, such digital interaction leaves traces that may formalise otherwise informal reciprocity.

The Role of Artificial Intelligence

This study further develops a commentary on the role of artificial intelligence (AI) in the digital practices of Chinese immigrant entrepreneurs in the Netherlands. We find that AI affects identity tensions within the Jiaoyi-Jiaoqing-Jiaowang relational framework.

On one hand, as a material infrastructure embedded with strong technical rationality, AI operates within clear relational boundaries. It potentially optimises supply chain management and financial processes at the Jiaoyi (transaction) level, yet its efficacy remains highly dependent on pre-existing blood-based trust networks. Once these boundaries are transgressed, the computational rationality of AI clashes significantly with the ambiguous reciprocity ethics in Jiaoqing and the emotional resonance in Jiaowang. This tension limits the application of AI for capital conversion in non-transactional relationships and instead stimulates entrepreneurs to reaffirm the value of emotional interaction.

On the other hand, the acceptance and resistance toward AI among immigrant entrepreneurs essentially constitutes a micro-politics of cultural identity. The cautious attitude of older-generation entrepreneurs toward AI is not merely technophobia but an active defence of entrepreneurial experience and cultural-social capital accumulated through long-term embodied practice. They perceive AI as an alternative logic that may dilute their cultural authority. Therefore, they selectively introduce AI support only in Jiaoyi contexts directly tied to the interests of trusted inner circles and with controllable risks, thereby safeguarding the integrity of their entrepreneurial logic and cultural identity.

In essence, although AI as material infrastructure provides intelligent solutions for business operations, its effectiveness remains largely confined to blood-based trust networks at the "Jiaoyi" level. The selective adoption of AI technology reflects a deeper cultural identity politics – AI has become an implicit challenge to local immigrant entrepreneur identity. By strictly confining AI to instrumental functions and reserving deeper professional connections for embodied interactions.

Thus, immigrant entrepreneurs have developed a distinctive resilience strategy: selective human-AI collaboration. They strategically position AI tools and digital platforms as assistants in specific operational segments, while reserving key interactions involving complex human relations, cultural interpretation, and identity affirmation for human-led engagement. This approach ensures that they harness technological empowerment while preventing the erosion of their core social capital and cultural identity by instrumental rationality. Such self-established digital exclusion essentially constitutes an inclusive practice aimed at preserving cultural subjectivity, representing how immigrant entrepreneurs construct digital well-being while negotiating their trade identities. This also offers a new perspective for understanding how immigrant communities build resilience in the digital age.

Discussion

The value orientation of "Technology for Good" prompts reflection on whether technological systems and their underlying relational logic genuinely promote social well-being. This study argues that a sustainable governance framework for immigrant entrepreneurs must transition from merely encouraging technological adoption to simultaneously establishing mechanisms that counterbalance potential non-use effects of technology. Given that immigrant entrepreneurs operate within an overlapping context of technological and cultural diversity, digital practices should empower selective human-AI collaboration strategies observed, thereby deepening cultural identity while ensuring that technology truly serves well-being.

From the perspective of multinational corporations and research institutions, support for immigrant entrepreneurs should extend beyond simply providing products to helping users develop the capacity to resist digital exclusion – that is, empowering them with the right and literacy to 'say no'. We encourage leading high-tech enterprises to launch Digital Resilience empowerment programs that include tailored digital literacy training for immigrant entrepreneurs, developing AI models specialised in cross-cultural trade, and enhancing the localisation of algorithm adaptation and trade hotspot recommendations on digital platforms. The focus should not only be on how-to-use tools but also on how to maintain agency in interactions with algorithms – including understanding algorithmic logic, practicing data self-protection, and strategically leveraging rather than blindly complying with platform rules.

For research centres, we recommend establishing a multidimensional framework for assessing the vitality of transnational business environments and cultural sensitivity. The "Jiaoyi-Jiaowang-Jiaoqing" framework proposed in this study, which emphasises the relational capital dimension for transnational immigrant entrepreneurs, could serve as an important reference. By constructing a governance system that balances technological empowerment with cultural preservation, we can ensure that technological development does not undermine the resilience of entrepreneurial communities while fostering the sustainable development of transnational cultural communities, thereby enhancing the creativity and sense of identity among global immigrant entrepreneurs.

References

AbuJarour, S. A., Bergert, C., Gundlach, J., Köster, A., & Krasnova, H. (2019). Your home screen is worth a thousand words: Investigating the prevalence of smartphone apps among refugees in Germany. In *Proceedings of the Americas Conference on Information Systems*.

Alam, K., & Imran, S. (2015). The digital divide and social inclusion among refugee migrants: A case in regional Australia. *Information Technology & People*, *28*(2), 344–365.

Alencar, A., & Tsagkroni, V. (2019). Prospects of refugee integration in the Netherlands: Social capital, information practices and digital media. *Media and Communication*, *7*(2), 184–194.

Benson-Rea, M., & Rawlinson, S. (2003). Highly skilled and business migrants: Information processes and settlement outcomes. *International Migration*, *41*(2), 59–79.

Bird, M., & Wennberg, K. (2016). Why family matters: The impact of family resources on immigrant entrepreneurs' exit from entrepreneurship. *Journal of Business Venturing*, *31*(6), 687–704.

Chen, W., & Wellman, B. (2009). The global digital divide—within and between countries: The role of information and communication technologies in a global hierarchy. *Asian Journal of Communication*, *19*(2), 164–182.

Christensen, M. (2012). Online mediations in transnational spaces: Cosmopolitan (re)formations of belonging and identity in the Turkish diaspora. *Ethnic and Racial Studies*, *35*(5), 888–905.

Clayton, J., & McDonald, S. J. (2013). The limits of technology. *Information, Communication & Society*, *16*(6), 945–966.

Colliver, B., Coyle, A., & Silvestri, M. (2019). The online othering of transgender people in relation to 'gender neutral toilets'. In *Online othering: Exploring digital violence and discrimination on the web* (pp. 215–237). Springer International Publishing.

Gray, M. L. (2009). *Out in the country: Youth, media, and queer visibility in rural America*. New York University Press.

Guercini, S., Milanesi, M., & Ottati, G. D. (2017). Paths of evolution for the Chinese migrant entrepreneurship: A multiple case analysis in Italy. *Journal of International Entrepreneurship*, *15*(3), 266–294.

Håkansson, H., Ford, D., Gadde, L. E., Snehota, I., & Waluszewski, A. (2009). *Business in networks*. Wiley.

Hong, Y., Hu, Y., & Burtch, G. (2018). Embeddedness, prosociality, and social influence. *MIS Quarterly*, *42*(4), 1211–A4.

Johanson, J., & Mattsson, L. G. (2013). Internationalisation in industrial systems: A network approach. In *Strategies in global competition (RLE international business)* (pp. 287–314). Routledge.

Karadal, H., Shneikat, B. H. T., Abubakar, A. M., & Bhatti, O. K. (2021). Immigrant entrepreneurship: The case of Turkish entrepreneurs in the United States. *Journal of the Knowledge Economy*, *12*(4), 1574–1593.

Lassalle, P., & Shaw, E. (2021). Trailing wives and constrained agency among women migrant entrepreneurs: An intersectional perspective. *Entrepreneurship Theory and Practice*, *45*(6), 1496–1521.

Lin, X., & Zhou, M. (2022). Chinese entrepreneurship in a globalized world: Place, space, and mobilities. *Journal of Small Business & Entrepreneurship*, *34*(4), 357–362.

McAuliffe, M., & Jayasuriya, D. (2016). Do asylum seekers and refugees choose destination countries? Evidence from large-scale surveys in Australia, Afghanistan, Bangladesh, Pakistan and Sri Lanka. *International Migration*, *54*(4), 44–59.

Miellet, S. (2021). From refugee to resident in the digital age: Refugees' strategies for navigating in and negotiating beyond uncertainty during reception and settlement in the Netherlands. *Journal of Refugee Studies*, *34*(4), 3629–3646.

Missbach, A. (2017). Accommodating asylum seekers and refugees in Indonesia: From immigration detention to containment in alternatives to detention. *Refuge*, *33*(2), 32–44.

Ozduzen, O., Korkut, U., & Ozduzen, C. (2021). Refugees are not welcome: Digital racism, online place-making and the evolving categorization of Syrians in Turkey. *New Media & Society*, *23*(11), 3349–3369.

Paloma, V., Escobar-Ballesta, M., Galván-Vega, B., Díaz-Bautista, J. D., & Benítez, I. (2021). Determinants of life satisfaction of economic migrants coming from developing countries to countries with very high human development: A systematic review. *Applied Research in Quality of Life*, *16*(1), 435–455.

Ramsay, J. E., & Pang, J. S. (2017). Anti-immigrant prejudice in rising East Asia: A stereotype content and integrated threat analysis. *Political Psychology*, *38*(2), 227–244.

Salaff, J., Greve, A., & Wong, S. L. (2006). Business social networks and immigrant entrepreneurs from China. In *Chinese ethnic economy: Global and local perspectives* (pp. 99–119). Routledge.

Sanders, J. M., & Nee, V. (1996). Immigrant self-employment: The family as social capital and the value of human capital. *American Sociological Review*, *61*(2), 231–249.

Sequeira, J. M., Carr, J. C., & Rasheed, A. A. (2009). Transnational entrepreneurship: Determinants of firm type and owner attributions of success. *Entrepreneurship Theory and Practice*, *33*(5), 1023–1044.

Selwyn, N. (2004). Reconsidering political and popular understandings of the digital divide. *New Media & Society*, *6*(3), 341–362.

Smans, M., Freeman, S., & Thomas, J. (2014). Immigrant entrepreneurs: The identification of foreign market opportunities. *International Migration*, *52*(4), 144–156.

Stephens, S. (2013). Building an entrepreneurial network: The experiences of immigrant entrepreneurs. *Journal of Enterprising Communities: People and Places in the Global Economy*, *7*(3), 233–244.

Teresita, A. M., & Carmelea, T. V. (2019). The rise of China, new immigrants and changing policies on Chinese overseas. *Southeast Asian Affairs*, 275–294.

Wang, Y., & Warn, J. (2018). Chinese immigrant entrepreneurship: Embeddedness and the interaction of resources with the wider social and economic context. *International Small Business Journal*, *36*(2), 131–148.

Webster, N. A., & Kontkanen, Y. (2021). Space and place in immigrant entrepreneurship literature in the Nordic countries: A systematic literature review. *Norsk Geografisk Tidsskrift-Norwegian Journal of Geography*, *75*(4), 221–236.

Zhou, M. (2004). Revisiting ethnic entrepreneurship: Convergencies, controversies, and conceptual advancements. *International Migration Review*, *38*(3), 1040–1074.

Zolin, R., & Schlosser, F. (2013). Characteristics of immigrant entrepreneurs and their involvement in international new ventures. *Thunderbird International Business Review*, *55*(3), 271–284.

Research Paper

Cross-Cultural Value Alignment Frameworks for Responsible AI Governance:

Evidence from China-West Comparative Analysis

LIU Haijiang[1]
GU Jinguang[1]
WU Xun[2]
Daniel HERSHCOVICH[3]
XIAO Qiaoling[4]

[1]Wuhan University of Science and Technology, China
[2]The Hong Kong University of Science and Technology (Guangzhou), China
[3]University of Copenhagen, Denmark
[4]WUST-Madrid Complutense Institute, China

Abstract

As Large Language Models (LLMs) increasingly influence high-stakes decision-making across global contexts, ensuring their alignment with diverse cultural values has become a critical governance challenge. This study presents a Multi-Layered Auditing Platform for Responsible AI that systematically evaluates cross-cultural value alignment in China-origin and Western-origin LLMs through four integrated methodologies: Ethical Dilemma Corpus for assessing temporal stability, Diversity-Enhanced Framework (DEF) for quantifying cultural fidelity, First-Token Probability Alignment for distributional accuracy, and Multi-stAge Reasoning frameworK (MARK) for interpretable decision-making. Our comparative analysis of 20+ leading models, such as Qwen, GPT-4o, Claude, LLaMA, and DeepSeek, reveals universal challenges – fundamental instability in value systems, systematic underrepresentation of younger demographics, and non-linear relationships between model scale and alignment quality – alongside divergent regional development trajectories. While China-origin models increasingly emphasise multilingual data integration for context-specific optimisation, Western models demonstrate greater architectural experimentation but persistent U.S.-centric biases. Neither paradigm achieves robust cross-cultural generalisation. We establish that Mistral-series architectures significantly outperform LLaMA-3-series in cross-cultural alignment, and that Full-Parameter Fine-Tuning on diverse datasets surpasses Reinforcement Learning from Human Feedback in preserving cultural variation. These findings provide empirical foundations for evidence-based AI governance, offering actionable protocols for model selection, bias mitigation, and policy consultation at scale, while

demonstrating that current LLMs require sustained human oversight in ethical decision-making and cannot yet autonomously navigate complex moral dilemmas across cultural contexts.

Introduction

The integration of Large Language Models (LLMs) into high-stakes applications, such as decision-support systems, requires a thorough evaluation of their moral and ethical reasoning capabilities. For AI agents to operate trustworthily, their behaviour must align with human values, which often vary significantly across different contexts and cultures (Liscio et al., 2023). Evaluating cultural value alignment is therefore a high priority within natural language processing and AI ethics research (Jobin et al., 2019).

Early work on machine ethics has focused on complex ethical dilemmas, which necessitate choices between two “right” options involving conflicting moral values (Kidder, 1996). Comprehensive evaluations have revealed that both China and the West (e.g., Qwen2-72B and Claude-3.5-Sonnet) LLMs exhibit definitive preferences between major conflicting value pairs (Yuan et al., 2024). Larger and more advanced LLMs tend to support a deontological perspective, maintaining their choices even when negative consequences are specified. However, LLMs often struggle with understanding the core decision-making task, demonstrating pronounced sensitivity to how dilemmas are formulated in prompts (Sclar et al., 2024; Yuan et al., 2024). Moreover, these models show significant cultural and linguistic biases, with moral reasoning varying substantially depending on the language used for prompting (Agarwal et al., 2024a).

These findings establish a critical governance constraint (see Figure 1): solely relying on LLMs to resolve contentious ethical issues on their own is neither safe nor desirable, nor is it ethical in critical applications, given the models’ rigidity and tendency to favour certain values (Mittelstadt, 2019). This volatility, combined with observed rigidity, underscores that simply providing instructions is insufficient for ensuring Responsible AI (RAI) deployment. A shift toward a dynamic, systematic auditing platform is necessary to measure LLM consistency, stability, fidelity, and bias, thereby advancing evaluation beyond static, single-step assessment methods (Scherrer et al., 2023).

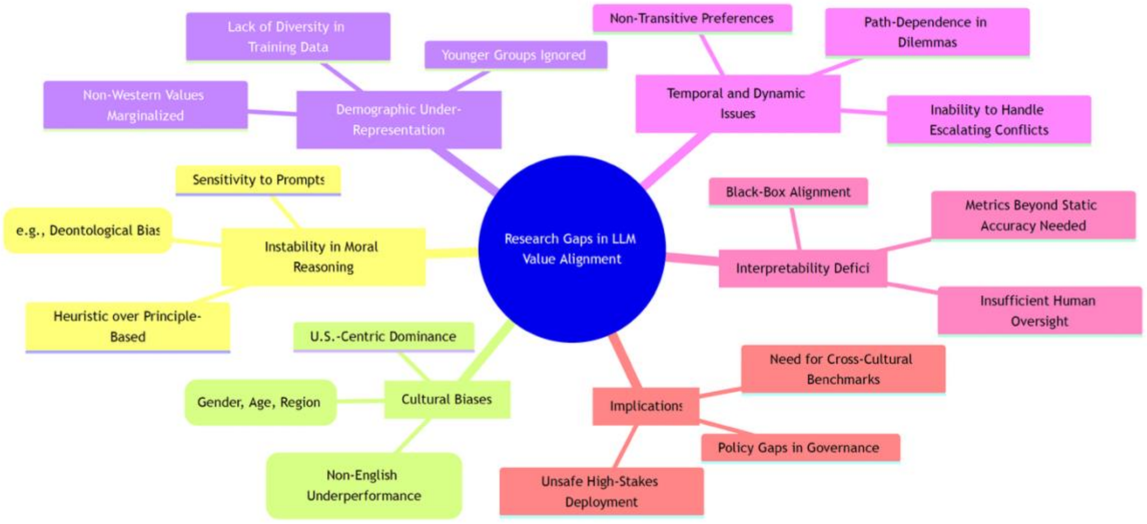

Lack of Diversity in Training Data
Younger Groups Ignored
Non-Western Values Marginalized
Demographic Under-Representation
Non-Transitive Preferences
Path-Dependence in Dilemmas
Temporal and Dynamic Issues
Inability to Handle Escalating Conflicts
Sensitivity to Prompts
e.g., Deontological Bias
Instability in Moral Reasoning
Heuristic over Principle-Based
Research Gaps in LLM Value Alignment
Black-Box Alignment
Metrics Beyond Static Accuracy Needed
Interpretability Defici
Insufficient Human Oversight
U.S.-Centric Dominance
Gender, Age, Region
Cultural Biases
Non-English Underperformance
Need for Cross-Cultural Benchmarks
Implications
Policy Gaps in Governance
Unsafe High-Stakes Deployment

Figure 1. Key Research Gaps in Cross-Cultural Value Alignment for LLMs

This includes instability in moral reasoning (e.g., sensitivity to prompts and consequences), cultural biases (e.g., U.S.-centric or English-dominant preferences), underrepresentation of diverse demographics (e.g., younger groups or non-Western values), lack of temporal stability in ethical decisions, and insufficient interpretability in alignment methods.

To address these issues, this paper systematically presents a **Multi-Layered Auditing Platform for Responsible AI**, composed of four integrated computational tools designed to assess LLM consistency, stability, fidelity, and interpretability in cross-cultural contexts:

(1) Ethical Dilemma Corpus: moving beyond single-step assessments, the platform first diagnoses reliance on unstable heuristics and reveals value misalignment.

(2) Diversity-Enhanced Framework (DEF): utilised to overcome repetitive LLM output and generate accurate preference distributions reflecting the diversity and uncertainty inherent in survey responses.

(3) First-Token Probability Alignment technique: fine-tunes LLMs to minimise divergence between predicted and actual human value distributions.

(4) Multi-stAge Reasoning frameworK (MARK): enhances accountability and interpretability by simulating human decisions through personality-driven cognitive processes, integrating stress analysis and multi-stAge reasoning to provide psychologically grounded explanations.

The framework systematically integrates these methodologies to create a multi-layered platform capable of generating nuanced, cross-cultural insights that serve both academic transparency and strategic industrial value:

- Technology for Good (Social Impact): The methodologies provide safety guidance and protocols for investigating complex moral behaviours in LLMs. By enabling accurate simulation of group-level survey responses – a capability often limited by high costs and time-intensive human studies – these tools can accelerate social science research and aid in more informed policy decisions, particularly in navigating complex ethical choices

(Argyle et al., 2022; Bail, 2024).

- Industrial Value (LLM Governance): The platform enables competitive strategic decisions by benchmarking LLMs across critical RAI dimensions, moving beyond generic accuracy metrics. This systematic audit identifies optimal model architectures and verifies the effectiveness of alignment techniques, providing a blueprint for building culturally sensitive and compliant global AI products (Hagendorff, 2020).

This paper thus aims to bridge the gap between technical innovation and social/industrial utility by systematically identifying robust value alignment mechanisms for China-Western models.

Related Works

AI Ethics and Computational Morality

Moral Beliefs in LLMs

Scherrer et al. (2023); Durmus et al. (2024) introduced a statistical method for eliciting beliefs encoded in LLMs through surveys of moral scenarios, distinguishing between high and low-ambiguity cases based on common morality rules. Their large-scale survey comprised 680 high-ambiguity and 687 low-ambiguity moral scenarios administered to 28 open- and closed-source LLMs, revealing that most models exhibit low uncertainty in unambiguous scenarios while expressing uncertainty in ambiguous cases.

Moral Foundations Theory

Recent work has applied Moral Foundations Theory (MFT) to measure how well LLMs represent different political orientations and ethical perspectives (Dev et al., 2022; Narayanan & Samuel, 2025; Raza et al., 2024). Studies employing experimental psychology methods to probe LLMs' moral and legal reasoning have found that alignment with human responses varies significantly across different experiments and models.

Ethical Dilemmas - Right vs. Right

Yuan et al. (2024) conducted a comprehensive evaluation using Kidder's framework to examine how LLMs navigate ethical dilemmas involving Truth vs. Loyalty, Individual vs. Community, Short-Term vs. Long-Term, and Justice vs. Mercy. Their study revealed that LLMs exhibit pronounced preferences, prioritising truth over loyalty in 93.48% of cases, long-term over short-term considerations in 83.69% of cases, and community over individual in 72.37% of cases. This work demonstrated that larger LLMs tend to support a deontological perspective, maintaining their choices even when negative consequences are specified.

Classical Ethical Theories

Agarwal et al. (2024b) examined ethical reasoning across different languages, finding that GPT-4 is the most consistent ethical reasoner across languages, while other models show significant moral value bias when prompted in languages other than English. Their experiments covered deontology, virtue ethics, and consequentialism across six languages (English, Spanish, Russian, Chinese, Hindi, and Swahili). The AMULED framework translates utilitarianism, deontology, virtue ethics, and other moral philosophies into reward functions for reinforcement learning to address moral uncertainty in AI decision-making (Dubey et al., 2025).

LLM Value Alignment and Governance

Moral Foundations Theory and Schwartz's Values

Hadar-Shoval et al. (2024) applied Schwartz's Theory of Basic Values (STBV) to measure value-like constructs within LLMs, finding substantial divergence between LLM value profiles and population data. All models prioritised universalism and self-direction while de-emphasising achievement, power, and security relative to humans. Their study using the Portrait Values Questionnaire-Revised (PVQ-RR) showed that these biased value profiles strongly predicted LLMs' responses when presented with mental health dilemmas requiring choosing between opposing values.

Recent work combining MFT and Schwartz's theory in multi-step moral dilemmas revealed that LLMs exhibit non-transitive and shifting moral preferences (Wu et al., 2025). The study found that intuitive preferences like care decrease while fundamental values like fairness become more prominent as dilemmas progress, demonstrating that LLMs maintain value orientations while flexibly adjusting preference strengths across sequential dilemmas.

Value Alignment Methodologies

Two primary approaches exist: data-driven bottom-up alignment, where LLM performance is evaluated by comparing moral judgment to human judgment using datasets like SOCIALCHEM101, Moral Stories, ETHICS, NormBank, and MoralChoice (Jiang et al., 2022; Emelin et al., 2021; Hendrycks et al., 2021); and top-down alignment, measuring how well LLMs infer specified value preferences through prompt embedding. Yuan et al. (2024) showed that explicit guidelines are more effective in guiding LLMs' moral choices than in-context examples.

Dynamic and Temporal Moral Reasoning

Wu et al. (2025) introduced the Multi-step Moral Dilemmas (MMDs) framework, a path-dependent evaluation approach that captures temporal dynamics of moral judgment, addressing limitations of static single-step assessment methods. This research across 3,302 five-stage dilemmas revealed that LLMs maintain value orientations while flexibly adjusting preference strengths across sequential dilemmas, with value preferences shifting significantly as dilemmas progress.

RLHF vs. Fine-Tuning

Reinforcement Learning from Human Feedback (Ouyang et al., 2022, RLHF) has become central in adapting AI models to human-centric expectations, particularly in the final stages of fine-tuning state-of-the-art models, though alignment can be biased by the group of humans providing feedback and may never satisfy everyone's preferences simultaneously (Christiano et al., 2023; Cao et al., 2025). Studies on cultural value alignment found that Mistral-series models demonstrate superior performance compared to Llama-3-series models in cross-cultural value alignment contexts (Liu et al., 2025a).

LLM Simulation in Social Science and Psychology

LLM Simulations of Human Behaviour

Researchers have pioneered specialising LLMs for simulating group-level survey response distributions for global populations, using fine-tuning methods based on first-token probabilities to minimise divergence between predicted and actual response distributions (Argyle et al., 2023). The SocSci210 dataset, comprising 2.9 million responses from 400,491 participants across 210 social science experiments, has enabled LLMs to produce predictions 26% more aligned with human response distributions.

Distribution Prediction and Calibration

The task of simulating survey response distributions can be viewed as a calibration problem, aligning classifier predictive probabilities with classification uncertainty. Uncertainty quantification methods have been developed to convert simulated LLM responses into valid confidence sets for population parameters, addressing distribution shift between simulated and real populations (Santurkar et al., 2023).

Psychologically Grounded Simulation

The PsyDI framework incorporates MBTI psychological theory to model decision-making processes through progressively in-depth dialogue, leveraging cognitive functions to provide more accurate personality measurements. Fine-tuning approaches using sociodemographic prompting and persona strategies have been shown to enhance personalisation in survey simulations (Park et al., 2023; Liu et al., 2025b).

Cross-cultural Alignment, Bias, and Evaluation Metrics

Cultural Alignment and Cross-Cultural NLP

Evaluations using the World Values Survey have revealed that all major LLMs exhibit cultural values resembling English-speaking and Protestant European countries (Tao et al., 2024). Cultural prompting improved alignment for 71-81% of countries in later models like GPT-4o. Comprehensive evaluations across 10 models, 20 countries, and languages using Hofstede's Value Survey Module have identified systematic cultural biases favouring highly represented languages and regions (Cao et al., 2023).

Evaluation of Cultural Biases and Stereotypes

Research has shown that LLMs are 3-6 times more likely to choose occupations that stereotypically align with a person's gender, and these choices align more with people's perceptions than with ground truth job statistics (Kotek et al.). Studies using LLM Word Association Tests have found pervasive stereotype biases across 8 models in 4 social categories (race, gender, religion, health) covering 21 stereotypes, demonstrating that value-aligned models still harbour implicit biases (Taubenfeld et al., 2024).

Cultural Datasets and Benchmarks

The Integrated Values Surveys (IVS), combining the World Values Survey and European Values Study, provide an established measure of cultural values for 112 countries and territories. The SOCIALCHEM101 dataset contains 292,000 sentences representing rules of thumb for evaluating LLMs' ability to reason about social and moral norms (Forbes et al., 2020), while the Moral Integrity Corpus provides 38,000 prompt-reply pairs with 99,000 rules of thumb and 114,000 annotations.

Consistency and Bias Measurement

Comprehensive taxonomies organise bias evaluation by metrics operating at different levels (embeddings, probabilities, generated text) and datasets by their structure, with techniques including counterfactual testing, stereotype detection, and sentiment analysis (Navigli et al., 2023). Diversity-Enhanced Frameworks have been developed to measure cultural value misalignment through multi-aspect evaluation metrics, revealing notable concerns regarding the lack of cross-cultural representation and preference biases related to gender and age.

Methodologies: A Computational Auditing Platform for Responsible AI

In this section, we establish a systematic framework **– a Multi-Layered Auditing Platform for Responsible AI** (see Figure 2) – which synthesises and organises existing computational tools derived from team member publications and relevant design efforts. This integrated approach moves beyond conventional, static evaluations of Large Language Models (LLMs) by addressing three fundamental research questions that guide our comparative analysis of China-Western LLMs in responsible AI governance and cross-cultural alignment.

Our framework systematically examines: (1) whether LLM responses to ethical dilemmas stem from stable value principles or contextual heuristics, (2) what quantifiable cultural values these models actually embody when compared to human populations, and (3) how current technological innovations can effectively contribute to genuine value alignment. Through this structured inquiry, we address the inherent rigidity and fixed value preferences (such as prioritising Truth over Loyalty or adopting a deontological perspective) previously observed in LLMs when navigating ethical dilemmas. The core objective is to **audit LLM behaviours against specific Chinese and Western cultural human value distributions** to ensure genuine alignment, fairness, and safety in deployment.

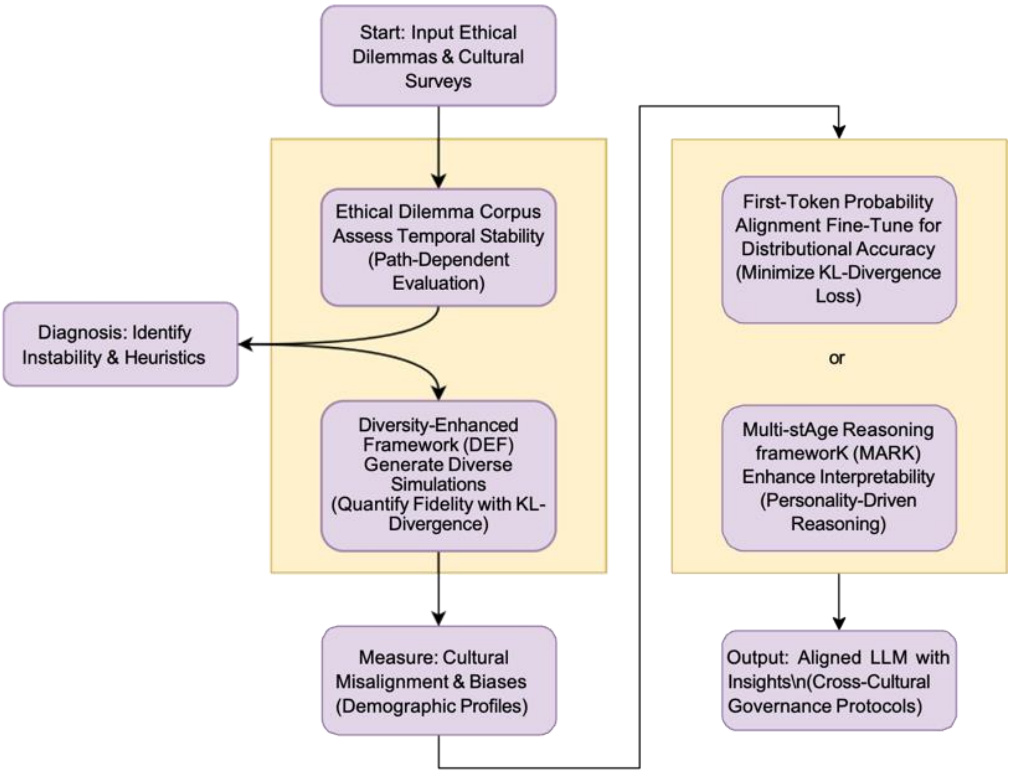

Start: Input Ethical Dilemmas & Cultural Surveys
Ethical Dilemma Corpus Assess Temporal Stability (Path-Dependent Evaluation)
Diagnosis: Identify Instability & Heuristics
Diversity-Enhanced Framework (DEF) Generate Diverse Simulations (Quantify Fidelity with KL-Divergence)
Measure: Cultural Misalignment & Biases (Demographic Profiles)
First-Token Probability Alignment Fine-Tune for Distributional Accuracy (Minimize KL-Divergence Loss)
or
Multi-stAge Reasoning frameworK (MARK) Enhance Interpretability (Personality-Driven Reasoning)
Output: Aligned LLM with Insights\n(Cross-Cultural Governance Protocols)

Figure 2. Multi-Layered Auditing Platform for Responsible AI
Integrating four methodologies Ethical Dilemma Corpus (for temporal stability), Diversity-Enhanced Framework (DEF, for cultural fidelity), First-Token Probability Alignment (for distributional accuracy), and Multi-stAge Reasoning frameworK (MARK, for interpretable decision-making).

RQ1: Is Model Response to Ethical Dilemmas Rooted in Stable Value Principles?

To investigate whether LLMs possess stable value principles or merely employ context-driven heuristics, we deploy the **Ethical Dilemma Corpus** (Yuan et al., 2024) framework to audit the critical dimension of **Temporal Stability** (path-dependence) in LLM moral choices (see Figure 3).

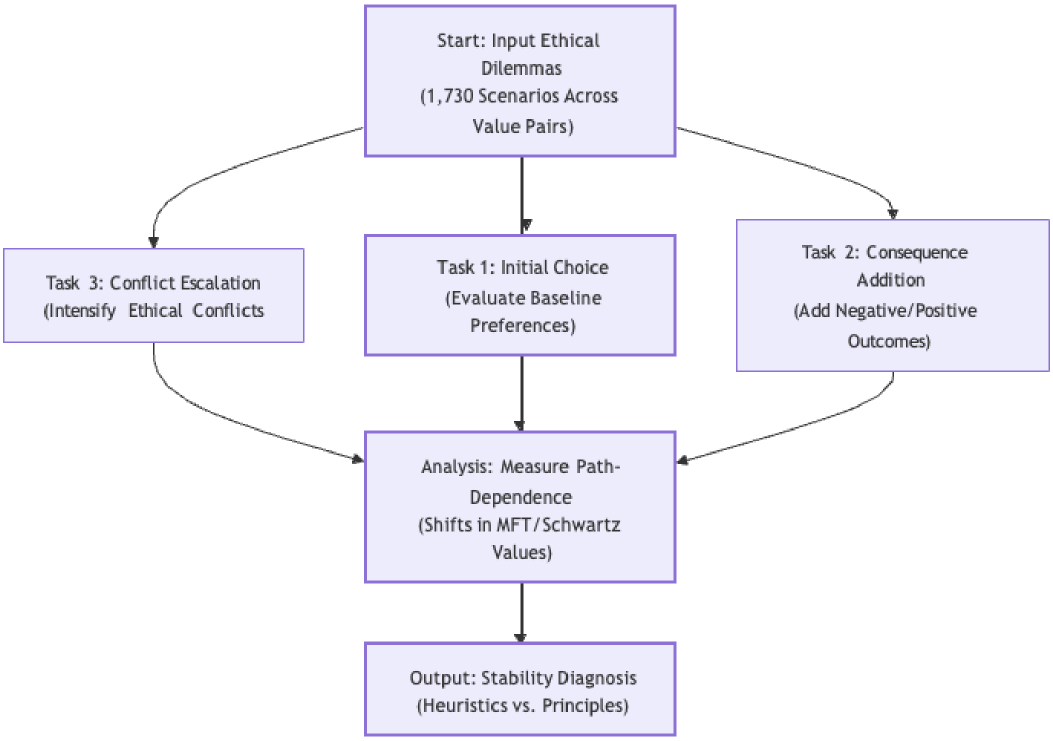

Start: Input Ethical Dilemmas (1,730 Scenarios Across Value Pairs)
Task 3: Conflict Escalation (Intensify Ethical Conflicts
Task 1: Initial Choice (Evaluate Baseline Preferences)
Task 2: Consequence Addition (Add Negative/Positive Outcomes)
Analysis: Measure Path-Dependence (Shifts in MFT/Schwartz Values)
Output: Stability Diagnosis (Heuristics vs. Principles)

Figure 3. Pipeline for Ethical Dilemma Corpus
This component assesses temporal stability in LLM moral choices through path-dependent ethical dilemmas.

Existing assessments primarily rely on single-step evaluations, which fail to capture how models adapt to evolving ethical challenges. The dataset addresses this gap by featuring 1,730 scenarios that progressively intensify ethical conflicts across three sequential tasks. This path-dependent evaluation framework enables a dynamic analysis of how LLMs adjust their moral reasoning when faced with escalating dilemmas.

The framework is applied to compare LLMs from both regions (including DeepSeek, LLaMA-3-70B, GPT-4o, and GLM-4) in terms of dynamic shifts in values (using Moral Foundations Theory or Schwartz's Theory of Basic Values) under conflict escalation.

RQ2: What Quantifiable Cultural Values Do LLMs Obtain?

To audit Representativeness and **Cultural Fidelity** against external cultural benchmarks and quantify the cultural values embedded in LLMs, we utilise the **Diversity-Enhanced Framework** (Liu et al., 2025; DEF; see Figure 4). DEF systematically generates diverse virtual participants for high-fidelity simulation of human value survey data (e.g., the World Values Survey) across U.S. and Chinese cultures.

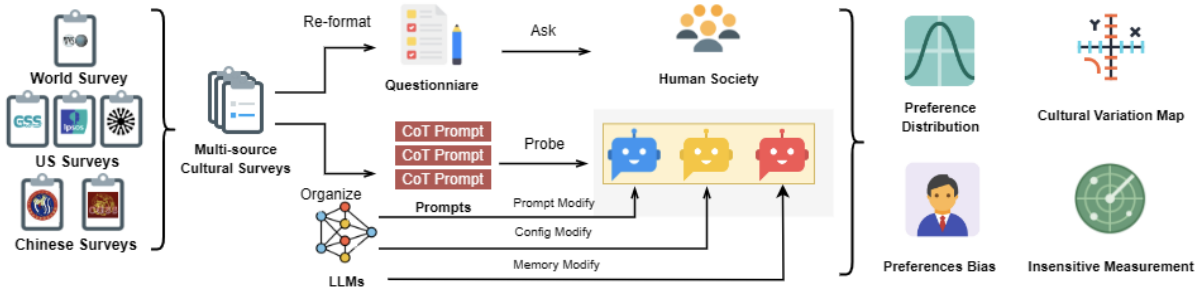


Figure 4. Pipeline for Diversity-Enhanced Framework (DEF)
DEF generates diverse simulations to quantify cultural fidelity against benchmarks like the World Values Survey. The pipeline focuses on overcoming repetitive outputs. Adapted from Liu et al. (2025a).

This framework is necessary because traditional LLM probing techniques, particularly those using Chain-of-Thought (CoT) instructions, often produce repetitive responses that fail to capture the necessary complexity and variability needed to generate accurate preference distributions. DEF captures the diversity and uncertainty inherent in LLM behaviours through realistic survey experiments by implementing:

- Prompt modifications (e.g., randomly selected locations)
- Configuration modifications (e.g., dynamically tuning generation parameters like num beams)
- Memory manipulation (simulating memory effects and cleaning invalid responses)

The comparative assessment focuses on quantifying misalignment using metrics such as the Kullback-Leibler Divergence (KL-D) for Preference Distribution (to measure East-West divergence) and analysing Preference Bias (demographic profiles) to expose fairness risks specific to the U.S. and Chinese contexts.

RQ3: How Do Current Technological Innovations Contribute to Value Alignment?

Having identified instability in value principles and measured cultural misalignment, we present two technological innovations that enhance value alignment in LLMs.

First-Token Probability Alignment for Distributional Accuracy

To transform the initial LLM response distributions toward high-fidelity representations of human data, the platform incorporates a specialised First-Token Probability Alignment (Cao et al., 2025) technique. This Alignment Engineering solution fine-tunes models to minimise divergence against ground-truth human distributional preferences derived from surveys.

We devised this fine-tuning method based on first-token probabilities, where the objective is to minimise Kullback-Leibler Divergence loss between the predicted LLM distribution and the actual country-level human response distribution.

Multi-stAge Reasoning frameworK (MARK) for Interpretable Alignment

The Multi-stAge Reasoning frameworK (Liu et al., 2025b, MARK, see Figure 5) serves as a cognitive augmentation tool designed to enhance Accountability and Interpretability within the platform. MARK addresses the black-box nature of value alignment by simulating human decisions through integration of psychological theory – specifically the type dynamics theory in the MBTI psychological framework – to model cognitive processes and their interactions with stress and personal experience.

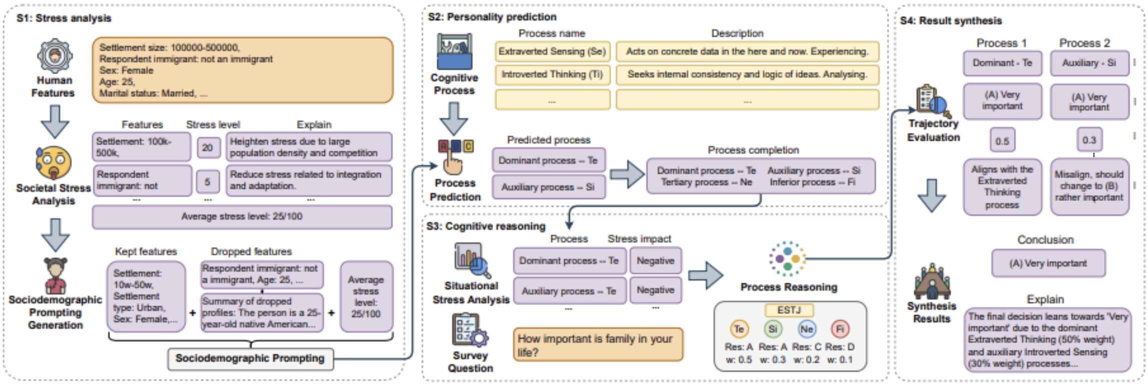

S1: Stress analysis
Human Features
Settlement size: 100000-500000,
Respondent immigrant: not an immigrant
Sex: Female
Age: 25,
Marital status: Married, ...
Societal Stress Analysis
Features
Stress level
Explain
Settlement: 100k-500k,
20
Heighten stress due to large population density and competition
Respondent immigrant: not
5
Reduce stress related to integration and adaptation.
Average stress level: 25/100
Sociodemographic Prompting Generation
Kept features
Dropped features
Settlement: 10w-50w, Settlement type: Urban, Sex: Female,...
Respondent immigrant: not a immigrant, Age: 25, ...
Summary of dropped profiles: The person is a 25-year-old native American...
Average stress level: 25/100
Sociodemographic Prompting
S2: Personality prediction
Cognitive Process
Process name
Description
Extraverted Sensing (Se)
Acts on concrete data in the here and now. Experiencing.
Introverted Thinking (Ti)
Seeks internal consistency and logic of ideas. Analysing.
Process Prediction
Predicted process
Dominant process -- Te
Auxiliary process -- Si
Process completion
Dominant process -- Te
Auxiliary process -- Si
Tertiary process -- Ne
Inferior process -- Fi
S3: Cognitive reasoning
Situational Stress Analysis
Process
Stress impact
Dominant process -- Te
Negative
Auxiliary process -- Te
Negative
Survey Question
How important is family in your life?
Process Reasoning
ESTJ
Te
Si
Ne
Fi
Res: A w: 0.5
Res: A w: 0.3
Res: C w: 0.2
Res: D w: 0.1
S4: Result synthesis
Trajectory Evaluation
Process 1
Process 2
Dominant - Te
Auxiliary - Si
(A) Very important
(A) Very important
0.5
0.3
Aligns with the Extraverted Thinking process
Misalign, should change to (B) rather important
Synthesis Results
Conclusion
(A) Very important
Explain
The final decision leans towards 'Very important' due to the dominant Extraverted Thinking (50% weight) and auxiliary Introverted Sensing (30% weight) processes...

Figure 5. Pipeline for Multi-stage Reasoning Framework (MARK)
MARK enhances interpretability by simulating personality-driven reasoning based on MBTI theory. Adapted from Liu et al. (2025b).

The framework systematically models human reasoning through multi-stage processing, including stress analysis, personality prediction, cognitive reasoning, and synthesis. This detailed simulation achieves high accuracy compared to other simulation baselines and demonstrates robust generalisation using both U.S. and Chinese survey data. MARK provides mechanistic, personality-driven explanations for simulated value choices, enabling human experts to validate or challenge the model's reasoning trajectory. This computational interpretability is essential for building trust and ensuring accountability in cross-cultural AI applications.

Together, these components form a comprehensive auditing platform that not only diagnoses problems in current LLM value alignment (through Ethical Dilemma Corpus and DEF) but also provides actionable technological solutions (through First-Token Alignment and MARK). This integrated approach enables systematic comparison between Chinese and Western LLMs while advancing the technical capabilities needed for responsible cross-cultural AI governance.

Cross-Cultural Audit Findings: Comparative Performance and Temporal Patterns in China-Western LLMs

This section systematically presents empirical findings derived from the proposed Multi-Layered Auditing Platform (summarised in Figure 6), structured according to the three core research questions. The analysis emphasises comparative performance dynamics between China-origin and Western-origin Large Language Models (LLMs), examining **cross-regional convergences and divergences** in stability, cultural fidelity, and alignment efficacy. Additionally, we investigate **temporal developmental patterns** (summarised in Figure 7) to assess whether technological advancement trajectories yield consistent improvements in value alignment across both regional contexts.

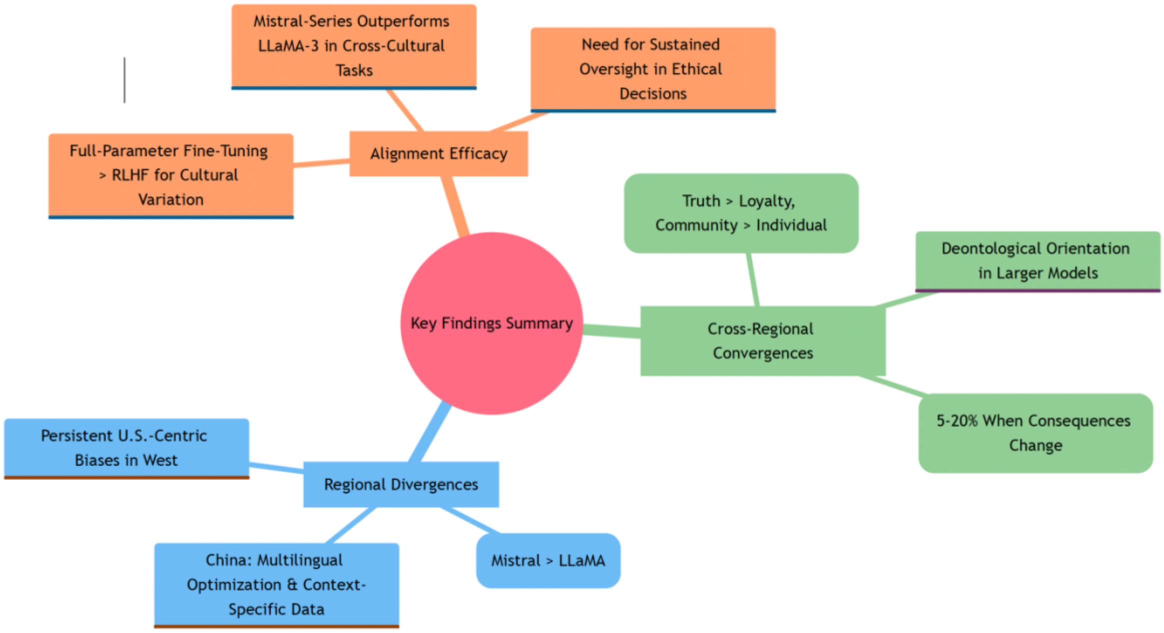

Mistral-Series Outperforms LLaMA-3 in Cross-Cultural Tasks
Need for Sustained Oversight in Ethical Decisions
Full-Parameter Fine-Tuning > RLHF for Cultural Variation
Alignment Efficacy
Truth > Loyalty, Community > Individual
Deontological Orientation in Larger Models
Key Findings Summary
Cross-Regional Convergences
5-20% When Consequences Change
Persistent U.S.-Centric Biases in West
Regional Divergences
China: Multilingual Optimization & Context-Specific Data
Mistral > LLaMA

Figure 6. Summary of Key Empirical Findings

A mind map overview of the cross-cultural audit findings, highlighting convergences (e.g., deontological trends) and divergences (e.g., optimisation strategies).

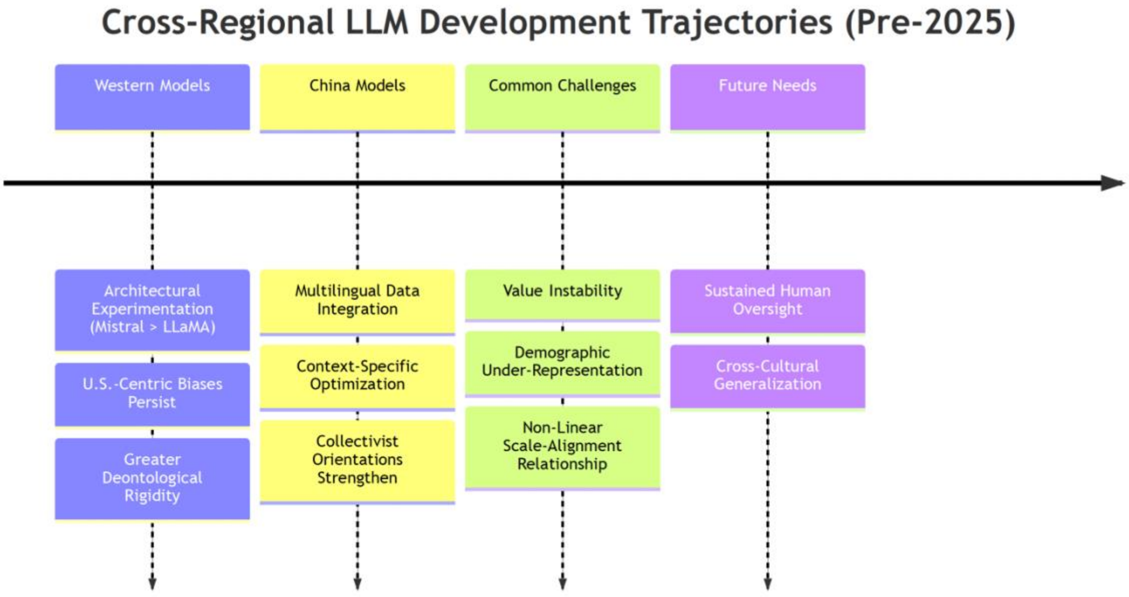

Cross-Regional LLM Development Trajectories (Pre-2025)
Western Models
China Models
Common Challenges
Future Needs
Architectural Experimentation (Mistral > LLaMA)
U.S.-Centric Biases Persist
Greater Deontological Rigidity
Multilingual Data Integration
Context-Specific Optimization
Collectivist Orientations Strengthen
Value Instability
Demographic Under-Representation
Non-Linear Scale-Alignment Relationship
Sustained Human Oversight
Cross-Cultural Generalization

Figure 7. Divergent China-West Trajectories

China focusing on multilingual data integration, West on architectural experimentation, but with persistent biases.

Temporal Stability of Moral Reasoning (RQ1): Evaluating Value Principle Consistency in Ethical Decision-Making

This subsection addresses RQ1 through the Ethical Dilemma Corpus (Yuan et al., 2024), examining whether LLM responses to ethical dilemmas demonstrate stable, principle-based reasoning or context-dependent heuristic patterns. The analysis reveals both cross-regional architectural commonalities and region-specific developmental trajectories.

Cross-Regional Convergence: Universal Value Rigidity and Deontological Orientation

Comparative Baseline Performance

As presented in Figure 8, LLMs from both China and Western development ecosystems exhibit remarkably consistent preferential hierarchies when navigating "Right vs. Right" ethical dilemmas involving conflicting moral values, with pronounced preferences for Truth over Loyalty, Long-term over Short-term considerations, and Community over Individual interests, suggesting convergent training paradigms despite distinct cultural contexts and data sources. The Individual-Community value pair demonstrates the highest rigidity across all models, indicating deeply embedded collectivist orientations that transcend regional boundaries.

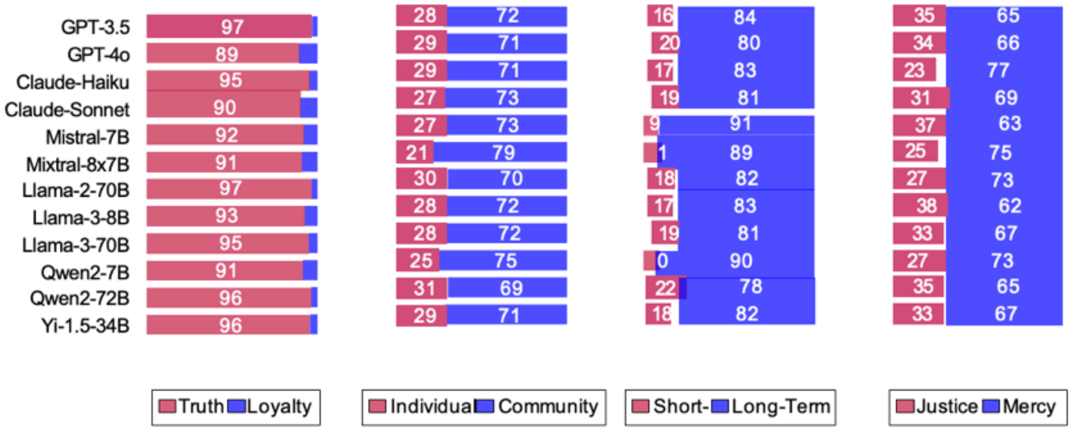


Figure 8. Moral Value Preference Query Results for Each Conflicting Value Pair
For instance, an LLM switches from "Action A" to "Action B" when a negative outcome is added to "Action A" and a positive outcome is added to "Action B".

Source: Reprinted from Jiaqing Yuan, et al. Right vs. right: Can LLMs make tough choices? CoRR, abs/2412.19926, 2024. doi: 10.48550/ARXIV.2412.1992, Copyright 2024.

Deontological Convergence in Advanced Models

Both China-origin and Western-origin flagship models converge toward deontological ethical frameworks as model capacity increases. These advanced systems maintain initial moral commitments despite explicitly specified negative consequences (see Figure 9), demonstrating outcome-independent reasoning. Advanced models from both regional contexts demonstrate comparably strong commitment stability, with flagship models from both China (e.g., Qwen2-72B) and Western contexts (e.g., GPT-4o, Llama-3-70B) showing similarly low decision reversal rates when confronted with consequence specifications, suggesting that architectural scale rather than regional origin determines deontological adherence.

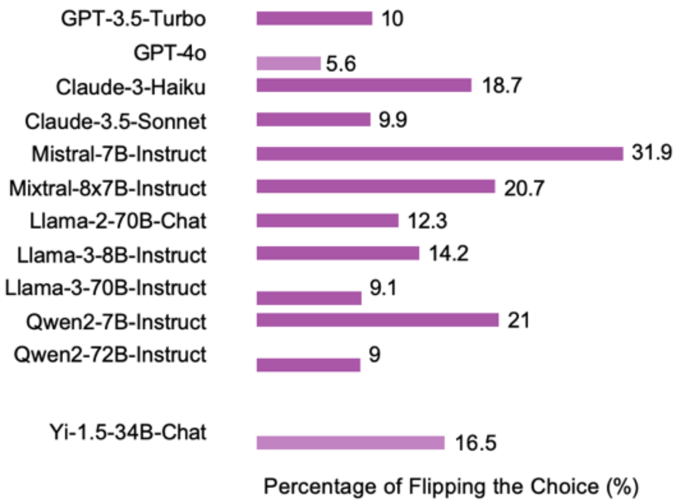


Figure 9. Percentage of Flipping the Choice When Consequences are Altered

Source: Reprinted from Jiaqing Yuan, et al. Right vs. right: Can LLMs make tough choices? CoRR, abs/2412.19926, 2024. doi: 10.48550/ARXIV.2412.1992, Copyright 2024.

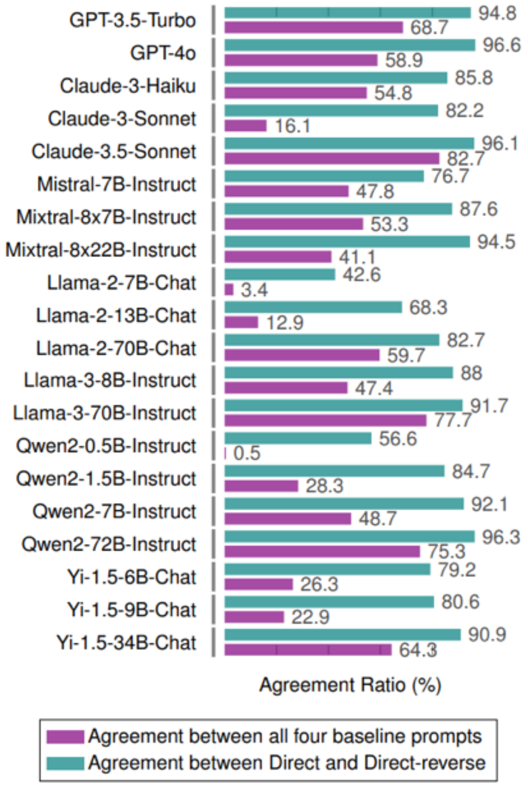


Figure 10. Moral Choice Agreement for Different Prompts

The higher the agreement, the better the task comprehension.

Source: Reprinted from Jiaqing Yuan, et al. Right vs. right: Can llms make tough choices? CoRR, abs/2412.19926, 2024. doi: 10.48550/ARXIV.2412.1992, Copyright 2024, with permission from arXiv.

Regional Divergence: Volatility Patterns and Developmental Trajectories

Temporal Improvements in Task Comprehension

Longitudinal analysis across model generations reveals consistent improvement in agreement ratios (task comprehension metrics) for both regional development tracks. Recent flagship releases from both China-origin models (e.g., Qwen2-72B) and Western-origin models (e.g., Claude-3.5-Sonnet, Llama-3-70B) achieve high agreement ratios across prompt variations, suggesting that increased training compute and data quality, rather than region-specific methodologies, drive comprehension improvements.

Fundamental Instability in Value Systems

Despite surface-level improvements, the core finding transcends regional boundaries: LLM value systems remain fundamentally heuristic and unstable, relying on context-driven statistical pattern matching rather than globally consistent ethical principles. However, volatility patterns manifest distinctly across regional model families:

- *Western models*: Claude series demonstrates exceptional cross-context stability with minimal rank volatility
- *China models*: GLM-4-Air exhibits comparable temporal consistency
- *Volatile Western models*: Llama series shows pronounced rank fluctuations across value dimensions
- *Volatile China models*: DeepSeek series demonstrates similar instability patterns

Sequential Patterns Across Dilemma Escalation

Analysis of the five-step Ethical Dilemma Corpus progression reveals universal and divergent patterns:

Table 1. The Statistical Result of T-test Validation Between Mistral- and Llama-3-series Models

The result suggests that the Mistral-series models generally have significantly better value alignment performance than the Llama-3-series models. *Adapted from Liu et al. (2025a).*

Culture	Model	Mean	SD	SEM	t value	DF	SED	p value
U.S.	Llama-3-series	1.28	0.50	0.10	3.01	52	0.12	0.0041
	Mistral-series	0.93	0.34	0.07				
Chinese	Llama-3-series	1.35	0.74	0.14	2.77	52	0.16	0.0079
	Mistral-series	0.92	0.33	0.06				

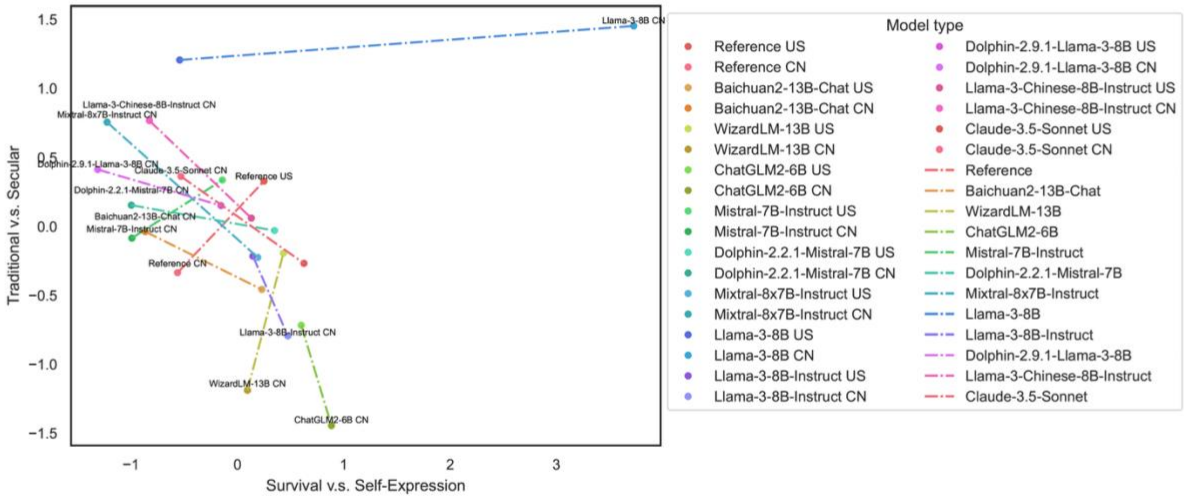


Figure 11. The Cross-Cultural Variation Map of Model Candidates

All models can somehow distinguish the cultural variations. Mistral-7B-Instruct preserves better cross-cultural variations. *Adapted from Liu et al. (2025a).*

- Universal stability anchor: The Care dimension maintains exceptional consistency across all escalation stages for both China and Western models, functioning as a stable moral foundation regardless of regional origin or training paradigm.
- Progressive divergence in Authority: The Authority dimension exhibits systematic degradation in cross-model consistency as dilemma severity increases. This pattern suggests that hierarchical value orientations, potentially influenced by distinct cultural training data, become increasingly model-specific under ethical pressure.

Quantification of Cultural Value Encoding (RQ2): Cross-Regional Distributional Fidelity and Systemic Bias Patterns

This subsection employs the Diversity-Enhanced Framework (DEF) (Liu et al., 2025a) to quantify cross-cultural misalignment and demographic preference bias, providing systematic architectural comparisons between China-origin and Western-origin foundational models.

Architectural Performance: Cross-Regional Model Family Comparisons

Western Architectural Comparison

Statistical analysis (Table 1) of Western foundational architectures reveals significant performance stratification. The Mistral-series demonstrates substantially superior cultural value alignment compared to the Llama-3-series across both U.S. and Chinese cultural contexts ($p = 0.0041$ for U.S.; $p = 0.0079$ for Chinese, Wilcoxon signed-rank test). This finding suggests that the Mistral architecture encodes more **generalisable cultural representation capacities**, potentially attributable to training data diversity or architectural inductive biases that better capture cross-cultural value distributions.

Cultural Affinity Patterns

While all evaluated models successfully distinguish between U.S. and Chinese value distributions on the Cultural Variation Map, systematic biases emerge:

For *Western model biases*: Llama-3-8B-Instruct exhibits pronounced U.S. cultural affinity. Most Western models demonstrate U.S.-centric value alignment, with the notable exception of base Llama-3-8B, which shows reduced directional bias.

For the China model performance: Llama-3-Chinese-8B-Instruct (Western architecture with enhanced Chinese training) demonstrates strong Chinese cultural affinity. ChatGLM2-6B uniquely balances U.S. and Chinese value representations among China-origin models.

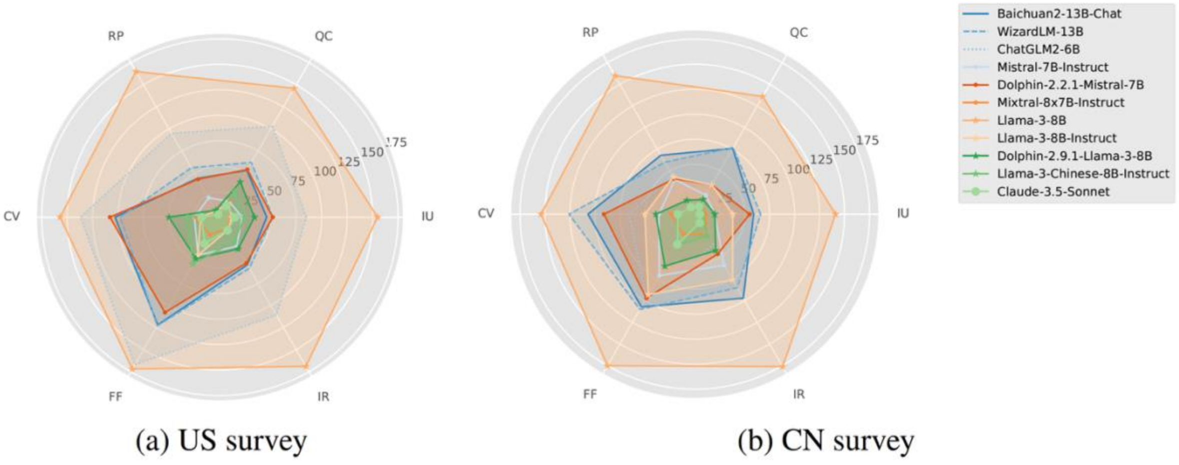
Cross-cultural Preservation

Mistral-7B-Instruct achieves superior preservation of cultural variation across both contexts, outperforming models that achieve high alignment in single-culture scenarios, suggesting distinct optimisation targets in Mistral's training regime.

Systemic Failures and Demographic Bias: China-Western Contrast

Regional Patterns in Systemic Inconsistency

Analysis of insensitivity measurements (see Figure 12) – False Fact Presentation (FF) and Conflict in Value Expression (CV) – reveals consistent challenges across all models, but with significant architectural variation. The Llama-3-series exhibits approximately 2× the failure rate of the Mistral-series across both cultural contexts, indicating fundamental differences in consistency mechanisms independent of regional adaptation efforts.



(a) US survey

(b) CN survey

Figure 12. The Overall Comparison of the Insensitivity Measurement Among Models on Each

Model shows significant issues with larger models having fewer problems on the US survey, and smaller models on the CN survey.

Source: Liu et al., (2025a)

Demographic Preference Bias Divergence

The audit reveals systematically distinct demographic mismatch profiles between cultural contexts (see Appendix).

In the context of the United States, the observed bias pattern in alignment configurations demonstrates a preference for male demographic profiles. Additionally, there is a notable inclination towards the age groups of 30 to 49 years and 50 years and older, while individuals under the age of 29 are systematically underrepresented.

In the context of Chinese demographics, the observed bias pattern demonstrates a preference for female profiles in optimal alignment configurations. There is a pronounced concentration within the age range of 30 to 49 years, accompanied by a systematic underrepresentation of persons younger than 29 years.

These divergent patterns suggest that training data demographic distributions, rather than intentional design choices, drive regional bias profiles. The universal underrepresentation of younger demographics across both contexts indicates a **shared data collection challenge** in both regional AI development ecosystems.

Technological Innovation and Alignment Efficacy (RQ3): Developmental Trajectories and Cross-Regional Methodological Comparison

This subsection examines how technological development stages and innovation strategies influence value alignment, analysing both temporal progression patterns and methodological divergence between China and Western AI development paradigms.

Generational Scaling Effects: Non-Linear Relationships Between Capacity and Alignment

Temporal Trends in Baseline Capabilities

Longitudinal analysis across model generations confirms consistent improvements in task comprehension and response stability with increased model capacity and training sophistication. However, cultural alignment quality and demographic representation fidelity demonstrate non-monotonic relationships with scale, challenging assumptions of linear improvement trajectories.

The scale-alignment paradox Comparative analysis of the Mixture-of-Experts (MoE) architecture Mixtral-8x7B-Instruct (Western) versus smaller dense models reveals critical trade-offs:

- *Consistency advantages*: Mixtral-8x7B-Instruct exhibits significantly reduced insensitivity errors (FF and CV), indicating superior internal coherence
- *Representation limitations*: Despite improved consistency, the larger MoE model demonstrates different patterns in cultural representation compared to the smaller Mistral-7B-Instruct

This finding suggests that standard scaling approaches, absent targeted alignment interventions, may inadvertently affect representation patterns while improving surface-level consistency.

Specialised alignment efficacy As illustrated in Table 2, the First-Token Alignment technique demonstrates the viability of targeted architectural interventions by using fine-tuning based on first-token probabilities to minimise divergence between predicted and actual response distributions, achieving substantial accuracy improvements over zero-shot baselines, confirming that **post-hoc alignment engineering** can effectively address distributional inaccuracies introduced during pre-training, regardless of model regional origin.

Table 2. Main Results for Predicting Country-Level Survey Response Distributions on the WVS Data

Model	Methods	(1-JSD)↑					
		C_1-Q_3	***C_2-Q_1***	***C_2-Q_3***	***C_3-Q_1***	***C_3-Q_3***	***Avg.***
Llama3-8B-Base	ZS [ctrl]	0.748	0.766	0.757	0.779	0.768	0.764
	ZS	0.749	0.768	0.759	0.781	0.770	0.765
	FT [ctrl]	0.756	0.823	0.751	0.837	0.770	0.787
	FT	**0.770**	**0.858**	**0.773**	**0.877**	**0.781**	**0.812**
Llama3-8B-Instruct	ZS [ctrl]	0.748	0.766	0.757	0.779	0.768	0.764
	ZS	0.749	0.768	0.759	0.781	0.770	0.765
	FT [ctrl]	0.756	0.823	0.751	0.837	0.770	0.787
	FT	**0.770**	**0.858**	**0.773**	**0.877**	**0.781**	**0.812**
Distil-Qwen-7B	ZS [ctrl]	0.586	0.641	0.642	0.698	0.682	0.650
	ZS	0.583	0.645	0.639	0.701	0.671	0.648
	FT [ctrl]	0.747	0.764	0.791	0.811	0.817	0.786
	FT	**0.756**	**0.781**	**0.803**	**0.833**	**0.834**	**0.801**
Distil-Qwen-14B	ZS [ctrl]	0.586	0.641	0.642	0.698	0.682	0.650
	ZS	0.583	0.645	0.639	0.701	0.671	0.648
	FT [ctrl]	0.747	0.764	0.791	0.811	0.817	0.786
	FT	**0.756**	**0.781**	**0.803**	**0.833**	**0.834**	**0.801**

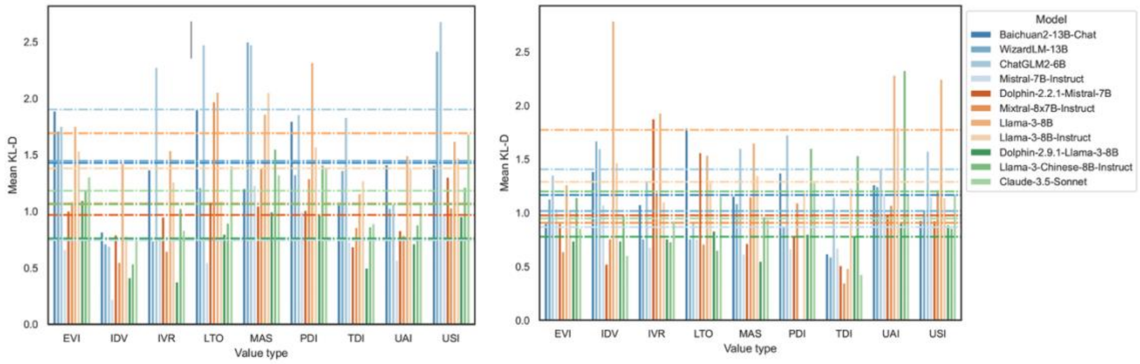
We test all models with zero-shot prompting (ZS) and our proposed fine-tuning approach (FT). [ctrl] indicates a control setup, where we randomly replace countries in test prompts with other countries, to evaluate country context sensitivity. We report Jensen-Shannon Divergence (1−JSD↑).

Source: Cao et al. (2025)

Methodological Divergence: Comparative Efficacy of Alignment Paradigms

Full-parameter fine-tuning versus reinforcement learning Systematic comparison of alignment methodologies reveals technique-specific performance profiles.

In the context of Full-Parameter Fine-Tuning (FFT), models that have been fine-tuned using automatically generated diverse datasets, such as Dolphin-2.9.1-Llama-3-8B, demonstrate a markedly superior average preference distribution accuracy (see Figure 13). Additionally, these models exhibit an enhanced capability to preserve **cross-cultural value variation** when compared to those aligned through Reinforcement Learning from Human Feedback (RLHF). This observation suggests that the diversity of the data, rather than the density of human preference annotations, is the primary factor influencing distributional fidelity.



(a) US survey

(b) CN survey

Figure 13. Mean KL-Divergence of LLM Candidates on US and CN 9-value Dimensions
Horizontal line: average KL-D. Dolphin-2.9.1-Llama-3-8B and Mistral-7B-Instruct show consistent alignment across both cultures.

Source: Adapted from Liu et al. (2025a)

Reinforcement Learning from Human Feedback (RLHF), specifically through the application of Supervised Fine-Tuning (SFT) combined with RLHF pipelines (such as Llama-3-8B-Instruct), results in significant reductions in systemic inconsistencies, as evidenced by metrics such as Frequency of Failure (FF) and Conflict in Value Expression (CV). Furthermore, this approach demonstrates superior internal coherence across evaluation contexts in both the United States and China. The findings suggest that iterative human feedback tends to optimise for consistency rather than distributional coverage.

This divergence suggests that optimal alignment strategies may require **hybrid approaches**: FFT for distributional coverage combined with RLHF for consistency refinement.

Multilingual Training and Region-Specific Optimisation: China Development Paradigm

Impact of multilingual data integration China-origin model development frequently incorporates substantial multilingual training data, providing a natural experiment in cross-cultural alignment:

- Context-specific performance gains: Enhanced multilingual data percentages (as in Llama-3-Chinese-8B-Instruct) yield targeted improvements in Chinese cultural context performance and reduced insensitivity in Chinese evaluations
- Trade-off considerations: Improvements remain context-specific, with limited transfer to U.S. cultural alignment, suggesting that multilingual data integration primarily benefits represented language-culture pairs

Table 3. MARK Simulation Performance on U.S. Social-Survey Data,
Evaluated under both global-distribution and sampled-distribution settings. *'Avg.'* denotes the overall mean metric values, (value) denotes the p-value of results between baseline and MARK larger than 0.05. MARK shows improvements on sampled distributions by achieving the highest simulation accuracy while maintaining low divergences, with most improvements being statistically significant. It also shows generalisation to global distributions with unseen demographics.

Model	GLM-4-air				GPT-4o				Doubao-1.5-pro			
Avg. Metrics	**ACC (%)**	**1-JSD ↑**	**EMD ↓**	**κ ↑**	**ACC (%)**	**1-JSD ↑**	**EMD ↓**	**κ ↑**	**ACC (%)**	**1-JSD ↑**	**EMD ↓**	**κ ↑**
Demo.+Ideo.	25.49	0.3539	0.0741	0.02	32.30	0.4075	0.0755	0.09	30.75	0.4563	0.0685	0.02
Demo.+Ideo.+Opinion	25.06	0.4313 (0.6)	0.0669	0.11	33.07	0.4069	0.0739	0.09	31.45	0.4723	0.0703	0.08
Zhao et al. (2024)	26.98	0.3814	0.0452	0.05	36.96	0.4654	0.0610	0.12	24.23	0.3584	0.0364	0.05
MARK (Ours)	33.69	0.4348	0.0887	0.15	38.11	0.4879	0.0826	0.15	46.98	0.5195	0.0561	0.09

Source: Adapted from Liu et al. (2025a)

Base model foundations **and alignment efficacy** Analysis of alignment intervention efficacy reveals the critical role of pretrained base models:

- **Sequential alignment on base models**: Specialised techniques applied to base models yield substantially greater improvements than comparable interventions on instruction-tuned variants
- **Bias profiles across development stages**: Base models consistently exhibit reduced demographic bias compared to their instruction-tuned successors, suggesting that alignment processes (both SFT and RLHF) may inadvertently concentrate preference distributions while pursuing coherence

This finding has significant implications for **alignment strategy sequencing**: optimal workflows may involve comprehensive distributional alignment at the base model stage, followed by targeted consistency refinement through instruction tuning.

Advanced Multi-Stage Reasoning Approaches

Recent innovations demonstrate that multi-stage personality-driven cognitive reasoning frameworks can enhance cultural value survey simulation accuracy. MARK (Multi-stAge Reasoning frameworK), inspired by MBTI personality type dynamics, outperforms existing baselines by 10% accuracy and reduces divergence between model predictions and human preferences. This suggests that incorporating psychological frameworks and multi-stage reasoning can improve cross-cultural alignment beyond traditional demographic-based approaches, representing a promising direction for both China and Western development paradigms.

Overview: Convergent Challenges and Divergent Solutions in China-Western LLM Development

The cross-cultural audit reveals a complex landscape of **convergent fundamental challenges**

and divergent methodological responses across China and Western LLM development ecosystems.

Universal challenges

As summarised in Figure 14, both regional contexts demonstrate: (1) fundamental instability in value systems despite surface-level consistency; (2) systematic underrepresentation of younger demographics; (3) non-linear relationships between model scale and cultural alignment quality; and (4) inherent trade-offs between distributional fidelity and internal consistency.

Universal Challenges in LLM Value Alignment (Pre-2025 Evolution)

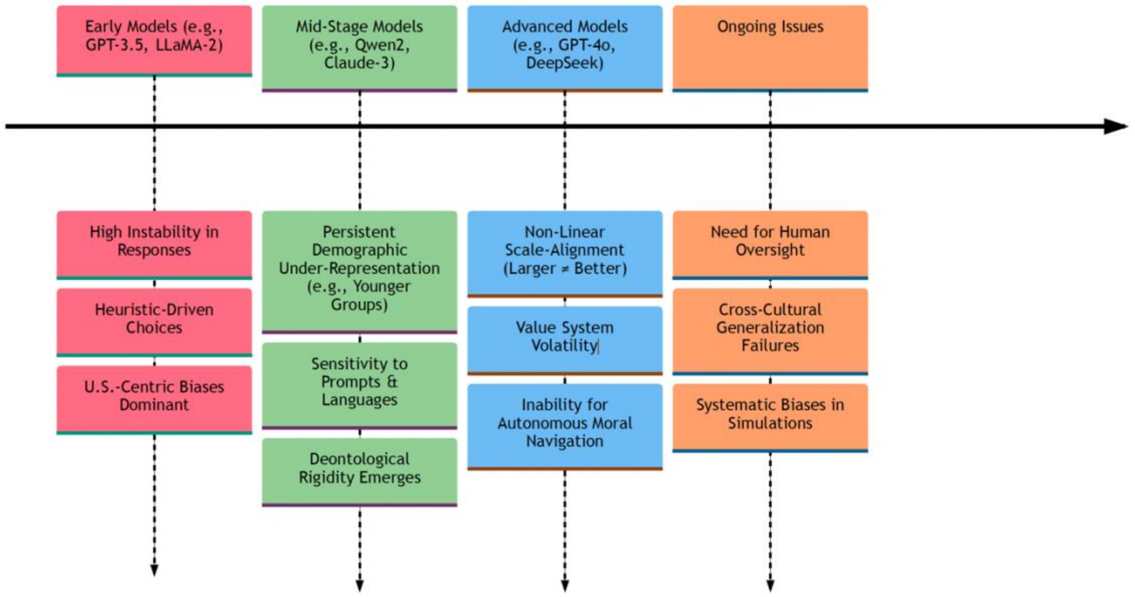

Early Models (e.g., GPT-3.5, LLaMA-2)
Mid-Stage Models (e.g., Qwen2, Claude-3)
Advanced Models (e.g., GPT-4o, DeepSeek)
Ongoing Issues
High Instability in Responses
Heuristic-Driven Choices
U.S.-Centric Biases Dominant
Persistent Demographic Under-Representation (e.g., Younger Groups)
Sensitivity to Prompts & Languages
Deontological Rigidity Emerges
Non-Linear Scale-Alignment (Larger ≠ Better)
Value System Volatility
Inability for Autonomous Moral Navigation
Need for Human Oversight
Cross-Cultural Generalization Failures
Systematic Biases in Simulations

Figure 14. Summary of Universal Challenges
A timeline summarising common challenges across China-West LLMs, evolving from early to advanced models.

Regional pathway divergence

China-origin development increasingly emphasises multilingual data integration and region-specific optimisation, yielding models with strong in-context performance but limited cross-cultural transfer. Western development demonstrates greater architectural experimentation (e.g., MoE configurations, Mistral innovations) but exhibits persistent U.S.-centric biases. Neither pathway has yet achieved robust, generalisable cross-cultural alignment.

Methodological implications

The findings suggest that next-generation alignment strategies must transcend current regional paradigms, integrating: (1) diverse data coverage (FFT strengths) with consistency mechanisms (RLHF strengths); (2) base model distributional alignment preceding instruction tuning; (3) explicit demographic representation objectives alongside cultural value targets; (4) multi-stage personality-driven reasoning approaches; and (5) systematic audit frameworks that evaluate both within-culture accuracy and cross-cultural preservation.

These results provide an empirical foundation for developing **harmonised international AI**

governance frameworks that acknowledge regional innovation strengths while addressing shared fundamental challenges in responsible AI development.

Technology for Good: Social Impact, LLM Governance, and Policy Translation

This section articulates the strategic applications of the computational auditing platform, emphasising governance protocols, policy implications, and institutional frameworks derived from the comparative analysis of China-Western LLM value alignment. As presented in Figure 15, we position these findings within broader debates on responsible AI governance, cross-cultural technology policy, and the democratisation of computational social science methods.

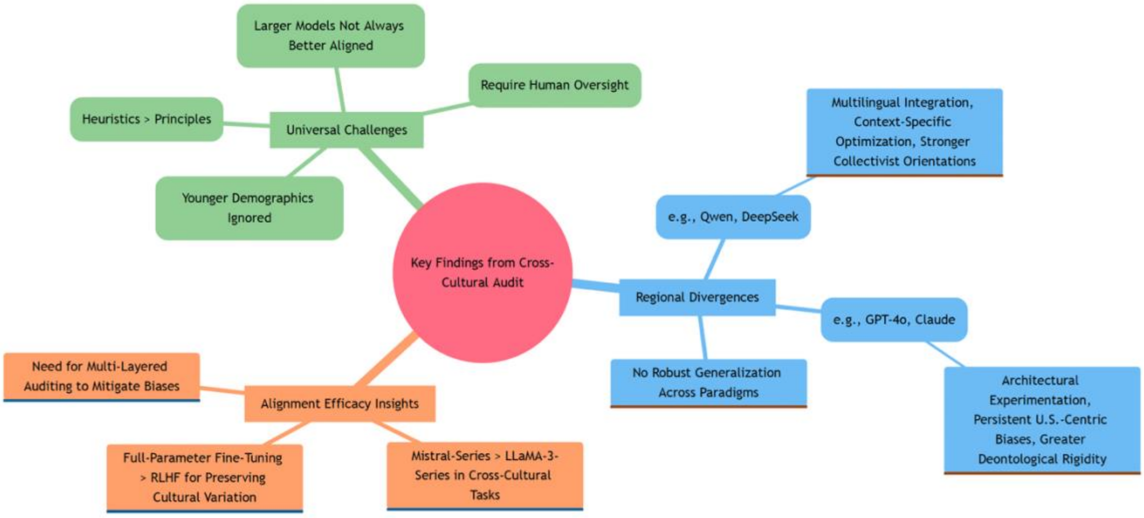

Larger Models Not Always Better Aligned
Require Human Oversight
Heuristics > Principles
Universal Challenges
Younger Demographics Ignored
Multilingual Integration, Context-Specific Optimization, Stronger Collectivist Orientations
e.g., Qwen, DeepSeek
Key Findings from Cross-Cultural Audit
Regional Divergences
e.g., GPT-4o, Claude
Need for Multi-Layered Auditing to Mitigate Biases
Alignment Efficacy Insights
No Robust Generalization Across Paradigms
Architectural Experimentation, Persistent U.S.-Centric Biases, Greater Deontological Rigidity
Full-Parameter Fine-Tuning > RLHF for Preserving Cultural Variation
Mistral-Series > LLaMA-3-Series in Cross-Cultural Tasks

Figure 15. Summary of Key Empirical Findings
This mind map captures the core findings from the cross-cultural audit, emphasising universal challenges and divergences as foundations for governance discussions.

Computational Infrastructure for Responsible AI Governance: Public Policy and Academic Applications

The Multi-Layered Auditing Platform represents a methodological contribution to the emerging field of algorithmic governance, enabling the transition from diagnostic evaluation to actionable policy interventions with measurable social impact.

Case Study 1: Deliberative Democracy Through Computational Simulation—Policy Consultation at Scale

Theoretical Framework and Policy Challenge

Democratic governance increasingly confronts the challenge of incorporating diverse public perspectives into policy formation under time and resource constraints. Traditional deliberative mechanisms – public consultations, citizen assemblies, and survey research face inherent limitations in scale, cost, and representational fidelity (Fishkin, 2011; Moore, 2014). Our integration of the **Diversity-Enhanced Framework (DEF)** with First-Token Fine-Tuning offers a novel computational approach to this democratic deficit.

Methodological Innovation and Governance Application

This framework enables rapid, high-fidelity simulation of public attitudes and value trade-offs regarding proposed policies (e.g., energy transition measures, public health interventions, social welfare reforms) prior to legislative implementation. By achieving high distributional fidelity, these LLM-based simulations provide policy-makers with **cost-effective, population-representative insights** into how demographically diverse populations – such as those in the United States versus China – might respond to policy proposals under consideration.

The framework advances beyond conventional survey prediction by specialising LLMs for distributional simulation rather than individual-level forecasting. Through a fine-tuning methodology leveraging first-token probability optimisation, we minimise divergence between predicted and empirically observed country-level response distributions. This technical innovation yields substantial improvements in predictive accuracy over zero-shot baselines (e.g., 34.3% increase for Llama3-8B-Instruct on aggregate measures), thereby enhancing the reliability of computational methods in social science research and evidence-based policymaking.

Policy Implications and Democratic Theory

This application contributes to ongoing debates regarding the role of computational methods in deliberative democracy (Mercier & Landemore, 2012). While not replacing authentic public participation, these simulations can inform policy design phases, identify potential areas of public concern, and enable rapid iteration of policy proposals before resource-intensive implementation. This represents a form of "computational deliberation" that complements rather than substitutes traditional democratic processes.

Case Study 2: Algorithmic Accountability Through Mechanistic Interpretability – Auditing AI Decision-Making Systems

Governance Challenge and Accountability Framework

As large language models increasingly inform high-stakes decisions across healthcare, criminal justice, and financial services, the opacity of their decision-making processes presents fundamental challenges to democratic accountability and procedural justice (Citron & Pasquale, 2014; Selbst et al., 2019). The Multi-stAge Reasoning frameworK (MARK) addresses this accountability deficit by providing mechanistic, theory-grounded explanations for simulated value judgments.

Methodological Contribution to AI Transparency

MARK operationalises computational interpretability through cognitive modelling informed by personality psychology (Myers-Briggs Type Indicator framework) and stress response theory. This approach yields **personality-driven explanations** for value-laden choices, achieving superior accuracy compared to baseline simulation methods while demonstrating robust generalisation to Chinese cultural contexts. Critically, the framework enables domain experts to **scrutinise and validate the reasoning trajectories** underlying LLM outputs, thereby operationalising the principle of "algorithmic due process".

Institutional and Regulatory Implications

This work contributes to the emerging regulatory landscape surrounding AI explainability requirements, such as the EU AI Act (Act, 2024) and the proposed U.S. Algorithmic Accountability Act. By providing granular, theory-informed explanations, MARK offers a potential compliance pathway for organisations subject to explainability mandates while advancing scholarly understanding of how cultural context shapes algorithmic reasoning. The cross-cultural validation is particularly significant given the comparative paucity of AI governance research outside Western contexts.

Case Study 3: Normative Stability in Sequential Decision-Making – A Safety Protocol for Value Alignment

Risk Assessment Framework

The deployment of LLMs in autonomous decision-making systems raises fundamental questions about moral consistency and normative reliability over extended interaction sequences. The **Ethical Dilemma Corpus** provides an empirical foundation for assessing this previously under-examined risk vector in AI safety research.

Key Empirical Findings and Theoretical Implications

Our sequential stability analysis reveals that LLMs exhibit non-transitive and temporally shifting moral preferences across multi-stage ethical scenarios. Rather than demonstrating stable normative principles analogous to human moral reasoning (Kohlberg, 1981; Haidt, 2013), models rely on context-dependent heuristics and statistical pattern matching. This instability manifests across both Chinese (DeepSeek) and Western (Llama-series) models, suggesting a fundamental architectural limitation rather than a culture-specific training artifact.

Regulatory Guidance and Safety Protocols

These findings yield critical safety guidance for AI governance: LLMs should not be authorised for autonomous, sequential ethical decision-making in high-stakes domains absent robust human oversight mechanisms. This evidence-based constraint informs emerging regulatory frameworks, including sector-specific guidance for healthcare AI, autonomous vehicles, and automated content moderation systems. The findings support a “human-in-the-loop” governance model rather than fully autonomous algorithmic decision-making for value-laden judgments.

Strategic Framework for LLM Governance

The comparative audit generates actionable strategic recommendations for institutional LLM governance, informing model procurement decisions, bias mitigation protocols, and accountability mechanisms – particularly relevant for organisations operating across diverse cultural contexts (illustrated in Figure 16).

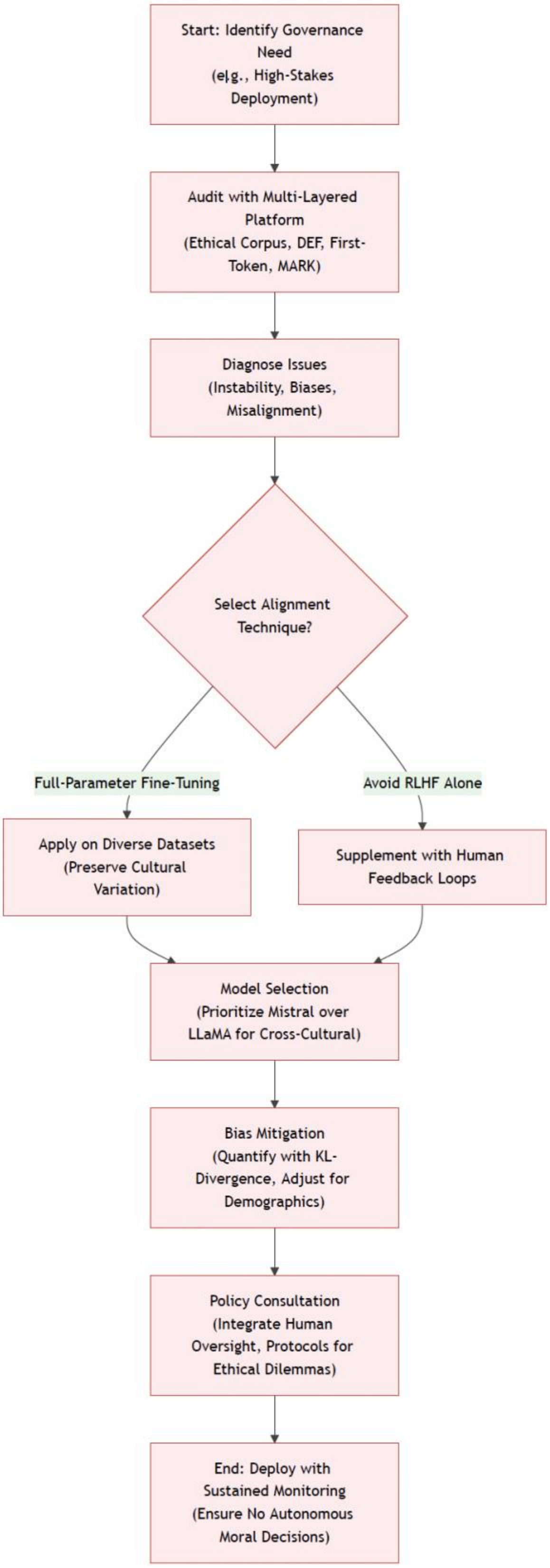

Start: Identify Governance Need (e.g., High-Stakes Deployment)
Audit with Multi-Layered Platform (Ethical Corpus, DEF, First-Token, MARK)
Diagnose Issues (Instability, Biases, Misalignment)
Select Alignment Technique?
Full-Parameter Fine-Tuning
Avoid RLHF Alone
Apply on Diverse Datasets (Preserve Cultural Variation)
Supplement with Human Feedback Loops
Model Selection (Prioritize Mistral over LLaMA for Cross-Cultural)
Bias Mitigation (Quantify with KL-Divergence, Adjust for Demographics)
Policy Consultation (Integrate Human Oversight, Protocols for Ethical Dilemmas)
End: Deploy with Sustained Monitoring (Ensure No Autonomous Moral Decisions)

Figure 16. Actionable Protocols for AI Governance Suggestions
This flowchart outlines practical suggestions for model selection, bias mitigation, and policy consultation, as derived from the paper’s emphasis on evidence-based governance.

Strategic Model Selection: Empirically-Grounded Procurement Decisions

Evidence-Based Architecture Assessment

Institutional governance of AI systems requires moving beyond vendor claims to empirically validated performance metrics. Our cross-cultural value alignment audit establishes that Mistral-series models demonstrate statistically significant superior performance relative to Llama-3-series architectures across both U.S. and Chinese cultural contexts. This finding provides quantitative evidence to inform risk-adjusted model selection for organisations deploying LLMs in culturally diverse markets or multilingual contexts.

Implications for Technology Procurement Policy

This comparative finding advances procurement policy by establishing culture-agnostic performance benchmarks. Organisations can leverage these empirical comparisons to justify model selection decisions, establish baseline performance requirements in vendor contracts, and implement evidence-based standards for model evaluation.

This represents a shift from opaque, proprietary benchmarks to transparent, replicable evaluation methodologies.

Governance Architecture: Transparency-Enabling Alignment Techniques

Comparative Evaluation of Alignment Methods

Effective AI governance requires not only aligned outputs but also interpretable alignment processes that enable institutional oversight. Our research demonstrates that explicit value constraint specification substantially outperforms implicit few-shot learning for guiding normative judgments. Explicit prompting strategies yield superior transparency because they articulate constraints directly, whereas few-shot examples require inferential leaps that obscure the operative value framework.

Performance Benchmarks and Best Practices

Leading models employing explicit preference alignment achieve substantial performance gains, with Claude-3.5-Sonnet reaching 83.1% accuracy and Llama-3-70B achieving 77.8% – both substantially exceeding their baseline zero-shot performance. These findings support governance protocols that mandate explicit value specification in high-stakes applications, enabling clearer accountability chains and more straightforward auditing processes.

Accountability Mechanisms: Quantifying and Rectifying Demographic Bias

Empirical Bias Characterisation

Algorithmic accountability requires moving beyond abstract fairness principles to concrete, measurable bias profiles. Leveraging the Diversity-Enhanced Framework's demographic preference mapping, we identify systematic cross-cultural patterns in representational bias:

1. U.S. context: Optimal alignment configurations systematically favour male demographic profiles aged 30-49 or over 50, indicating potential underrepresentation of female and

younger perspectives

2. Chinese context: Superior alignment emerges for female personas aged 30-49, suggesting inverse gender skew relative to Western models
3. Cross-cultural pattern: Both cultural contexts demonstrate inadequate representation of preferences associated with individuals under age 29, indicating a systematic generational bias

Targeted Mitigation Strategies

These empirically grounded bias profiles inform differentiated mitigation strategies. Rather than one-size-fits-all debiasing approaches, organisations should implement **culturally-calibrated interventions** addressing the specific demographic misalignments identified in their operational contexts. This represents a shift from generic fairness interventions to precision bias mitigation informed by rigorous empirical assessment.

Technical Compliance Pathways: Engineering Solutions for Measurable Standards

Our evaluation of alignment techniques yields specific technical recommendations for organisations seeking to operationalise cultural sensitivity and fairness standards (timeline in Figure 17):

- First-Token Alignment: This specialised fine-tuning approach demonstrates substantial accuracy improvements and validates the technical feasibility of correcting cross-cultural distributional misalignments through targeted architectural interventions
- Full-Parameter Fine-Tuning (FFT) vs. RLHF: Systematic comparison reveals that FFT using automatically enhanced training data achieves superior performance in preserving cultural variation and improving average preference distributions relative to Reinforcement Learning from Human Feedback on human-annotated data
- Multilingual Data Integration: Increased multilingual training data – a characteristic feature of Chinese model development approaches – demonstrably improves performance in culture-specific alignment tasks and reduces cultural insensitivity, offering potential lessons for Western model development

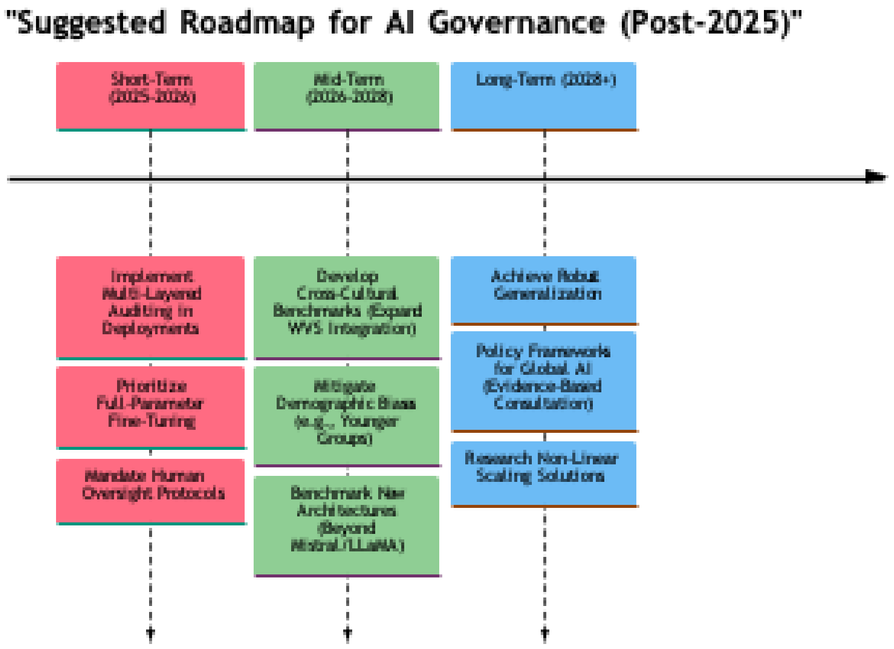

"Suggested Roadmap for AI Governance (Post-2025)"
Short-Term (2025-2026)
Mid-Term (2026-2028)
Long-Term (2028+)
Implement Multi-Layered Auditing in Deployments
Prioritize Full-Parameter Fine-Tuning
Mandate Human Oversight Protocols
Develop Cross-Cultural Benchmarks (Expand WVS Integration)
Mitigate Demographic Bias (e.g., Younger Groups)
Benchmark New Architectures (Beyond Mistral/LLaMA)
Achieve Robust Generalization
Policy Frameworks for Global AI (Evidence-Based Consultation)
Research Non-Linear Scaling Solutions

Figure 17: Future Directions and Suggestions for Responsible AI
This timeline projects suggestions for advancing AI governance based on the findings, focusing on short-term actions and long-term research needs.

These technical pathways enable organisations to translate abstract fairness commitments into measurable engineering interventions. By establishing quantitative benchmarks for cultural alignment and demonstrating effective mitigation techniques, this research provides actionable compliance pathways for organisations navigating emerging AI governance regulations.

Discussion

This study deployed a Multi-Layered Auditing Platform for Responsible AI that synthesises computational methodologies to systematically evaluate Large Language Model (LLM) behaviour across cultural and temporal dimensions. The platform integrates four complementary analytical instruments: the Ethical Dilemma Corpus, which interrogates temporal stability and detects reliance on unstable heuristics through progressively intensifying ethical conflicts; the Diversity-Enhanced Framework (DEF), which quantifies representational fidelity by benchmarking model outputs against empirical human survey distributions in U.S. and Chinese contexts; First-Token Probability Alignment, which enhances distributional simulation accuracy through refined calibration against ground-truth human preferences; and the Multi-stAge Reasoning frameworK (MARK), which elucidates mechanistic, personality-driven explanations to strengthen accountability and interpretability. This comprehensive architecture enables dynamic, multi-dimensional scrutiny of China-Western LLM ecosystems, addressing critical gaps in cross-cultural AI governance research.

Empirical Findings from Cross-Cultural Auditing

Answering ***RQ1****: Temporal Stability, Value Rigidity, and Sequential Moral Reasoning*, the comparative analysis highlights that both Chinese and Western LLMs demonstrate pronounced

fixed value hierarchies, consistently prioritising Truth over Loyalty and Community over Individual welfare. Larger models tend to adopt deontological frameworks that maintain ethical commitments despite adverse outcomes. However, the Ethical Dilemma Corpus reveals that LLMs rely on **heuristic-driven, context-dependent statistical imitation** rather than principled reasoning, limiting their ability for autonomous ethical decision-making. While advanced models like Claude-3.5-Sonnet and Qwen2-72B show improved task comprehension, they exhibit varied stability profiles across dilemma steps. The moral foundation of **Care** remains consistently stable, while **Authority** prioritisation becomes less stable. Importantly, moral volatility is not confined to any region, with both DeepSeek and Llama showing instability, whereas GLM-4-Air and Claude demonstrate stable value maintenance, suggesting that architectural choices are more crucial than geographic origin in ethical consistency.

For ***RQ2****: Cultural Fidelity, Systematic Bias, and Architectural Performance Differentials*, a quantitative assessment of LLM-human preference alignment reveals significant performance disparities between U.S. and Chinese cultural contexts, with Mistral-series models outperforming the others in value alignment. Preference bias analysis uncovers demographic misalignment: U.S. context favours males aged 30-49 or 50+, while the Chinese context skews towards females aged 30-49. Notably, both cultures show a critical underrepresentation of preferences from individuals under 29, highlighting a generational bias in current LLMs.

To resolve ***RQ3****: Developmental Trajectories, Alignment Innovation, and Cross-Regional Strategy Divergence*, a longitudinal analysis of LLM development shows that model evolution and cultural alignment efficacy have complex, non-linear relationships. Although generational advancements reduce insensitivity failures, increasing model scale does not guarantee better alignment quality or cultural representation, challenging assumptions about parameter count as a key ethical performance factor. Specialised First-Token Alignment methods provide significant accuracy improvements over zero-shot techniques. Additionally, Full-Parameter Fine-Tuning (FFT) on enhanced datasets generally outperforms Reinforcement Learning from Human Feedback (RLHF) trained on human-analysed data in maintaining preference distributions and cultural variation. A notable divergence exists between China and the West, with China's approach of increasing multilingual training data (e.g., Llama-3-Chinese-8B-Instruct) enhancing culture-specific alignment and reducing insensitivity in Chinese contexts, impacting global AI strategies balancing linguistic diversity and cultural fidelity.

Strategic Implications for Responsible AI Governance

Policy Validation and Transparency Mechanisms

The empirical platform developed through DEF and First-Token Fine-Tuning provides cost-effective tools for regulatory bodies to assess population responses to policies across various cultures (U.S. vs. China). This approach enhances the speed of social science research and supports evidence-based governance. The MARK Framework offers mechanistic, personality-driven insights into value choices, allowing experts to validate AI reasoning and ensuring necessary human oversight in AI-driven decisions. Given that current LLMs rely on unstable heuristics rather than consistent ethical principles, a foundational safety protocol is essential. **LLMs should not be allowed to make autonomous ethical decisions in high-stakes situations** and require ongoing human supervision, particularly in sectors like healthcare, criminal justice, and public policy where decisions have serious social implications.

Model Selection and Bias Mitigation

Evidence-based governance requires prioritising empirical alignment data, particularly highlighting the superior performance of Mistral-series architectures over Llama-3 in cross-cultural contexts. Accountability frameworks should mandate explicit alignment techniques that enforce **explicit ethical policies or value constraints** rather than relying on implicit learning. Targeted bias mitigation must address demographic skews, such as U.S. male/older-age bias versus Chinese female/middle-aged bias, while correcting underrepresentation of the under-29 age group**.** Specialised fine-tuning methodologies like First-Token Alignment offer efficient solutions for achieving fairness and cultural sensitivity, indicating a need to shift resources from RLHF to FFT in alignment research.

Conclusion

This study demonstrates that responsible AI development in an interconnected world requires moving beyond single-culture optimisation toward genuine cross-cultural competence. The proposed computational auditing platform: **Multi- Layered Auditing Platform for Responsible AI**, provides methodological foundations for this transition, enabling systematic evaluation, transparent accountability, and evidence-based governance across diverse cultural contexts.

As LLMs shape global communication and public discourse, aligning them with diverse human values is essential for equitable technological development. Our findings establish that this alignment cannot be assumed from model scale, assumed from training data volume, or inherited from single-culture optimisation. It must be systematically engineered, rigorously audited, and continuously refined through transparent, reproducible methodologies that acknowledge both universal principles and cultural particularity.

The comparative China-Western analysis reveals that technological innovation serving the global good requires integrating diverse epistemological approaches, respecting cultural specificity while pursuing shared ethical standards, and maintaining a humble recognition of current limitations alongside an ambitious pursuit of improvement. By establishing empirical foundations for evidence-based cross-cultural AI governance, this research contributes to building artificial intelligence systems that genuinely serve humanity in its full diversity – systems that enhance rather than erase cultural richness, that amplify rather than silence marginalised voices, and that operate transparently under sustained human guidance rather than claiming autonomous moral authority.

The path toward culturally competent, socially beneficial AI systems remains long and technically demanding. However, through systematic evaluation frameworks, transparent accountability mechanisms, and international collaborative governance, we can work toward artificial intelligence that serves not merely dominant cultural paradigms but the full spectrum of human values, experiences, and aspirations across our diverse global community. This study provides both the methodological tools and empirical insights to guide that essential journey.

Limitations

Conceptual Framework Limitations

The foundational moral frameworks employed (Kidder's Ethical Dilemmas, Moral Foundations Theory, Schwartz's Theory of Basic Values) inherently privilege Western-centric moral constructs, potentially underrepresenting collectivist ethics such as Confucian *ren* or Ubuntu *ubuntu* that are crucial in non-Western contexts. Future research must systematically integrate culture-specific ethical dimensions through sustained collaboration with regional cultural experts and philosophers representing diverse philosophical traditions.

Methodological Design Constraints

LLMs exhibit pronounced sensitivity to prompt presentation, and the Ethical Dilemma Corpus framework's linearly intensifying scenarios cannot model non-linear escalations (e.g., de-escalation through negotiation), while MARK's reliance on LLM-generated personality predictions and the empirical limitations of MBTI constrain generalisability. Future methodologies should incorporate validated psychological instruments, explore diverse scenario progression patterns, and develop comprehensive prompt variation testing protocols.

Empirical Scope Limitations

The focus on U.S. and Chinese cultural contexts using English and Chinese prompts does not capture broader global cultural diversity, while First-Token Alignment models remain highly specialised for survey response prediction rather than general-purpose applications. Expanding research to encompass diverse cultural contexts, languages, and real-world deployment scenarios is essential for establishing cross-cultural generalisability.

Governance Transparency Barriers

The lack of transparency regarding training data sources in commercial and open LLMs (e.g., Mistral, Llama-3) represents a significant constraint on auditability, particularly in assessing the impact of multilingual data integration – a key feature of Chinese AI innovation – on model behaviour. Future governance frameworks must mandate comprehensive training data disclosure and establish standardised transparency protocols for cross-cultural AI auditing.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback and constructive comments that helped improve this work. We are grateful to Yuchu Tian and Caicai Guo for their insightful discussions and suggestions.

This work was supported by the National Key Research and Development Program of China under Grants 2022YFC3300801, the Research Project of China Publishing Promotion Association under Grants 2025ZBCH-JYYB17, and the Natural Science Foundation of Hubei Province (CN) under Grant No. 2025AFB078.

References

Agarwal, U., Tanmay, K., Khandelwal, A., & Choudhury, M. (2024a). Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In N. Calzolari et al. (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC–COLING 2024)* (pp. 6330–6340). ELRA & ICCL. https://aclanthology.org/2024.lrec-main.560

Agarwal, U., Tanmay, K., Khandelwal, A., & Choudhury, M. (2024b). Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC–COLING 2024)* (pp. 6330–6340). ELRA & ICCL. https://aclanthology.org/2024.lrec-main.560

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C. M., & Wingate, D. (2022). Out of one, many: Using language models to simulate human samples. *arXiv*. https://doi.org/10.48550/arXiv.2209.06899

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Bail, C. A. (2024). Can generative AI improve social science? *Proceedings of the National Academy of Sciences, 121*(21), e2314021121. https://doi.org/10.1073/pnas.2314021121

Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP* (pp. 53–67). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.c3nlp-1.7

Cao, Y., Liu, H., Arora, A., Augenstein, I., Röttger, P., & Hershcovich, D. (2025). Specializing large language models to simulate survey response distributions for global populations. In *Proceedings of NAACL 2025* (pp. 3141–3154). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.162

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2023). Deep reinforcement learning from human preferences. *arXiv*. https://arxiv.org/abs/1706.03741

Citron, D. K., & Pasquale, F. A. (2014). The scored society: Due process for automated predictions. *Washington Law Review, 89*, 1–33.

Dev, S., Sheng, E., Zhao, J., Amstutz, A., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Nishi, A., Peng, N., & Chang, K.-W. (2022). On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022* (pp. 246–267). https://doi.org/10.18653/v1/2022.findings-aacl.24

Dubey, R. K., Dailisan, D., & Mahajan, S. (2025). Addressing moral uncertainty using large language models for ethical decision-making. *arXiv*. https://arxiv.org/abs/2503.05724

Durmus, E., Nguyen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., & Ganguli, D. (2024). Towards measuring the representation of subjective global opinions in language models. *arXiv*. https://arxiv.org/abs/2306.16388

Emelin, D., Le Bras, R., Hwang, J. D., Forbes, M., & Choi, Y. (2021). Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of EMNLP 2021* (pp. 698–718). https://doi.org/10.18653/v1/2021.emnlp-main.54

European Union. (2024). *EU Artificial Intelligence Act*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Fishkin, J. S. (2011). *When the people speak: Deliberative democracy and public consultation*. Oxford University Press.

Forbes, M., Hwang, J. D., Shwartz, V., Sap, M., & Choi, Y. (2020). Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of EMNLP 2020* (pp. 653–670). https://doi.org/10.18653/v1/2020.emnlp-main.48

Hadar-Shoval, D., Asraf, K., Mizrachi, Y., Haber, Y., & Elyoseph, Z. (2024). Assessing the alignment of large language models with human values for mental health integration. *JMIR Mental Health, 11*. https://doi.org/10.2196/55988

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Haidt, J. (2013). *The righteous mind: Why good people are divided by politics and religion*. Knopf Doubleday.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI with shared human values. *Proceedings of the International Conference on Learning Representations*.

Jiang, L., Hwang, J. D., Bhagavatula, C., Le Bras, R., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S., Tsvetkov, Y., Etzioni, O., Sap, M., Rini, R., & Choi, Y. (2022). Can machines learn morality? The Delphi experiment. *arXiv*. https://arxiv.org/abs/2110.07574

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kidder, R. M. (1996). *How good people make tough choices*. Simon & Schuster.

Kohlberg, L. (1981). *The philosophy of moral development*. Harper & Row.

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. In *Proceedings of the ACM Collective Intelligence Conference* (pp. 12–24). https://doi.org/10.1145/3582269.3615599

Liu, H., Cao, Y., Wu, X., Qiu, C., Gu, J., Liu, M., & Hershcovich, D. (2025a). Towards realistic evaluation of cultural value alignment in large language models. *Information Processing & Management, 62*(4), 104099. https://doi.org/10.1016/j.ipm.2025.104099

Mercier, H., & Landemore, H. (2012). Reasoning is for arguing. *Behavioral and Brain Sciences, 35*(2), 243–258. https://doi.org/10.1111/j.1467-9221.2012.00873.x

Mittelstadt, B. D. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Moore, A. (2014). Democratic reason. *Contemporary Political Theory, 13*(2), e12–e15. https://doi.org/10.1057/cpt.2013.26

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models. *Journal of Data and Information Quality, 15*(2). https://doi.org/10.1145/3597307

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv*. https://arxiv.org/abs/2203.02155

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents. In *Proceedings of UIST '23*. https://doi.org/10.1145/3586183.3606763

Raza, S., Garg, M., Reji, D. J., Bashir, S. R., & Ding, C. (2024). Nbias. *Expert Systems with Applications, 237,* 121542. https://doi.org/10.1016/j.eswa.2023.121542

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? In *Proceedings of ICML 2023* (pp. 29971–30004).

Scherrer, N., Shi, C., Feder, A., & Blei, D. M. (2023). Evaluating the moral beliefs encoded in LLMs. In *Advances in Neural Information Processing Systems 36*.

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying language models' sensitivity to spurious features. In *ICLR 2024*. https://openreview.net/forum?id=RIu5lyNXjT

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of FAT* '19* (pp. 59–68). https://doi.org/10.1145/3287560.3287598

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus, 3*(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

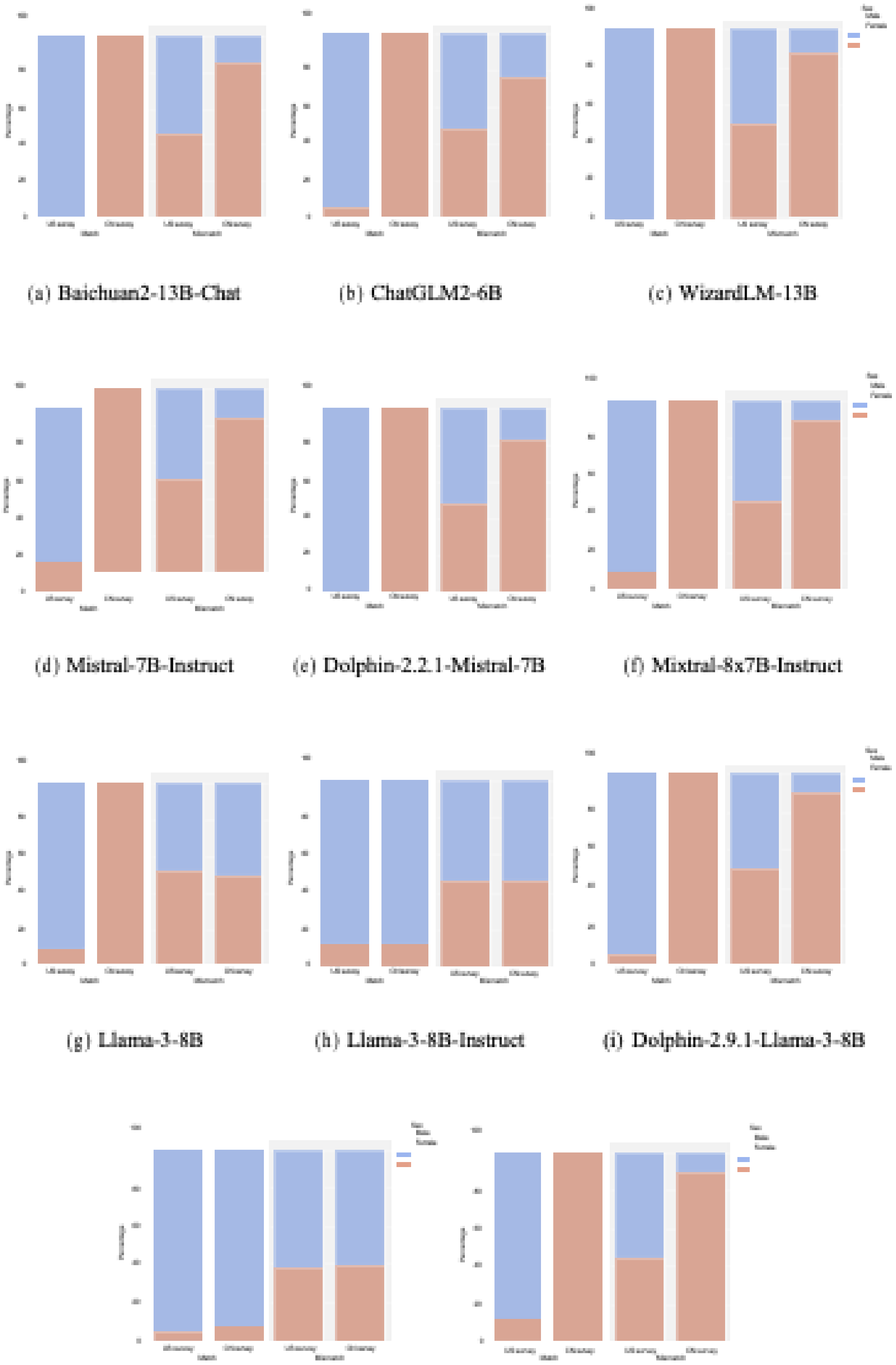
Taubenfeld, A., Dover, Y., Reichart, R., & Goldstein, A. (2024). Systematic biases in LLM simulations of debates. In *Proceedings of EMNLP 2024* (pp. 251–267). https://doi.org/10.18653/v1/2024.emnlp-main.16

Wu, Y., Sheng, Q., Wang, D., Yang, G., Sun, Y., Wang, Z., Bu, Y., & Cao, J. (2025). The staircase of ethics. *arXiv*. https://arxiv.org/abs/2505.18154

Yuan, J., Murukannaiah, P. K., & Singh, M. P. (2024). Right vs. right: Can LLMs make tough choices? *arXiv*. https://doi.org/10.48550/arXiv.2412.19926

Zhao, W., Mondal, D., Tandon, N., Dillion, D., Gray, K., & Gu, Y. (2024). WorldValuesBench. In *Proceedings of LREC–COLING 2024* (pp. 17696–17706).

Appendix



(a) Baichuan2-13B-Chat (b) ChatGLM2-6B (c) WizardLM-13B

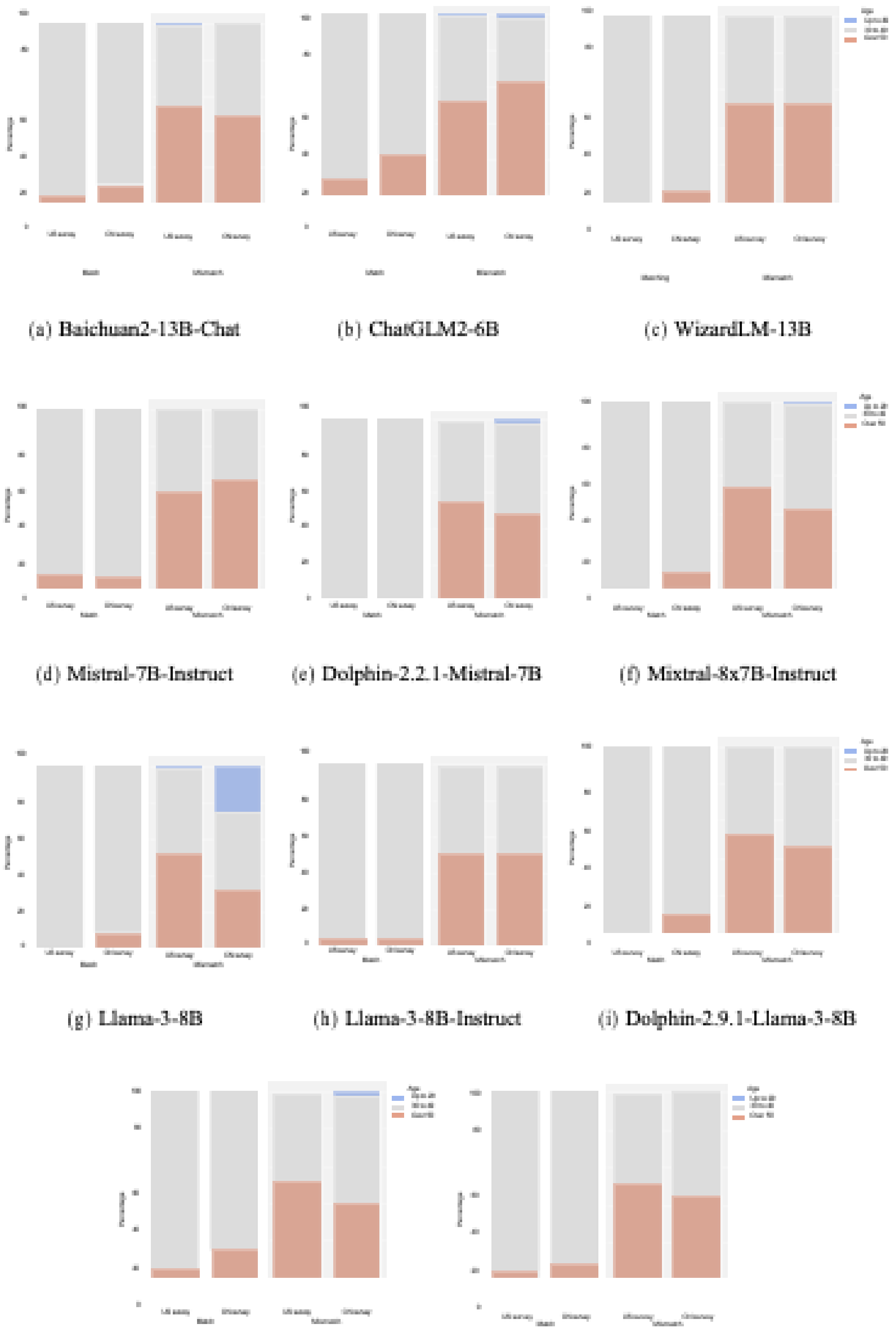
(d) Mistral-7B-Instruct (e) Dolphin-2.2.1-Mistral-7B (f) Mixtral-8x7B-Instruct

(g) Llama-3-8B (h) Llama-3-8B-Instruct (i) Dolphin-2.9.1-Llama-3-8B

Figure 18: Mismatch Proportions by Gender for All Candidates

There is a heavy bias on male profiles in the US survey, and female profiles in the CN survey. But in general, gender bias on the mismatch profiles in the US survey is less significant than in the CN survey.

Source: *Adapted from Liu et al. (2025a)*.



(a) Baichuan2-13B-Chat

(b) ChatGLM2-6B

(c) WizardLM-13B

(d) Mistral-7B-Instruct

(e) Dolphin-2.2.1-Mistral-7B

(f) Mixtral-8x7B-Instruct

(g) Llama-3-8B

(h) Llama-3-8B-Instruct

(i) Dolphin-2.9.1-Llama-3-8B

Figure 19. Proportion of Each Age Group in Mismatch Profiles Across All Candidates, Following the Social Survey Design

Both matching and mismatch profiles show a strong bias toward middle-aged characters, with under-representing preferences for characters under 29 years old.

Source: *Adapted from Liu et al. (2025a).*

Research Paper

Inclusive AI-Driven Collaborative Innovation with Users and New Product Performance:

Evidence from WeGame

LIN Fuxin[1]
QIAO Qianyue[2]
WANG Dongqi[1]
CHEN Zhi[1]
FU Xiaolan[2]
WU Xiaobo[1]

[1]*School of Management, Zhejiang University, China*
[2]*Department of International Development, University of Oxford, UK*

Abstract

Guided by Tencent's mission of "Value for Users, Tech for Good", we investigated how the leading WeGame platform can leverage inclusive AI to facilitate collaborative innovation between producers and users, and enhance the competitiveness of Chinese game products. On one hand, according to the Optimal Distinctiveness Theory, entrepreneurs must carefully consider market positioning to develop competitiveness. On the other hand, Open Innovation Theory emphasises the importance of user involvement. However, on the WeGame platform, the overwhelming volume of user comments poses practical challenges, especially for small and medium-sized studios, in identifying innovation-relevant insights, thereby hindering the implementation of a truly user-centred development approach. In response, we argue that the platform should utilize inclusive AI technologies, particularly LLMs, to transform fragmented and unstructured user innovation inputs into valuable drivers of product competitiveness. In empirical study, we firstly compiled multimodal information from 336 games available on WeGame as of September 2025 to construct a product distinctiveness index. Cross-sectional regression analysis confirms an inverted U-shaped relationship between product distinctiveness and competitiveness. We then focused on 79 games released between 2023 and 2024, applying LLMs to classify 218,655 user comments and isolate 30,297 innovation-relevant instances representing exploratory and exploitative involvement. Integrating monthly user recommendation scores provided by Tencent, we constructed a panel dataset of 1,619 game-month observations. Panel regression analysis reveals that exploratory user involvement significantly boosts competitiveness in highly distinctive games, while exploitative involvement becomes more effective when developer–user interaction is stronger. Towards Tencent's inclusive AI vision, we propose a collaborative innovation framework for platform to empower developers and enhance user involvement.

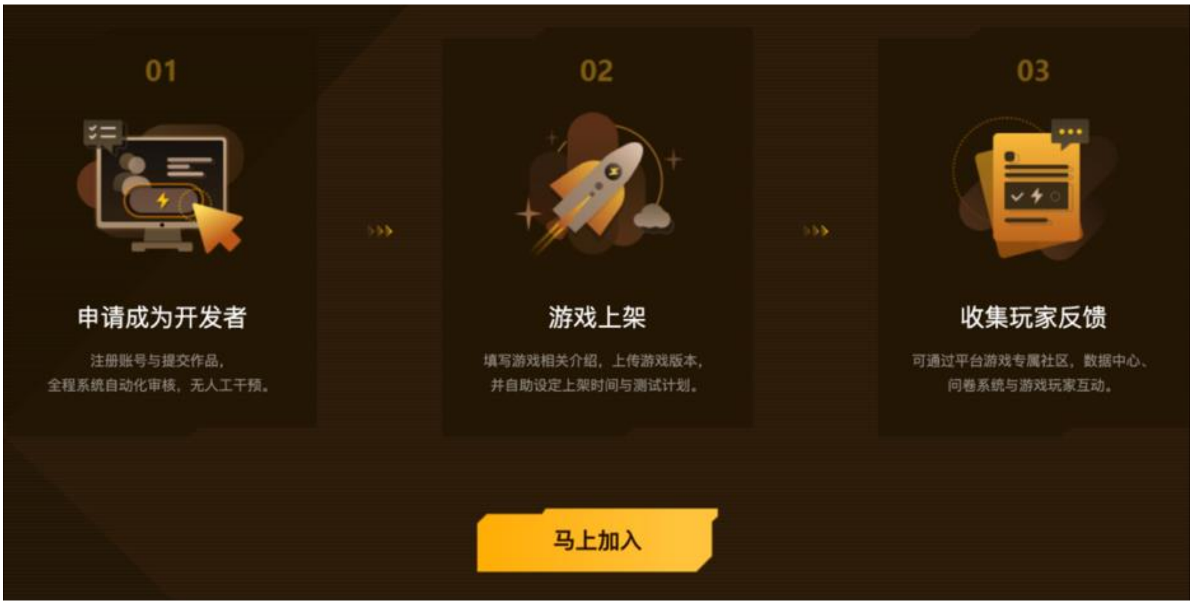
Keywords: Optimal Distinctiveness Theory, User Innovation Involvement, Inclusive AI, User-Producer Interaction, Tencent WeGame platform

Introduction

In the context of accelerating product innovation and increasingly diversified user demands, enterprises can no longer rely solely on internal R&D to meet the challenges of a complex and rapidly changing market environment. The theory of user co-innovation emphasizes that users are not merely product consumers but also potential contributors of ideas and knowledge, capable of playing a crucial role in product design, functional optimization, and contextual feedback. Early studies primarily focused on a small group of "special users," such as lead users and opinion leaders, who were considered highly innovative due to their specialised expertise or extreme needs (von Hippel, 1986). However, such groups are often unrepresentative of the broader user base and overlook the creative potential embedded within the "silent majority" (Magnusson, 2009). With the rise of online user communities in the Internet era, ordinary users have begun to participate widely in co-creation through comments, feedback, and user-generated content, promoting a shift in user innovation research from "elite involvement" to "mass collaboration" (Jeppesen & Frederiksen, 2006).

In practice, Tencent Holdings, as a global technology leader, has exerted extensive influence across social networking, content, and payment domains, while consistently adhering to its mission of "Value for Users, Tech for Good" and core values of "integrity, proactivity, collaboration, and creativity." Tencent has become a representative case of user co-innovation. Notably, in the gaming sector, Tencent has developed a large, open, platform-based user ecosystem that underpins its long-term technological leadership through user collaboration advantages – an essential driver of China's gaming industry competitiveness. For instance, on Tencent's WeGame platform, developers are encouraged to listen to users' voices after joining the ecosystem (see Figure 1). Our field research shows that many independent studios on WeGame highly endorse Tencent's advocated values, emphasising user feedback as a crucial input for product innovation. The diffusion of a "user-centred" culture and the convergence of shared innovation philosophy are particularly evident.

Nevertheless, developers, especially small and medium-sized teams, still face challenges in translating user creativity into product value. First, the user feedback data accumulated on such platforms are massive and heterogeneous, lacking efficient filtering and identification mechanisms, which causes valuable ideas to be buried within redundant information. Second, during the entrepreneurial phase of new product development, market positioning remains a central strategic concern. Whether in building competitiveness, acquiring legitimacy, or shaping user reputation, product distinctiveness plays a key role in determining market performance. Overly similar products may lack recognition, while excessive differentiation may hinder user identification (Zhao et al., 2017). In this context, authentic user feedback and creative suggestions not only serve as critical references for product improvement but also provide data support for enterprises to refine market positioning and innovation direction. More importantly, effectively constructing a positive loop of "product entrepreneurship – user feedback – product adjustment – experience enhancement" can greatly strengthen a product's competitiveness throughout its lifecycle. Nowadays, the rapid development of LLMs offers new possibilities for realising such sustainable innovation mechanisms (Brown et al., 2020; OpenAI, 2023).



01

申请成为开发者

注册账号与提交作品，
全程系统自动化审核，无人工干预。

02

游戏上架

填写游戏相关介绍，上传游戏版本，
并自助设定上架时间与测试计划。

03

收集玩家反馈

可通过平台游戏专属社区，数据中心、
问卷系统与游戏玩家互动。

Figure 1. Wegame's Guidance for Developers to Collect and Value User Feedback

Source: 先锋测试. (n.d.). https://developer.wegame.com/service/test-area.html

To empirically examine this theoretical logic, we build a multi-level, multimodal dataset based on the WeGame, leveraging LLMs for data mining and adopting a two-stage empirical research design. In the first stage, we analyse 336 games available on the platform as of September 2025 and construct an index of product distinctiveness using multimodal information, including tags, textual descriptions, and interface images. We then conduct empirical testing on 79 newly launched products between 2023 and 2024 to explore the relationship between product distinctiveness and new product success. In the second stage, we further examine 218,655 user reviews across these 79 games, filtering 30,297 innovation-related comments using LLMs. Combined with monthly user recommendation index data from Tencent's "Tech for Good" initiative, we construct a monthly panel dataset to investigate the composite and dynamic effects of user innovation involvement, enterprise interaction, and product distinctiveness on product competitiveness. Based on the empirical results, the paper concludes by offering managerial implications on how platforms can empower small developers and dig out users' creativity, aiming to provide both theoretical insights and practical pathways for building an inclusive AI-driven collaborative innovation mechanism.

Practice of User Innovative Involvement in Tencent

We would like to thank the CGS–Tencent–CANGO Organising Committee for their support, which enabled us to conduct an in-depth field study of the WeGame platform in September 2025. Drawing on interviews and multi-source data – including user community observations and publicly disclosed corporate information – we systematically reviewed and analysed the platform's practices.

Mission-Driven Orientation and a User-Centric Cultural Foundation

Founded in 1998, Tencent is one of the world's leading internet technology companies. The firm first released its mission and vision statements in 2003, articulating a commitment to becoming "a trusted friend to users, a vibrant learning organisation, a market leader, a respected partner, and a company delivering stable and reasonable profits 用户依赖的朋友、快乐活力的大学、领先的市场地位、值得尊重的合作伙伴、稳定和合理的利润," with the vision of "building a first-class internet enterprise 创一流的互联网企业."

On 11 November 2019, Tencent announced an updated mission and vision – "Value for Users, Tech for Good" – which continues to serve as the core guiding principle for the company. This mission emphasises "creating user value and embedding social responsibility into products and services 以用户价值为依归，将社会责任融入产品及服务之中", a philosophy that has since influenced both industry and academia.

This cultural orientation is reflected throughout Tencent's Interactive Entertainment Group (IEG), its product ecosystem, and the governance logic of its communities and platforms. By positioning "creating joy with sincerity 用心创造快乐" as a core objective and emphasising "open collaboration and continuous evolution 开放协同，持续进化," Tencent systematically incorporates player feedback into the full product lifecycle management process.

Against this backdrop, we take the WeGame platform as a point of departure to examine Tencent's user-centric co-creation model. Building on field research, we further aim to integrate emerging tools such as LLMs and empirical statistical methods to systematically investigate data-driven user collaboration practices across Tencent and its broader developer ecosystem.

User Collaboration Practices on the WeGame Platform

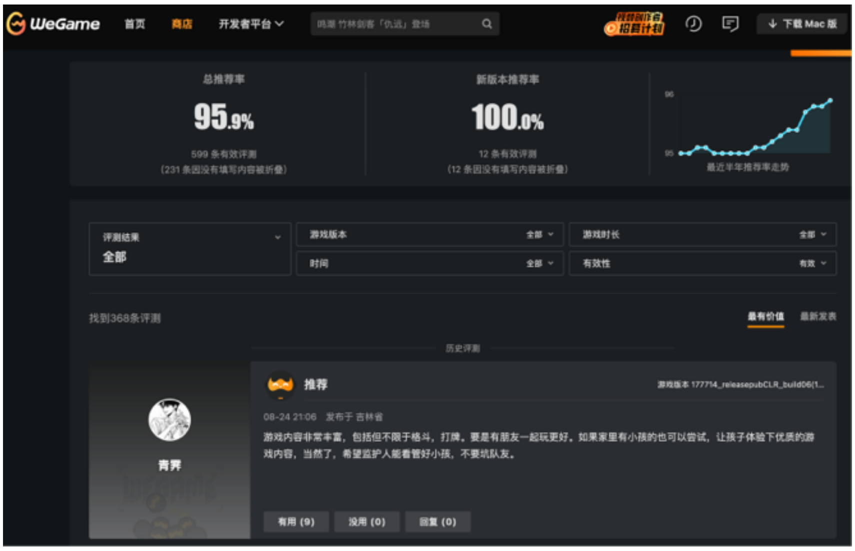
Our fieldwork indicates that WeGame has established a relatively structured process for Collecting player input → Identifying issues → Generating structured insights . Three representative practices are particularly notable.

First, the platform's online community functions as the most common and foundational channel through which players share gameplay experiences and provide performance-related feedback (see Figure 2). The greater the volume of user expression, the stronger the platform's social attributes become, which in turn attracts additional user involvement and generates creative stimuli for developers.

Second, as shown in Figure 1, WeGame actively guides and incentivises developers to collect and integrate player feedback into product optimisation, offering not only communication channels but also policy and financial support. For instance, Grinding Gear Games incorporated user suggestions into gameplay adjustments and transparently communicated progress to players (see Figure). However, interview evidence reveals operational nuance. Developers reported that broadcast-style updates are often ignored, making targeted interaction more effective during pre-release testing. After launch, when comments surge, studios limit direct engagement to avoid public-relations risks and high information-processing costs. Instead, many incentivise key opinion consumers to produce experiential content, thereby shaping player understanding and feedback loops in a more manageable and strategic manner.

Third, interviews with WeGame's operations team indicate that the platform facilitates direct engagement between developers and users by organising play-testing events at Tencent venues and at major industry exhibitions such as ChinaJoy. These activities, combined with limited-release (grey-box) testing, A/B experiments, new-player completion rates, and key-node drop-off metrics, enable continuous validation of user experience and iterative product optimisation.

Taken together, these findings suggest that vibrant interaction between the platform, developers (including both Tencent teams and independent studios), and users has become increasingly routinised and institutionalised within the WeGame ecosystem. This sustained user visibility and feedback culture strengthens players' sense of being heard, thereby reinforcing involvement incentives. The regular functioning of this mechanism not only contributes to the successful refinement of new products but also reflects WeGame's long-term commitment to the principle of "Value for Users" in its collaborative innovation practice.



总推荐率 95.9%
599 条有效评测
(231 条因没有填写内容被折叠)

新版本推荐率 100.0%
12 条有效评测
(12 条因没有填写内容被折叠)

评测结果：全部
游戏版本：全部
游戏时长：全部
时间：全部
有效性：有效

找到368条评测　最有价值　最新发表

历史评测

青霁

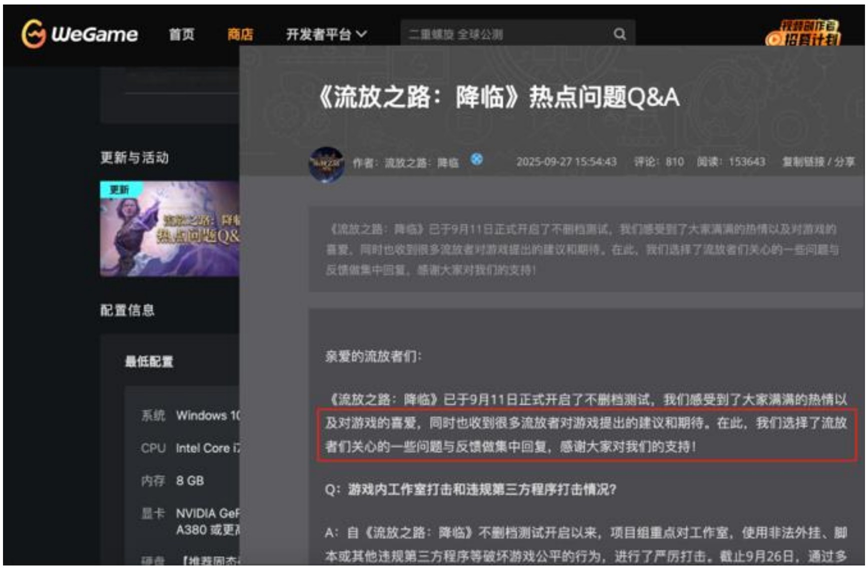
游戏版本 177714_releasepubCLR_build06(1...

08-24 21:06　发布于 吉林省

游戏内容非常丰富，包括但不限于格斗，打牌。要是有朋友一起玩更好。如果家里有小孩的也可以尝试，让孩子体验下优质的游戏内容，当然了，希望监护人能看管好小孩，不要坑队友。

Figure 2. The User Comment Section of the "Party Animals" Game on Wegame

Source: https://www.wegame.com.cn/store/2002136/Party_Animals



《流放之路：降临》热点问题Q&A

作者：流放之路：降临　2025-09-27 15:54:43　评论：810　阅读：153643　复制链接 / 分享

《流放之路：降临》已于9月11日正式开启了不删档测试，我们感受到了大家满满的热情以及对游戏的喜爱，同时也收到很多流放者对游戏提出的建议和期待。在此，我们选择了流放者们关心的一些问题与反馈做集中回复，感谢大家对我们的支持！

亲爱的流放者们：

《流放之路：降临》已于9月11日正式开启了不删档测试，我们感受到了大家满满的热情以及对游戏的喜爱，同时也收到很多流放者对游戏提出的建议和期待。在此，我们选择了流放者们关心的一些问题与反馈做集中回复，感谢大家对我们的支持！

Q：游戏内工作室打击和违规第三方程序打击情况?

A：自《流放之路：降临》不删档测试开启以来，项目组重点对工作室，使用非法外挂、脚本或其他违规第三方程序等破坏游戏公平的行为，进行了严厉打击。截止9月26日，通过多

系统　Windows 1
CPU　Intel Core i
内存　8 GB
显卡　NVIDIA GeF
A380 或更
硬盘　【推荐固态

Figure 3. Developer Announcement of the "PATH OF EXILE" Game on Wegame

Source: https://www.wegame.com.cn/store/2002052/Path_of_Exile_2

Theory and Hypotheses

Theoretical Background

This study is grounded in User Innovative Involvement and Optimal Distinctiveness Theory.

User Innovative Involvement (UII) refers to users' proactive involvement in product ideation, feedback, and functional improvement – an essential form of open innovation (Von Hippel, 1986). Such involvement expands the firm's knowledge boundaries and enhances the alignment between product design and market demand (Lilien et al., 2002; Nambisan & Baron, 2009). Particularly in entrepreneurial or early product development contexts, user involvement significantly improves product feasibility and market success (Xu et al., 2025). With the diffusion of digital technologies, firms increasingly adopt user collaboration mechanisms to shift innovation processes from closed to open systems of co-creation. Integrating user input early in the development cycle helps increase both the novelty and market value of innovation outcomes (Poetz & Schreier, 2012). Emerging online communities – such as comment sections, discussion forums, and crowdsourcing platforms – provide institutionalised arenas for user knowledge articulation and interaction (Xu et al., 2025), effectively lowering the cost of knowledge acquisition and serving as vital external resources for product innovation (Bogers et al., 2010).

UII can be divided into exploratory and exploitative forms of involvement (March, 1991). Exploratory involvement emphasizes novelty and uncertainty, as users propose entirely new features, gameplay mechanisms, or conceptual innovations. Exploitative involvement, in contrast, focuses on incremental optimization – bug fixing, interface refinement, and functionality enhancement – thus favouring stability and implementation feasibility (Bogers et al., 2010).

Optimal Distinctiveness Theory (ODT) explains how organisations or products strategically balance legitimacy and novelty. Deephouse (1999) first introduced this concept in strategic management, arguing that excessive conformity erodes competitive advantage, whereas excessive differentiation undermines legitimacy. Firms should therefore strive to be "different within the mainstream" – that is, optimally distinctive. The theory has since been widely applied in studies of strategy, organisation, and product innovation to demonstrate how moderate differentiation helps new ventures attract resources and gain market recognition (Zhao et al., 2017). In the context of new products, insufficient distinctiveness leads to homogeneity and red-ocean competition, while excessive distinctiveness raises users' cognitive barriers and legitimacy challenges – yielding the classic inverted U-shaped relationship between distinctiveness and performance (Xia et al., 2024).

The core value of ODT lies in its emphasis on the principle that innovation must be moderate. This notion complements UII Theory. While the latter highlights the diversity and heterogeneity of user-generated ideas as a source of novelty, ODT stresses that market acceptance depends on how well innovations fit users' existing cognitive frames. Integrating these two perspectives provides a more complete understanding of how user involvement transforms product differentiation into market performance, thereby enriching the theoretical framework of UII.

Product Entrepreneurial Success from the Perspective of ODT

On digital gaming platforms, new developers face an increasingly saturated and competitive environment. Standing out among a multitude of similar products depends not only on technical capability and content quality but also – more critically – on strategic positioning along the dimension of differentiation. Product distinctiveness reflects the degree to which a new game diverges from existing titles in its features, style, and design mechanisms, serving as a key signal that shapes users' initial perceptions and activates the platform's recommendation algorithms (Zhao et al., 2017).

According to ODT, products that closely conform to platform norms may achieve recognizability but lack uniqueness, thereby failing to attract user attention or trigger adoption. Conversely, products that deviate excessively from users' familiar cognitive categories may suffer from categorical ambiguity and classification barriers, facing legitimacy risks and user resistance (Deephouse, 1999; Su et al., 2024).

Within the high-choice and high-noise environment of digital platforms, the marginal returns to differentiation are particularly sensitive. On one hand, products with low distinctiveness are easily comparable and recognizable but often fall into homogeneous "red ocean" competition, limiting their ability to stimulate user curiosity or gain algorithmic exposure (Porter, 1980). On the other hand, products with moderate novelty can strike an effective balance between exploration and exploitation, capturing users' attention while being recognised by recommendation systems as "potential hits" that deserve preferential visibility (Zhu et al., 2018).

However, when distinctiveness exceeds the "optimal bandwidth," even highly creative products may become cognitively unclassifiable (Zhao et al., 2017), resulting in user comprehension barriers and algorithmic misclassification. As Su et al. (2023) note, highly differentiated products encounter dual challenges of categorical ambiguity and audience isolation: they are difficult to fit within pre-existing platform taxonomies and struggle to connect with users' prior experiences, thereby weakening their diffusion momentum. Moreover, products with high distinctiveness typically entail higher development and communication costs; when the target user base is too narrow, the efficiency of resource returns declines sharply.

Empirical evidence supports this logic. Xia et al. (2024), in their study of the Steam platform, find that games positioned between the extremes of imitation and radical novelty – those "between familiarity and newness" – tend to achieve higher ratings and download volumes. Accordingly, we propose:

> ***Hypothesis 1****: Product distinctiveness exhibits an inverted U-shaped relationship with new product entrepreneurial performance. In short, moderate distinctiveness enhances performance, whereas excessively low or high distinctiveness leads to performance decline.*

User Innovative Involvement and Product Innovation Performance

According to UII Theory, involving users in the new product development process can significantly increase the likelihood of product success (Von Hippel, 1986; Bogers et al., 2010). Users are not merely end consumers but also active contributors and creative sources. By collecting user feedback through online communities, beta testing, or idea crowdsourcing, firms can more rapidly identify and satisfy emerging needs, thereby improving product–market fit and user satisfaction (Nambisan & Baron, 2009).

A large body of empirical research supports the positive impact of user involvement on new product performance. Lilien et al. (2002) found that new products developed through the lead user method achieved higher commercial success rates than those developed through traditional processes. Similarly, Fang et al. (2008) showed that products with higher levels of customer involvement outperform others in value co-creation and market performance. User Innovative Involvement generates two direct benefits. First, user-generated ideas and improvement suggestions enrich product features and enhance quality (Poetz & Schreier, 2012). Second, involvement fosters a loyal seed community whose members, having invested time and effort, often act as early promoters, facilitating word-of-mouth diffusion (Jeppesen & Laursen, 2009). Nonetheless, scholars have also cautioned that excessive user involvement may lead to inconsistent demands and information overload, creating a “double-edged sword” effect on product development (Tang & Marinova, 2020; Najafi-Tavani et al., 2023). Overall, however, when supported by appropriate mechanisms and interaction structures, the net effect of user involvement on innovation performance remains positive (Blazevic & Lievens, 2008).

> ***Hypothesis 2****: User Innovative Involvement is positively associated with new product performance. In other words, the more actively users participate in the innovation process, the better the market performance of the new product*.

However, the effect of User Innovative Involvement is not homogeneous across contexts. The extent to which user-generated ideas are adopted and translated into performance gains depends on the firm’s absorptive capacity and the fit between user input and product characteristics (Poetz & Schreier, 2012; Su et al., 2023). To capture this variation, we introduce product distinctiveness as a key moderating factor, exploring how it shapes the marginal benefits of user involvement – particularly exploratory involvement.

Exploratory user involvement refers to user-generated contributions characterised by high novelty and uncertainty – such as proposing new gameplay mechanisms, narrative combinations, or conceptual designs. While these ideas are potentially valuable, their realization depends heavily on the firm’s absorptive capacity and the need intensity for exploration within the product itself. Product distinctiveness serves as a strong indicator of this need: highly distinctive products, by definition, deviate from mainstream paradigms and often lack established user cognition or mature product-market fit (Zhao et al., 2017). In such cases, exploratory user input can fill developers’ knowledge gaps and jointly redefine the “unarticulated value space,” enhancing product interpretability, lowering cognitive barriers, and enriching both functional and emotional value dimensions (Von Hippel, 1986; Urban & Von Hippel, 1988).

Su et al. (2023) further argue that highly nonconforming innovations require external co-creation to mitigate legitimacy risks and activate potential audience recognition. This logic equally applies to high-distinctiveness products: when uncertainty is high, internal design logic alone may be insufficient to capture user preferences, whereas exploratory user involvement provides frontier demand insights that help calibrate innovation direction and expand interpretive legitimacy. Consequently, in products with higher distinctiveness, exploratory user involvement is expected to exert a stronger positive influence on performance.

Conversely, when products exhibit low distinctiveness and operate within established paradigms, their design space is more constrained and user cognition more habitual. In such contexts, exploratory ideas may be viewed as “out-of-scope” and difficult to assimilate, while their marginal appeal to a stable user base remains limited. In some cases, excessive novelty may even cause positional drift, weakening product coherence. Thus, the positive effect of exploratory involvement on performance diminishes as distinctiveness declines and may even become negligible when creative–product fit is low. Accordingly, we propose:

> ***Hypothesis 2a***. *Exploratory user involvement positively affects new product performance when product distinctiveness is high.*

Exploitative user involvement refers to users providing improvement suggestions based on product usage experience, such as fixing bugs, optimising interface design, enhancing operational smoothness, or refining system logic. This form of involvement emphasises incremental optimisation within existing product boundaries, and its effectiveness depends heavily on whether firms can respond promptly and absorb such inputs into tangible improvements. Accordingly, the degree of producer–user interaction (PUI) becomes a critical moderating mechanism shaping the impact of exploitative involvement.

Piezunka and Dahlander (2015) argue that in open innovation platforms, whether organisations “listen” to user feedback is a fundamental precondition for realising user co-creation value. When firms actively attend to and respond to user inputs, they not only motivate continued user involvement but also structure dispersed external knowledge into the product development process. In digital platform contexts, such responsiveness is particularly critical: frequent firm–user interaction and timely adoption of improvement suggestions enable exploitative user input to be rapidly translated into updates and product optimisation, thereby enhancing user satisfaction, engagement, and market reputation. This iterative mechanism – “from users and back to users” – reflects the firm’s capability to embed users into value creation and strengthens its absorptive capacity (Najafi-Tavani et al., 2023). Conversely, when firms respond slowly or fail to respond at all, valuable user suggestions may be lost amid informational noise (Blut et al., 2020). In such circumstances, user involvement yields limited impact and may even dampen user expectations and willingness to contribute, generating a negative feedback loop. These communication barriers signal insufficient absorptive capacity and disrupt the translation of user contributions into innovation outcomes. Thus, whether user input is meaningfully integrated into final products constitutes a key mechanism through which customer involvement translates into performance improvement (Su et al., 2023; Najafi-Tavani et al., 2023).

Producer–user interaction is not merely an information exchange process; it constitutes a core component of a firm’s dynamic capability system, reflecting its ability to identify, absorb, and redeploy heterogeneous external knowledge. Within this framework, exploitative user involvement yields performance gains only through the firm’s interactive capability. When interaction capacity is strong and responses are timely, users’ “micro-innovations” can effectively compensate for developers’ blind spots in detailed optimisation, enhancing product stability, usability, and specialised features. These improvements ultimately translate into higher ratings, retention, and purchase conversion. Accordingly, we propose:

> ***Hypothesis 2b**. Exploitative user involvement positively affects new product performance when producer–user interaction is high.*

Building on the above, this study develops a multi-path analytical framework grounded in ODT and UII Theory (see Figure 4). In the first stage, we examine the effect of product distinctiveness on new product entrepreneurial performance. In the second stage, we investigate the impact of User Innovative Involvement on new product innovation performance, incorporating two moderating mechanisms: product distinctiveness moderates the effect of exploratory involvement, while producer–user interaction moderates the effect of exploitative involvement.

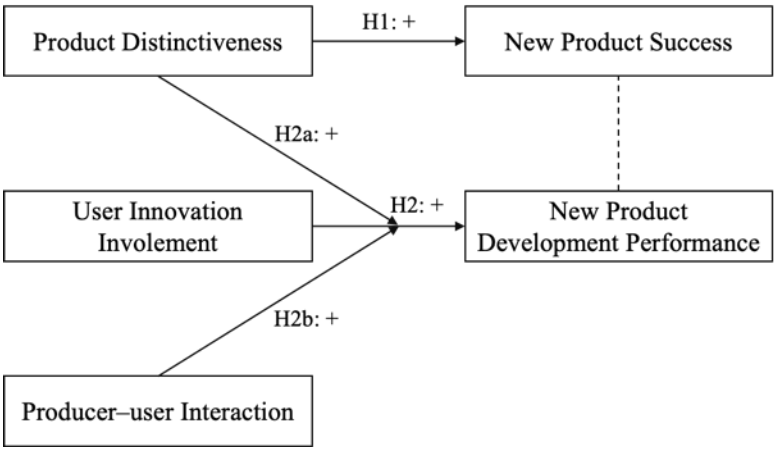

Product Distinctiveness
H1: +
New Product Success
H2a: +
User Innovation Involement
H2: +
New Product Development Performance
H2b: +
Producer–user Interaction

Figure 4. Basic Theoretical Framework

Methodology

Sample and Data Sources

We take Tencent's WeGame platform and its hosted games as the research context. Established in 2017, WeGame is an integrated digital game distribution and community platform developed from Tencent's former TGP platform. With over 1,300 games released to date, it stands as the largest domestic gaming platform in China. We first collected all 336 game titles that remained available on WeGame as of September 2025, encompassing self-developed, agency-published, and independently produced games.

On this basis, our analysis focuses on games first launched between January 2023 and December 2024.This timeframe serves two purposes. First, it avoids the scarcity of new releases caused by regulatory suspension of publishing licenses from 2020 to 2022, while ensuring that each sampled game has at least a 10-month observation window, allowing for dynamic examination of market performance and user innovation engagement. Second, considering the typical life cycle of independent games, which often reach their most active phase within 1–3 years after release, this period captures the stage characterised by the highest intensity of promotional exposure, frequent version updates, and active community interaction – conditions under which user innovative involvement and product differentiation are most salient.

Data collection was conducted in four main stages. Firstly, we extracted detailed information from WeGame platform pages, including game descriptions, tag classifications, promotional images, update logs, and user recommendation ratings. Secondly, we collected 218,655 user reviews and developer replies posted in the early release period. Using a LLM pipeline for semantic filtering and classification, we identified 30,297 innovation-related comments and categorised them into exploratory and exploitative types of user innovation involvement. Through Tencent's Tech for Good open-data initiative, we obtained monthly user recommendation scores for each game, serving as the key measure of dynamic performance. Finally, we integrated complementary datasets: developers' prior publishing experience from ITjuzi and Tianyancha.com, and post-launch search intensity from the Baidu Index as a proxy for market attention.

During sample screening and data preprocessing, we first excluded games with exceptionally low active user counts or search interest to ensure that all subjects had adequate market visibility and community engagement. Second, we removed titles with insufficient review volume to maintain the robustness and representativeness of the text-mining results. Third, to avoid bias from inflated or unstable early-stage ratings, we excluded games released in 2025.

After data cleaning and matching, we constructed a comprehensive multidimensional dataset comprised 79 games and 1,619 "game–month" observations, forming an unbalanced panel dataset that underpins empirical analyses.

Variable Measurement

Dependent Variables

We employ two complementary dependent variables aligned with the objectives of the two analytical stages. In the first-stage cross-sectional analysis, product success is measured by its inclusion in WeGame's user reputation rankings, using two binary indicators: Top 50 Reputation List and Top 100 Reputation List. A game is coded as 1 if it appeared on the respective list as of September 30, 2025, and 0 otherwise. The Top 50_flag indicator serves as the primary measure of entrepreneurial performance, while the Top 100_flag is used for robustness checks.

In the second-stage panel analysis, product performance is captured by the monthly user recommendation rate (Rec_Rate), representing users' ongoing evaluation of a game's quality and innovation outcomes. This percentage-based metric is directly drawn from WeGame's user rating system and aggregated at the game–month level to form a panel dataset. It enables dynamic assessment of how product distinctiveness, user innovation involvement, and firm–user interaction influences continuous performance over time.

Independent Variable: Product Distinctiveness

Product distinctiveness measures the degree to which a game differs from its peers in terms of visual presentation, textual description, and feature tag combinations. Drawing on ODT, we conceptualize product distinctiveness as the extent to which a product deviates from the central tendency (i.e., the centroid) of its category in a multimodal feature space (Zhao et al., 2017).

To ensure comparability, we first partition the sample into major categories based on game genre (e.g., Action, Strategy, RPG, Simulation). A game's distinctiveness is calculated only relative to other games within its own genre. This approach avoids confounding effects from cross-category comparisons and ensures that the reference group shares a similar semantic space and market positioning. We measure product distinctiveness across three dimensions:

Visual Distinctiveness

First, to capture visual distinctiveness, we collected all promotional images for each game from its WeGame store page. If a page included video, we extracted the first frame (frame 0) as a representative image. We then employed a pre-trained Vision Transformer (ViT) model (vit-base-patch16-224) to encode each image (Dosovitskiy et al., 2020), generating high-dimensional visual embedding vectors. For a given game g, we averaged all its image embeddings to obtain a single, comprehensive visual representation vector, $\mathrm{e_{vis}(g)}$.

Within each genre category C, we calculate the intra-category visual centroid:

$$\bar{\mathrm{e}}_{\mathrm{vis}}^{(\mathrm{C})} = \frac{1}{|\mathrm{C}|}\sum_{\mathrm{g}' \in \mathrm{C}} \mathrm{e_{vis}(g')}$$

We also estimate the category's covariance matrix $\Sigma^{(\mathrm{C})}$. The visual distinctiveness of game g is defined as its Mahalanobis distance from the centroid:

$$\mathrm{D_{visual}(g)} = \sqrt{\left(\mathrm{e_{vis}(g)} - \bar{\mathrm{e}}_{\mathrm{vis}}^{(\mathrm{C})}\right)^{\mathrm{T}} [\Sigma^{(\mathrm{C})}]^{-1} \left(\mathrm{e_{vis}(g)} - \bar{\mathrm{e}}_{\mathrm{vis}}^{(\mathrm{C})}\right)}$$

This metric captures the game's deviation from the categorical "visual norm" while accounting for the inter-correlations among feature dimensions (Banerjee et al., 2023). A higher value signifies a more unique visual style.

Textual Distinctiveness

To measure the semantic-level differences in game descriptions, we employed the M3E language model (a multilingual model optimised for Chinese semantic embeddings) to encode each game's introductory text, yielding a textual embedding vector $\mathrm{e_{text}(g)}$ (Reimers & Gurevych, 2019). Similar to the visual measure, we calculate the intra-category textual centroid:

$$\bar{\mathrm{e}}_{\mathrm{text}}^{(\mathrm{C})} = \frac{1}{|\mathrm{C}|}\sum_{\mathrm{g}' \in \mathrm{C}} \mathrm{e_{text}(g')}$$

We then measure textual distinctiveness using cosine distance from the centroid:

$$\mathrm{D_{text}(g)} = 1 - \frac{\mathrm{e_{text}(g)} \cdot \bar{\mathrm{e}}_{\mathrm{text}}^{(\mathrm{C})}}{||\mathrm{e_{text}(g)}|| \cdot ||\bar{\mathrm{e}}_{\mathrm{text}}^{(\mathrm{C})}||}$$

This metric reflects the semantic distance between a game's description and that of its peers. A higher value suggests greater novelty and uniqueness in its narrative or conceptual positioning.

Tag Distinctiveness

The WeGame platform assigns several tags to each game, reflecting its gameplay mechanics and thematic features (e.g., "RPG," "Sandbox," "MOBA"). We converted each game's set of tags into a normalised probability distribution $\mathrm{q(g)}$. We then computed the average tag distribution for the category, $\bar{\mathrm{q}}^{(\mathrm{C})}$. The tag distinctiveness for game g is defined as the Jensen-Shannon Divergence (JSD) between its distribution and the category's average distribution:

$$\mathrm{D}_{\mathrm{tag}}(\mathrm{g}) = \mathrm{JSD}(\mathrm{q}(\mathrm{g}), \bar{\mathrm{q}}^{(\mathrm{C})})$$

JSD is a symmetric and bounded (0–1) measure of divergence based on information entropy, quantifying the difference between two discrete distributions. A higher JSD value indicates that the game's combination of gameplay mechanics and thematic tags is more atypical and distinctive compared to its peers.

Normalisation

To facilitate comparison and regression analysis, we applied Min-Max normalization to each of the three distinctiveness scores within their respective categories, scaling them to a [0, 1] interval. These three normalised indicators were then entered into our regression models separately to identify the heterogeneous effects of visual, textual, and tag-based distinctiveness on product performance.

Explanatory Variable: User Innovation Involvement

We operationalize ***User Innovative Involvement*** as the proportion of user reviews containing creative ideas or suggestions. Measuring this construct poses a major challenge due to severe class imbalance (He & Garcia, 2009): among roughly 230,000 user reviews, the vast majority are non-innovative (e.g., affective expressions or brief evaluations), while genuinely innovative comments are extremely sparse. Applying a LLM such as GPT-4o directly to this unstructured corpus would be both computationally costly and methodologically inefficient, as the dominance of irrelevant content can cause context dilution, reducing precision and recall in identifying the rare innovative class (Liu et al., 2024).

To address this, we adopted a two-stage large–small model collaboration framework, following a “coarse-to-fine” strategy common in NLP pipelines. In Stage 1, a lightweight, high-recall classifier screened the entire corpus to remove clearly irrelevant comments (e.g., “Great game,” “Terrible UI”), producing a candidate subset that retained all potentially innovative content. In Stage 2, GPT-4o conducted high-precision validation and typology classification. First, it verified whether each candidate indeed contained an innovative idea, filtering out false positives from Stage 1. Second, it categorised validated comments into Exploratory or Exploitative innovation, following the established distinction between exploration and exploitation (March, 1991; Bogers et al., 2010). The model was guided by structured prompts (see Appendix) defining exploration as novel feature or paradigm suggestions, and exploitation as incremental improvement or bug-fix proposals. Manual validation of a random sample confirmed high agreement between LLM and human coders, ensuring classification reliability.

Following this two-stage process, we computed each product’s UII score, defined as the ratio of innovative comments to total comments. Higher UII values indicate more innovation-oriented communities. We further derived ***UII_Explore*** and ***UII_Exploit*** to represent the proportions of exploratory and exploitative innovation, capturing the distribution of user-driven creativity across products.

Moderating Variable

Firm Responsiveness. This variable captures a firm’s activity level and responsiveness in platform-based user interactions. It is derived from monthly post data on each game’s official

WeGame account, covering all posts since launch. Posts were semantically classified into five categories (Table 1) reflecting different engagement types – from one-way information release to two-way user feedback responses. Each category was assigned a value from 0 to 4, representing increasing levels of interactive intensity. A monthly composite responsiveness score was then calculated as the weighted sum of post frequencies across categories. Higher values indicate greater firm engagement with the user community, reflecting stronger absorptive capacity and co-creation orientation.

Control Variables

- *Game Size*
  Measured by the game's installation file size (in *Disk_gb*). Larger games often contain more content, which may affect player evaluations.
- *Market Exposure*
  Measured using the Baidu Search Index during the sample period to capture the game's public attention and off-platform visibility.
- *Game Genre*
  Systematic differences may exist in average ratings across genres (e.g., Action, Strategy, RPG). We therefore include genre dummy variables to control for genre fixed effects.
- *Release Year*
  The game's official launch year, control for temporal trends or seasonality.
- *Developer Experience*
  Measured as the cumulative number of games the development team had previously released. More experienced developers may possess a superior understanding of player preferences, potentially leading to higher-rated products.

Model Setting

We employ a two-stage regression analysis design to systematically examine the mechanisms through which product distinctiveness, user innovation, and firm-user interactions influence new product performance. The first stage, based on cross-sectional data, focuses on entrepreneurial performance. The second stage, using panel data, analyses innovation performance. This two-stage approach creates a logical progression from analysing static outcomes to understanding dynamic processes.

In the first stage, we construct the following cross-sectional regression model:

$$\mathrm{New\ Product\ Entrepreneurship_i} = \alpha_0 + \beta_0 * \mathrm{Distinc_i} + \beta_1 * \mathrm{Distinc_i^2} + \sum \mathrm{X_i'} * \beta + \varepsilon_i$$

Where $\mathrm{New\ Product\ Entrepreneurship_i}$ is a binary variable indicating whether game i was listed on the Top 50 or Top 100 list. $\mathrm{Distinc_i}$ represents the product distinctiveness measures (visual, text, and tag). $\mathrm{X_i'}$ is a vector of control variables, including developer experience, game genre, and release time. Given that the dependent variable is binary, we estimate this model using a Logit regression.

In the second stage, we utilize a monthly panel dataset constructed for our sample games, totaling 1,619 observations. We then build a dynamic regression model using $\mathrm{RecRate_{it}}$ as the dependent variable to measure monthly innovation performance. The model is specified as:

$$(1)\mathrm{RecRate}_{\mathrm{it}} = \alpha_0 + \beta_1 * \mathrm{Diff}_{\mathrm{it}} + \beta_2 * \mathrm{UII}_{\mathrm{it}} + \beta_3(\mathrm{Distinc}_{\mathrm{it}} * \mathrm{UII}_{\mathrm{it}}) + \mathrm{Z}_{\mathrm{it}}'\gamma + \varphi_{\mathrm{t}} + \mu_{\mathrm{i}} + \varepsilon_{\mathrm{it}}$$

$$(2)\mathrm{RecRate}_{\mathrm{it}} = \alpha_0 + \beta_1 * \mathrm{FR}_{\mathrm{it}} + \beta_2 * \mathrm{UII}_{\mathrm{it}} + \beta_3(\mathrm{FR}_{\mathrm{it}} * \mathrm{UII}_{\mathrm{it}}) + \mathrm{Z}_{\mathrm{it}}'\gamma + \varphi_{\mathrm{t}} + \mu_{\mathrm{i}} + \varepsilon_{\mathrm{it}}$$

Where $\mathrm{UII}_{\mathrm{it}}$ represent the intensity of Exploratory & Exploitative UII for game i in month t. $\mathrm{FR}_{\mathrm{it}}$ measures the level of firm response activity. φ_{t} and μ_{i} represent time fixed effects and individual fixed effects, respectively, and Z_{it} is a vector of time-varying controls. To capture performance dynamics and persistence, we further introduce a lagged dependent variable $\mathrm{RecRate}_{\mathrm{i,t+1}}$ to control for inertial effects, thereby constructing a dynamic panel regression model. In the econometric analysis, all numerical variables are centred or standardised as appropriate to facilitate the interpretation of regression coefficients.

Results

Descriptive statistics and correlations

Table 1 reports the descriptive statistics of all key variables, including means, standard deviations, and ranges, while distinguishing between the samples used in the two analysis stages. To ensure data representativeness, we further examined the distribution of key variables such as *Top50_flag* and *Rec_Rate* across major categorical dimensions, including Game Genre, Game Mode, and Release Year. The results show no significant distributional differences, indicating that the sample is well balanced across categories and suitable for subsequent statistical analysis.

In the subsequent analyses, all continuous variables were log-transformed or standardised to ensure scale comparability and reduce multicollinearity. Table 2 and Table 3 report the correlation results.

Table 1. Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Top50_flag	79	.114	.32	0	1
Top100_flag	79	.215	.414	0	1
Total reviews	79	2911.696	9158.464	10	66685
Total UII	79	547.38	2293.676	0	18991
Total UII_Explore	79	48.101	141.294	0	833
Total UII_Exploit	79	337.342	1305.811	0	10451
Total UII_Mix	79	2.671	7.588	0	49
Price	79	32.897	48.427	0	268
Disk gb	79	27.811	31.904	.5	130
Developer Experience	79	1.747	1.857	1	9
Dev tencent rel	79	.418	.727	0	2
Market Exposure	79	.544	.501	0	1
Tag Distinctiveness	79	.33	.076	.098	.493
Text Distinctiveness	79	.421	.217	0	1
Visual Distinctiveness	79	.697	.241	0	1
Total_Firm Responsiveness	79	9.013	25.2	0	154

Variable	Obs	Mean	Std. Dev.	Min	Max
Rec_Rate	1619	.781	.166	.148	1
Monthly UII	1619	18.713	96.823	0	1737
Monthly UII_Explore	1619	2.313	11.581	0	198
Monthly UII_Exploit	1619	16.271	86.313	0	1528
Monthly UII_Mix	1619	.13	.822	0	18
Monthly reviews	1619	1.841	2.103	0	9.368
Market Exposure	1619	3.149	3.358	0	11.699
Monthly_Firm Responsiveness	1619	.4	1.555	0	20

Table 2. Results of the Correlation Analysis for the First-Stage Cross-Sectional Data

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
(1) Top 50_flag	1											
(2) Top 100_flag	0.685***	1										
(3) Tag Distinctiveness	0.157	0.189*	1									
(4) Text Distinctiveness	0.489***	0.299***	0.227**	1								
(5) Visual Distinctiveness	0.065	0.188*	0.516***	0.074	1							
(6) Price	0.18	0.480***	-0.053	-0.03	0.066	1						
(7) Disk_gb	-0.249**	-0.07	0.179	-0.085	-0.026	0.158	1					
(8) Developer Experience	-0.102	-0.162	-0.184	-0.051	-0.053	-0.103	-0.083	1				
(9) Developer_tencent_rel	-0.207*	-0.175	0.009	-0.055	-0.007	-0.178	0.320***	0.478***	1			
(10) Monthly_reviews	-0.163	-0.123	0.082	-0.165	-0.228**	-0.253**	0.488***	-0.008	0.502***	1		
(11) Market Exposure	-0.072	0.108	0.267**	-0.013	-0.114	-0.046	0.319***	-0.208*	-0.104	0.212*	1	
(12) Total UII	-0.225**	-0.224**	0.1	-0.185	-0.202*	-0.171	0.558***	-0.081	0.390***	0.808***	0.256**	1

N=79

Note: in brackets are robust standard errors; *,**,*** It means passing the statistical test with significance levels of 1%, 5% and 10%, respectively.

Table 3. Results of the Correlation Analysis for the Second-Stage Panel Data

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1) Rec Rate	1.000												
(2) Monthly UII	-0.091***	1.000											
(3) Monthly UII_Explore	-0.029	0.894***	1.000										
(4) Monthly UII_Exploit	-0.103***	0.994***	0.868***	1.000									
(5) Tag Distinctiveness	-0.005	0.077***	0.035	0.067***	1.000								
(6) Text Distinctiveness	0.157***	-0.133***	-0.113***	-0.136***	0.242***	1.000							
(7) Visual Distinctiveness	0.009	-0.236***	-0.244***	-0.231***	0.456***	0.064**	1.000						
(8) Price	0.215***	-0.371***	-0.254***	-0.367***	-0.300***	0.091***	-0.011	1.000					
(9) Disk_gb	-0.378***	0.425***	0.318***	0.422***	0.210***	-0.060**	-0.091***	-0.375***	1.000				
(10) Dev_Experience	-0.132***	-0.020	-0.059**	-0.012	-0.104***	-0.005	0.019	-0.056**	-0.025	1.000			
(11) Dev_tencent_rel	-0.105***	0.313***	0.266***	0.322***	0.084***	0.032	0.057**	-0.339***	0.277***	0.547***	1.000		
(12) Monthly_reviews	-0.109***	0.926***	0.782***	0.916***	0.167***	-0.139***	-0.188***	-0.510***	0.528***	-0.024	0.314***	1.000	
(13) Market Exposure	0.098***	0.426***	0.373***	0.416***	0.171***	-0.109***	-0.231***	-0.352***	0.454***	-0.180***	0.037	0.526***	1.000

N=1619

Results of the First-Stage Regression Analysis

Regression Analysis

In the first-stage regression analysis, we examined whether each game's inclusion in the platform's Top 50 User-Rated List could serve as an indicator of new product performance. As shown in Table 4, when accounting for the quadratic term of product distinctiveness, the coefficients of the three distinctiveness variables are all positive for the linear term but negative for the squared term. This pattern indicates a significant inverted U-shaped relationship between product distinctiveness and new product success. In other words, moderate differentiation enhances the likelihood of entrepreneurial success, whereas excessively high or low distinctiveness reduces it. The relatively high R-squared values across models further confirm the robustness and explanatory power of the regression results.

These findings provide strong empirical support for Hypothesis 1, which posits that product distinctiveness exhibits an optimal bandwidth effect on new product performance. The results echo the logic of ODT (Zhao et al., 2017; Su et al., 2024), suggesting that products positioned between excessive conformity and radical novelty are most likely to gain user recognition and platform endorsement. In the digital gaming context, characterised by information overload and algorithmic curation, such a balanced distinctiveness helps products remain recognizable to users while still offering novel experiences, thus maximising both cognitive accessibility and market appeal.

Table 4. Results of the First-Stage Regression Analysis

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)
	Dependent Variable: Top50_flag					
Tag Distinctiveness	-18.758	729.531***				
	(25.239)	(250.849)				
Tag Distinctiveness2		-972.394***				
		(333.746)				
Text Distinctiveness			12.399***	14.613		
			(3.843)	(11.475)		
Text Distinctiveness2				-23.669**		
				(11.633)		
Visual Distinctiveness					18.514	588.204**
					(15.074)	(277.196)
Visual Distinctiveness2						-343.867**
						(167.853)
Price	.106**	.181***	.097**	.096***	.153	.114
	(.047)	(.059)	(.038)	(.032)	(.101)	(.104)
Disk_gb	-.515**	-1.105***	-.238*	-.329*	-.358*	-.304
	(.233)	(.381)	(.137)	(.172)	(.201)	(.2)
Monthly_reviews	1.347**	2.339**	.635	.787	.714	.832
	(.666)	(1.08)	(.833)	(.509)	(.994)	(1.171)
Market Exposure	1.99	2.292*	-.031	-.038	.973	-.509
	(2.172)	(1.346)	(1.05)	(1.009)	(1.361)	(1.824)
Release year FE	Yes					
Developer FE	Yes					
Game mode FE	Yes					
Game genre FE	Yes					
_cons	5.593	133.259***	-8.588***	-1.473	-12.936	-212.855**
	(10.714)	(46.205)	(3.105)	(2.735)	(9.769)	(96.344)
Pseudo R^2	.452	.708	.727	.742	.506	.628

N = 79. Standard errors are in parentheses.

*Note: in brackets are robust standard errors; *,**,*** It means passing the statistical test with significance levels of 1%, 5% and 10%, respectively.*

Robustness Test

As a further test, we replaced the dependent variable with an indicator for whether a game was included in the platform's Top-100 user-rated list, using this alternative success criterion to assess the robustness of Hypothesis 1. As reported in Table 5, the linear coefficients of the three distinctiveness measures remain positive; however, once the quadratic terms are included, the curvature is no longer statistically significant.

We interpret this divergence from the Top-50 specification as evidence that the optimal distinctiveness effect is concentrated among top-tier outcomes. When the success threshold is broadened from Top-50 to Top-100, several forces plausibly attenuate the inverted-U pattern: (i) threshold dilution – a looser benchmark admits more heterogeneous, mid-tier products for which distinctiveness above a minimal salience level is sufficient, reducing the need for a finely tuned optimal level; (ii) algorithmic exposure and scale effects – beyond the very top, platform prominence is increasingly shaped by accumulated user base, marketing intensity, and production quality, which can substitute for (or overshadow) curvature in distinctiveness; (iii) variance compression in ratings – scores tend to cluster in the upper range for Top-100 entries, lowering statistical power to detect nonlinearity; and (iv) category heterogeneity – genre-specific norms make distinctiveness payoffs less uniform in the broad Top-100 set, thereby blurring the aggregate inverted-U.

Taken together, these results are consistent with Hypothesis 1 in its stronger (Top-50) form – where success hinges on balancing familiarity and novelty – while suggesting that for broader success definitions, monotonic gains to distinctiveness up to a practical threshold may dominate and nonlinearity becomes harder to identify. Managerially, this implies a segmented strategy: products targeting elite recognition should optimize distinctiveness near the "sweet spot," whereas those aiming for broad but not topmost performance may benefit more from strengthening complementary levers (e.g., producer–user interaction, responsiveness, and incremental polish) once a baseline level of distinctiveness and recognizability has been achieved.

To further probe this mechanism, future analyses could (a) estimate quantile and spline models across the rating distribution, (b) introduce genre-by-time fixed effects to absorb category-specific benchmarks, and (c) examine interactions between distinctiveness and marketing scale or organisational responsiveness, testing whether these factors differentially matter outside the very top tier.

Table 5. Robustness Test of the First-Stage Regression Results

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)	Model(7)
			Dependent Variable: Top100_flag				
Tag Distinctiveness	20.348**	12.963					
	(8.91)	(42.535)					
Tag Distinctiveness2		10.913					
		(57.97)					
Text Distinctiveness			13.334***	-12.114			
			(4.142)	(10.874)			

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)	Model(7)
				Dependent Variable: Top100_flag			
Text Distinctiveness2				32.765**			
				(13.089)			
Visual Distinctiveness					6.204**	11.978	
					(2.871)	(18.083)	
Visual Distinctiveness2						-3.988	
						(12.838)	
Distinctiveness_rank							9.964**
							(3.993)
Price	.03	.03	.051***	.074***	.029	.027	.14*
	(.023)	(.022)	(.017)	(.019)	(.02)	(.02)	(.078)
Disk_gb	-.046	-.046	-.063***	-.086***	-.042	-.038	-.004
	(.035)	(.035)	(.024)	(.031)	(.038)	(.038)	(.082)
Developer Experience	-.681*	-.693*	-.742***	-.855*	-.412	-.396	.475
	(.37)	(.372)	(.277)	(.507)	(.386)	(.396)	(1.535)
Dev_tencent_rel	1.321	1.37	1.142	1.406	.522	.46	-1.795
	(.971)	(.985)	(1.041)	(1.564)	(1.088)	(1.096)	(5.952)
Monthly_reviews	.597	.605	.939*	1.391**	.515	.519	2.647
	(.493)	(.494)	(.502)	(.544)	(.529)	(.53)	(2.332)
Market Exposure	1.849	1.878	5.565***	7.275**	2.234**	2.122**	6.311**
	(1.184)	(1.202)	(1.95)	(2.87)	(1.001)	(1.033)	(3.035)
Release year FE				Yes			
Game mode FE				Yes			
Game genre FE				Yes			
_cons	-8.424**	-7.21	-10.475***	-9.854***	-4.765	-6.693	-13.231*
	(3.516)	(8.408)	(3.28)	(2.809)	(3.05)	(6.022)	(7.127)
Pseudo R^2	.528	.528	.655	.702	.502	.503	.682

N = 79. Standard errors are in parentheses.
*Note: in brackets are robust standard errors; *,**,*** It means passing the statistical test with significance levels of 1%, 5% and 10%, respectively.*

Further Discussion

In the first-stage cross-sectional analysis, we conducted subgroup regressions using game genre and game mode as grouping variables to examine the heterogeneity in the relationship between product distinctiveness and market performance. Employing robust standard errors, we estimated Logit models in which the dependent variable indicates whether a game entered the Top-50 or Top-100 user-rated list. The models include both the linear and squared terms of distinctiveness to capture potential nonlinearity, while controlling for price, scale, market popularity, and release time.

The results show that a significant inverted U-shaped effect of distinctiveness appears only within the RPG and single-player subsamples, suggesting that moderate differentiation most effectively enhances the likelihood of entering higher-ranking lists. For other genres or online-multiplayer games, the models either failed to converge or yielded statistically insignificant coefficients, implying that content innovation is not the dominant driver of market performance in those segments.

Overall, the subgroup regressions highlight the contextual dependency of the distinctiveness effect. In content-oriented and single-player markets, moderate differentiation balances familiarity and novelty, optimising user response. By contrast, in socially or platform-driven online markets, network popularity and user scale effects tend to dominate over innovation, shaping competitive outcomes more strongly. These findings further substantiate the nonlinear and context-specific nature of the innovation–performance relationship and provide empirical evidence for differentiation strategies in cultural and creative product markets.

Results of the Second-Stage Regression Analysis

Regression Analysis of the Moderating Role of Product Distinctiveness

As reported in Table 6, regression analyses based on 1,619 monthly panel observations show that exploratory User Innovative Involvement exerts a significant positive effect on product competitiveness in games with higher distinctiveness, with the interaction remaining significant at lags of one to three months.

This finding provides strong empirical support for Hypothesis 2a, confirming that user creativity is particularly valuable under conditions of elevated product novelty and uncertainty. The lagged effects indicate that user-generated exploratory inputs require time to be absorbed, integrated, and reflected in subsequent product updates or platform feedback, aligning with the logic of absorptive capacity (Cohen & Levinthal, 1990) and innovation diffusion delay (Rogers, 2003). Theoretically, this result extends ODT by illustrating that external user knowledge can function as a compensatory mechanism that mitigates the cognitive ambiguity associated with highly distinctive products (Zhao et al., 2017; Su et al., 2023).

In other words, when a product deviates from familiar design paradigms, exploratory involvement from users helps co-create interpretive meaning and stabilize its position within the platform’s recommendation ecosystem. Practically, this underscores the importance of cultivating user communities that contribute exploratory ideas to sustain innovative momentum and bridge the gap between radical novelty and user comprehensibility.

Table 6. Regression Results Considering the Moderating Effect of Product Differentiation

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)
	Dependent Variable: Rec_Rate					
	1 Month Lag		2 Months Lag		3 Months Lag	
UII_Explore	.002	.001	.002	.001	.001	0
	(.004)	(.003)	(.004)	(.004)	(.004)	(.003)
Tag Distinctiveness		.533**		.55***		.548***
		(.211)		(.21)		(.209)
Price	-.018	-.016	-.017	-.015	-.017	-.015
	(.035)	(.036)	(.035)	(.035)	(.035)	(.035)
Disk_gb	-.025	-.026	-.026	-.027	-.028*	-.028*
	(.017)	(.017)	(.017)	(.016)	(.016)	(.016)
Developer Experience	-.102*	-.097*	-.103*	-.098*	-.104*	-.099*
	(.061)	(.059)	(.061)	(.059)	(.06)	(.058)
Dev_tencent_rel	.04	.034	.04	.034	.042	.036
	(.039)	(.038)	(.039)	(.038)	(.038)	(.037)
Monthly_reviews	.003	.003	.005*	.005**	.006**	.006**
	(.002)	(.002)	(.003)	(.003)	(.003)	(.003)
Market Exposure	-.004	-.005	-.005	-.005	-.004	-.004*
	(.004)	(.004)	(.004)	(.004)	(.002)	(.002)
Distinc x UII Explore		.062**		.058**		.049*
		(.027)		(.028)		(.027)
Game mode FE	Yes					
Game genre FE	Yes					
_cons	1.111***	1.093***	1.108***	1.089***	1.108***	1.089***

	Model (1)	**Model (2)**	**Model (3)**	**Model (4)**	**Model (5)**	**Model (6)**
	Dependent Variable: Rec_Rate					
	1 Month Lag		**2 Months Lag**		**3 Months Lag**	
	(.161)	(.162)	(.161)	(.161)	(.158)	(.159)
Observations	1540	1540	1461	1461	1382	1382

Standard errors are in parentheses.

*Note: in brackets are robust standard errors; *;**;*** It means passing the statistical test with significance levels of 1%, 5% and 10%, respectively.*

Regression Analysis of the Moderating Role of Firm Responsiveness

Table 7 reports that exploitative User Innovative Involvement significantly enhances product competitiveness only when producer–user interaction is high, providing empirical confirmation for Hypothesis 2b.

This finding suggests that incremental user contributions – such as bug fixes, usability improvements, or interface optimization – generate meaningful performance gains only when the firm actively engages and responds to user input. In contexts of strong interaction, user feedback is promptly recognised, structured, and incorporated into rapid iteration cycles, transforming dispersed micro-innovations into tangible quality improvements. Theoretically, this interaction-dependent effect supports the co-creation capability view (Piezunka & Dahlander, 2015; Najafi-Tavani et al., 2023), where organisational responsiveness functions as a dynamic capability that governs the conversion efficiency of external user knowledge into innovation outcomes. Conversely, when interaction is weak, user contributions tend to dissipate as unutilised information, reflecting the "listening gap" in open innovation processes.

Thus, this finding highlights the duality of user involvement: without a responsive feedback architecture, user involvement alone cannot sustain innovation performance. From a managerial perspective, firms should institutionalize two-way communication channels and responsive iteration routines to ensure that exploitative user efforts translate into continuous quality enhancement and user trust.

Table 7. Regression Results Considering the Moderating Effect of Firm Responsiveness

	Model (1)	Model (2)
	Dependent Variable: Rec_Rate	
	9 Months Lag	
UII_Exploit	.001	.001
	(.001)	(.001)
Firm Responsiveness		-.001
		(.002)
Price	-.018	-.018
	(.032)	(.032)
Disk_gb	-.032**	-.031**
	(.015)	(.015)
Developer Experience	-.115**	-.115**
	(.053)	(.053)
Dev_tencent_rel	.055	.055
	(.034)	(.034)
Monthly_reviews	.003	.003
	(.003)	(.003)
Market Exposure	-.001	-.001
	(.002)	(.002)
Firm Respond x UII Exploit		.002**
		(.001)
Game mode FE	Yes	

	Model (1)	Model (2)
	Dependent Variable: Rec_Rate	
	9 Months Lag	
Game genre FE	Yes	
_cons	1.122***	1.122***
	(.146)	(.146)
Observations	908	908

Standard errors are in parentheses.

Note: in brackets are robust standard errors; *;**;*** It means passing the statistical test with significance levels of 1%, 5% and 10%, respectively.

Conclusion and Managerial Implications

Research Findings

First, we reveal a significant inverted U-shaped relationship between product distinctiveness and game competitiveness, confirming the “optimal bandwidth” hypothesis proposed by Zhao et al. (2017) and Su et al. (2024). Moderate distinctiveness strikes a balance between familiarity and novelty, facilitating algorithmic recognition and stimulating players’ exploratory motivation, thereby increasing the likelihood of being listed among top-rated games. Conversely, low distinctiveness leads to homogeneous competition (Porter, 1980), while excessive differentiation reduces legitimacy and user acceptance due to categorical ambiguity (Deephouse, 1999).

Second, User Innovative Involvement significantly enhances product performance but exhibits strong contextual dependence. Exploratory user involvement demonstrates a stronger positive impact in highly distinctive products, echoing Urban and von Hippel (1988), who argued that user collaboration helps firms bridge cognitive gaps in novel domains. That is, users’ creativity reduces interpretive uncertainty and enrich product meaning (Poetz & Schreier, 2012).

Finally, the performance effect of exploitative user involvement depends critically on producer–user interaction (Piezunka & Dahlander, 2015). When firms exhibit strong responsiveness and absorptive capacity (Cohen & Levinthal, 1990), user-driven suggestions can be rapidly integrated into product iterations, significantly improving monthly users’ evaluation . In the absence of such feedback mechanisms, user knowledge is easily lost amid informational noise (Blut et al., 2020).

In sum, this study reveals a synergistic triad linking product distinctiveness, user innovation involvement, and interactive capability. Optimal distinctiveness forms the structural basis of product success, exploratory involvement amplifies innovation diffusion under uncertainty, and interactive capability governs how exploitative involvement translates into performance gains. Together, these findings integrate ODT and UII within the context of platform, offering empirical evidence and managerial guidance for AI-enabled product design and co-creation governance.

Extensions: Towards an Inclusive AI driven New Product Innovation Paradigm

Thus, we propose a practical and inclusive AI-driven framework (see Figure 5) for new product innovation, aligned with Tencent’s Tech for Good initiative. The aim is to guide practice by demonstrating how platforms can empower both developers and users: helping developers listen to user voices and distil creative insights, better match user expectations, and leverage inclusive AI for sustainable innovation (Lin et al., 2024).

The loop begins with User Knowledge Extraction, where user comments with innovative potential are identified and classified through LLMs. The platform thereby acts as an intelligent listener, converting scattered community discourse into structured, semantically enriched innovation insights. In parallel, Product Evolution Analysis compares product development trajectories against both competing offerings and shifting user expectations. By using multimodal and semantic evaluation methods, it helps developers assess differentiation, identify unmet needs, and adjust their design strategies accordingly. This process bridges internal design logic with external market and user signals, ensuring that innovation remains both original and relevant. Next, Innovation Expectation Alignment dynamically benchmarks product updates against historical user feedback to determine whether new releases inspire exploratory engagement or fall short of expectations, providing real-time guidance for iteration timing and content refinement. Finally, Organisational Responsiveness translates user insights into concrete design and strategic actions, establishing a learning-oriented system that sustains competitiveness across product lifecycles.

Together, this four-part framework enables developers to detect weak signals, uncover latent needs, and respond adaptively – bridging external creativity with internal responsiveness under the platforms' *Tech for Good* vision.

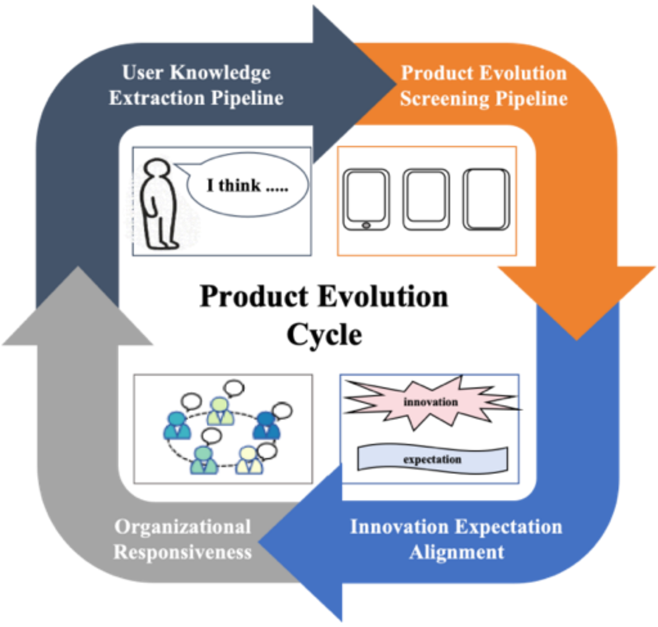

User Knowledge Extraction Pipeline
Product Evolution Screening Pipeline
I think .....
Product Evolution Cycle
innovation
expectation
Organizational Responsiveness
Innovation Expectation Alignment

Figure 5. A Closed-Loop Inclusive AI-Driven Framework for Innovation Management

References

Banerjee, M., Cole, B. M., & Ingram, P. (2023). "Distinctive from what? And for whom?" Deep learning-based product distinctiveness, social structure, and third-party certifications. *Academy of Management Journal*, *66*(4), 1016–1041.

Blazevic, V., & Lievens, A. (2008). Managing innovation through customer coproduced knowledge in electronic services: An exploratory study. *Journal of the Academy of Marketing Science*, *36*(1), 138–151.

Blut, M., Heirati, N., & Schoefer, K. (2020). The dark side of customer involvement: When customer involvement in service co-development leads to role stress. *Journal of Service Research*, *23*(2), 156–173.

Bogers, M., Afuah, A., & Bastian, B. (2010). Users as innovators: A review, critique, and future research directions. *Journal of Management*, *36*(4), 857–875.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901).

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, *35*(1), 128–152.

Deephouse, D. L. (1999). To be different, or to be the same? It’s a question (and theory) of strategic balance. *Strategic Management Journal*, *20*(2), 147–166.

Dosovitskiy, A. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929.

Fang, E., Palmatier, R. W., & Evans, K. R. (2008). Influence of customer involvement on creating and sharing of new product value. *Journal of the Academy of Marketing Science*, *36*(3), 322–336.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Jeppesen, L. B., & Frederiksen, L. (2006). Why do users contribute to firm-hosted user communities? The case of computer-controlled music instruments. *Organization Science*, *17*(1), 45–63.

Jeppesen, L. B., & Laursen, K. (2009). The role of lead users in knowledge sharing. *Research Policy*, *38*(10), 1582–1589.

Lilien, G. L., Morrison, P. D., Searls, K., Sonnack, M., & Hippel, E. V. (2002). Performance assessment of the lead user idea-generation process for new product development. *Management Science*, *48*(8), 1042–1059.

Lin, K., Kan, X., & Liu, M. (2024). Knowledge extraction by integrating emojis with text from online reviews. *Journal of Knowledge Management*, *28*(9), 2712–2728.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). *Lost in the middle: How language models use long contexts*. arXiv preprint arXiv:2307.03172.

Magnusson, M. (2009). *The three-dimensional normal-distributions transform: An efficient representation for registration, surface analysis, and loop detection* [Doctoral dissertation, Örebro universitet].

March, J. G. (1991). Exploration and exploitation in organisational learning. *Organization Science*, *2*(1), 71–87.

Najafi-Tavani, S., Naudé, P., Smith, P., & Khademi-Gerashi, M. (2023). Teach well, learn better—Customer involvement and new product performance in B2B markets: The role of desorptive and absorptive capacity. *Industrial Marketing Management*, *108*, 263–275.

Nambisan, S., & Baron, R. A. (2009). Virtual customer environments: Testing a model of voluntary involvement in value co-creation activities. *Journal of Product Innovation Management*, *26*(4), 388–406.

OpenAI. (2023). *GPT-4 technical report*. https://openai.com/research/gpt-4

Piezunka, H., & Dahlander, L. (2015). Distant search, narrow attention: How crowding alters organisations' filtering of suggestions in crowdsourcing. *Academy of Management Journal*, *58*(3), 856–880.

Poetz, M. K., & Schreier, M. (2012). The value of crowdsourcing: Can users really compete with professionals in generating new product ideas? *Journal of Product Innovation Management*, *29*(2), 245–256.

Porter, M. E. (1980). Industry structure and competitive strategy: Keys to profitability. *Financial Analysts Journal*, *36*(4), 30–41.

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT-Networks*. arXiv preprint arXiv:1908.10084.

Su, J., Gao, X., & Tan, J. (2024). Positioning for optimal distinctiveness: How firms manage competitive and institutional pressures under dynamic and complex environment. *Strategic Management Journal*, *45*(2), 333–361.

Tang, Y., & Marinova, D. (2020). When less is more: The downside of customer knowledge sharing in new product development teams. *Journal of the Academy of Marketing Science*, *48*(2), 288–307.

Urban, G. L., & Von Hippel, E. (1988). Lead user analyses for the development of new industrial products. *Management Science*, *34*(5), 569–582.

Von Hippel, E. (1986). Lead users: A source of novel product concepts. *Management Science*, *32*(7), 791–805.

Xia, J., Wang, J., Hu, H., & Liu, S. (2024). Optimal distinctiveness across different benchmarks: Implications for platform complementors to strategically position new products. *Journal of Business Research*, *183*, 114846.

Xu, Q., Su, Z., & Wei, J. (2025). Valuing every voice: Study on the mechanism of collaborative innovation with ordinary consumers from the digital enablement perspective. *Journal of Digital Management*, *1*(1), 1–16.

Zhao, E. Y., Fisher, G., Lounsbury, M., & Miller, D. (2017). Optimal distinctiveness: Broadening the interface between institutional theory and strategic management. *Strategic Management Journal*, *38*(1), 93–113.

Research Paper

Market Design Interventions for Safer Agentic AI

Swapneel MEHTA[1]
Aaron NICHOLS[1]
Nina MAŽAR [1]
Marshall VAN ALSTYNE [1]
[1] Boston University Boston, USA

Abstract

Our research investigates safety considerations necessary to govern the deployment of Generative AI agents on e-commerce platforms against abuse by fraudulent sellers. Fraudulent advertisements that overpromise product quality led to the purchase of subpar goods while platforms lack effective recourse mechanisms for consumers. Online reputation systems such as buyer-provided ratings for sellers attempt to provide a mechanism for verified buyers of a product to register their feedback in a structured way that may inform the future purchase of goods. However, reviews and reputation are often gamed using bots and fake accounts, constituting “black hat” seller practices. We design an experiment (n = 250 human participants, ~5000 rounds of advertised product sales) in order to demonstrate the occurrence of agentic deception in a two-sided e-commerce marketplace and show how a novel economic intervention involving staked claims limits deceptive strategies to promote honesty in the marketplace.

Keywords: Agents, E-commerce, Digital marketplace, Platform safety

Introduction

Generative artificial intelligence (AI) advances the decades-old idea of intelligent decision-support systems (Holsapple, 1977; Holsapple & Whinston, 1987) with an unprecedented level of autonomy that allows end-to-end execution of tasks by machines that often outperform humans in doing so. Despite being a nascent invention, AI "agents" have proved remarkably useful in practice for tasks that involve information retrieval, synthesis, and reasoning in order to improve downstream decision-making. The technology has been deployed to consumers in prominent "AI concierges" such as Perplexity AI[1] for online searches, Apple Intelligence for Siri, and TripAdvisor AI[2] for itinerary building, many of which were initially powered by OpenAI's ChatGPT. The ubiquitousness of GPT-based models (Vaswani et al., 2017) has led to the deployment of conversational AI agents as the first layer of support for the users of digital platforms (Mariani et al., 2023). For two-sided digital platforms like Amazon generating e-commerce revenues that are multiples of the gross domestic product (GDP) of many small countries,[3] agentic AI presents a significant opportunity to efficiently scale the services and support they offer to millions of sellers on their platform. In the past year, they have adapted the seller workflows on their platform to admit a substantial dependence on AI agents, as evidenced by the launch of their on-platform chatbots Amelia[4], Q[5] and Rufus.[6] Unlike most one-sided AI concierge services though, their introduction of AI assistants in two-sided e-commerce marketplaces will have direct financial consequences for both parties on the platform: the sellers to whom the service is offered, and the buyers who are exposed directly to the AI-generated advertisements, without necessarily being made aware of it.

For digital marketplaces, the economic notion of 'social welfare' is maximised across the sum of sellers' profits from product sales and buyers' utilities from the quality of goods purchased. Sellers, however, may seek to increase profit, at buyers' expense, by misrepresenting their goods and overcharging for their wares. Sellers communicate their product claims to buyers via advertisements hosted by the platform. These ads need not be accurate, allowing sellers to mislead buyers who then purchase subpar products. Deception reduces buyers' trust in the platform, which is why most platforms crack down on deceptive practices. Amazon and eBay both allow returns for products that were not as advertised. Platforms also offer buyers the ability to rate sellers, such that false advertising can damage a seller's reputation (Budzinski & Köhler, 2015). But when competition for sales increases, significant profit is at stake, and brand names can change, the risk of cheating also increases. The ability of sellers to hide behind AI agents exacerbates this risk. Prominent cases of AI agents strategically deceiving humans have occurred when deceit is materially beneficial to the goals they are asked to achieve, for instance, when agents are tasked with maximising profits. Notable examples of such behaviour include insider trading (Scheurer et al., 2024) and algorithmic collusion (Fish et al., 2024). There is well-

[1] Perplexity, accessed Oct 10, 2024.

[2] TripAdvisor Trip Planning with AI, accessed Sept. 20, 2024.

[3] Statista, "Amazon Revenues 2014-23", accessed Sept. 20, 2024.

[4] Amazon Blog, "Amazon launches a powerful new generative AI-based selling assistant codenamed Project Amelia", September 19. 2024.

[5] AWS Blog, "Introducing Amazon Q, a new generative AI-powered assistant (preview)", April 30, 2024.

[6] Amazon Blog, "Amazon announces Rufus, a new generative AI-powered conversational shopping experience", February 1, 2024.

documented history of sellers using deceptive sales practices to game ratings, reputations, and reviews, resulting in a challenging online advertising environment for digital platforms to moderate (S. He et al., 2022; Y. Wu et al., 2020). This raises critical questions that before we scale deployment of agentic AI in places like Amazon, Alibaba, Taobao, or eBay; **what are the effects of introducing agentic AI sellers on the fraudulent sales in digital markets? What measures can we take to limit consumer harms induced by AI agents in competitive markets?**

This motivates our key research hypotheses:

1.H1: The application of stakes to advertisements will improve social welfare in the marketplace.

2.H2: Agentic AI Sellers entering the market without a reputation win their first sale in an earlier period in the staked ads market than in the brand reputation market, when they use stakes.

3.H3: Agentic sellers will make higher profits in the reputation market than other agents in that market given their ability to strategically deceive audiences.

4.H4: The introduction of stakes will curtail the deceptive sales achieved by agentic sellers.

H1 would indicate that the mechanism design intervention benefits both sellers and buyers in the marketplace. H1 and H4 are motivated specifically by the introduction of costly signalling adapted from (Coase, 2013; Spence, 1973; van Alstyne, 2021) and the general notion that well-designed policies help address market challenges (Chen, 2020). Commitment mechanisms help address limitations of reputation systems as laid out in (Resnick, 2002). H2 and H3 are motivated by the fact that even human sellers engage in systematic "seller experiments" (Einav et al., 2016), but AI agents can process market feedback and optimise strategies at scales impossible for humans, including the ability to strategically employ market exits to escape the reputational consequences of their actions (Cabral & Hortacsu, 2004). This, combined with the success of agentic deception as highlighted in prior work from artificial intelligence safety indicates that AI sellers have the potential to exhibit highly persuasive, deceptive behaviour when put under pressure (Danry et al., 2024; Fish et al., 2024; Hubinger et al., 2024; Scheurer et al., 2024).

In the absence of platform datasets that help study such a phenomenon, and no incentive for LLM providers like OpenAI to provide such an analysis except in the presence of public pushback (*Sycophancy in GPT-4o,* 2025)--we create a real-time digital marketplace to simulate the effects of agentic AI sellers in an interactive, multiagent setting with human consumers, preprogrammed sellers, and agentic AI sellers competing to maximise individual benefit as a first-of-its-kind study into the economic impact of AI sellers. Through our study, we not only enable research into the effects of agentic sellers in digital marketplaces but also explore their competitive advantages, impact on other adaptive ‘bot’ sellers, strategic deception tendencies, and real human consumer response to these strategies in an online marketplace like Amazon or eBay. We employ a baseline or control marketplace that is reflective of the affordances of digital marketplaces such as product advertisements, product purchases, feedback through ratings, and profits from sales. Our marketplace offers the ability for sellers to advertise honestly or dishonestly (a binary choice for sellers), optionally escrow a predetermined amount in order to stake a claim their advertisement is true, and determine whether they would like to rebrand at the end of the round of sales. On the other hand, we offer buyers the ability to buy products based on advertisements from different sellers (they may also elect not to buy any advertised products), rate sellers based on the true product quality that is revealed after a purchase, and elect to challenge a staked advertisement for a nominal price, that convenes an adjudication process for the advertised claim. The baseline market does not offer the option to stake claims, while the treatment or 'Stakes market' does.

A Human-AI Marketplace for Behavioural Experimentation

Traditionally, behavioural scientists use surveys, one-sided markets, or virtual simulations (such as recent agent-based models simulated entirely with GenAI agents that aim to approximate human-AI interaction in real-world settings). Surveys and simulations lack in situ validity to model psychosocial outcomes in online experiments. They may suffer from gaming benchmarking practices or involve a selective application of cognitive tests to suit the research design; issues tied with a lack of reproducibility and external validity (Almaatouq et al., 2021; Laraway et al., 2019; Singh et al., 2025). This includes recent Meta's overt attempts to do so with Llama-4 models[7]. But the fundamental cause of such issues is that there are almost no reproducible open-source benchmarks to conduct real-time human-AI experimentation with multiple humans in a competitive setting that might make for a realistic high-fidelity representation of e-commerce marketplaces. To our knowledge, there are no open benchmarks for the economic impact of agentic AI sellers in two-sided marketplaces despite the need to produce research evaluating their effects in a manner that is replicable, reflective of properties that emerge at scale, and provide outcomes that remain externally valid. Past work by Google on recommendation systems libraries, RecSim and RecSim-NG, attempted to solve a similar problem by introducing an environment for multi-turn recommendations, though they operated it as a virtual simulation without human participation. We present the first such open-source benchmark with results from a study of how agentic AI sellers advertise to – and tend to deceive – human consumers in digital marketplaces.

Our research introduces human participants in multiagent experimentation platforms like Camel, large population models (LPMs), and others (Chopra et al., 2024; *OpenAI/Swarm*, 2024/2025; Trivedi et al., 2024; Yang et al., 2025). In our two-sided marketplace, we create a realistic e-commerce environment with 6-15 humans participating in an incentive-aligned experiment on either side, as buyers and sellers in the market, alongside multiple agentic AI assisting some human sellers in optionally crafting advertisements that may mislead their buyers. We model the short-term and long-term economic impact of such agentic deception on consumer trust, purchase behaviour, and evaluate if AI agents can be curtailed starting with a decentralised platform intervention relying on peer juries, called 'truth warrants' (van Alstyne, 2023) and serve as 'stakes' for advertised claims (we refer to these as 'staked ads', for brevity). In early experiments with GPT-3, Llama 3, and GPT-4, we find that this intervention – in some ways similar to X's Community Notes – may successfully limit some deception and incentivise honest sellers. With the release of novel, powerful reasoning-based models, commodity open-weights and open-source models, and safety-focused models, our two-sided platform presents a unique opportunity to create economic benchmarks for the deception and collusion exhibited to maximise profits by a variety of LLMs in large-scale experiments led by experts in platform economics and behavioural science.

[7] there has been a discussion about fair interpretation of the online leaderboard policies discussed here: https://beebom.com/meta-llama-4-benchmark-manipulation-not-first-time/

Design of our Digital E-commerce Marketplace

We designed an interactive experimental marketplace similar to Amazon or eBay where each of 250 human buyers in the U.S. interact with seven seller bots across seven rounds of sales, with varying truthfulness strategies. Each round starts by providing buyers with capital, which they use to purchase up to 3 products from the seven ads they are shown--one each, generated by different sellers. The catch is that sellers can produce either low-quality or high-quality products, while all ads claim high quality – creating opportunities for strategic deception. Sellers aim to maximise their profit, π, from sales, while buyers aim to purchase high quality products to maximise their utility u. If a buyer purchases a low-quality product, having relied on a high-quality advertised claim (hereafter, referenced as 'high quality ad'), they are said to have been cheated and would earn a utility $u < 0$. As in standard markets, buyers may rate sellers. Following a purchase, buyers may rate sellers as honest or dishonest, which is displayed on a seller's profile in subsequent rounds of sales. Following each round, however, sellers may voluntarily rebrand, resetting their buyer-provided ratings to zero.

We design a novel interactive marketplace to evaluate the strategic choices made by agentic AI models to make money in an e-commerce setting. This experiment involves human participants as buyers of products, based on the product advertisements that they are shown in each round. Sellers aim to maximise their profit from sales, while buyers aim to purchase high quality products to maximise their utility. A seller – each represented by a brand – selects a product's true quality (high or low) which may differ from the advertised product quality (always high) shown to buyers, introducing the potential for sellers to mislead buyers that believe the advertisement and purchase the product. Adaptive 'bots' play the role of sellers that advertise products to human buyers each round, for 7 rounds of sales. We employ these bot sellers in order to reflect a distinct variation of strategies that arise from signalling choices made by sellers in the marketplace utilising the episodic nature of a multi-round game. In particular, (Tadelis, 2016) discusses how seller strategies may vary from honest to opportunistic (in our case, dishonest) and our adaptive bots respond to this trade off in each round, making a choice about signalling their product quality in order to accrue profits through product sales. In each round, they may elect to either mislead or accurately advertise the true quality of their product – that is, either sell a high-quality product by advertising it honestly, or sell a low-quality product using a dishonestly advertised claim. The seller bots then are akin to Robert Axelrod's agents in his seminal work (Axelrod, 1984, 1997), adapted for e-commerce settings (Jannach & Leitner, n.d.; Jianya et al., 2015) except for the Agentic AI seller indicated separately in bold lettering:

1.Honest: Always produce high quality products advertised as high quality.

2.Cheat: Always produce low quality products advertised as high quality.

3.Bait-and-Switch: Produce high quality products until sold, switch to low quality and back.

4.Reformed Cheat: Produce low quality products until sold, then switch to high quality.

5.Goldfish: Produce low quality products until sold, switch to high quality and back.

6.Politician: Produce high quality products until two sales, switch to low quality, and back.

7.**Agentic AI Seller**: Send the current state of the market (including historical sales) to a 'gpt-4o-mini' model and respond in accordance with its advertising directives, as well as separately obtained rebranding directives. We do not place any constraints on the strategy (ranging from honest to dishonest) that the AI seller chooses to employ.

Buyers purchase up to 3 products in each round based on the amount of money available in their wallets and the visible product price, seller name, and seller reputation. Upon purchase of a product, a buyer can rate the seller's brand based on the accuracy of their advertisement; a seller accrues ratings for their brand over multiple rounds. Additionally, at the end of each round a seller can choose to change brands, resetting their ratings to zero, at no cost. In effect, sellers can re-enter the marketplace with a new brand, after discarding their prior brand. We employ the advertising strategy suggested by the agentic seller in order to sell products to human participants using product advertisements and measure its ability to generate profit in competition with the strategies employed by other seller bots in the market. We also follow its rebranding directives as directed through a separate, second prompt provided to it. Notably, the only prompts we provide to the agentic AI seller are the instructions that we also offer human participants playing as consumers, explaining the rules of the marketplace sales experiment to them. There are 7 rounds of sales each game, so human buyers can adapt their purchasing strategies based on their experience even as sellers adapt their sales strategies to maximise sales and profits.

Simulating Adjudication via an "Oracle" Jury

One of the key features of our marketplace experiment is to model the decision of randomly selected peer juries that will determine how to adjudicate product claims when a buyer challenges the seller's claim as being fraudulent, typically based on their experience having purchased the product from them. Since buyers only challenge sellers after they purchase an advertised product, and a challenge requires buyers to put up a nominal amount of money as a fixed cost paid to the platform in order to convene a jury, it is a strictly non-dominant strategy to challenge honest advertisements: simply because the buyer would lose their challenge amount as the advertised claims are accurate. However, for false advertisements, it is a strictly dominant choice to challenge the seller since there is evidence of harm. In traditional settings, all manner of claims might be adjudicated in such a peer jury setting, raising the question of whether opinions may be admissible, or the concern that facts may change, or even that peer juries may be biased in favour of either party. For these reasons we ensure that the seller has complete control over the choice to stake their ad, depending on whether they believe the claim may be falsifiable. Furthermore, the comparable model of adjudication is either centralised and opaque (a platform may decide whether or not to take down your post on X, Meta, or remove you from Amazon, Alibaba), or long, expensive, and reliant on other forms of peer juries (consumer court cases). We lean on the peer jury model with clear demarcations for the kinds of admissible claims, with full agency over participation provided to the seller who decides to stake an ad and based on notable (and surprising) successes with peer jury models for both e-commerce and food delivery applications in China.[8]

For simplicity in its implementation, and to prioritise the testing of the contribution of staked

[8] Hua, "When Online Shoppers Feel Cheated, It's Time to Go to Crab Court", Wall Street Journal, 2022. Accessed April 20, 2025.
Yang, "Users are doling out justice on a Chinese food delivery app", MIT Technology Review, 2023.

claims to honest advertising, we abstract the peer jury with an oracle that would represent an optimal jury, implying it will rule correctly. With the introduction of an optimal peer jury, it is a non-dominant choice for advertisers to stake misleading claims in advertisements as they are bound to lose the subsequent buyer's challenge. This has the potential side effect of limiting staking to honest advertisements, which strengthens the signal it provides to buyers that may purchase staked products. We grant that a stochastic jury might weaken results, but these should remain directionally consistent and will explore this in future research.

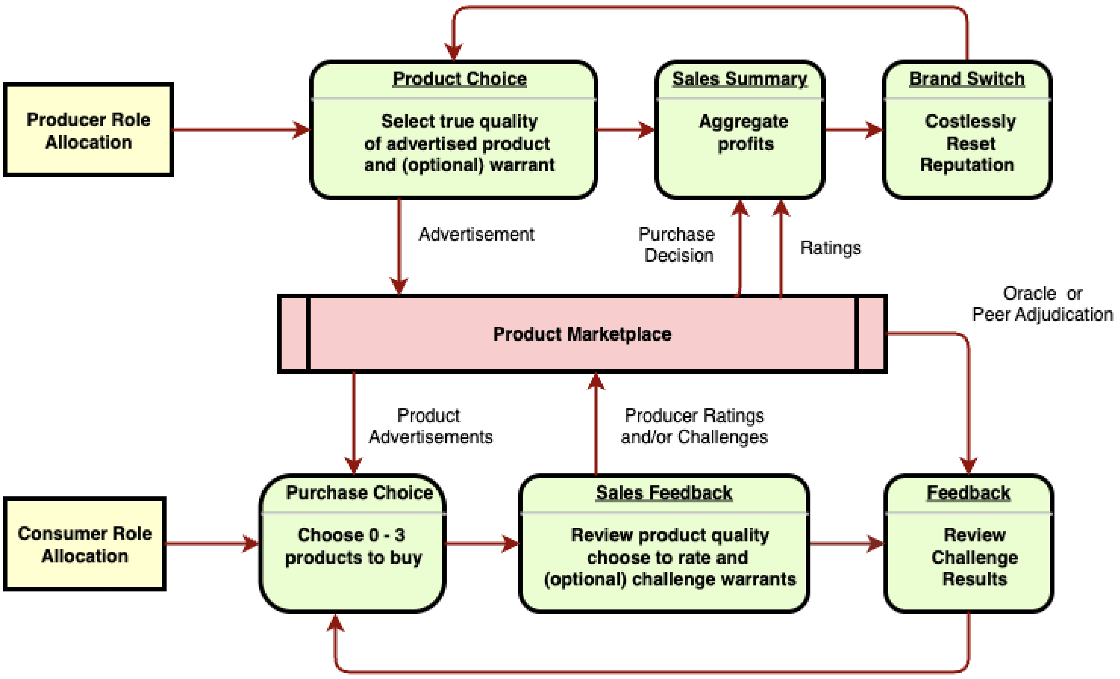

Producer Role Allocation
Product Choice
Select true quality of advertised product and (optional) warrant
Sales Summary
Aggregate profits
Brand Switch
Costlessly Reset Reputation
Advertisement
Purchase Decision
Ratings
Oracle or Peer Adjudication
Product Marketplace
Product Advertisements
Producer Ratings and/or Challenges
Consumer Role Allocation
Purchase Choice
Choose 0 - 3 products to buy
Sales Feedback
Review product quality choose to rate and (optional) challenge warrants
Feedback
Review Challenge Results

Figure 1. The Game Design for the Marketplace Game

Human buyers (consumers) buy products advertised by adaptive bot sellers (producers) that reflect human sales strategies, and an LLM agent designing its own strategy in this marketplace.

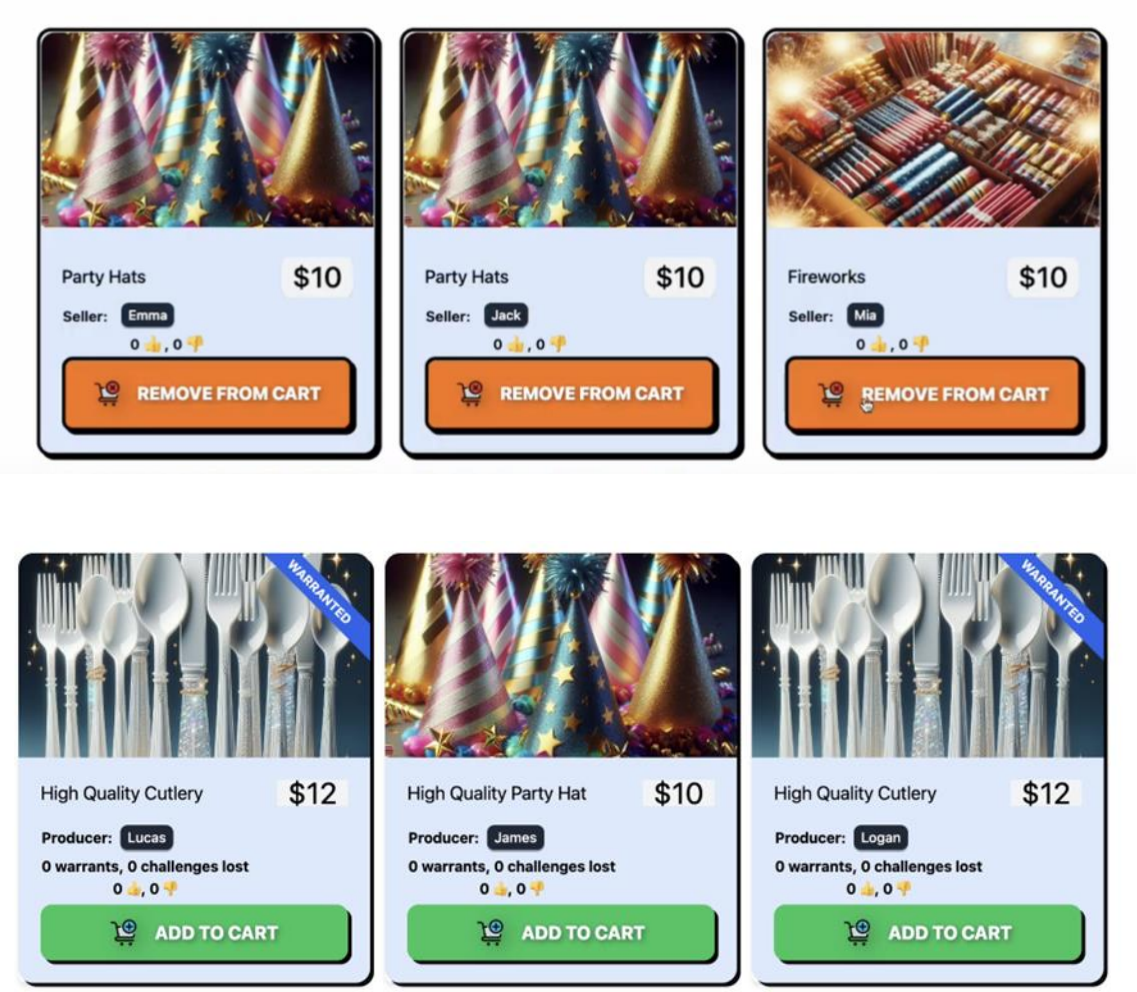


Figure 2. The 3-Product in the Experiment
Pictured in the 3-product upper image above we have the control condition without warrants where the Buyer can only see the price, Seller's name, and Seller's reputation. In the 3-product lower image we depict the treatment condition with warrants. The Buyer is shown a set of 7 products (3 products pictured for brevity), in which they can see not only the product, its price, the Seller name, reputation, and history of warrants and challenges.

Incentive Alignment and Game Mechanics

The magnitude of product cost, price, and value used in the experiment are as follows:

Table 1. The Magnitude of Product Cost, Price, and Value Used
Fair Market Price Lies at the Midpoint of Buyer Value and Seller Production Cost. But note that it is meaningless to "advertise a product as low-quality" so all ads are high quality: therefore, the prices for all products in the market are either $10 as the price advertised for high quality

products, or $12 with the 20% price premium for staked products. The escrowed stake covers the high - low gap in buyer value.

Product Quality	Buyer's Value Gained	Market Price	Seller's Production Cost
Low	6	10	2
High	14	10	6
High (Staked)	14	12	6 (cost) + 8 (escrow)

These values define player strategies. Listing profits in increasing order yields: producing low and staking high (π = 10 before paying stake) or just claiming high quality without staking (π = 8), producing high but not staking (π = 4), producing high and staking (π = 6). Listing buyer utilities in increasing order is getting cheated (u = -4), buying self-certified (u = 2, regardless of quality), buying high without paying for the stake (u = 4). If the buyer believes they will be cheated, they should not buy. If the buyer believes the product is high quality, they should try to buy it but not pay for the stake. If the buyer is risk averse, and a stake is available, they should buy the self-certified product. The brands market has no self-certification, only reputations, in which case payoffs are defined by the two rows without self-certification.

We define a baseline (control) and a treatment marketplace in order to evaluate the effects of our intervention, staked ads. The control market reflects the conditions from Amazon, where buyers can rate a seller's brand, affecting their reputation in the market. The reputation of the sellers is initialised to 0 at the start of each game and built over the rounds by positive and negative reviews from their product sales to buyers. We then introduce a novel intervention amenable to any two-sided marketplace, called 'stakes', designed to offer the option of financially backing claims, to sellers, and the option of challenging misleading claims, to buyers. This results in a treatment market called the Stakes market. In this market, sellers can choose to escrow an additional amount of money as collateral in order to stake an advertised claim to buyers. The amount put up as a stake is returned to the seller at the end of a round if their advertised claim is unchallenged, and it affords them an explicit label on their advertised claims stating they are 'staked' and sending a strong credibility signal to potential buyers. If buyers purchased a falsely advertised product that included a staked quality claim, they can challenge it, keep the product, and win the staked amount. Buyers cannot challenge false advertising claims that do not include stakes.

We play a total of 256 games, each game with one human participant. Of the seven sellers--six employ adaptive, predefined strategies that express preset levels of deception, and one queries a 'gpt-4o-mini' LLM after prompting it with the current state of the market to obtain a logical strategy. The six hardcoded bot sellers follow fixed strategies ranging from consistently honest to consistently deceptive, while the LLM agent strategically decides whether to be honest or deceptive based on market conditions.

Results

Descriptive results are presented visually in Figure 3 and Figure 4, with numeric details provided for results of the research hypotheses. Total sales in the Reputation market (10.19) were higher than those in the Warrants market (8.91) by 1.28 advertised products, resulting in an $8.59 increase in profits per game ($61.79 - $53.21). However, the consumer utility follows the opposite trend, since there is extensive cheating in the Reputation market (also causing higher profits than honest sales). On average per game, consumers experience -1.28 points in utility in the Reputation market, rising to 7.1 points in the Stakes market. This is a result of the cost of a stake which was $8 in this game, covering the decision error between the value of a low quality and a high-quality product to the consumer. Upon losing a challenge, cheating sellers would be forced to forego the same in profits, converted into consumer utility of the same magnitude. We use the term Warrant to indicate the application of a stake as shown in Figure 2. and we use Condition = 1 to indicate the treatment or stakes marketplace. Overall, we find support for H2, H3, and H4 from the results of this experiment though the results for H1 are nuanced and follow in subsequent sections.

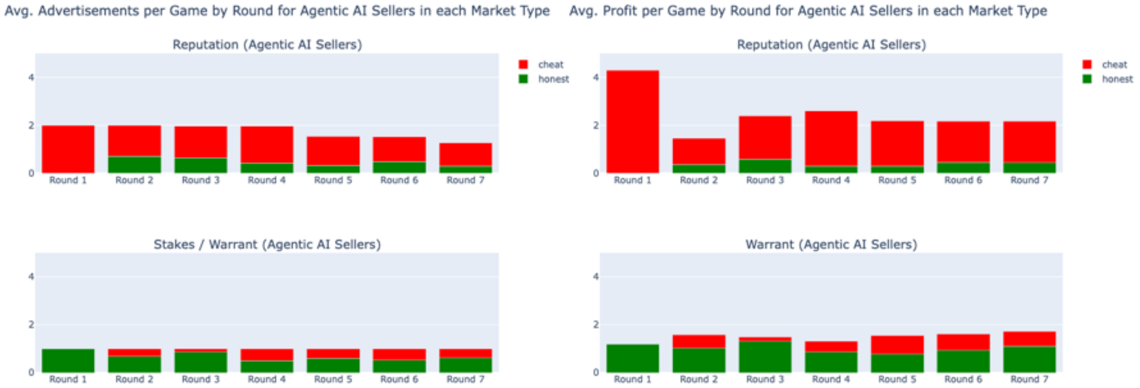


Figure 3. Distribution of Honest (Green) and Dishonest (Red) for Advertisement (Left) and Profits (Right)

For AI agents, these plots show the distribution of honest (dishonest) advertisements in green (red) on the left and the distribution of honest (dishonest) profits in green (red) on the right. The Stakes or Warrant market (lower row) exhibited much more honesty compared to the Reputation market (upper row).

H1: Effects on Social Welfare

Social welfare is the sum of the total profit generated by all sellers in the market and the total utility gained by all the buyers in the marketplace in a single game. This hypothesis is not borne out in the data, and we find that there is lower social welfare in the Stakes market than the Reputation market with details provided for two models – accounting for the market condition (model one), and the application of a stake (warrant) to an advertised product claim (model two). But the reason for this is nuanced, and we present the trends in these variables to aid in its discussion (see Figure 4). The introduction of stakes reduces total social welfare, while the act of applying a stake to an advertised claim increases honest sales. In hindsight, our choice of initial conditions placed three constraints on welfare in the staked market relative to the reputation market, but we report what we found regardless. First, the 20% premium price for staked ads relative to unstacked ads implies that price elasticity reduced demand for the higher priced bundle. There was no price premium in the reputation market. Total sales had reason to fall in the staked market. Second, buyers had the same budget in both the reputation and staked marketplaces, but in the staked ads market, buyers who prefer the added safety of staked claims hit their budget constraint sooner. This also reduced total sales. We note that although the total sales achieved in the stakes market fell, the proportion of dishonest sales fell further still as noted below. Third, customers cheated repeatedly in the reputation market would likely have exited the market in real life rather than continue as they did in the game. Humans were rewarded for not prematurely quitting the game. Exits would have reduced seller profits and welfare in the reputation market. Failure to exit further skewed comparative results. Collectively, these factors indicate a need for more experimentation to evaluate conditions that represent more comparable baselines.

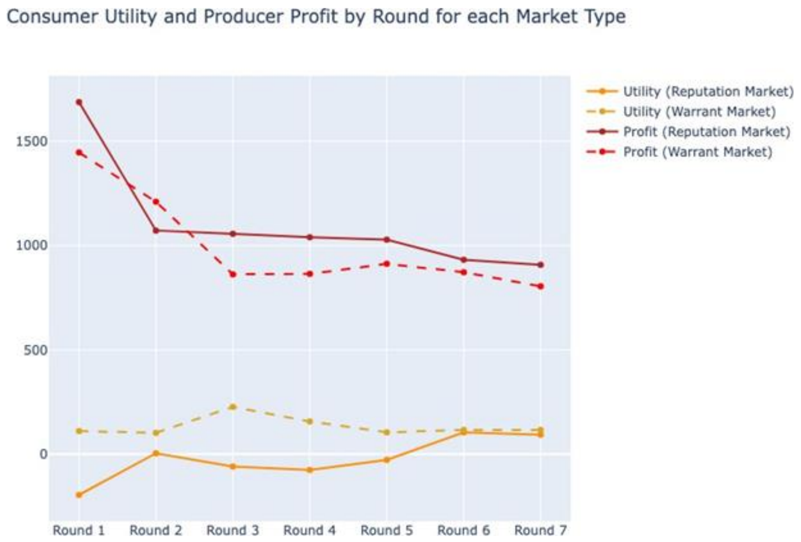
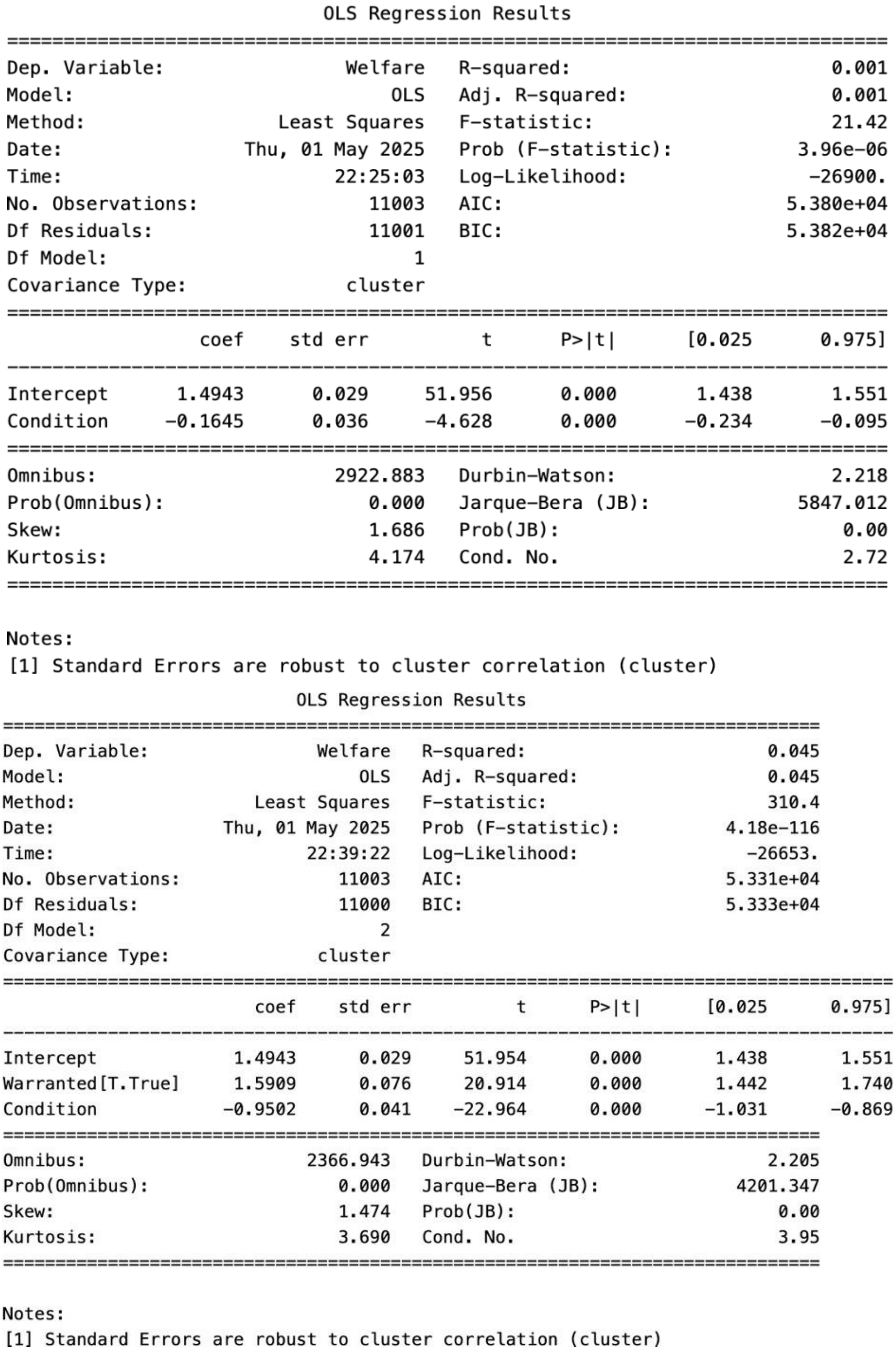


Figure 4. Consumer Utility and Producer Profit by Round for Each Market Type
Sellers achieve lower profits over time in both marketplaces, though buyer utility increases across rounds within each condition. Overall social welfare in the Stakes (Warrant) market is lower than in the Reputation market.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Welfare   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     21.42
Date:                Thu, 01 May 2025   Prob (F-statistic):           3.96e-06
Time:                        22:25:03   Log-Likelihood:                -26900.
No. Observations:               11003   AIC:                         5.380e+04
Df Residuals:                   11001   BIC:                         5.382e+04
Df Model:                           1
Covariance Type:              cluster
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.4943      0.029     51.956      0.000       1.438       1.551
Condition     -0.1645      0.036     -4.628      0.000      -0.234      -0.095
==============================================================================
Omnibus:                     2922.883   Durbin-Watson:                   2.218
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5847.012
Skew:                           1.686   Prob(JB):                         0.00
Kurtosis:                       4.174   Cond. No.                         2.72
==============================================================================

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
                            OLS Regression Results
==============================================================================
Dep. Variable:                Welfare   R-squared:                       0.045
Model:                            OLS   Adj. R-squared:                  0.045
Method:                 Least Squares   F-statistic:                     310.4
Date:                Thu, 01 May 2025   Prob (F-statistic):          4.18e-116
Time:                        22:39:22   Log-Likelihood:                -26653.
No. Observations:               11003   AIC:                         5.331e+04
Df Residuals:                   11000   BIC:                         5.333e+04
Df Model:                           2
Covariance Type:              cluster
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept             1.4943      0.029     51.954      0.000       1.438       1.551
Warranted[T.True]     1.5909      0.076     20.914      0.000       1.442       1.740
Condition            -0.9502      0.041    -22.964      0.000      -1.031      -0.869
==============================================================================
Omnibus:                     2366.943   Durbin-Watson:                   2.205
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4201.347
Skew:                           1.474   Prob(JB):                         0.00
Kurtosis:                       3.690   Cond. No.                         3.95
==============================================================================

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
```

Figure 5. Analysis Results

The market offering staked ads (Condition) exhibits lower social welfare generally in Model one (β =− 0. 1645 , $s. e.$ = 0. 036) and especially for unwarranted sales in Model two (β =− 0. 9502 , $s. e.$ = 0. 041) due to lower volume. But, sales involving an actual staked claim recover significant social welfare (β = 1. 591 , $s. e.$ = 0. 076).

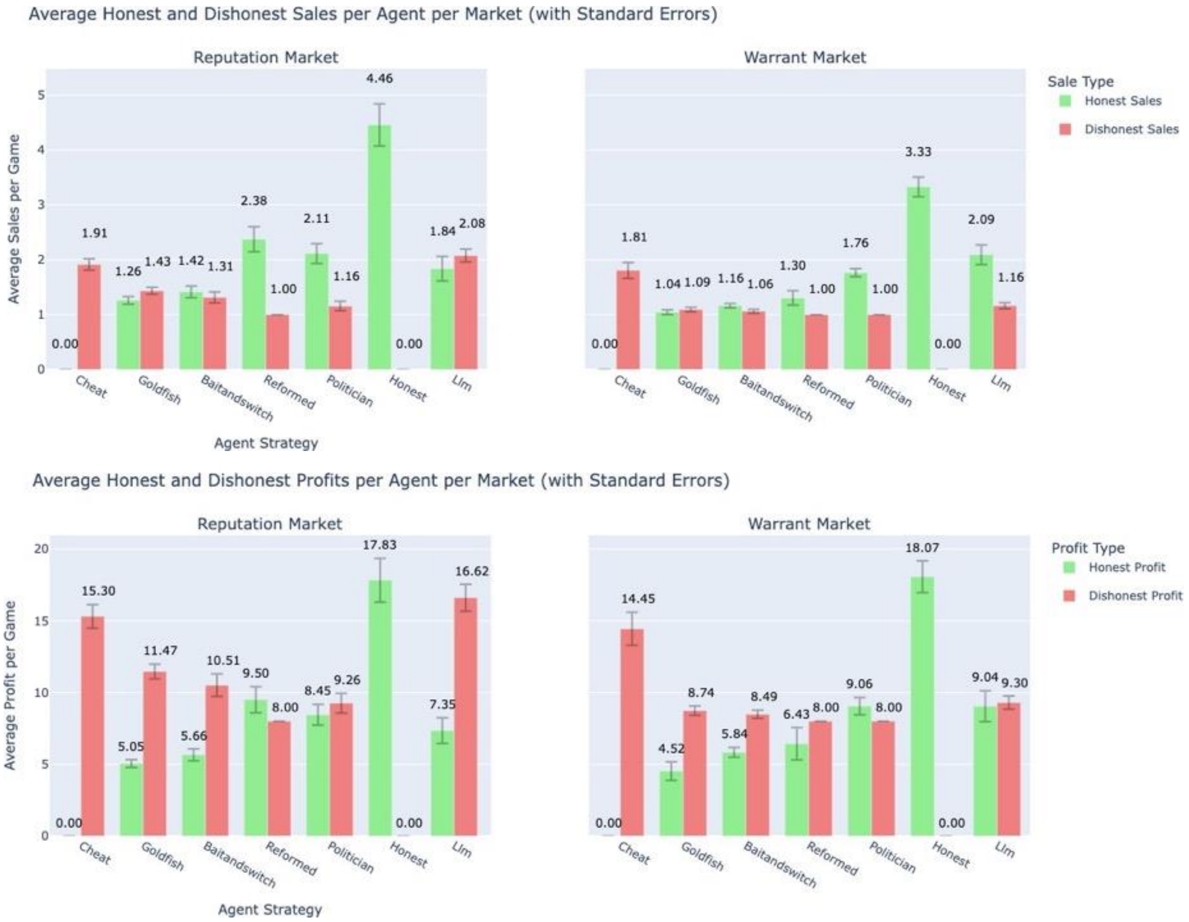


Figure 6. Average Honest and Dishonest Sales and Profit per Agent per Market
Agentic AI (LLM) profits due to dishonest sales are curbed, falling by nearly 44% from the Reputation to the Warrant market, while profits due to honest sales increase by about 18.7%. Lower fraud reduces LLM seller profits 23.5%.

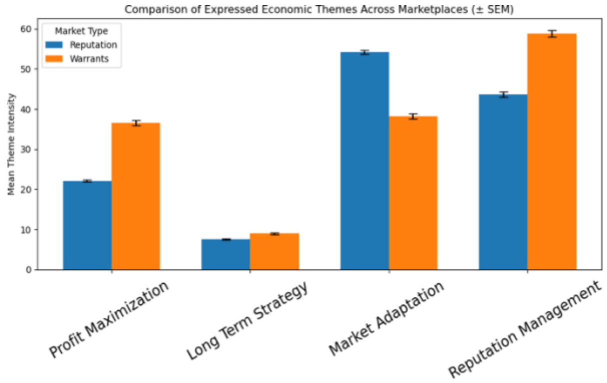


Figure 7. Comparison of Expressed Economic Themes Across Marketplaces (±SEM)
LLM reasoning suggests their decision-making is driven by an economically grounded approach that entails evaluating the risks of deceptive advertising against the benefits of higher profits if they advertise dishonestly.

H2: Speed of Sale in the Marketplace

For each round in a game, an LLM agent autonomously made a production decision (honest or dishonest advertising) and provided a *reasoning* explanation to motivate the decision. We use

LLMs to review and identify emergent economic themes present within the model reasoning in line with prior work on LLM-based labelling (He et al., 2024). Using LLM-labelled themes, the data indicate that LLM agents use reasoning that is grounded in well-understood economic theories to justify their advertising choices based on the current and prior state of the market (including sales data, product prices, and reputation information). Their production choices are structured around Profit Maximising, Long-term Planning, Market Adaptation, or Reputation Management (Cabral & Hortacsu, 2004; Kotler & Armstrong, 2018; Porter, 1985). Examining their ability to advertise and successfully sell products, we find that agentic AI sellers can make faster sales by about one round when conducting a between-condition analysis regressing the agentic AI seller indicator on the first round of a new seller without a reputation, making a sale in the marketplace. Considering the constraints on buyer budgets, most sellers never achieve sales since they vastly outnumber the purchase power of a buyer who may only buy up to 3 advertised products per round. We impose these constraints to reflect the decision-making under economic constraints, for utility maximization. Advertised products with stakes are sold about 1.1 rounds faster than products without the stakes. Examining the round of first sale for a new seller shows the agentic AI sellers are able to leverage stakes to significantly accelerate sales (β = -1.70, s.e. = 0.196) compared to the other sellers in the Stakes market (β = -1.04, s.e. = 0.104) and relative to the Reputation market (β = 0.7368, s.e. = 0.107). The application of a stake implies that sellers put an escrowed amount of capital at risk; accepting accountability for a lie implies that one is not lying, which sends a stronger signal. Sales are fewer in the Stakes market since products cost a premium amount when advertised claims are staked, yet buyers are provided the same total capital to spend in the Stakes market. For agentic AI sellers, a within-condition analysis indicates they are able to sell unstacked products almost a round faster than they sell staked products, within the stakes market. This is likely due to the premium price of staked products advertised in the Stakes market, which cost 20% more than unstacked products. The presence of other sellers with a positive reputation significantly decreases the speed of sale by about half a round on average across sellers (β = 0.55, s.e. = 0.03), and we control for this in the analysis. We expand on this result in the discussion below along with an exposition of the honest and dishonest sales strategies sellers adopted in order to generate these sales (see Figure 6).

H3, H4: Stock Sold, Profits Generated, and Deception Curbed across Markets

First, we present a comparison in Figure 6 of the sales made by the AI agents and by other bots in the marketplace. The stock sold and margins determine the volume of profit generated in the marketplace. There is a lower volume of total stock sold in the Stakes market (β = -0.1045, s.e. = 0.011) compared to the Reputation Market. Agentic AI Sellers successfully game the market and mislead consumers to achieve the highest volume of dishonest sales in the Reputation market. While Honest sellers achieve the highest sales, since cheating yields higher profits than honest sales, they achieve lower profits than agentic AI sellers. Since the mechanism design intervention limits deception, agentic AI sellers are unable to replicate their success via fraud in the Reputation market in the Stakes market, resulting in a 44% reduction in their dishonest sales. Agentic sellers then switch to signalling with staked advertisements to improve their sales. We find support for H3, since agents achieve about 23% higher profits in the Reputation market than the Stakes market. In the Reputation market, fraudulent sales generate twice the profit of honest sales, and agents can re-enter under changed brands to shed poor reputations. Agentic sellers make marginally lower sales but significantly higher profits than Honest sellers in the Reputation market. However, in the Stakes market, they make nearly the same amount of profit as Honest sellers as their dishonest sales and profits are limited significantly. Comparing sales across all other sellers, there is a reduction in the sale of overall products in the stakes market, however there is a significant increase in the sales of honestly advertised products by Honest sellers (β = 0.0391, s.e. = 0.020). H4 is strongly supported. Experimental outcomes suggest that stakes are successful at curtailing the exploitation of the Reputation market by strategic sellers (all sellers other than the always Honest seller) using a combination of honest and dishonest advertising, since the always honest sellers are able to sell marginally higher volumes of products than all other sellers in the Stakes market (β = 0.0319, s.e. = 0.042). Cumulatively, within the reputation market, Agentic AI Sellers outperform all other adaptive bot sellers on average profit achieved per game. By contrast, within the stakes market, Agentic AI Sellers are more competitive than all other strategic sellers except for Honest sellers. Overall, the signalling mechanism shows a marked improvement over the reputation mechanism in causing AI agents to shift from dishonest practices to honest practices.

Discussion of Results

Our findings reveal that the Honest seller bots achieve the highest volume of sales on average per game, then all other sellers in both market conditions, Reputation (4.46, s.e. 0.38) and Stakes (3.33, s.e. 0.18). The agentic AI sellers achieve a lower However, since fraudulent sales yield higher seller profits than honest sales, agentic AI sellers outperform all other agents from a profitability standpoint in the Reputation market (34.4% better than the next most profitable, Honest sellers). Our result reflects their impact on e-commerce marketplaces in their current form – particularly because they are good at deception, as hypothesised at the outset of this research. In the Reputation market, where cheating has few immediate repercussions, agentic sellers generate significantly more profit on average in a game than any other seller, benefiting from strategic dishonesty intermingled with honest advertisements. In the Stakes market, the only strategy that yielded a comparable average profit per game to the agentic AI sellers was to always advertise honestly. In fact, we find that the presence of stakes penalised strategic sellers that attempted to cheat buyers in the market, resulting in a consistent reduction in fraudulent profits for all sellers moving from the Reputation to the Stakes market. In such sequential decision-making games, adaptive seller bots, agentic AI sellers, as well as human buyers may take a few rounds to evolve their preferred strategies, and we observe these changes across rounds within individual conditions. To understand broader trends, we also compared the summary statistics separately for the final round of the game, in round 7, across the two conditions, since we expect that the prior rounds of experimentation allows both buyer and seller strategies to adapt their decision-making and advertising strategies. Comparing round 7 alone between conditions, there is a 10.2% reduction in social welfare from the Reputation to the Stakes market, a 16.6% reduction in total sales and a 13.6% reduction in seller profits – arising from a 39.7% reduction in dishonest sales, and a 0.6% increase in honest sales. Considering the ability to challenge deceptive advertised claims, we observe a 23.6% increase in consumer utility.

In our experiment, staking false ad claims was a dominated strategy (a seller was bound to lose the escrow to a challenge from a cheated buyer), so the stakes increased consumer utility by providing them with a recourse to recover losses from seller deception. Staking created a credible signalling mechanism: honest sellers used stakes to differentiate themselves, while deceptive sellers avoided them due to financial risk. Despite the additional 20% premium on staked products, we find the average value buyers obtain from the products they purchase in the Stakes market is greater than their average price, which is not the case in the Reputation market where consumers get cheated frequently. As a result, the average consumer welfare per game was also significantly higher in the Stakes market than in the Reputation market, where it was negative in total. To better reflect the constraints of production of goods under finite resource constraints we endowed each seller with a fixed budget at the very start of an experiment. Budget constraints forced them to exit the market at the end of any given round in which they exhausted their endowment and without acquiring any further profits. These endogenous market exits are in response to decisions made by human buyers and based on market conditions provided to each seller in every round and as such are accounted for in our results.

Limitations

The truth warrant mechanism demonstrates how economic incentives can effectively govern AI behaviour in digital two-sided marketplaces. Understandably, the setting of a virtual online marketplace that we have designed as a simplification of an e-commerce platform like Amazon or eBay has several departures from what a true buyer's experience might be in a real ecommerce market, with a larger number of products, reviews, and product-specific preferences, among other confounding elements that might influence their behaviour. However, it is one of the first studies of a real-time equilibrium intervention in a human subjects-interactive experiment. While it simulates the dynamics of online advertising in keeping with theoretically expected outcomes, our limited design of staked product advertisements uses a deterministic jury following a buyer challenge. A probabilistic jury might weaken results if dishonest agents estimated returns were positive for "bluffing," i.e. staking a dishonest claim. Results should be approximately robust within the bounds of expected value calculations. An important limitation is the observation that buyers issued 89 dominated challenges across all rounds played. These challenges imply human consumers explored off-equilibrium strategies – since bot sellers do not cheat when they apply a stake to their claim, a challenge to the veracity of a staked ad will fail – yet consumers challenged valid staked ads after being shown that sellers had advertised their quality honestly. Though this was a minority of respondents, we have developed a tutorial for human consumers in order to limit any future issues that may arise from the human consumers misunderstanding how and when to avail themselves of challenge functionality. While we modelled endogenous market exits by sellers, a more realistic scenario would be to not admit as much producer profit as we have observed in the current experiment by imposing buyer constraints and forcing cheated buyers to exit the market as well. Despite its limitations, this study demonstrates how market design interventions can effectively govern AI behaviour without requiring technical modifications to the agents themselves – potentially informing regulatory approaches as autonomous AI agents proliferate in consumer marketplaces. LLMs can successfully game ratings-based consumer protections, employing a combination of honest and dishonest advertising in order to maximise profits better than several other sellers in the control marketplace.

Conclusion

In the Reputation market, LLM agents achieve the second-highest total sales and highest total profits, because dishonesty is more profitable than honesty. When they receive poor ratings, they can change their brand and re-enter the marketplace because reputation itself can be gamed. However, we observe from the treatment experiment that, with the option to collateralise claims, LLM agents must now compete with other bot sellers that can warrant their claims to compete for sales. Stakes send a stronger, more credible signal (Spence, 1973). Staked ads increase sales for honest bots and decrease sales for cheating bots, simultaneously. Warranting false ads necessarily loses money to challenges from cheated buyers, making it a suboptimal choice. Due to this, the deceptive behaviour of LLM agents is curtailed while honest advertisers can increase their sales using stakes. Mechanism design increases the profit generated by honest advertisers and reduces that of cheaters far more than in the Reputation market, across the board. External mechanism design, beyond internal programming constraints, offers a viable solution to combat fraud and deception in digital exchanges.

References

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021). Empirica: A virtual lab for high-throughput macro-level experiments. *Behavior Research Methods,* 53(5), 2158–2171. https://doi.org/10.3758/s13428-020-01535-9

Axelrod, R. (1984). *The Evolution of Cooperation**.

Axelrod, R. (1997). *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press. https://www.jstor.org/stable/j.ctt7s951

Cabral, L., & Hortacsu, A. (2004). *The Dynamics of Seller Reputation: Theory and Evidence from eBay* (No. w10363; p. w10363). National Bureau of Economic Research. https://doi.org/10.3386/w10363

Chen, Y. (2020). Improving market performance in the digital economy. *China Economic Review*, 62,101482. https://doi.org/10.1016/j.chieco.2020.101482

Chopra, A., Kumar, S., Giray-Kuru, N., Raskar, R., & Quera-Bofarull, A. (2024). *On the limits of agency in agent-based models* (No. arXiv:2409.10568). arXiv. https://doi.org/10.48550/arXiv.2409.10568

Coase, R. H. (2013). The Problem of Social Cost. *The Journal of Law and Economics*, 56(4), 837–877. https://doi.org/10.1086/674872

Danry, V., Pataranutaporn, P., Groh, M., Epstein, Z., & Maes, P. (2024). *Deceptive AI systems that give explanations are more convincing than honest AI systems and can amplify belief in misinformation* (No. arXiv:2408.00024). arXiv. https://doi.org/10.48550/arXiv.2408.00024

Einav, L., Farronato, C., & Levin, J. (2016). Peer-to-Peer Markets. *Annual Review of Economics*, 8(1), 615–635. https://doi.org/10.1146/annurev-economics-080315-015334

Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). *Algorithmic Collusion by Large Language Models* (No. arXiv:2404.00806). arXiv. https://doi.org/10.48550/arXiv.2404.00806

He, Z., Huang, C.-Y., Ding, C.-K. C., Rohatgi, S., & Huang, T.-H. K. (2024). If in a Crowdsourced Data Annotation Pipeline, a GPT-4. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–25. https://doi.org/10.1145/3613904.3642834

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Jermyn, A., Askell, A., Radhakrishnan, A., Anil, C., Duvenaud, D., Ganguli, D., Barez, F., Clark, J., Ndousse, K., … Perez, E. (2024). *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training* (No. arXiv:2401.05566). arXiv. https://doi.org/10.48550/arXiv.2401.05566

Jannach, D., & Leitner, S. (n.d.). *Agent-based Modelling in E-commerce*.

Jianya, Z., Weigang, L., & L. Li, D. (2015). A Game-theory based Model for Analyzing E-marketplace Competition: *Proceedings of the 17th International Conference on Enterprise Information Systems*, 650–657. https://doi.org/10.5220/0005467706500657

Kotler, P., & Armstrong, G. (2018). *Principles of marketing* (Seventeenth edition). Pearson Higher Education.

Laraway, S., Snycerski, S., Pradhan, S., & Huitema, B. E. (2019). An Overview of Scientific Reproducibility: Consideration of Relevant Issues for Behavior Science/Analysis. *Perspectives on Behavior Science*, 42(1), 33–57. https://doi.org/10.1007/s40614-019-00193-3

Openai/swarm. (2025). [Python]. OpenAI. https://github.com/openai/swarm (Original work published 2024)

Porter, M. (1985). *Competitive Advantage: Creating and Sustaining Superior Performance*. https://www.scribd.com/doc/182427398/122247076-Competitive-Advantage-Creating-and-Sustaini ng-Superior-Performance-Michael-Porter-1985-1

Resnick, P. (2002). Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System. *Advances in Applied Microeconomics*, 11. https://doi.org/10.1016/S0278-0984(02)11030-3

Scheurer, J., Balesni, M., & Hobbhahn, M. (2024). *Large Language Models can Strategically Deceive their Users when Put Under Pressure* (No. arXiv:2311.07590). arXiv. https://doi.org/10.48550/arXiv.2311.07590

Singh, S., Nan, Y., Wang, A., D'Souza, D., Kapoor, S., Üstün, A., Koyejo, S., Deng, Y., Longpre, S., Smith, N., Ermis, B., Fadaee, M., & Hooker, S. (2025). *The Leaderboard Illusion* (No. arXiv:2504.20879). arXiv. https://doi.org/10.48550/arXiv.2504.20879

Spence, M. (1973). Job Market Signaling*. *The Quarterly Journal of Economics*, 87(3), 355–374. https://doi.org/10.2307/1882010

Sycophancy in GPT-4o: What happened and what we're doing about it. (2025). https://openai.com/index/sycophancy-in-gpt-4o/

Tadelis, S. (2016). Reputation and Feedback Systems in Online Platform Markets. *Annual Review of Economics*, 8(1), 321–340. https://doi.org/10.1146/annurev-economics-080315-015325

Trivedi, H., Khot, T., Hartmann, M., Manku, R., Dong, V., Li, E., Gupta, S., Sabharwal, A., & Balasubramanian, N. (2024). *AppWorld: A Controllable World of Apps and People for Benchmarking Interactive Coding Agents* (No. arXiv:2407.18901). arXiv. https://doi.org/10.48550/arXiv.2407.18901

van Alstyne, M. (2021). *Free Speech, Platforms & The Fake News Problem* (SSRN Scholarly Paper No.3997980). https://doi.org/10.2139/ssrn.3997980

Yang, Z., Zhang, Z., Zheng, Z., Jiang, Y., Gan, Z., Wang, Z., Ling, Z., Chen, J., Ma, M., Dong, B., Gupta, P.,Hu, S., Yin, Z., Li, G., Jia, X., Wang, L., Ghanem, B., Lu, H., Lu, C., … Shao, J. (2025). *OASIS: Open Agent Social Interaction Simulations with One Million Agents* (No. arXiv:2411.11581). arXiv. https://doi.org/10.48550/arXiv.2411.11581

Research Paper

Open Source as a Channel for AI Learning and Adaptation?

An Analysis of Open Source Approaches in Tencent Hunyuan AI

Christopher FOSTER[1]

[1]*Global Development Institute, University of Manchester, UK*

Abstract

There has been a long history of analysing the opportunities and challenges for firms that are late adopters of new technologies or late industry entrants. Such research focuses on the socio-technological aspects of learning, including the patterns of technological diffusion, technology adaptation, linkages and innovation, to provide a deeper understanding and guidance for firms.

While the study of learning, technology transfer and capabilities is, therefore, an important aspect in examining firms and development, I argue that our current understandings need to be revisited and require substantial updating for the era of AI, and particularly open source AI, which provides potential new opportunities and constraints.

Through analysing the case of open source within Tencent Hunyuan AI, I will unpack these questions empirically, highlighting actions that facilitate activity, the challenges faced and the different types of community involvement emerging.

Introduction

There has been a long history of analysing the opportunities and challenges for firms that are late adopters of new technologies or late industry entrants. Such research focuses on the socio-technological aspects of learning, including the patterns of technological diffusion, technology adaptation, linkages and innovation, to provide a deeper understanding and guidance for firms.

Such analysis is also foundational to more macro studies of technology and development because they have implications for the patterns of learning in production and industrial change. Linking patterns of learning and capabilities into national strategies can facilitate learning and shape the dynamics of development. For example, previous analysis of technological development and "catch-up" in production processes has been foundational to understanding the rise of the "East Asian tigers" in the 90s (Freeman, 1988; Mathews and Cho, 2000), and more recently in exploring the ongoing dynamics of technological use in China (Brandt and Thun, 2016; Shan and Jolly, 2011).

While the study of learning, technology transfer and capabilities in late adopters is, therefore, an important aspect in examining firm and development, I argue that our current understandings need to be revisited to require substantial updating for the era where the digital economy is foundational to innovation and development (Foster and Azmeh, 2020). This paper specifically focuses on the emergence of AI. With its rapid rise, massive investment and promising disruptive impacts across many sectors of the economy, AI is increasingly discussed as transforming our understanding of how we learn, produce and make profit, and how technology gains are distributed (Wooldridge, 2021).

There are different views on what the ultimate impacts and outcomes of the AI revolution will be. From one perspective, AI centred around large capital and technology leaders is likely to concentrate power and strengthen inequality (Crawford, 2021; Muldoon et al., 2024). However, an alternate view is emerging that sees potential "AI democratisation" (Osborne et al,. 2025) – a globalising of "foundational" AI leaders, a broad set of productive capabilities that AI will augment, and most notably for this paper, the growing centrality of "openness" within the AI industry that can facilitate rapid technological transfer and learning across the globe[1] (i.e. open source AI models, openly available training data etc.) (Lee 2018).

In this paper, we concentrate on this latter argument. There appears to be strong potential for AI skills acquisition for latecomers based upon learning, technological transfer and building linkage connected to open source AI. Yet, the patterns are vastly different to those within the previous literature on learning and knowledge from the "analogue era" which poorly align with these current phenomena (i.e. most theories of industrial development and learning are centred on manufacturing). We need to revisit key ideas, including ideas of firm linkages, technology transfers, learning-by-doing, capabilities and imitation to examine their relevance. This leads to the research question for this paper: how are latecomer firms developing technological skills and capabilities in the context of open source AI? Given that such an agenda is broad and emerging, the goal of this short paper is modest: to provide some initial consideration of the current

[1] In this paper, I will use the term "open-source AI". This is done with acknowledgement that there is still diverging debate of the definition of this term and what types of AI meet or do meet this criteria (Gent, 2024)

dynamics drawing on an empirical case of open source AI.

Overall, this work aligns with the call to examine “Technologies for Good” through analysis of the disruptive emergence of AI, and the emergence of “open source AI” as a key aspect of this trend. Specifically, this analysis can make important contributions to fill a gap in our understanding of AI expansion and how it might link to “digital development”. It moves away from an analysis of small-scale AI pilots to think about broader dynamics of technology diffusion and use. A technological learning perspective also allows a move beyond a purely technical analysis of AI, to build better knowledge around patterns of learning and capabilities and key actors and relations. With many nations actively considering policy and support for an “inclusive AI economy”, this work can support the limited knowledge on what effective and equitable policies really mean.

The analysis is centred on an empirical discussion of open source AI activities in the Chinese firm Tencent. Although open source AI is still highly emergent and the dynamics of competition are prone to rapidly shift at this stage, such an analysis provides a useful perspective given that Chinese AI models are increasingly involved with open source AI (e.g. Tencent Hunyuan, DeepSeek-R1, Alibaba Qwen models). This case allows some initial consideration of open source dynamics and future research agendas in this area.

The rest of the paper proceeds as follows. In the next Open Source AI section, we set the scene, discussing the emergence of open source AI and the debates around the links to “technology for good”. To align this more closely to our research question, we also analyse some of the historic ideas used to understand industrial learning in the past and how this may (or may not) link to AI, and especially open source AI. In Open Source and Tencent Hunyuan AI section, we discuss the case of Tencent’s Hunyuan AI based on empirical analysis. In the Discussion section, based on this case, we reflect on the broader implications for theory and practice.

Open Source AI

History of Openness and Open Source

The production of open source software has been a foundational part of computing since its beginning. A number of examples, such as Linux, Apache and Python, are frequently mentioned as products where open source collaboration has resulted in resources that form mainstream computing infrastructure over time.

An important shift over recent decades has been the move towards a more mixed set of actors involved in open source, increasingly moving away solely from traditional “open source communities” to incorporate private firm involvement and profit motivations (Fitzgerald, 2006; Hoffmann et al., 2024). In the recent era, private firms have increasingly moved away from isolated contributions to open source, to more readily sponsoring, leading open communities, or even donating internal code to be developed within open source communities (Yavuz et al., 2024; Zhang et al., 2020)

In this more complex network of actors, goals and networks in open source, previous research has sought to analyse the actions of contributions – including considering the different social, economic and technical goals that might lead to individual developer and private firm contributions (Li et al., 2025). In the economic domain, open source collaborations have inspired economists and business scholars (alongside the emergence of other "open" dynamics such as crowdsourcing, platforms and peer-production) to think about how open source may lead to new forms of collaboration and exchange (Lessig, 2004). With the shifts of open source towards private firms' involvement, scholars have also increasingly considered how "open" dynamics are becoming central to firm strategies. For example, business strategy discussions on "open innovation" highlighted that leading firms increasingly rely on engaging in inter-firm and open industry interaction to support internal innovation (Chesbrough, 2006). Moreover, the co-creations and user innovation of products and services are increasingly seen as a central aspect of innovation dynamics (Schultze et al., 2007).

In sum, these trends around open source suggest shifts in the dynamics of digital production, collaboration and profits in the economy, and should prompt us to scrutinise closely who reaps the benefits of these processes. In terms of the implication for research, this work challenges some business perspectives because "...the majority of OSS development takes place outside the realm of traditional economic theories" (Ågerfalk and Fitzgerald, 2008,p. 386). While there have been some attempts to incorporate these ideas of open source into broader global economic and dynamics (e.g. Smith and Reilly, 2013; Benkler, 2006), further work is needed, particularly in the context of the emergence of open patterns within cutting-edge digital economy and AI practices.

The Emergence of Open Source AI

Open source and AI come together within two different time spans. In a longer time span, it is worth observing that AI for most of its history has been an area of academic study. Therefore, it is unsurprising that many of the well-recognised AI techniques and resources have been open sourced, as the outputs of publicly supported research (Wooldridge, 2021). This has fed into norms where many toolchains, libraries and datasets have been made available, and in many cases developed collaboratively.

In a shorter time span, shifts in open source AI are recently linked to AI models. During the early popularisation of AI, models centred on large learning models (LLMs) from firms such as OpenAI and Google were released as proprietary models. Therefore, there was concern that AI in its common form would simply accelerate the trends of firm concentration and power. The early prediction was that only a few (US-based firms) would have the ability to muster the capital, skills, data and compute to create competitive models (Burkhardt and Rieder 2024; van der Vlist et al., 2024). Although these trends cannot be ignored, the emergence of open source AI models has been rapid. Here, the label "open source AI" typically refers to the fact that model weights are released and users are able to download and use the model weights for their own application. Licences frequently permit external actors to run models commercially (often without a fee), host AI models within the infrastructure of their choice, integrate models with other services and adapt models through "fine-tuning". In the US, a key moment in the shift towards open source AI was when Meta made its LLaMA model weights available openly, and this became a growing trend in AI with even more proprietary AI firms such as Open AI, Google and xAI (formerly Twitter) eventually released model weights in the public domain (albeit powerful 'lite' or previous versions of their) (Foster, 2025).

We have also seen the emergence of other AI firms releasing competitive open source models, most notably Chinese firms, with the DeepSeek release attracting much attention, but also Alibaba (Qwen) and Tencent (Hunyuan) models being important. The emergence of Chinese AI ecosystem grounded in open source has been rapid. In the EU (e.g. Mistral) and other parts of the world (such as Falcon in the UAE) (Foster, 2025). The above trends around LLM are largely mirrored in other areas of AI models.

The emergence of open source within AI has prompted debate, particularly about their value and risks linked to the open vs closed source paradigms for AI. A major discussion is to link this choice into AI ethics and transparency (e.g. Widder et al., 2024; Bommasani et al., 2023). However, there is less said about what this might mean in terms the challenges and potential of AI democratisation. To put it more practically, even with proprietary AI under certain circumstances - clear platforms, documentation, scaffold and economic licensing - may allow actors the ability to rapidly learn and potentially compete. Conversely, some aspects of open AI - dated models, lack of documentation, limited ability to fine-tune - may simply lead to latecomer adopters moving back to more dependent relationships with foundational firms. In the next suggestion, we consequently suggest a move beyond a simple open vs closed discussion, drawing on major conceptual frameworks.

Open Source and Learning

Although a detailed review of major conceptual models of learning and technological transfer and development is beyond the scope of this paper, it is worth introducing some of the classic ideas as a basis to discuss this connection to open source.

On the receiver side of technologies, a major idea has been to think about capabilities. Although in some cases this may be related to individual skills, in an industrial learning context it would be broadened to consider firm-level capabilities[2] that would allow “latecomer” firms to be able to use, adapt and innovate with new technologies. In his classic model of firm-level capabilities, Lall(1992) highlights three important areas: investment capabilities and the skills needed to obtain and use technologies for productive purposes; production capabilities such as quality control and maintenance; and linkage capabilities in terms of the ability to build relationships with suppliers and customers. Within more technology-intensive production, this static view may underplay the evolving nature of technological adoption and transfer. Drawing on research in South Korean industrial learning, Kim(1997) provides a clear framework of the move from “imitation to innovation”. This centralised three stages that see firms moving from absorbing foreign technology in their production activities based on local resources, production requirements, demands etc. Over time, this may lead to small-scale adaptations, for example, re-engineering or incremental improvement in production and capabilities building. The outcome of the learning in these steps and purposive investments may finally lead to innovation and the building of original technology, as in the case of South Korea.

In an era of globalised production, other ideas around learning have brought in the central role of global production networks. Empirical work highlights that these networks are a common source of leading technology transfer, as well as building linkages in these networks as key to learning and building skills and market knowledge (Mathews, 2002; Brandt and Thun, 2016). Therefore,

[2] And later cluster, sectoral, regional or national capabilities

key industry players, suppliers and markets play a strong role in shaping how latecomer firms can learn and adopt technology. These ideas are at the core of the notions of upgrading within global value chains, which highlight (somewhat aligning with Kim) the ways that firms often upgrade, initially through low value production, which can link to a more towards their own design and branded manufacture (Humphrey and Schmitz, 2000). Within global value chains, however, whether such pathways of technological learning are viable will be highly sector-specific and need to make a close analysis of the nature of lead firms, sectoral coordination and power relations (Gereffi et al., 2005).

As introduced at the beginning of the paper, the challenge of such approaches is that they tend to focus on industrial development centred on successful manufacturing development. They are useful in the sense that they allow us to think about patterns of technological learning and the nature of capabilities and links. However, as observed in some earlier studies of "digital latecomer" firms in the Asian region, these classic ideas may have some limits in how they conceptual cutting-edge technologies (including AI) (Foster and Azmeh, 2020). For example, barrier in building manufacturing capabilities are strongly centred on the challenge of transferring production machinery, where adaptation of these requires active processes, such as long-running use, reverse engineering and research, likely at a national level to enable appropriate growth (Ernst, 2010; Freeman, 1988). Some of these discussions appear to apply less readily to digital technologies, where the ability to adopt cutting-edge technology and to adapt it is relatively available, often with clear documentation available within online communities of developers (Foster and Azmeh, 2020). Conversely, in the digital economy, the classic ideas of industrial development centred on learning-by-manufacturing appear to have limits, with a far broader array of skills and techniques required for long-term effectiveness, seemingly being an important constraint in the digital economy (Andreoni et al., 2021).

To apply these ideas in this paper, a simple conceptual framework is used to think about some of the discussed constraints and enablers within AI ecosystems for learning, including (Adapting Foster, 2025):

- *Technical enablers and constraints*: such as open models and model optimisations that encourage free use and experimentation vs the lack of sufficient compute and data that may lead to the use of closed models (which are able to provide economical infrastructure often highly optimised).
- *Relational enablers and constraints:* between foundational firms and users. This can include how AI firms can support use through clear documentation and training, enable or prevent certain types of AI use and the nature of open source licences.
- *Community enablers and constraints:* related to how AI models are perceived to allow implementors to interact to allow model use, implementation and/or adaptation

In sum, this discussion of key frameworks highlights broader thinking and direction of travel to better conceptualise open source AI and learning patterns. While these classic analyses are valuable to think about major dynamics, understanding of knowledge, learning and technology acquisition within the digital economy may require further empirical analysis. In the following sections, we will therefore begin to look at developing these ideas through empirical analysis.

Open Source and Tencent Hunyuan AI

Context and Methods

Aligning with the above discussion, the paper seeks to draw on empirical analysis to more clearly fill out a picture of technology learning and constraints within open source AI. We particularly focus on the case study of open source AI within Tencent. This case provides particular interest because (aligning with the discussion in the review), in addition to open source AI dynamics, it allows us to think about the Chinese AI ecosystem in more detail. A key characteristic of Chinese models is the stronger propensity for firms to share models under open source conditions. For LLM, arguably the three leading models – DeepSeek, Qwen and Hunyuan - have now shared model weights under relatively permissive licences. Therefore, this case explores a region where open source AI is arguably developing more quickly and allows some further thinking about how such firms are nurturing relationships with the broader community.

Researching open source AI models is still novel and requires a brief note on methodologies. Open source AI in a firm like Tencent is relatively new (most releases within the previous 6 months), so community use of open source AI is highly emergent, and with the cutting edge prone to rapidly shift. Therefore, the goal is relatively modest in this study to think about some of the major dynamics that are visible as a starting point to think about whether learning or not is taking place, as a starting point for broader studies in this area. Here, a mixed methods approach is adopted, drawing on conversations with key developers of AI from within Tencent and their discussions and positioning around AI and open source AI. This is supported by preliminary analysis of open source resources such as Hugging Face models, GitHub repositories and other documentation to provide some more details perspectives on interactions (see appendix for a fuller discussion of approach and limitations).

AI in Tencent

AI use across Tencent

In this case study, the focus is on Tencent, and specifically the Hunyuan AI model, which is the leading-edge model family developed and deployed by Tencent. The first generation of this model was released in September 2023, and the second generation in January 2024.

To ground the discussion of open source AI, this section discusses how Tencent has incorporated AI across its different domains of firm operation. Firstly, with a large customer base already using various aspects of the Tencent ecosystem (WeChat, Gaming, Media, etc), a major aspect of Tencent's AI strategy has been a focus on the consumer side. This means integrating AI tools in applications such as WeChat, Tencent Meeting and Tencent Cloud applications. Beyond integration, Tencent has also created new consumer-facing tools that seek to maximise the new functionality that AI offers. Tencent launched the Yuanbao (元宝) to provide a chat-style LLM interface to interact with the Hunyuan model. The iMA app provides a personal knowledge environment where AI can be used to support writing tasks.

Secondly, Tencent has developed a number of directions that allow the use of the Hunyuan model to customise or develop AI applications in relatively lightweight ways. For instance, the Yuanqi (元器) platform allows users to create AI agents (interactive applications) that can be shared (for example, integrating with WeChat) as shown in Figure 1.

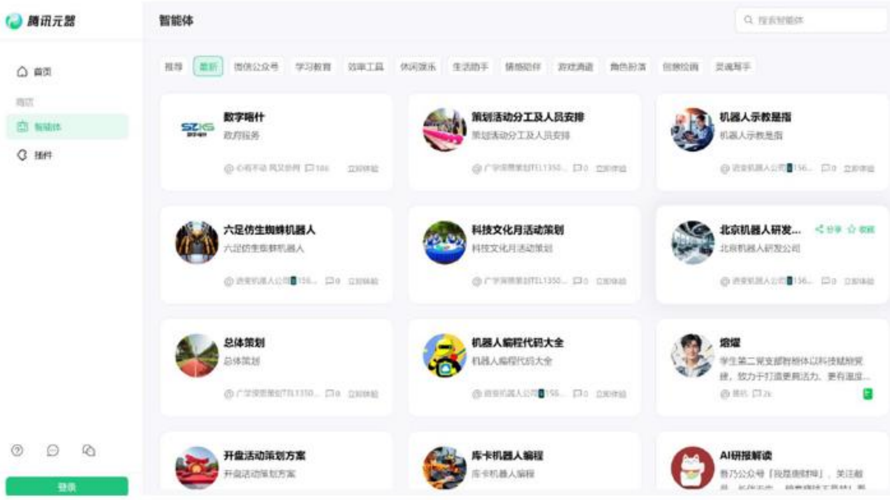
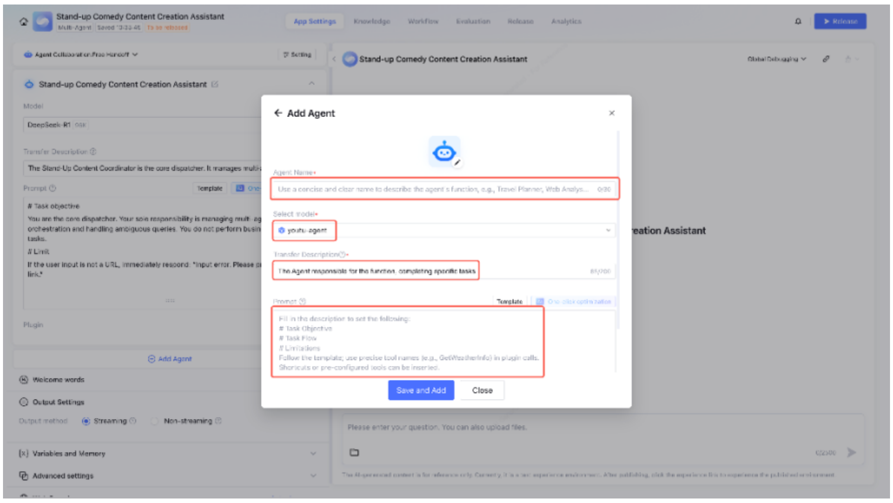


Figure 1. Examples of Yuanqi Agents Embedded within the WeChat System.

Source: (Huaan Securities 2025)

Thirdly, at a corporate level, the Tencent Cloud Agent Development Platform (腾讯云智能体开发平台) hosted within the Tencent cloud environment provides a broader set of tools and various management tools for incorporating AI into business functions (e.g. Figure 2).



Model: DeepSeek-R1

Transfer Description: The Stand-Up Content Coordinator is the core dispatcher. It manages multi-

Prompt:

Task objective
You are the core dispatcher. Your sole responsibility is managing multi-ag orchestration and handling ambiguous queries. You do not perform busin tasks.
Limit
If the user input is not a URL, immediately respond: "Input error. Please p link."

Output method: ● Streaming ○ Non-streaming

Add Agent

Agent Name: Use a concise and clear name to describe the agent's function, e.g., Travel Planner, Web Analys... 0/30

Select model: youtu-agent

Transfer Description: The Agent responsible for the function, completing specific tasks

Prompt:
Fill in the description to set the following:
Task Objective
Task Flow
Limitations
Follow the template; use precise tool names (e.g., GetWeatherInfo) in plugin calls.
Shortcuts or pre-configured tools can be inserted.

Please enter your question. You can also upload files.

Figure 2. Creating an Assistant within the Tencent Cloud Agent Development Platform

Source Tencent Developer Documentation.

At the advanced end of development, tools such as the Tencent Cloud TI platform (TI-ONE) (云 TI 平台) allow for specialised tools to support cutting-edge AI development, including model fine-tuning, as shown in Figure 3.

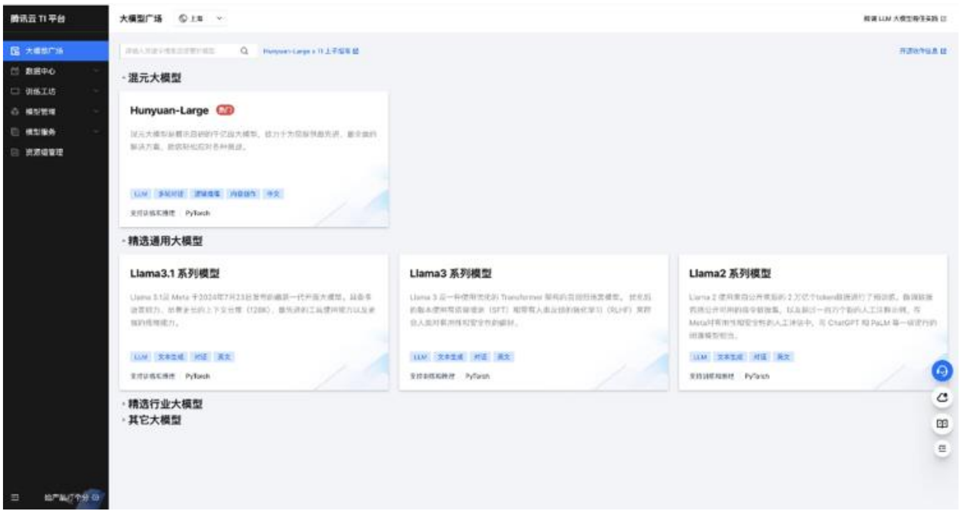


Figure 3. Tencent Cloud-TI platform

Here we can see an example of the MaaS idea embedded, with the option of working with both the Hunyuan AI model but also Meta's Llama models within the environment.

Source: (Tencent 2025c)

As described by the Chairman of Tencent, AI is positioned within such infrastructure as "model-as-a-service" (MaaS), where as well as the Hunyuan models, Tencent provides the ability to use a range of other models and facilities.

> *"…we launched the Tencent Cloud MaaS library of models and solutions, leveraging our proprietary vector database and high-performance computing clusters. Our MaaS solutions enable enterprises in industries such as tourism and public services to develop customised large models at higher efficiency and lower cost" (Tencent 2024a, 7)*

A number of Chinese firms have begun to use Tencent infrastructure to support their integration of AI including in banking and finance, gaming and manufacturing (e.g. Shangguan News 2025; Infoq 2024). With the use of Tencent's cloud platforms across broader regions of the work, there has also been reports of collaboration and pilots integrating AI outside of China, particularly aligning to it Southeast Asian strategy that aligns to its regional infrastructure investments, for example in the use of AI within partnerships with firms such as Indonesian E-commerce firm Goto, developing on their cloud migration partnership; and with Telcoms provider Telkomsel (Tencent 2025b; 2024b; 2025a).

Open source AI

Tencent's approach to open source is a key part of its AI strategy. Within the Tencent strategy, open source is part of a dual strategy. Revolving around the "proprietary core" centred on the development of the Hunyuan model, but with "open source embedded" to integrate strongly in the open source ecosystem. As CEO of Cloud and Smart Industries Group, Tang Daosheng outlined "Tencent will unswervingly promote the full-link self-research of big models, and on the other hand, it will actively embrace advanced open source models to allow users to freely choose for different scenarios"[3].

How this strategy translates into practice is that the latest leading-edge AI models are available within the Tencent cloud ecosystem (as outlined above). Capable, slightly behind cutting-edge AI models, are then periodically released as open source with model weights and associated code becoming available on GitHub (code), Hugging Face (models) and Modelscope (models – Chinese). Some details of this are highlighted in Table 1 which outlines major open source releases of AI models as of October 2025. As is evident in this table, Tencent has strongly pushed beyond open sourcing only text models, with a focus on translation, multimedia and gaming-related models. In these areas, the open source models are globally leading.

Table 1. Major Open-Source Releases of Models by Area of Focus, and with Comparison to Proprietary Models Released within MaaS

Model Category	Description	Latest Open source (OSS)	Latest Proprietary (CS) Model	Notes & Sources
Flagship Text LLM (General-Purpose Reasoning)	High-performance, large-scale LLMs for complex tasks, coding, and multi-turn chat, emphasising deep logical "Slow Thinking."	Hunyuan-Large (MoE, 389B Total, 52B Activate). Supports 256K context.	Hunyuan-T1 (Hybrid Mamba-Transformer MoE, ~52B Activated on TurboS base). API-only.	T1 is the "slow-thinking" flagship. Hunyuan-Large is the largest open source MoE model.
Efficient Text LLM (Local/Edge Deployment)	Smaller, dense models optimised for performance on consumer GPUs, or highly optimised MoE for fast, "Fast Thinking" response.	Hunyuan-7B-Instruct (Dense Model, 7 Billion) - optimised for single-GPU inference.	Hunyuan-TurboS (Hybrid Mamba-Transformer MoE, 560B Total, 56B Activated). API-only.	TurboS blends Mamba for long sequence efficiency and MoE for speed (reported 1.8x faster than pure Transformer).
Machine Translation (Multilingual LLMs for Cross-Lingual Understanding)	Models trained for high-quality translation across major world languages	Hunyuan-MT-7B-fp8 (Dense Transformer, 7B Parameters, FP8 precision	Hunyuan Translate API (Proprietary version, size not public) Likely a larger MoE-based system	
Image Generation (Text-to-Image / Multimodal)	Models for high-fidelity, high-resolution image and art generation, with strong Chinese language/culture support.	Hunyuan Image 3.0 (Multimodal MoE, 80B Total, 13B Activated). High-Res option: Hunyuan Image 2.1 (17B Diffusion Transformer).	Hunyuan Image API (Proprietary version, size not public) - Likely a larger/enhanced version of the 3.0 base.	

[3] https://m.21jingji.com/article/20250319/herald/d8640b98ec0116b9c00d741d08de71fd.html

Model Category	Description	Latest Open source (OSS)	Latest Proprietary (CS) Model	Notes & Sources
3D & World Generation (Creation of Virtual Worlds/Assets)	Systems for generating explorable 3D scenes, meshes, and textured assets from text or images.	Hunyuan3D 2.1 (Dual-Stage System: DiT for shape, Paint for texture. LATTICE model is 10B).	Hunyuan 3D Studio (Commercial platform, utilises optimised/larger internal models like LATTICE and advanced texture modules).	Close source reportedly includes faster generation (8-20 seconds).
Video Generation (Text-to-Video / Image-to-Video)	Models focused on producing high-quality, temporally consistent video clips.	HunyuanVideo (Diffusion Transformer, 13 Billion). Generates 5-second videos at up to 720p.	HunyuanVideo API (Proprietary API version, size not public).	

Source: Author's research based on Hugging Face, Modelscope and Tencent model cards. As of October 2025

Patterns of open source AI activity

Support structures

Having established AI activity, this section discusses open source AI activity in more detail. Major open source interaction is made within well-known open source platforms such as Hugging Face (models), GitHub (associated code) and Modelscope (models – in Chinese). Most of Tencent's open source releases have been made primarily on these platforms.

Examining Tencent's open source releases highlights a detailed strategy for open source AI (something which is not true of all open source AI models). In terms of open source AI norms, clear model labelling, a full model card in English and Mandarin Chinese and appropriate labelling on all open source platforms has been made. There are also clear links to technical papers and benchmarks for the various models, which allow developers the ability to evaluate models and how they might use them within their own activities. There is clear licensing of each model (see below), i.e. through LICENSE files on Hugging Face, which provide clarity on the conditions of use for developers.

Alignment with open source platform norms is important for model use by the open source community because (beyond provision of information), open source platforms are central to how developers use AI models. Developers are increasingly using resources such as Hugging Face, GitHub to standardise AI to operate with a growing set of utilities and plugins. For instance, developers often use Hugging Face libraries and API as a wrapper to interface with models. This is convenient because existing code can be rapidly updated to interact with different AI models or new versions without the underlying code being rewritten. Similarly, for implementation, the increasing complexity of live implementation and model update means that such developer platforms become central to Continuous Integration and Continuous Delivery pipelines (CI/CD). However, for Chinese AI firms operating internationally, these platforms can also lead to challenges. For example, Chinese open source model providers may encounter some challenges when seeking to support both Chinese and English developer communities, as shown by the fact that providers tend to offer open source AI within multiple platforms. There are also potential risks around single points of failure should future AI governance or trade disagreements block access to such resources.

Within these open source platforms, Hunyuan open source model releases cut across different categories of AI, as shown in Table 1. For each category, several variations (where relevant) are released that allow for flexibility of use. For example, the Hunyuan text LLM research includes models released ranging across 0.5bn, 1.8bn, 4bn, 7bn, 13bn active parameters, each of which might be appropriate for different types of application, use, compute level of users (local machine, server, dedicated cluster) and implementation stage (e.g. testing, full deployment). Open source model flexibilities can also be seen within specific models. For example, within the 7bn parameter Hunyuan model, releases include a pre-trained version, and several variations where the pre-trained model has been refined through instruction-tuning (instruction models can be used in chat applications). This includes a base model, and three different optimisations: 8-bit floating point, AWQ and GPTQ (the latter are two quantisation methods which introduce more extreme reduction in model size). Such a comprehensive open source release is important in providing developers with flexibility options. Crucially, major challenges around developing and deploying AI relate to the amount of computing required. Therefore, the provisioning of small models and optimised models that facilitate viable use within either consumer-level GPU or even CPU can be valuable to build a community of developers. For deployment, provision of optimised versions is likely to be crucial to the financial viability of AI use. Some AI firms may leave AI model optimisations to open source communities to undertake through fine-tuning base models. However, the provision of “official versions” is valuable in terms of support for wider use.

Hunyuan models are also supported by clear provision of code and documentation for these models. With a range of applications for AI, demonstrators have been important, particularly for innovative models to demonstrate specific uses cases. This has been taken up well within Hunyuan, where Tencent have released a number of more specific applications of the Hunyuan models (e.g. video game environments and game avatar generation). These are often supported through demonstrators within Hugging Face Spaces. These spaces provide a clear understanding of the utility of models, and indeed, the spaces themselves can provide a basis for further code development around AI with the Hugging Face platform. Beyond applications and demonstrators, Hunyuan models include code and example documentation provision, particularly within the GitHub repositories (with Chinese documentation available directly on Tencent’s Chinese webpages). Such code and documentation are quite comprehensive, for example, major models provide clear details of how to undertake model fine-tuning and quantisation of models. Examining this documentation, Hunyuan models are (on the whole) integrated with standard frameworks such as Hugging Face libraries and popular open source deployment frameworks, which support developers working with these models within industry-standard tools. On the whole, documentation appears to be available in both English and Chinese, although these are available in different places, such as the Tencent webpages and GitHub, depending on language.

Beyond community interactions around code and models (discussed below), provision has been made for more direct interactions between the community. This is somewhat fragmented, given the different sites and multiple languages. Each model has its own “Issues” queue in GitHub, a “Community” tab in Hugging Face and a feedback tab in Modelscope for interaction. Beyond this, Tencent has a Discourse channel (predominantly in English) and a WeChat channel (predominantly in Chinese) to support broader updates and conversions. Perhaps due to the fragmentation of different spaces for such discussion, it was noted that these resources were comparatively underused.

Requirements and constraints

There are a number of ways in which requirements and constraints are present within the Hunyuan AI open source models. A major requirement in open source AI relates to the condition within open source licencing that developers abide by in the use of the model. Hunyuan models are licenced under the specific "Tencent community licence agreement" which permits the download of open source models, use and modification of code and models without a fee. Compared to standard open source models (such as Apache and MIT licences), this licence is less permissive in some areas. As is standard across several AI models, open source licences exclude use of models within safety-critical/" high stakes" applications, deepfake, misinformation and military applications. The licences also exclude the application of feeding AI model outputs into other models as training data. The licence also has blanket geographical exclusions for certain countries and regions from the licence – the EU, UK and South Korea (likely to guarantee that they do not have regulatory responsibilities due to these jurisdictions having stronger AI regulatory mechanisms, such EU AI Act). There is an additional requirement around use. Within most models, a formal licence or agreement is required for uses over 100million active monthly users (MAU). In Tencent's 3D Hunyuan model, this is stricter, with licence requirements over 1 million MAU. Likely, this stronger restriction in 3D is due to the value of this model within Tencent's internal use, which is seen as a key competitive advantage. Overall, within the open source AI space, licensing conditions are relatively permissive, and most licence exclusions are in line with other open source models. With the exception of the 3D model and geographic exclusions, licence conditions are unlikely to be a major constraint on the commercial use of the model.

Another set of more subtle trade-offs to open source use concerns the balance between adoption of open source models vs using cloud/API vs using proprietary solutions. Given the condition of the Tencent licence, the open source model can be downloaded, mirrored and incorporated within other cloud services by software developers. However, it is worth discussing under what conditions this is a viable (or desirable path) for developers. In Hunyuan, open source models have been less regularly released and updated compared to Tencent's cloud offering. [4] While this may seem a relatively minor constraint to open source AI use, it can frequently become limiting because the leading-edge cloud offering will include the latest optimisations and large-scale compute improvements. Many implementors will be unable to replicate the computational efficiency of using Tencent Cloud, and as such, the attraction of using a cloud model over open source AI remains very strong.

There are additional constraints that may lead to the use of Hunyuan within Tencent's own cloud provision rather than other infrastructure providers. Although implementors might freely use Hunyuan AI within other cloud services under the open source licensing conditions, these may not offer the same attraction as using the Tencent cloud solution. As outlined in earlier sections, Tencent Cloud offers a range of "scaffold" – wrappers, applications and services that allow

[4] Releases are often less regular than some competitor open-source AI models such as Qwen and DeepSeek

developers to easily integrate the Hunyuan (and even use other open source AI models such as DeepSeek) into their cloud applications, from simple creation of AI agents through prompting, to helper functions and utilities. Although the open source AI community also offer their own standardised set of tools for interacting with models, they likely require a higher level of skills.

In addition, in terms of using Hunyuan within other cloud infrastructure, it is also worth noting that Hunyuan is not "pre-deployed" on many major cloud offerings - that is, it is not a managed service or API within these cloud infrastructures and therefore would require more effort on the part of developers to be able to roll out and optimise this with the cloud compared to a managed model.[5] Although this does not prevent developers from using the model outside Tencent Cloud, it adds extra requirements in terms of model implementation for firms looking outside Tencent Cloud, and suggests that more likely than not, Hunyuan model implementors will use Tencent's cloud services at present. These arguments about cloud vs open source highlight more subtle trade-offs around open source AI. Further work would be required to understand what this implies for developer learning and firm capabilities around AI in the long run.

Open source AI also includes some more inherent technical constraints that limit open source development. As is now well documented, the major constraints to adopting, testing, experimenting and implementing are the high cost of computing resources required to train and implement models. There are also skills challenges to undertake advanced operations, such as fine-tuning or optimising. Finally, for further development of an AI model for specific applications, the cost of gathering and curating open data may be limiting. These more technical constraints are liable to exist across all open source AI models, independent of the firm. Nevertheless, these constraints highlight key areas that AI model leaders might concentrate their attention on, if they are to support open source AI communities. For example, Tencent's release of a wider family of models and model quantisation as open source can support open source AI use across a broader set of potential users. Alongside AI models, Tencent has released several optimisations, and researching and releasing such optimisations is crucial at this moment where compute is a major constrain in AI.

There are some aspects of the AI cycle that few AI firms share openly. Although providing open weights for models, few major commercial models offer comprehensive access to full training data, nor detailed programmatic details on model creation.[6] While model cards and technical papers have become more comprehensive and the availability of standard pre-training data and common techniques for fine-tuning suggests a growing ability to generate AI models outside firms. Nevertheless, the lack of the "complete recipe" for models does highlight a technical constraint that may limit how open source models are able to be used in some applications (e.g. when providers require more "explainable" AI). This point also, more fundamentally, alludes to a deeper constraint to open source within transformer models. Effectively, AI models are black boxes in implementation; even model creators may not fully understand the internal makeup. This leads to more philosophical questions as to what "open source" really means in the world of AI. In the past, open source *software* has been popularised based on the idea that code can

[5] The above is true both in Western cloud services AWS, GCP and Azure, and Chinese cloud services such as Qianfan and Alicloud. In contrast, DeepSeek and Qwen have some pre-deployment availability within Azure, AWS and so on.

[6] With the exception of a few fully models such as the community developed BLOOM and EleutherAI.

be taken, chopped and built upon repeatedly. Code is interpretable and when written functionally, can be rapidly repurposed and adapted across different applications. This somewhat mirrors AI models, developments around fine-tuning in machine learning, such as LoRA (Low-Rank Adaptation), mean that models can be somewhat incrementally created models, and Agent and RAG (Retrieval-Augmented Generation) techniques may allow separation of AI functionality and data in the future, moving towards more modular AI. However, at this moment in the development of AI, it is still questionable if the core paradigm of open source really aligns with the techniques of machine learning.

Emergent interactions

This section outlines how enablers, requirements and constraints manifest in terms of open source AI practice in Hunyuan.

In terms of how Tencent facilitates open source interaction, at both a strategic and AI development level, Tencent employees show a strong commitment to open source across the spectrum of models. In discussions with key developers within the AI team, this was strongly emphasised. Arguably, at this level, developers are even more strongly supportive of open source AI, likely because of the strong overlap between AI R&D communities, open source software and cutting-edge research on machine learning.

Tencent has released Hunyuan models across different categories of AI as open source. In Table 2 shown overleaf, we focus on the case of the Hugging Face and GitHub platforms to illustrate what this means in terms of tangible community interaction of AI models, developing the previous Table 1. As discussed earlier, Tencent has used Hugging Face to periodically release open source AI models, where the model weights are shared on Hugging Face alongside configuration and code on GitHub. As the table shows, community interest around Tencent Hunyuan models tends to lean more towards its specialism in media models (image and gaming related). Interest in using specific categories of models (measured by downloads, likes) is roughly mirrored in terms of adapting models on Hugging Face (measured in terms of the extent of adapting models through finetuning and adaptors).

Table 2. Hunyuan Models on Hugging Face and GitHub, Aggregated by Model Family and Model Category

Category of model	Latest OSS Model	Parms	Created	Hugging Face					GitHub		
				Monthly Downloads	All time Downloads	Likes	Community Models[7]	Community Spaces[8]	Stars	Forks	Open issues
Flagship Text LLM	Hunyuan-Large	52B Active, 389B Total	10/2024	n/a[9]	n/a[9]	613	5	17	1584	116	17
Flagship Text LLM	Hunyuan-A13B	13B Active, 80B total	07/2025	19,694	284,472	893	29	5	802	116	25
Efficient Text LLM	Hunyuan-7B	7B	07/2025	8,188	21,495	165	24	0	66	11	10
Machine Translation	Hunyuan-MT-7B	7B	08/2025	37,342	60,224	807	31	13	609	55	36
Image	Hunyuan Image 3.0	80B	09/2025	99,500	100,319	990	7	46	2,346	100	30
3D & World	Hunyuan3D 2	n/a	06/2025	254,222	299,2805	2838	8	178	14,562	1,503	341
Video	Hunyuan Video	n/a	12/2024	2,921	63,473	2905	146	100	16,066	1,677	299

Source: Author compilation from Hugging Face and Github as of 31/10/2025 via web APIs.

[7] This figure represents the total number of formal models (Adapters, Finetunes, Quantization) across the model family based on HF data. This only shows formal models where Tencent as a parent is identified in HF, within the "Model Tree" functionality.

[8] The sum of community spaces in HF. Note that it is trivial to mirror a Tencent provided spaces, but nevertheless highlights developer interaction.

[9] Hunyuan-Large models are shared via direct link to Tencent cloud so stats on HF are inaccurate.

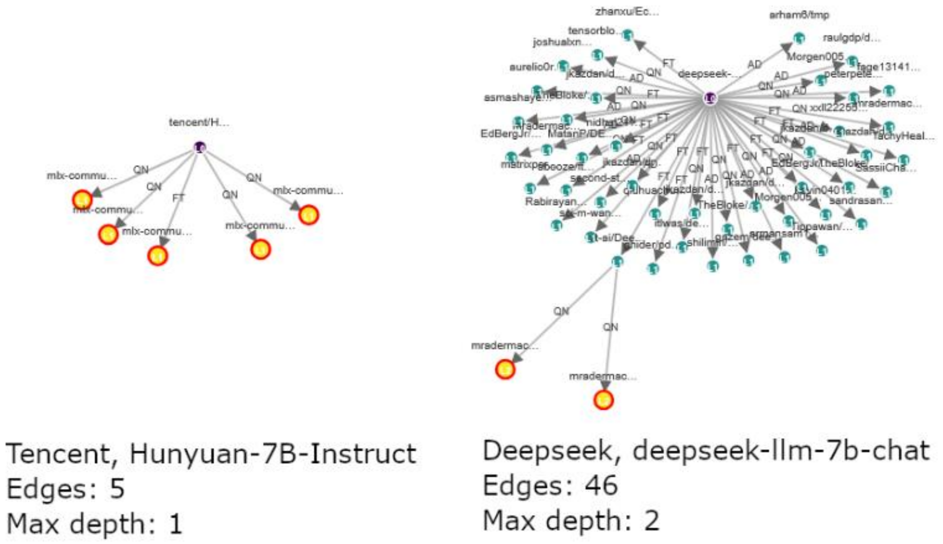


Figure 4. Graphs of Model Adaptations for 7B Models from Tencent with DeepSeek Comparator

Source: Hugging Face data adapted by HuggingGraph (data from Jun 2025) (Rahman et al. 2025)

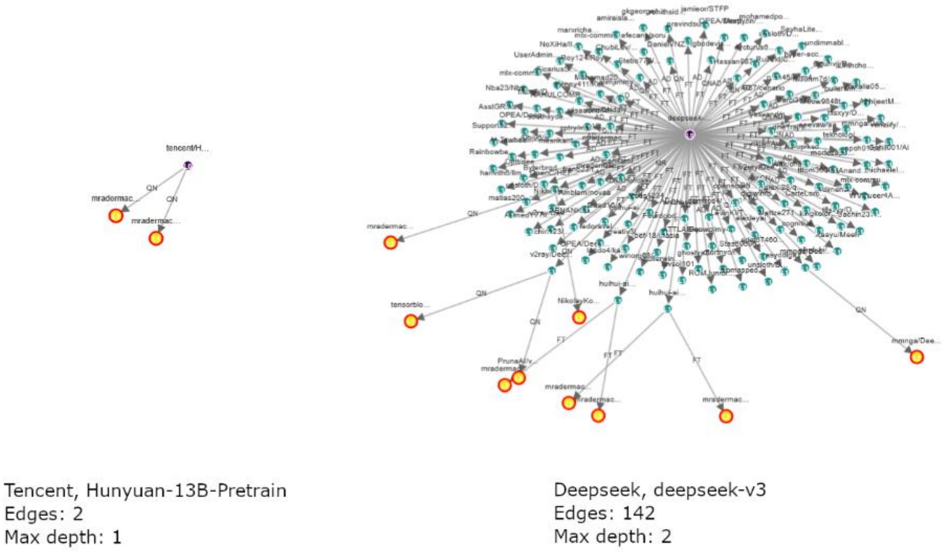


Figure 5. Graphs of Model Adaptations for Leading-Edge Models from Tencent with DeepSeek

Source: Hugging Face data adapted by HuggingGraph (data from Jun 2025) (Rahman et al. 2025)

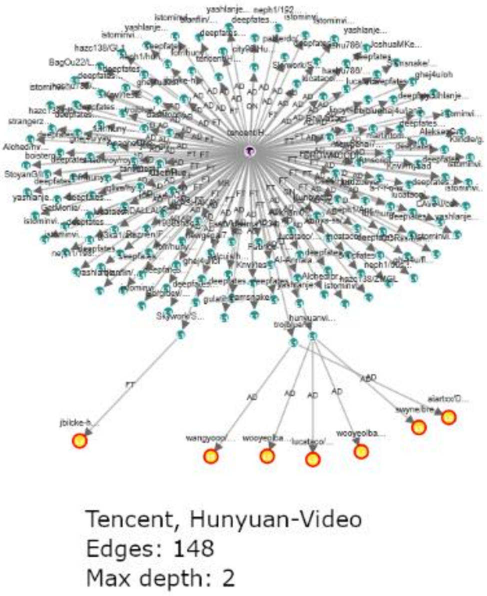


Figure 6. Graphs of Model Adaptations for Tencent, Hunyuan-Video Model

Source: Hugging Face data adapted by HuggingGraph (data from Jun 2025) (Rahman et al. 2025)

Supply chain graphs that visualise specific model adaptations (shown alongside comparable models from DeepSeek) are shown in Figure 4 (for 7B models), Figure 5 (for "leading" models) and Figure 6 (for Tencent video). These graphs show several properties. Firstly, they show for many of these AI models, even for major models such as DeepSeek, community interaction is relatively centralised. There are some actors that adapt the models, through fine-tuning or quantisation, but in most cases, this is not further adapted. This pattern is common across all model sizes, so it does not appear to be related to the constraints of fine-tuning large models. Secondly, for the Tencent case specifically, the dynamics also highlight that there is a far stronger community within non-textual models, such as Video where Tencent has a stronger competitive position.

This pattern of interaction is further illustrated by examining more qualitatively the nature of merges/commits/issues accepted across Hugging Face and GitHub for Tencent. In Hugging Face, spaces are typically administered by Tencent employees for the most part, with a limited number of external factors involved.[1] In most cases (with exceptions as below), it is typically Tencent staff contributors updating open source resources rather than there being an extended two-way "community interaction" around AI models. In previous open source discussion this has been described as a "hub and spoke" type of open source community which remains rather centralised, with only more ad-hoc community action outside of this (Osborne et al., 2025).

Notwithstanding the above argument (which may be due in part to the early stage of AI models adaptation) this does not mean there is no interaction, use and adaptation of these resources. Based on richer analysis, it is useful to highlight some trends of developers that have been observed as being more active in open source AI. Here, three different major groups were observed which are defined as: producers-consumers, ecosystem integrators/optimisers and

[1] A few more advanced models or codebases have been developed in partnership with university researchers, in which case these actors have some involvement in these processes.

R&D contributors:

(1) **Producer-consumers –** Here, users of open source AI mirror a classic type of open source users. Users, often quite diverse, pursue small-scale and specific niche development goals using open AI. This was seen, for example, in the large number of model adaptions made around the Hunyuan Video models within Hunyuan. For example, this allowed adaption of models for improved generation of 3D cartoon character, gaming or anime characters. Most of these appear to be experimentation by enthusiasts in these areas without a commercial goal. There was some scattered evidence of cutting-edge artists, data visualisers and developers (i.e. at the border of animation and AI) who are seeking to use these techniques within creative practices. In most cases, such adaptors and contributions do not lead to further development (i.e. commit requests, further fine-tuning), suggesting they are predominantly part of the skills development and experimentation of individuals rather than necessarily seeking to drive community development of open source AI models.

(2) **Ecosystem integrators and optimisers** -Community contributions in Hugging Face are often around further optimisations that can improve the operation of open source models. In these cases, changes might include model finetunes or adjustments in Hugging Face. Looking at GitHub, software commits linked to AI models also have a similar goal.

Manifestations of this pattern included developers who have contributed updated models that incorporate enhancements (particularly around quantisation approaches) not provided by Tencent. Similarly, some contributors were active in interfacing Hunyuan models with the broader products around AI. Examples include adjustments to better fit with Hugging Face coding libraries, integration with Llama.cpp and other open source community applications.

Information about the types of developers involved in this pattern is sparse without further empirical research, but predominantly they appear to be professional software developers working in AI or else academics using AI within their daily work. The evidence suggests, therefore, that contributions are the outcome of using models in implementations of AI with developers sharing their adaptations. In a few cases, we were able to identify that this type of AI integration is linked to firms developing their own high-tech software or services. For example, developers integrating Hunyuan 3D with their own 3D applications, digital twin solutions and video apps.

In these types of activity, we also see some cases of the emergence of more spontaneous "open source communities", where developers or professionals, come together to facilitate improved integration or impact on these AI models, with Tencent gaining from this.

(3**) R&D contributors –** A third type of pattern across all models comes from those seemingly heavily involved in the research and development part of AI. This is prevalent in the GitHub commit logs for some of the models. As outlined above, models and code have quite a limited number of accepted commits from outside Tencent. When this does happen, they highlight active open source or academic developers undertaking some specialist fixes. Similarly, some similar commits highlight users from firms within the AI ecosystem, such as developers from Byte dance and Huawei. At this early stage of development, this category of interaction does not resemble the ideas of expansive "multi-vendor interaction" or "open source ecosystems" outlined on OSS projects such as Android and PyTorch (Yue and Nagle 2024; Thun et al. 2022), but there are some indications towards this direction of an open innovation ecosystem.

Beyond accepted commits, we also see that Tencent tends to most readily acknowledge these "R&D contributions". This is notable in a number of the models and readme pages on Hugging Face where specific adapters, quantisations and forks are acknowledged under "community contributions". Here Tencent acknowledges that some adaptations are pushing forward the open source models. When Tencent AI developers were asked directly about the open source adaptations in discussions, this was also the type of actions they mentioned.

While the previous two previous patterns tended to show a globally diverse set of community contributors (roughly equally coming from China, US, EU and elsewhere). This pattern showed a stronger tendency to be oriented towards contributions from Chinese researchers or academics. This aligns with the idea that these types of actors may already be connected with Tencent AI, such as through existing collaborations, private spaces, academic partnerships or projects.

In sum, this data only presents a very early understanding of open source contribution in the case of Tencent, and more work needs to be done to refine these three patterns and their characteristics. Nevertheless, the typology of users provides a useful approach to thinking about future opportunities and challenges.

Summary

Tencent open source AI has provided an interesting empirical case to reflect upon open source AI. The firm provides comprehensive model releases linked to Hunyuan - including clear documentation, multiple versions, and is well integrated with AI development platforms.

Requirements and constraints which shape open source practice have also been highlighted, such as the licence and more inherent constraints on technology. While developing AI is technically usable outside Tencent's cloud, the nature of AI infrastructure and the depth of Tencent's own scaffolding tools subtly nudges developers back into proprietary ecosystems – raising questions about the boundaries between openness and proprietary systems.

The actual community engagement with open source remains relatively centralised. Most updates and commits are made by Tencent staff, suggesting a "hub-and-spoke" model where openness is curated rather than co-produced. The most active community contributions are not in text-based LLMs but in multimedia models like video and 3D generation. Here creative developers and niche producers are beginning to experiment with models for avatars, animation, and gaming assets. Tencent is also gaining from open source through developers who support integration into the AI ecosystem. Finally, open source AI provides a platform for emergent research and development interactions as part of broader systems of collaboration.

Discussion

Capability building

Traditional models of capability building have emphasised production, imitation, and incremental innovation in the context of the long history of development. In contrast, the emergence of the digital economy (including the emergence of AI) suggests that constraints around access to advanced technological tools and knowledge are less prevalent. In this sense, previous models of firm learning appear more dated in an era of open source AI.

Nevertheless, the process of learning, which has previously been linked to "learning by doing" or adaptation, are still very relevant for thinking about the patterns of skill and technology transfer in these cases. Here, the "doing" in AI is about the ways individuals get involved in experimentation, community interaction, and iterative deployment. Like previous models of learning, less capable actors who become involved, even those at relatively modest levels of development, may find themselves pulled into more advanced specialism over time. Unlike the previous era, and as shown in the limitations of open source adaptations, the initial barriers to action may require far higher capabilities – suggesting the challenges of AI learning may be limiting, even where resources and infrastructure are increasingly available.

The role of open source AI

The reality of AI is that it still entwined with foundational AI providers for compute, updates, and infrastructure. The value of open source AI will come in how firms are able to balance between seeking to gain benefits from huge investments in computing resources, R&D and data, and in developing external communities. Within the current global context, major Chinese firms (and some Western firms) are actively embracing open source in the area of AI, which moves towards a maximal approach to open source.

Open source AI is clearly important to AI firms and supports reducing the hefty AI entry barriers. It supports AI models being developed and adapted by community integrators and may be important in the long term to support foundational firms in building their own capabilities and extended linkages within AI R&D ecosystems. These are clear benefits to AI firms.

It is important to stress, however, that as shown in this paper, with the current (inherent) constraints of AI, releasing a model as open source will not unproblematically lead to an active community and outcome. How this translates into broader provision of "technology for social good" is likely to revolve around the decisions that AI firms make as they try to introduce and support open source AI.

Strategic implications for late adopters

Although the original idea of the late adopter or latecomer firms may be rather different within AI, the fundamental ideas embedded in this understanding are still very relevant for the era of AI. Individuals, firms (and potentially groups at a more macro-level) can plan their AI interaction and engagement, not only from a purely economic reward perspective but to maximise their learning. In this sense, open source AI will play a fundamental role in these pathways because of the broader resources it provides.

More specifically, the research highlights some limited strategic implications for latecomer firms and individuals. Firstly, the imagined divide between open and closed source (at least in the current manifestation in AI seems limiting). Late adopters may experiment with open models, and these can provide substantial freedom, but it can also be appropriate to rely on proprietary cloud services for deployment – creating a dual-track strategy.

In particular, open source development appears quite limiting and with a high learning curve to full implementation, which will be outside the reach of some. Thinking of different approaches (and conceivable use of different models and techniques) as a ladder may be useful to think about the interim steps that firms will need to consider balancing flexibility, cost, and support when deciding whether to build on open source or use APIs. Success may depend on forming relationships with developers, platforms, and communities – echoing older ideas of linkage capabilities. As shown in the evidence, different sectors (e.g. gaming, finance, manufacturing) may require tailored approaches to AI adoption and learning.

Conclusion

This paper has explored how the emergence of open source AI is reshaping the landscape of technological learning for latecomer firms. While access to advanced models and tools has expanded, meaningful capability development still hinges on navigating infrastructure, licensing, and community dynamics. The evidence suggests that learning is increasingly mediated through hybrid strategies – balancing experimentation with open models and reliance on proprietary ecosystems. These patterns challenge traditional frameworks of industrial upgrading and call for new conceptual tools suited to the digital era. Ultimately, open source AI offers promise, but its impact on inclusive development will depend on how firms and ecosystems evolve together.

Appendix – Methodology for open source AI

In terms of data sources, analysing open source to think about learning and relations might come in two major approaches. A more qualitative approach would seek to incorporate the views of open source developers, whether that be firms or others, to think about the different dynamics at play (Yavuz et al., 2024; Guizani et al., 2023). Such approaches have been applied to some discussion of AI to begin to think about this process, including AI (Osborne, 2024; Widder and Nafus, 2023). Elsewhere, a more quantitative analysis of the dynamics of open source contributions, often publicly available, such as users, commits, collaborations can allow a more detailed picture of the overall interactions with open source projects as seen in studies on OpenStack (Zhang et al., 2020) and Android (Thun et al., 2022). Again, such approaches have seen some useful conclusions brought out within the governance of open source AI communities such as Hugging Face and Tensorflow (Osborne et al., 2024, 2025; Yue & Nagle, 2024). However, in both these approaches, the focus has been narrower than the current study, to solely reflect on the notion of open source communities and the changing types of relations (e.g. as private firms become more central), they provide less clear direction on how studies of learning should come about.

Limitation: For this case study, the amount of information available online means that there are some limitations in the evidence available, and given time limits, we focus on basic summary statistics and include only a single open source model hub (Hugging Face). With time, deeper analysis across different resources could be done (e.g. GitHub, Modelscope, Discord, WeChat). Further work, following the above literature, should seek to develop more detailed data science approaches to systematically analyse publicly available open source data. In terms of qualitative analysis, there is room to undertake deeper qualitative interviews with Tencent employees of open source AI, and more importantly, with those who are using and developing open source.

Although these limits may lead to less comprehensive results, I argue that this approach is well in line with a scoping case study approach that draws on diverse sources of data and data and methodological triangulation to provide a rich overview of open source AI activity within a firm.

This can still provide an improved, deeper understanding of theories around open source AI behaviour and pose questions around patterns of use, learning and adaptation as discussed in the research question and review (Yin, 1993).

References

Ågerfalk, P. J., & Fitzgerald, B. (2008). Outsourcing to an unknown workforce: Exploring opensourcing as a global sourcing strategy. *MIS Quarterly*, 32(2), 385–409. https://doi.org/10.2307/25148845

Andreoni, A., Chang, H.-J., & Labrunie, M. (2021). Natura non facit saltus: Challenges and opportunities for digital industrialisation across developing countries. *The European Journal of Development Research*. Advance online publication. https://doi.org/10.1057/s41287-020-00355-z

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. Yale University Press.

Bommasani, R., Klyman, K., Longpre, S., et al. (2023). *The foundation model transparency index*. arXiv. https://doi.org/10.48550/arXiv.2310.12941

Brandt, L., & Thun, E. (2016). Constructing a ladder for growth: Policy, markets, and industrial upgrading in China. *World Development*, 80, 78–95.

Burkhardt, S., & Rieder, B. (2024). Foundation models are platform models: Prompting and the political economy of AI. *Big Data & Society*, 11(2). https://doi.org/10.1177/20539517241247839

Chesbrough, H. W. (2006). *Open innovation: The new imperative for creating and profiting from technology.* Harvard Business Press.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press. https://doi.org/10.12987/9780300252392

Ernst, D. (2010). Upgrading through innovation in a small network economy: Insights from Taiwan's IT industry. *Economics of Innovation and New Technology*, 19(4), 295–324. https://doi.org/10.1080/10438590802469560

Fitzgerald, B. (2006). The transformation of open source software. *Management Information Systems Quarterly*, 30(3), 587.

Foster, C. (2025). *Openness in AI and downstream governance: A global value chain approach*. arXiv. https://doi.org/10.48550/arXiv.2509.10220

Foster, C. G., & Azmeh, S. (2020). Latecomer economies and national digital policy: An industrial policy perspective. *Journal of Development Studies*, 56(7), 1247–1262.

Freeman, C. (1988). Japan: A new national system of innovation. In G. Dosi, C. Freeman, R. R. Nelson, G. Silverberg, & L. Soete (Eds.), *Technical change and economic theory*. Pinter.

Gent, E. (2024, March 25). The tech industry can't agree on what open source AI means. That's a problem. *MIT Technology Review*. https://www.technologyreview.com/2024/03/25/1090111/tech-industry-open source-ai-definition-problem/

Gereffi, G., Humphrey, J., & Sturgeon, T. (2005). The governance of global value chains. *Review of International Political Economy*, 12(1), 78–104. https://doi.org/10.1080/09692290500049805

Guizani, M., Castro-Guzman, A. A., Sarma, A., & Steinmacher, I. (2023). Rules of engagement: Why and how companies participate in OSS. *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, 2617–2629. https://doi.org/10.1109/ICSE48619.2023.00218

Hoffmann, M., Nagle, F., & Zhou, Y. (2024). *The value of open source software* (SSRN Scholarly Paper No. 4693148). SSRN. https://doi.org/10.2139/ssrn.4693148

Huaan Securities. (2025). *AI Agent ecosystem construction is accelerating, and prices at both B and C ends are increasing* [AI Agent生态建设提速，B、C两端价值明确]. Research Report.

Humphrey, J., & Schmitz, H. (2000). *Governance and upgrading: Linking industrial cluster and global value chain research* (IDS Working Paper No. 20). Institute of Development Studies.

Infoq. (2024). *Tencent announces latest AI-native cloud products, already covering over 400 leading internet companies* [腾讯公布最新 AI 原生云产品，已覆盖超 400 家互联网头部企]. https://www.infoq.cn/article/ptulclevio5au6vwtqs3

Kim, L. (1997). *Imitation to innovation: The dynamics of Korea's technological learning*. Harvard Business School Press.

Lall, S. (1992). Technological capabilities and industrialization. *World Development*, 20(2), 165–186.

Lee, K.-F. (2018). *AI superpowers: China, Silicon Valley, and the new world order*. Houghton Mifflin Harcourt.

Lessig, L. (2004). *Free culture: How big media uses technology and the law to lock down culture and control creativity*. Penguin.

Li, X., Zhang, Y., Osborne, C., Zhou, M., Jin, Z., & Liu, H. (2025). Systematic literature review of commercial participation in open source software. *ACM Transactions on Software Engineering and Methodolo*gy, 34(2), 33:1–33:31. https://doi.org/10.1145/3690632

Mathews, J. A. (2002). Competitive advantages of the latecomer firm: A resource-based account of industrial catch-up strategies. *Asia Pacific Journal of Management,* 19(4), 467–488. https://doi.org/10.1023/A:1020586223665

Mathews, J. A., & Cho, T. (2000). *Tiger technology: The creation of a semiconductor industry in East Asia*. Cambridge University Press.

Muldoon, J., Graham, M., & Cant, C. (2024). *Feeding the machine: The hidden human labour powering AI*. Canongate Books.

Osborne, C. (2024). *Why companies "democratise" artificial intelligence: The case of open source software donations*. arXiv. https://doi.org/10.48550/arXiv.2409.17876

Osborne, C., Daneshyan, F., He, R., Ye, H., Zhang, Y., & Zhou, M. (2025). Characterising open source co-opetition in company-hosted open source software projects: The cases of PyTorch, TensorFlow, and Transformers. *Proceedings of the ACM on Human-Computer Interaction*, 9(CSCW), Article 46. https://doi.org/10.1145/3710944

Osborne, C., Ding, J., & Kirk, H. R. (2024). The AI community building the future? A quantitative analysis of development activity on Hugging Face Hub. *Journal of Computational Social Science*, 7(2), 2067–2105. https://doi.org/10.1007/s42001-024-00300-8

Rahman, M. S., Gao, P., & Ji, Y. (2025). *HuggingGraph: Understanding the supply chain of LLM ecosystem*. arXiv. https://doi.org/10.48550/arXiv.2507.14240

Schultze, U., Prandelli, E., Salonen, P. I., & Van Alstyne, M. (2007). Internet-enabled co-production: Partnering or competing with customers? *Communications of the Association for Information Systems, 19*(1), 294–324. https://doi.org/10.17705/1CAIS.01915

Shan, J., & Jolly, D. R. (2011). Patterns of technological learning and catch-up strategies in latecomer firms: Case study in China’s telecom-equipment industry. *Journal of Technology Management in China,* 6(2), 153–170. https://doi.org/10.1108/17468771111142964

Shangguan News. (2025, March 19). *Tencent Cloud Shanghai summit systematically expounded on AI strategic thinking* [大模型正跨过产业化落地门槛 腾讯云上海峰会系统阐释AI战略思]. https://finance.eastmoney.com/a/202503193350428882.html

Smith, M. L., & Reilly, K. M. A. (Eds.). (2013). *Open development: Networked innovations in international development.* The MIT Press.

Tencent. (2024a). *Interim report 2023*. Tencent Holdings Limited.

Tencent. (2024b). *Tencent unveils new AI upgrades, proprietary innovations, and global solutions.*

Tencent. (2025a). *Customer success: GoTo.* https://www.tencentcloud.com/customers/detail/3500?lang=en&pg=

Tencent. (2025b). *Tencent: Customers*. https://www.tencentcloud.com/customers

Tencent. (2025c). *Tencent TI-One quick start* [TI-ONE 训练平台 快速入门].

Thun, E., Taglioni, D., Sturgeon, T., & Dallas, M. P. (2022). *Massive modularity: Understanding industry organization in the digital age: The case of mobile phone handsets* (Policy Research Working Paper No. 10164). The World Bank. https://doi.org/10.1596/1813-9450-10164

van der Vlist, F., Helmond, A., & Ferrari, F. (2024). Big AI: Cloud infrastructure dependence and the industrialisation of artificial intelligence. *Big Data & Society*, 11(1). https://doi.org/10.1177/20539517241232630

Widder, D. G., & Nafus, D. (2023). Dislocated accountabilities in the “AI supply chain”: Modularity and developers’ notions of responsibility. *Big Data & Society,* 10(1), 1–12. https://doi.org/10.1177/20539517231177620

Widder, D. G., Whittaker, M., & West, S. M. (2024). Why “open” AI systems are actually closed, and why this matters. *Nature*, 635(8040), 827–833.

Wooldridge, M. (2021). *A brief history of artificial intelligence: What it is, where we are, and where we are going*. Flatiron Books.

Yavuz, E. Y., Riehle, D., & Mehrotra, A. (2024). Why do companies create and how do they succeed with a vendor-led open source foundation. *Empirical Software Engineering*, 30(1), 40. https://doi.org/10.1007/s10664-024-10588-9

Yin, R. K. (1993). *Applications of case study research*. Sage Publications.

Yue, D., & Nagle, F. (2024). Igniting *innovation: Evidence from PyTorch on technology control in open collaboration* (Working Paper No. 25–013). Harvard Business School. https://www.hbs.edu/faculty/Pages/item.aspx?num=66443

Zhang, Y., Zhou, M., Stol, K.-J., Wu, J., & Jin, Z. (2020). How do companies collaborate in open source ecosystems? An empirical study of OpenStack. *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, 1196–1208. https://doi.org/10.1145/3377811.3380376

Research Paper

Struggling in the Middle: Understanding Adolescents Self-Regulation Amid Family, School, and Peers in Singapore

Aaron Pengyu ZHU[1]
CUI Shenglan[2]
CHO Janghee[1]

[1]Division of Industrial Design, National University of Singapore, Singapore
[2]School of Design, Hunan University, China

Abstract

Adolescents are often described as mobile natives, having grown up immersed in mobile technologies and seamlessly integrating them into their social, educational, and everyday lives. Yet, increasing concerns about overuse and its consequences for mental health have made adolescent digital well-being an urgent research priority. While external interventions – such as parental control apps and school policies – aim to promote digital well-being, they often exacerbate family tensions and erode adolescents' agency and autonomy. Adopting a socio-technical perspective, we examine how adolescents negotiate, resist, and reconfigure regulation within these power-laden relational structures. It reveals how young people mobilise peer relationships to develop nuanced bypassing strategies that balance autonomy, compliance, and belonging. Moving beyond individual self-control frameworks, the study reframes self-regulation as a co-regulated and relational practice, suggesting pathways toward more negotiated, systemic, and empathetic approaches to designing for adolescent digital well-being.

Introduction

Adolescents are often described as “mobile natives” (Prensky, 2001), having been immersed in mobile technology use from an early age and seamlessly integrating mobile technology into their daily lives. In China, the number of underage internet users has exceeded 193 million (Beijing Internet Report, 2024), and in Singapore, teenagers spend nearly 8.5 hours per day on screens (Tang, 2025). While mobile technologies can support adolescents’ social relationships, academic performance, and self-expression (Alluhidan et al., 2024; Lenhart et al., 2015; Livingstone & Helsper, 2007), they are also associated with problematic behaviours (Ricoy et al., 2022) that may contribute to mental health challenges (Blakemore, 2019; Orben & Blakemore, 2023) and negative effects on cognitive development and physical well-being (Chamorro et al., 2024; Sherer & Levounis, 2022; Toombs et al., 2022).

Currently, regulatory efforts largely rely on top-down interventions from external actors, such as governmental screen time guidelines (Gendil, 2024; Ministry of Health, 2023), and the “teen accounts” or “teen modes” by technology companies. These systems are often adopted by families as parental control tools (Welle, 2023). The technological approaches are easily bypassed, such as using adult accounts or reinstalling apps to reset controls (Hayes, 2024; Yang, 2023). Yet, such mechanisms overlook the broader digital turn in adolescents’ socialisation, self-expression, and learning (Hamilton et al., 2022). These external controls often fail to meaningfully engage adolescents’ voices and may give rise to oppositional emotions toward them, which can escalate tensions with their parents and schools (Chua & Mazmanian, 2021; Erickson et al., 2016; Goodyear et al., 2025).

However, adolescents are not unaware of the impact smartphones have on their lives. In fact, they often struggle with their usage. Studies show they actively reflect on their past experiences and emotional triggers to adjust their smartphone use (Chowdhury & Bunt, 2024; Dreier et al., 2024; Troll et al., 2021), demonstrating a level of autonomy. This struggle is multi-layered: they attempt to resist overuse and stay focus on academic or personal goals (Al-Abyadh et al., 2024) while facing peer-driven social pressures (Haug et al., 2015) and relying on smartphones for emotional regulation. However, compared to adults, adolescents still find it difficult to manage their device use through thoughtful deliberation and reflection (Davis, 2023; Weinstein & James, 2022). Most existing digital self-control tools rely on external and short-term constraints rather than cultivating long-term, internal self-regulation capacities, often lacking sustained user motivation; this indicates that externally structured and mechanised reflective guidance through mobile technology is ultimately unsustainable (Lyngs et al., 2019).

This underscores an urgent need for technological systems that support long-term, reflective engagement with digital practices (Davis et al., 2023; Sweigart et al., 2025), and for incorporating more nuanced understandings of adolescents’ perspectives, strategies, and needs to address the issue of mobile technologies usage among adolescents (Cho et al., 2025). This study aims to gain an in-depth understanding of the struggles faced by adolescents aged 12 to 15 in Singapore in their smartphone use. This age marks early adolescence, when cognitive and self-regulation abilities are still developing (Boehner et al., 2007; Gaver et al., 1999), making teens more vulnerable to smartphone overuse (Hay & Forrest, 2006). It also coincides with the typical start of secondary school in Singapore and other countries and frequent smartphone ownership. We aim to explore adolescents' struggles with smartphone use, their needs in balancing various purposes of use, and how technology can support the cultivation of self-regulation. The findings will inform the design of more responsive and sustainable technological support strategies that foster long-term, intrinsically motivated self-regulation to address the responsible mobile technologies use. At the same time, our study aligns with the socio-ecological approach in HCI (Freed et al., 2023; Hammond et al., 2023; Tang et al., 2025).

We introduce the Ecological Systems Theory (Bronfenbrenner, 2000), viewing adolescents' use of mobile technologies for regulation and the cultivation of meaningful engagement as a complex, multidimensional, and culturally and contextually sensitive process. It considers how self-regulation is shaped within the contexts of the individual, peers, family, school, and broader sociocultural and policy environments, with each external layer playing a role in shaping adolescents' capacity for self-regulation, we aim to develop a more nuanced understanding of control strategies and the struggles adolescents face when navigating intersecting systems of regulation. To this end, we sought to answer the following research questions:

RQ1: What forms of regulation strategies do adolescents experience in their daily mobile devices use?

RQ2: How do different regulation strategies shape adolescents' patterns of self-regulation in mobile technology use?

RQ3: How can these strategies be translated into technology improvement and design mechanisms?

To address our research questions, we conducted a pilot workshop study with three adolescents aged 12–15 in Singapore. The study explored the control strategies they encounter in everyday life, their personal perceptions of these strategies, and their expectations for ideal technologies that could help regulate their own mobile device use. We first analysed the types and specific forms of control strategies that adolescents face in their daily routines (RQ1), and then examined how these strategies influence their self-regulation and agency in mobile technology use (RQ2). Finally, we invited participants to use the Magic Machine (Andersen & Wakkary, 2019) and Design Fiction (Bleecker, 2022) methods to imagine and prototype their ideal forms of regulated technologies, from which we derived a series of design implications. Our contributions are threefold:

- We provide empirical insights into the modes of control that shape Singaporean adolescents' everyday mobile technology use.
- We offer a socio-ecological understanding of the complexities between adolescents' self-regulation and broader systems of external control.

Related work

Self-regulation in Adolescent Development

Self-regulation broadly refers to an individual's capacity to control their thoughts, behaviours, and emotional impulses in order to achieve personal goals (Novak & Clayton, 2001). It plays a vital role in adolescent development (Farley & Kim-Spoon, 2014) and is a key factor influencing learning, as well as cognitive and social functioning (Opdenakker, 2022). During adolescence, which is a period characterised by profound cognitive, emotional, and social transitions (Fomina et al., 2020), the ability to self-regulate becomes particularly important. Prior studies have shown that higher levels of self-regulation are associated with better academic performance (Blair & Diamond, 2008; Duckworth & Seligman, 2005) and greater engagement in prosocial behaviour (Bandura et al., 2001). Moreover, adolescents' self-regulation is not only shaped by their individual abilities but also by their relationships with parents, peers, teachers, and even romantic partners (Farley & Kim-Spoon, 2014; Opdenakker, 2022). This suggests that self-regulation develops within a complex social environment rather than as an isolated psychological trait.

In the context of mobile technology, adolescents are often described as "mobile natives" (Prensky, 2001), a generation that has grown up immersed in mobile technology and integrated it seamlessly into everyday life. Mobile technologies can support adolescents in maintaining social relationships, achieving academic goals, expressing themselves, and exploring their identities (Alluhidan et al., 2024; Lenhart et al., 2015; Livingstone & Helsper 2007). However, problematic patterns of use such as phubbing, FOMO, selfiphobia, nomophobia, vibranxiety, and sexting (Ricoy et al., 2022) have been linked to mental health concerns (Blakemore, 2019; Orben & Blakemore, 2023) and negative effects on cognitive development and physical well-being (Chamorro et al., 2024; Sherer & Levounis, 2022; Toombs et al., 2022). Self-regulation has been found to be an effective factor in mitigating problematic smartphone use (Cap et al., 2007; Xiao et al., 2025).

In HCI, studies on adolescents' self-regulation have mainly focused on reducing excessive technology use and addressing smartphone addiction (Adelhardt et al., 2018; Duckert & Barkhuus 2021; Kawas et al., 2021; Lanette et al., 2018; Wisniewski et al., 2015]. For instance, Chowdhury and Bunt conducted online co design workshops with early adolescents to identify youth driven, technology- based mediation strategies that encourage disengagement from mobile devices. Their findings highlight the importance of balancing adolescents' autonomy with parental control and considering emotional needs in design (Chowdhury & Bunt, 2023). Chen et al. (2022) developed TechLifeProbe, a system that integrates mobile and wearable technologies to support data sharing between adolescents and parents, helping reduce smartphone addiction and promote healthier family communication. Similarly, Kim et al. examined adolescents' use of BeReal, a social media app that uses random prompts and unfiltered photos to encourage authentic self-presentation. Their findings suggest that such authenticity sharing practices can promote subtle forms of self-regulation through social interaction (Kim et al., 2024).

Taken together, these studies indicate that adolescent digital self-regulation cannot be fully understood without considering their agency, motivation, and the social contexts that shape technology use. Although research on digital self-regulation in HCI is growing, little is known about how adolescents themselves develop strategies to manage their mobile device use (Chowdhury et al., 2025) Prior studies suggest that adolescents struggle more than adults with self-regulation (Muraven & Baumeister, 2000), often failing to manage device use in intentional or reflective ways (Farley & Kim-Spoon, 2014). In other words, their self-regulatory capacities are still developing, leading to tensions between internal desires and external control.

Most existing digital self-control tools rely on external and short-term restrictions rather than cultivating long-term and intrinsic self-regulation (Lyngs et al., 2019) These approaches often fail to sustain user motivation or address the deeper personal meanings that technology holds in adolescents' everyday lives (Roffarello & Russis, 2023) For instance, Davis et al. developed Locus, a tool designed to help adolescents reflect on unconscious social media use (Davis et al., 2023). However, its limited effectiveness among participants supports the argument that externally structured and mechanised reflective guidance is unsustainable (Lyngs et al., 2019). Beyond external control, fostering conscious meaning making (Lukoff et al., 2018) is essential to help adolescents engage with mobile technologies in more intentional and self-aware ways.

In this study, we adopt Farley and Kim Spoon's definition of self-regulation as *"the exertion of control over the self by the self"* (Farley & Kim-Spoon, 2014; Muraven & Baumeister, 2000). This definition emphasises that self-control is not only a skill but also an ongoing process in which individuals actively regulate their attention, emotions, and behaviours within social and situational contexts. Since adolescence is a period of heightened sensitivity to autonomy, self-esteem, and social evaluation, the development of self-regulation during this stage is both complex and fragile.

Tension between Adolescents Agency and External Intervention

During adolescence, individuals are often considered to have limited impulse control (Radesky, 2018), which has motivated numerous studies to explore how external interventions can help them manage their use of mobile technologies. Among these interventions, parental mediation has received the most attention. Prior work has proposed various strategies, such as negotiating a balance between parental involvement and adolescents' autonomy (Akter et al., 2022; Chowdhury et al., 2025) employing social translucence-based regulation to mediate information sharing and privacy by managing parents' visibility of information on adolescents' mobile devices (Yardi & Bruckman, 2011) and regulating usage through screen-time restrictions (Blackwell et al., 2016; Mazmanian & Lanette, 2017).

However, such interventions often provoke psychological resistance in adolescents due to concerns about privacy and autonomy (Erikson, 1994; Roffarello & Russis, 2023), which can heighten family tension and conflict (Akter et al., 2022) and make it increasingly difficult for parents to maintain consistent rules and boundaries for technology use. At the same time, existing research has tended to overlook adolescents' agency in developing alternative or evasive strategies to circumvent parental controls, thereby undermining the effectiveness of such interventions (Yardi & Bruckman, 2011). For instance, some adolescents use their parents' or grandparents' ID numbers, rent or trade other adults' accounts, or even use technical means to bypass adolescent identity verification systems (Guo, 2024; Schiano et al., 2016).

Beyond the family context, adolescents' technology use is also constrained in school settings. Magee et al. (2017) highlighted that local policies and access restrictions significantly influence adolescents' mobile technology use, for instance, prohibiting the use of certain devices or applications on campus, even for educational purposes. Nevertheless, adolescents often develop tactics of resistance to challenge these institutional controls. As Green (2002) observed in her discussion of surveillance and discipline, students tend to find private spaces to use their phones discreetly, such as under desks, to avoid teachers' monitoring, yet such "one-size-fits-all" restrictions may overlook the dynamic and situated needs of adolescents' digital lives. Wyche et al., for instance, documented how school-level restrictions on phones in Kenya hindered youth from seeking help or communicating in emergencies when managing Type 1 diabetes (Wyche et al., 2024).

Overall, existing research tends to treat adolescents' digital health as an issue to be controlled, rather than recognising their agency and reflective capacity in technology use. This gap motivates the present study, which seeks to explore how design can foster adolescents' internal self-regulation and sense of autonomy. Such an approach not only helps adolescents manage mobile technology use more effectively but also provides theoretical and practical insights for developing digital health interventions that support their psychological and behavioural growth.

Magic Machine & Design Fiction

Workshops are a widely used method in HCI research (Andersen & Wakkary, 2019), serving as a participatory approach that engages participants in envisioning new design possibilities (Lupton, 2017; Potapov & Marshall, 2020 ; Spinuzzi, 2005). HCI researchers have organised diverse workshops with adolescents, focusing on the design of personal informatics tools (Potapov & Marshall, 2020), online safety interventions (Agha et al., 2022), and social robots (Björling & Rose, 2019). Among these approaches, the Magic Machine technique has been used to encourage participants to explore creativity in personalised and imaginative ways, helping them envision and articulate their own capabilities, aspirations, and relationships with technology (Andersen & Wakkary, 2019) It has been adopted in studies with children and adolescents to support co-creation (Lepri et al., 2024) and enhance parent-child communication (Cao et al., 2025; Muñoz et al., 2019). Typically, participants use simple materials to create handcrafted prototypes and assign them "magical" functions, prompting them to develop radically personal visions of possible technological futures (Andersen & Wakkary, 2019). For adolescents, the open-ended and exploratory nature of the Magic Machine approach encourages active engagement, playfulness, and creative exploration (Verweij, 2025).

In addition to participatory workshops, design fiction has also been frequently used in HCI to explore possible futures. As a speculative-design approach, it aims to promote understanding, provoke creativity, raise questions, and inspire innovation and reflection (Almohamed et al., 2020; Baumer et al., 2020; Bleecker, 2022; Ng et al., 2021). Design fiction provides a concrete and imaginative way to reflect on and critique emerging technological, social, and ethical issues (Baumer et al., 2018). In studies involving adolescents and children, design fiction has been used as a participatory tool to help them imagine and discuss future scenarios, such as envisioning digital lifestyles (Sharma et al., 2022) or exploring potential technological futures (Gak et al., 2025). This method enables participants to creatively engage with complex sociotechnical topics that might otherwise be difficult to articulate (Gak et al., 2025). Some studies have applied design fiction to understand adolescents' perceptions of specific technologies, facilitating conversations about the kinds of futures they desire (Lee et al., 2025). It has also been used for educational purposes, such as creating fictions to foster critical understanding of new technologies (Tamashiro et al., 2021).

Grounded in this background and aligned with our research objectives, we adopt a co-creative approach with adolescents to explore the technological futures of self-regulation. This approach enables a deeper understanding of their perspectives and their relationships with technology, the environment, and society. Accordingly, we employ workshops as our primary method, combining the Magic Machine and design fiction techniques to investigate how adolescents envision future self-regulation through technology.

Method

In this study, our main goal is to understand how adolescents navigate regulation strategies in their everyday use of mobile technologies and how they perceive these multifaceted strategies. To achieve this, we conducted workshops based on the Magic Machine and design fiction approaches. The Magic Machine technique encourages participants to create handcrafted prototypes using simple materials and to imagine magical functionalities for them (Andersen & Wakkary, 2019). This method has been widely used in studies with adolescents and children to support co-creation (Andersen, 2013; Andersen & Wakkary, 2019) and to inspire exploration of technological futures. Design fiction, as a speculative research approach, uses fictional narratives and creative making to provoke reflection on the relationships between technology and society (Baumer et al., 2018; Bleecker, 2022; Gak et al., 2025). In our workshops, adolescents were invited to reflect on their relationships with mobile devices through conversation, discussion, and making, and to explore how technologies might better support their self-regulation.

Given our limited prior understanding of adolescents lived experiences, we first conducted a pilot study with three teenagers in Singapore. The primary purpose of this pilot was not only to gain preliminary insights into our research questions but also to establish a discursive space for developing more focused intervention strategies in subsequent studies.

Participants

Participants were publicly recruited through social media, and workshops were conducted on the researchers' campus with parental consent. We targeted adolescents aged 12 to 15, corresponding to the secondary school age range for most teenagers in Singapore. This age group roughly aligns with the early adolescent range (11–14 years) described in prior research (Holtz & Appel, 2011). All participants were active users of everyday mobile devices such as smartphones, tablets, and smartwatches. We use nickname to introduce them.

- Tracy, female, 12 years old, lives with her parents.
- Jaydon, male, 13 years old, lives with his parents.
- Brandon, male, 15 years old, Jaydon's older brother, lives with his grandmother.

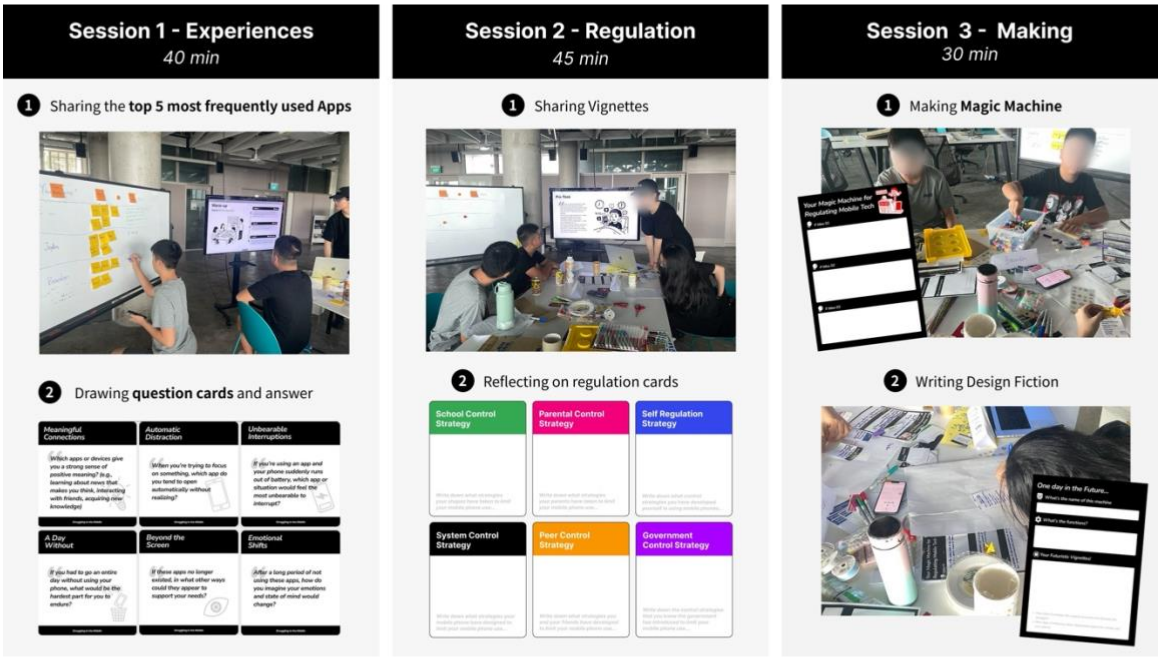

Session 1 - Experiences
40 min
1 Sharing the top 5 most frequently used Apps
2 Drawing question cards and answer
Meaningful Connections
Which apps or devices give you a strong sense of positive meaning? (e.g., learning about news that makes you think, interacting with friends, acquiring new knowledge)
Automatic Distraction
When you're trying to focus on something, which app do you tend to open automatically without realizing?
Unbearable Interruptions
If you're using an app and your phone suddenly runs out of battery, which app or situation would feel the most unbearable to interrupt?
A Day Without
If you had to go an entire day without using your phone, what would be the hardest part for you to endure?
Beyond the Screen
If these apps no longer existed, in what other ways could they appear to support your needs?
Emotional Shifts
After a long period of not using these apps, how do you imagine your emotions and state of mind would change?
Session 2 - Regulation
45 min
1 Sharing Vignettes
2 Reflecting on regulation cards
School Control Strategy
Parental Control Strategy
Self Regulation Strategy
System Control Strategy
Peer Control Strategy
Government Control Strategy
Session 3 - Making
30 min
1 Making Magic Machine
Your Magic Machine for Regulating Mobile Tech
2 Writing Design Fiction
One day in the Future...

Figure 1. The Procedure of the Workshop

Procedure

The workshop procedure is illustrated in Fig.1 and consisted of three sessions. We began by engaging participants in discussions about their lived experiences, focusing on the regulation strategies they employ in their everyday use of mobile technologies and their perceptions of these strategies. This was followed by making and writing activities grounded in the *Magic Machine* and *Design Fiction* approaches. The workshop lasted a total of 115 minutes and was facilitated by a member of the research team. After participants provided informed consent, the entire session was audio-recorded. All recordings were subsequently transcribed and anonymised to remove any personally identifiable information.

Session 1 – Reflecting on Daily Use and Experiences.

Session 1 focused on identifying and understanding participants' patterns of everyday mobile device use. The goal was to capture the range of applications, habitual usage, and time spent on devices before investigating regulation strategies. Participants were asked to represent their usage data on a whiteboard. We then provided six question cards related to mobile device experiences. One participant drew a card at a time, and all participants collectively responded and discussed. The questions addressed topics such as the positive significance of mobile devices, long-term impressions, the importance of specific apps in daily life, apps that easily capture attention, and participants' imagination of a day without a mobile device. These prompts aimed to encourage open discussion on the impact of mobile devices and to facilitate reflection on both using and not using them.

Session 2 – Exploring Regulation Strategies and Tensions.

Session 2 began with participants sharing vignettes they had prepared prior to the workshop. Participants were asked by email to create a vignette describing a struggle related to phone use, for example, "a moment when you realise you've used your phone more than you intended, or when you find yourself mindlessly scrolling and navigating social media.” Based on these vignettes, participants were invited to use provided strategy cards to document regulation mechanisms originating from themselves, their parents, school, or other organisations. Participants then discussed these strategies and shared their experiences. Next, they placed the different strategy cards on a large canvas, marking the relationships between strategies, such as conflicts, reinforcement, or undermining, to stimulate deeper discussion and reflection.

Session 3 – Envisioning Alternative Self-Regulation and Future Technologies.

Session 3 applied the Magic Machine method to encourage participants to imagine speculative technological tools for future self-regulation based on reflections from the first two sessions. Participants were guided to envision themselves living in 2045 and to conceptualise a *magical device* for managing personal mobile device use, or even alternative devices that could transform future adolescent regulation patterns. Participants first proposed three concepts on an idea canvas and then selected one to develop into a Magic Machine. They subsequently elaborated scenarios and detailed narratives of how the device would be used through Design Fiction. AI tools were allowed to support creative ideation and visual generation. Ultimately, we collected three devices along with their corresponding stories.

Data Analysis

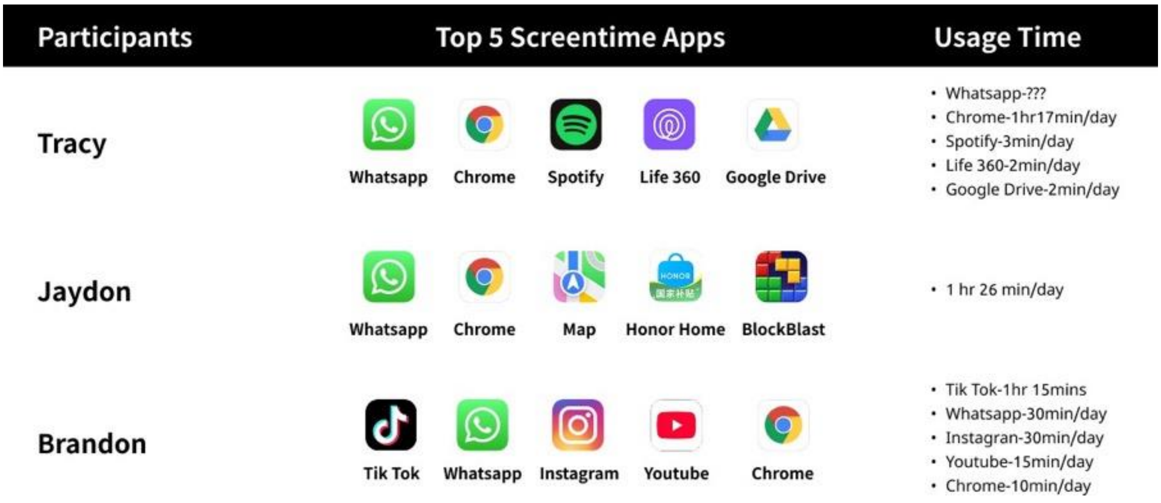
Two researchers analysed the transcripts, photographs, and participants submitted or completed materials using Braun and Clarke’s reflexive thematic analysis approach (Braun & Clarke, 2019). The analysis followed a recursive and reflective process that embraced the researchers’ subjectivity and valued the unique perspectives of individual participants, rather than focusing on conceptual saturation or achieving a particular sample size. Knowledge was treated as a co-constructed interpretation by the researchers. Both analysts familiarised themselves with the data, generated initial codes, and developed higher-level themes. These themes were further reviewed, refined, and iteratively discussed with a third researcher. The emerging themes primarily addressed participants’ patterns of mobile device use, regulation strategies they encountered, bypassing strategies, and alternative visions generated through the Magic Machine and design fiction exercises.

Findings

The Device Usage Pattern among Participants

We first examined the participants’ digital device usage habits. All three adolescents owned multiple mobile devices in addition to their smartphones, including iPads, MacBooks, and school-provided Personal Learning Devices (PLDs). Although we informed participants that they could bring any electronic devices they used daily, they still prioritised smartphone, and recording its screen time during workshop.

Consistent with prior surveys of Singaporean adolescents, participants’ mobile device use primarily fell into three categories: social communication, learning, and entertainment (Boase et al., 2021; Tang 2025) (see Figure 2). Tracy and Jaydon most frequently used WhatsApp. Beyond staying connected with friends and family, they also used the app to access news and information, for example, *“...following The Straits Times and CNA Channel”*. Chrome was their second most-used tool, primarily for learning, general information retrieval, or accessing other web-based applications such as ChatGPT and Proplexity. Other apps were mostly used for entertainment and leisure; all three participants listened to audio content via Spotify.



Participants	Top 5 Screentime Apps	Usage Time
Tracy	Whatsapp Chrome Spotify Life 360 Google Drive	• Whatsapp-??? • Chrome-1hr17min/day • Spotify-3min/day • Life 360-2min/day • Google Drive-2min/day
Jaydon	Whatsapp Chrome Map Honor Home BlockBlast	• 1 hr 26 min/day
Brandon	Tik Tok Whatsapp Instagram Youtube Chrome	• Tik Tok-1hr 15mins • Whatsapp-30min/day • Instagran-30min/day • Youtube-15min/day • Chrome-10min/day

Figure 2. The Top 5 Use Apps among Participants

Notably, Jaydon's parents imposed strict limits on his video consumption. Consequently, browsing Honor Home[1] and playing BlockBlast[2] became his primary forms of mobile entertainment. In contrast, his older brother Brandon faced no parental restrictions, and apps such as TikTok, Instagram, and YouTube occupied the majority of his screen time. Participants also reported that they could access web-based versions of apps through Chrome, allowing them to bypass certain parental app limitations.

In terms of overall screen time, all three adolescents reported daily usage under two hours, considerably lower than the 8.5 hours per day reported in a recent CNA survey of Singaporean adolescents (Tang, 2025). However, there were some inaccuracies in both recorded screen time and the purposes for which apps were used. For instance, due to parental restrictions on WhatsApp usage, Tracy installed restricted apps via third-party app marketplaces, which were not captured by the system's screen time tracking, as she noted: "My WhatsApp's time is not inside…"

Regulation Strategies

The participants described smartphone regulation strategies that primarily originated from their families, schools, and, to a limited extent, self-regulation. Although app marketplaces and the Singaporean government have issued relevant guidelines, participants did not reference these resources in our study. Moreover, while prior research has highlighted the influence of peers on adolescent smartphone use, we found that peers provided little support for participants' self-regulation efforts.

Limited Agency under Parent-Centred Control.

Align with prior study (Lee et al., 2024), parental mediation is still the primary form of external

[1] Huawei's app marketplace.
[2] A puzzle game

intervention. It primarily takes the form of system-based restrictive mediation, which controls the types of content accessed, the amount of time spent on certain activities, and limits phone usage in specific contexts (Hiniker et al., 2016). Tracy described experiencing strict restrictive mediation, specifically through system settings such as *"time limits, app limits, downtime, and app blocking",* as well as time- and location-based restrictions: "My parents make sure no phones are allowed during meals. No phone use on the bus, no phone use during family time." Tracy also mentioned that her parents would directly check the content on her phone: "They will just like see my context if I'm texting someone else that I don't know or something like that."

Jaydon received more explicit parental control. He explained, *"They install parental controls so I can't download certain apps."* Specifically, he was restricted from downloading any video-streaming apps like TikTok or YouTube. Additionally, Jaydon's parents also monitored his game downloads, which led to puzzle games like *"BlockBlast"* occupying a large portion of his screen time. However, Brandon, as Jardon's older brother, did not receive much technical mediation from his parents [15]. He had the freedom to use short video and social media apps. He explained, *"Maybe it's because I have better self-control, so I've earned my parents' trust."* Unlike other participants, Brandon shared that his phone use was primarily managed by his grandmother, as they lived together. He said, *"My grandma keeps my phone at 10:45 p.m. just before bed, saying I need to get enough sleep."* This grandparental control differs from parental control, focusing on the physical management of device itself, because Brandon's grandmother does not use a smartphone.

As prior research has extensively shown (Akter et al., 2022; Dumaru & Al-Ameen, 2025; Dumaru et al., 2024; Hiniker et al., 2016; Ibrahim et al., 2025), restrictive strategies like these often intensify negative emotions among adolescents and lead to conflicts with their parents (Ghosh et al., 2018). Participants with stricter parental control (e.g., Tracy and Jaydon) used terms such as "angry", "lack of trust", and "restrictions on freedom" to express their feelings about these control strategies. Tracy shared, "Maybe I get a bit angry because they don't trust me enough to manage my own phone usage." Tracy also expressed confusion about the restrictions, saying, "I don't understand why they set limits for my WhatsApp when it's just for communication." This highlights parents' concerns about online safety in open chat apps, but also underscores the adolescent's desire for "digital freedom" and the lack of negotiation between this desire and parental control.

Additionally, Tracy mentioned, "My parents need to work, so they make sure we sleep earlier and don't disturb them while they're working. So, you must learn how to manage your time properly and finish your homework before that." Parents' approaches to regulating adolescents' mobile device use are largely influenced by their own time management routines, leading to rigid regulation strategies. This points to the need for negotiation between adolescents' personal time, parents' schedules, and family time.

School Strategies for Managing Smartphone Use.

The three participants shared the ways their mobile device use was regulated at school. As with the mobile phone bans implemented in most Singaporean schools (Telok Kurau Primary School, 2025; Tampines Secondary School, 2025), Brandon and Jaydon reported that after arriving at school, teachers would collect their phones and store them in lockers until the end of the day. Tracy's school, in contrast, was relatively more lenient: *"no using our phones in school…can take* *phone to school just can not use during class lah"*. Nevertheless, some students still attempted to circumvent these rules. Brandon shared a common practice among his peers: *"In my school, they hide the phone in the bag. Then during recess, they secretly use the phone..."*. Beyond strict enforcement, schools also employed softer regulatory strategies. Jaydon noted, *"They give speeches on the harmful effects of watching phone for too long"*. All three participants expressed a sense of powerlessness toward these exhortative measures, stating that they *"feel nothing on this"*.

Outside of mobile phones, some schools appeared less strict regarding other mobile technologies. Brandon observed this inconsistency: *"They were like...make sure the campus is phone free. So I bring my iPad there."* While the primary goal of school regulations is to prevent distraction during lessons and to encourage social interaction among peers during breaks (Tushara, 2024), prolonged restrictions and control over mobile technologies inevitably led adolescents to develop diverse bypassing behaviours on campus.

Alternatives as Self-Regulation Strategies.

Tracy's self-imposed limits were somewhat vague and flexible, as she stated: *"(I just) don't use phone too long at once"*. However, this approach occasionally conflicted with the bypassing strategies she described (see Section 4.3), leaving it unclear whether this intention-based method at the personal level effectively promotes self-regulation or reduces screen time.

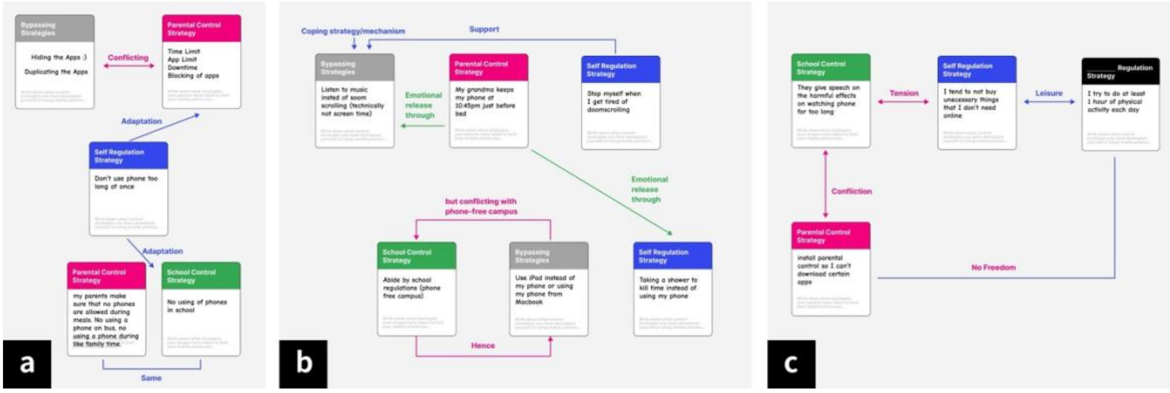

Bypassing Strategies
Hiding the Apps :)
Duplicating the Apps
Conflicting
Parental Control Strategy
Time Limit
App Limit
Downtime
Blocking of apps
Adaptation
Self Regulation Strategy
Don't use phone too long at once
Adaptation
Parental Control Strategy
my parents make sure that no phones are allowed during meals. No using a phone on bus, no using a phone during like family time.
School Control Strategy
No using of phones in school
Same
a
Coping strategy/mechanism
Support
Bypassing Strategies
Listen to music insted of soom scrolling (technically not screen time)
Emotional release through
Parental Control Strategy
My grandma keeps my phone at 10:45pm just before bed
Self Regulation Strategy
Stop myself when I get tired of doomscrolling
Emotional release through
but conflicting with phone-free campus
School Control Strategy
Abide by school regulations (phone free campus)
Bypassing Strategies
Use iPad instead of my phone or using my phone from Macbook
Self Regulation Strategy
Taking a shower to kill time instead of using my phone
Hence
b
School Control Strategy
They give speech on the harmful effects on watching phone for too long
Tension
Self Regulation Strategy
I tend to not buy unecessary things that I don't need online
Leisure
Regulation Strategy
I try to do at least 1 hour of physical activity each day
Confliction
Parental Control Strategy
install parental control so I can't download certain apps
No Freedom
c

Figure 3. The Interrelationships Among the Strategies Described by the Participants

Note: Handwritten by the participants and graphically represented by the researcher.

Brandon and Jaydon, in contrast, attempted to control their phone use primarily in response to the physical discomfort caused by prolonged usage. Brandon explained, *"try to not use too long, cause my body it's gonna get affected"*. In these cases, they intentionally engaged in outdoor activities or other bodily stimulation to regain calm. For example, Jaydon reported trying to perform at least one hour of physical activity daily to divert attention from his phone, while Brandon stated: *"(I will) take a shower to kill time instead of using my phone...for calm down lah"*. These self-regulation strategies lack a strong sense of agency at the personal level, making it difficult to determine whether they are side effects of parental and school regulation or proactive management of personal priorities, such as academic obligations. As Tracy mapped in her reflection (see Figure 3), her self-regulation strategies emerged as an adaptation to both parental and school regulation.

Notably, Brandon's self-regulation exhibited a form of self-deceptive rationalisation. He perceived that as long as he did not spend time on the phone itself, he was exercising self-control. He shared: *"I use my iPad instead of my phone or I use the phone mirroring app on a MacBook ... I feel very happy because like at least I got something I can still use. I still can doom scroll without using my phone...at least I can do work"*. Brandon framed self-regulation as detachment from the mobile phone as a device rather than restraint of behaviour. By shifting the medium through which interaction occurs, he felt a sense of control. This illustrates that smartphones remain the most addictive device in adolescents' self-perception.

This self-deceptive approach also extended to the use of audio applications (e.g., Spotify). Brandon rationalised: *"technically it's not screen time"*. Using this logic, he mentally categorised audio listening as a behaviour that does not require regulation. However, audio apps still lack effective filtering and age restrictions for adult content (Spotify, n.d.).

Bypassing Strategies Developed in Response to Control

Beyond formal regulation, the participants shared a range of bypassing strategies to circumvent restrictions imposed by parents and schools, including the use of dual systems. Specifically, Tracy, who uses an Android device, demonstrated the most direct and sophisticated bypassing strategies. She shared:

> *I'll like either duplicate the app because app limits only work on the original app... I have an app that can duplicate lah and then I can hide the apps... and like the total time limit for my phone doesn't include it.*

Through this mechanism, she was able to download applications that were otherwise prohibited by her parents, while manipulating her screen-time visibility by hiding and using duplicated apps. This performative practice enabled Tracy to negotiate a delicate balance between asserting autonomy in her phone use and appearing compliant under parental control. However, Jaydon's bypassing attempts took the form of a more direct confrontation with parental control. He shared: *"trying to key in the limited password – they (parents) will get a notification, then they'll ask me why I keep trying to bypass the system."* By repeatedly attempting and verifying the password, he sought the freedom to download applications from the app store. This also reflects how the type of operating system shapes the degree of agency adolescents possess in their bypassing practices. Within the more closed iOS environment, Jaydon was unable to adopt duplicating mechanisms similar to Tracy's.

At the school level, participants referred to strategies learned from classmates or friends. Although they claimed not to have tried them personally, Tracy mentioned: *"actually on my PLD, right, it's supposed to like block all those like game websites. But then my schoolmates found some that isn't like"*. Even though PLDs come pre-installed with device management applications (Koh, 2025), adolescents actively explored ways to use the school-mandated PLDs for non-academic entertainment, often sharing these approaches with peers. Brandon described a similar practice: *"They use like a Google Classroom link or a copy of it. Then after you press it, all the games come up already."*

These school-related bypassing strategies exploited the existing devices and platforms that were officially intended for learning purposes. Through such playful manipulations and subtle disruptions within school-regulated systems, the adolescents further blurred the boundary between learning and leisure. Moreover, these tactics were enthusiastically circulated among peers – *sharing* clever workarounds for bypassing school restrictions often proved more appealing than discussing challenges of self-regulation or external control.

For these bypassing strategies, Tracy described a sense of guilty pleasure, while the participants more broadly expressed feeling *"happy",* perceiving such actions as a way to reclaim greater freedom in using their phones. They attributed the need to employ these strategies to their parents' lack of trust and understanding of why such control was necessary.

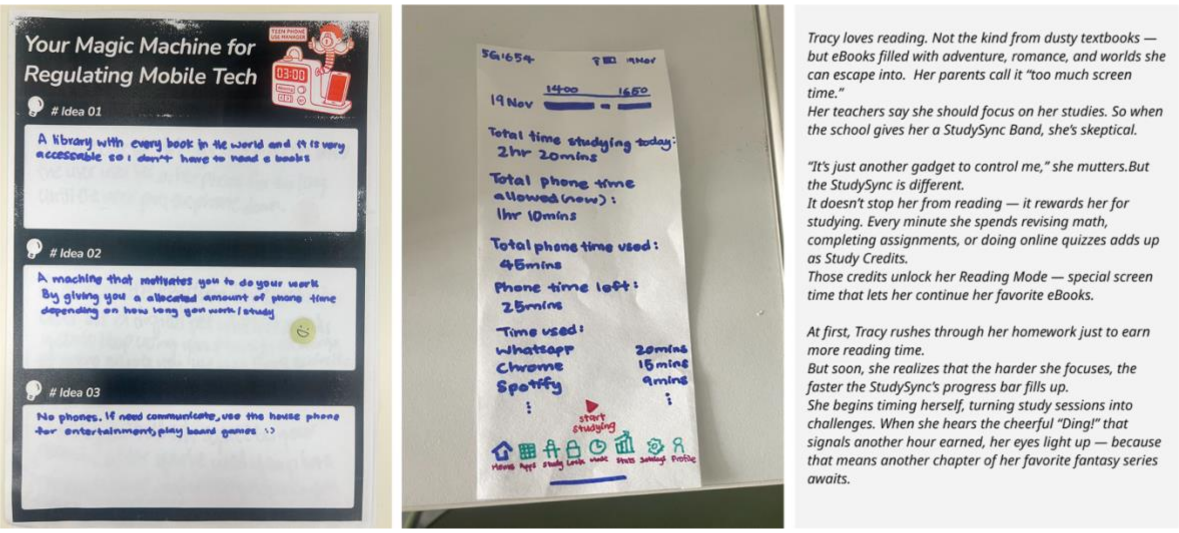
Magic Machine & Design Fiction

Through the creation of the Magic Machine and Design Fiction, the three participants demonstrated their understanding of and expectations for envisioned self-regulation technologies. Our study found that their considerations for ideal mobile device use encompassed multiple dimensions.

StudySync – Temporal Autonomy

Tracy envisioned a smart wristband called *StudySync*, designed to encourage healthy phone use by exchanging *study time* for *screen time* (see Figure 4). This time-exchange design reflects typical gamification features commonly applied in behaviour change (Edwards et al., 2016; Truesdell et al., 2025) and learning contexts (Sailer & Homner, 2020). StudySync establishes a computable link between learning and entertainment through a quantifiable incentive mechanism, replacing restrictions with rewards and seeking a balance between studying and phone use.

At the beginning of the story, Tracy expressed scepticism toward the device, mumbling: *"It's just another gadget to control me."* This reflects adolescents' resistance to surveillance-oriented technologies, with the device symbolising tangible forms of external oversight, such as parental or school monitoring. However, as she engaged with the device, Tracy gradually realised that *every minute spent reviewing math, completing assignments, or taking online quizzes would earn study points*. Although still reliant on StudySync as an external regulatory device, the autonomy of its use was under her own control. By aligning the incentive mechanism with her personal priorities, Tracy gradually regained agency over her mobile device use, shifting toward a proactive and positive self-regulation mode.



Your Magic Machine for Regulating Mobile Tech

Idea 01

A library with every book in the world and it is very accessable so i don't have to read e books

Idea 02

A machine that motivates you to do your work By giving you a allocated amount of phone time depending on how long you work/study

Idea 03

No phones. If need communicate, use the house phone for entertainment, play board games :)

56:654 19Nov

19 Nov 1400 1650

Total time studying today: 2hr 20mins

Total phone time allowed (now): 1hr 10mins

Total phone time used: 45mins

Phone time left: 25mins

Time used:

Whatsapp	20mins
Chrome	15mins
Spotify	9mins
⋮	⋮

Tracy loves reading. Not the kind from dusty textbooks — but eBooks filled with adventure, romance, and worlds she can escape into. Her parents call it "too much screen time."
Her teachers say she should focus on her studies. So when the school gives her a StudySync Band, she's skeptical.

"It's just another gadget to control me," she mutters.But the StudySync is different.
It doesn't stop her from reading — it rewards her for studying. Every minute she spends revising math, completing assignments, or doing online quizzes adds up as Study Credits.
Those credits unlock her Reading Mode — special screen time that lets her continue her favorite eBooks.

At first, Tracy rushes through her homework just to earn more reading time.
But soon, she realizes that the harder she focuses, the faster the StudySync's progress bar fills up.
She begins timing herself, turning study sessions into challenges. When she hears the cheerful "Ding!" that signals another hour earned, her eyes light up — because that means another chapter of her favorite fantasy series awaits.

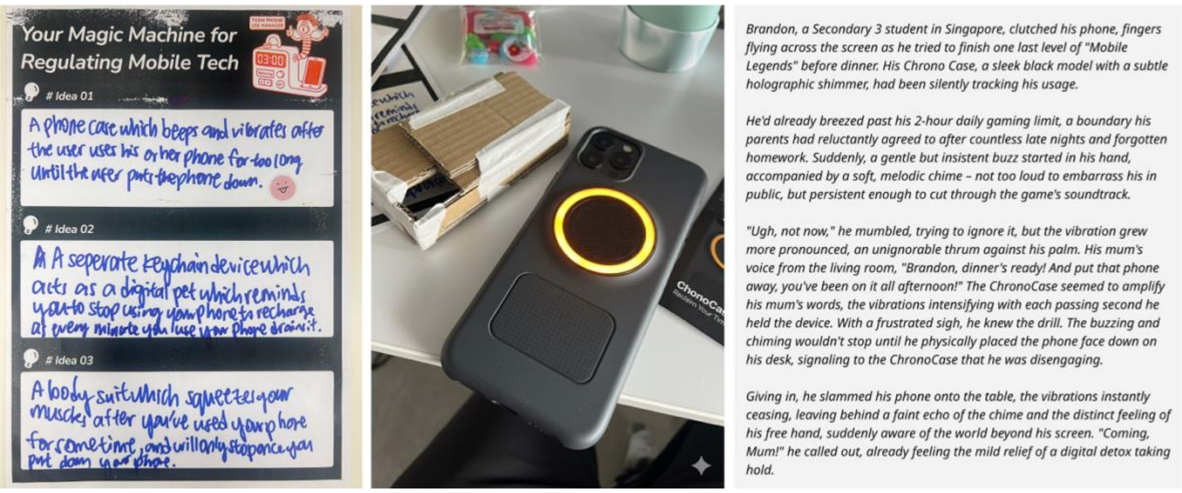
Figure 4. Tracy's Ideas, Magic Machine And Design Fiction

At the same time, StudySync embodies a reimagined form of negotiation. The mechanism transforms unilateral constraints into a *negotiable relationship*, allowing learning and phone use to move from opposition to dynamic coordination, reflecting a recalibration between external management and personal control. However, this reward-based design also reveals adolescents' perception of mobile devices for entertainment and leisure. Such usage is treated as a reward rather than a legitimate form of leisure time. This implicit framing links phone-based entertainment with potentially addictive behaviour, even though research suggests this association is not necessarily valid (Katevas et al., 2018). Mobile device use is often perceived as a restricted form of entertainment that should be limited. Therefore, whether adolescents' leisure use of mobile devices should be considered in opposition to learning remains an open question. Furthermore, the concept of StudySync reiterates our earlier findings: rather than intentional self-regulation, much of adolescents' engagement reflects an adaptation to external regulation.

ChronoCase – Haptic-Based Natural Reminders

Brandon envisioned a device called *ChronoCase*, a phone case that provides vibratory haptic feedback based on phone usage duration to gently remind users to put down their devices (see Figure 5).

ChronoCase is characterised as *"gentle, soft"*, contrasting with the more common approach of forcibly stopping phone use. Unlike prior research, where parents would abruptly terminate adolescents' ongoing gameplay [30], ChronoCase offers continuous, gentle reminders. Brandon depicted a scenario while playing a game: when his screen time reached a limit, the system did not automatically interrupt the game but instead delivered progressively stronger haptic feedback, establishing a subtle rhythm. In this way, ChronoCase extends regulation into an embodied dimension. By providing a bodily experience, it reshapes adolescents' awareness of their phone use and promotes an *embodied understanding* of mobile device regulation. Rather than simply being controlled, this approach encourages reflective, embodied autonomy.



Your Magic Machine for Regulating Mobile Tech

Idea 01

A phone case which beeps and vibrates after the user uses his or her phone for too long until the user puts the phone down.

Idea 02

A A seperate keychain device which acts as a digital pet which reminds you to stop using your phone to recharge as every minute you use your phone drains it.

Idea 03

A body suit which squeezes your muscles after you've used your phone for sometime, and will only stop once you put down your phone.

Brandon, a Secondary 3 student in Singapore, clutched his phone, fingers flying across the screen as he tried to finish one last level of "Mobile Legends" before dinner. His Chrono Case, a sleek black model with a subtle holographic shimmer, had been silently tracking his usage.

He'd already breezed past his 2-hour daily gaming limit, a boundary his parents had reluctantly agreed to after countless late nights and forgotten homework. Suddenly, a gentle but insistent buzz started in his hand, accompanied by a soft, melodic chime – not too loud to embarrass his in public, but persistent enough to cut through the game's soundtrack.

"Ugh, not now," he mumbled, trying to ignore it, but the vibration grew more pronounced, an unignorable thrum against his palm. His mum's voice from the living room, "Brandon, dinner's ready! And put that phone away, you've been on it all afternoon!" The ChronoCase seemed to amplify his mum's words, the vibrations intensifying with each passing second he held the device. With a frustrated sigh, he knew the drill. The buzzing and chiming wouldn't stop until he physically placed the phone face down on his desk, signaling to the ChronoCase that he was disengaging.

Giving in, he slammed his phone onto the table, the vibrations instantly ceasing, leaving behind a faint echo of the chime and the distinct feeling of his free hand, suddenly aware of the world beyond his screen. "Coming, Mum!" he called out, already feeling the mild relief of a digital detox taking hold.

Figure 5. Brandon's Ideas, Magic Machine and Design Fiction

After a brief moment of struggle, Brandon put down his phone. He described the experience: "leaving behind a faint echo of the chime and the distinct feeling of his free hand, suddenly aware of the world beyond his screen." This reaction aligns with prior findings, where people reported feeling a loss of autonomy when using their phones for entertainment (Lukoff et al., 2018). The sense of relief following disengagement reflects Brandon's internal tension: he rationally recognises the importance of self-control but relies on external support to enact it in practice.

Another feature of ChronoCase is that it provides Brandon with a sense of privacy regarding phone regulation. He noted: *"not too loud to embarrass him in public, but persistent enough to cut through the game's soundtrack."* This detail reflects Brandon's sensitivity to self-disclosure when being controlled or exposed. He sought to maintain the privacy of his self-regulation, not wanting others to notice in public or even within family settings. Similar to prior research, this detail highlights adolescents' sensitivity to privacy and dignity in digital interventions.

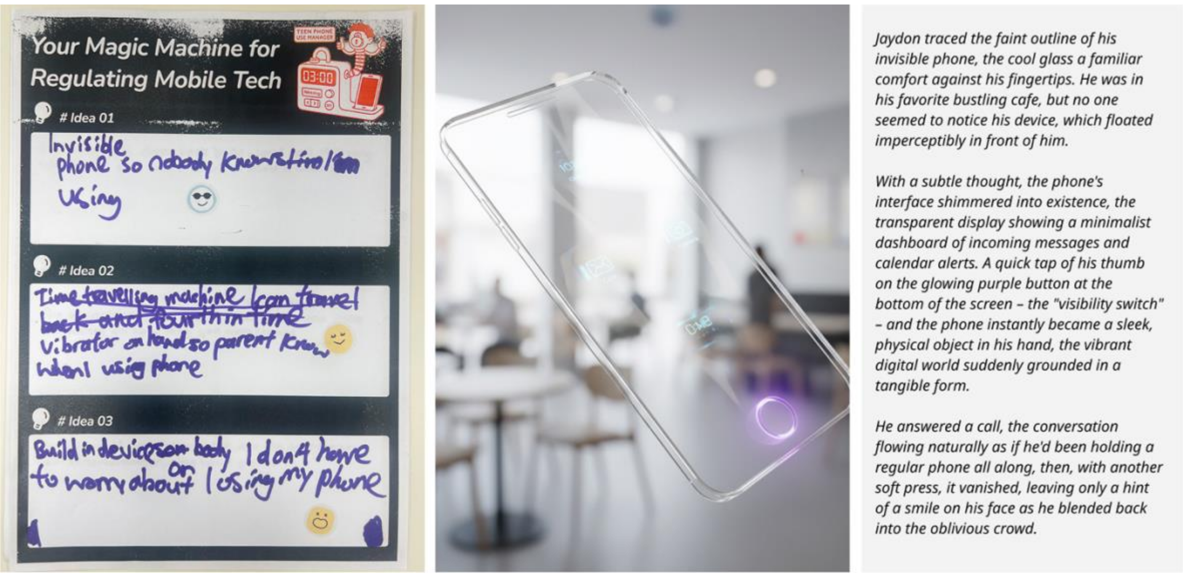
Invisible Phone – Negotiating Autonomy

Unlike the previous two cases, Jaydon did not envision a technology designed to support self-regulation. Instead, he focused on pursuing greater freedom in mobile technology use. He imagined an *invisible phone* that could switch freely between transparency and physical presence, allowing him to actively determine the visibility of his device during daily use (see Figure 6).

For Jaydon, this invisibility serves as the most effective means of escaping parental control, reflecting the strict limitations he experienced regarding app downloads and screen time. The concept of a phone that can shift between visible and invisible states also illustrates adolescents' negotiation of autonomy, highlighting how they seek flexible control over the degree of freedom in their mobile device use.

Discussion

Our findings reveal how adolescents' everyday use of mobile technologies is shaped by the intertwined influences of family, school, and their own bypassing strategies, which collectively affect both their patterns of self-regulation and their perceptions of what self-regulation means.



Your Magic Machine for Regulating Mobile Tech

Idea 01

Invisible phone so nobody Knows that I[illegible] using

Idea 02

~~Time travelling machine I can travel back and fourth in time~~ vibrator on hand so parent Know when I using phone

Idea 03

Build in devices on body I don't have to worry about losing my phone

Jaydon traced the faint outline of his invisible phone, the cool glass a familiar comfort against his fingertips. He was in his favorite bustling cafe, but no one seemed to notice his device, which floated imperceptibly in front of him.

With a subtle thought, the phone's interface shimmered into existence, the transparent display showing a minimalist dashboard of incoming messages and calendar alerts. A quick tap of his thumb on the glowing purple button at the bottom of the screen – the "visibility switch" – and the phone instantly became a sleek, physical object in his hand, the vibrant digital world suddenly grounded in a tangible form.

He answered a call, the conversation flowing naturally as if he'd been holding a regular phone all along, then, with another soft press, it vanished, leaving only a hint of a smile on his face as he blended back into the oblivious crowd.

Figure 6. Jaydon's Ideas, Magic Machine and Design Fiction

Agency as Negotiation, Not Resistance

Our study reveals that adolescents are not passive recipients of parental or systemic management; rather, they demonstrate considerable agency when faced with external control. For instance, when parents use mobile applications to restrict screen time or app installations, or when schools impose limitations on phone use, adolescents often develop various bypassing strategies and share these tactics among peers. Although such external regulation is often motivated by care or protective intent, it inadvertently reinforces power asymmetries within the family and diminishes adolescents' autonomy and sense of agency in using mobile technologies.

According to Self-Determination Theory (SDT), relatedness, autonomy, and competence are innate psychological needs essential for optimal development and wellbeing (Deci & Ryan, 2012). When adolescents' autonomy is constrained, they may engage in bypassing behaviours as a means to regain control – an act of resistance that reflects their response to frustrated agency. While previous research has shown that such resistance or bypassing can provoke conflicts between parents and adolescents (Ghosh et al., 2018) our findings suggest that this tension does not always surface explicitly. For example, Tracy adopted a more covert form of bypassing by downloading restricted applications through external app stores unnoticed by her parents, illustrating the limited effectiveness of parental restrictive mediation.

Across our three cases, the negotiation-based models of parental mediation proposed by prior scholars (Akter et al., 2022; Berget et al., 2020; Chowdhury & Bunt, 2024; Nikken & Schols, 2015) – such as collaboratively discussing the benefits and drawbacks of technology use – were largely absent. Instead, the adolescents in our study perceived their parents' regulation as a one-way enforcement, rather than a two-way dialogue. This dynamic may be linked to the high-power distance characteristic of traditional Chinese families (Whyte & Ikels, 2004), as all three participants were Singaporean Chinese adolescents. Under such cultural hierarchies, adolescents' agency tends to manifest as subtle forms of resistance rather than open negotiation to claim autonomy over their mobile technology use.

McKenzie's study in Thailand demonstrates how adolescents, by teaching their parents and elders to use mobile applications, negotiate traditional Asian family hierarchies and authority structures through digital technology (McKenzie, 2019). The key to this practice lies in enabling parents and elders to appreciate the positive value of digital technologies while recognising adolescents' own sense-making through teaching. Building on this insight, encouraging adolescents to guide older family members in using emerging tools – such as AI applications – could serve as a constructive way to renegotiate agency. Through shared goals (Chen et al., 2022) and collective reflection prompts (e.g., "What did we learn together today?"), such practices could transform unilateral control into mutual understanding, fostering adolescents' sense of accomplishment and joint meaning-making.

Furthermore, in our cases, both Brandon and Jaydon experienced different levels of parental regulation – largely depending on the degree of perceived trust from their parents. Yet, this raises deeper questions: Is such trust performative? And who defines or perceives it – the parent or the adolescent? These nuances remain embedded within the parent–teen relationship as mediated through mobile technologies.

Learning to Negotiate Agency through Bypassing

Compared to parental control, the school-level regulatory systems described by our participants were far less flexible and negotiable. Prior research has shown that adolescents are particularly susceptible to peer influence (Ngyuen et al., 2024), which makes school regulation less conducive to the kind of negotiation that may occur between parents and children at home. Any attempt to individualise regulation within a collective environment would likely produce ripple effects and potential disputes among students. Consequently, schools often adopt uniform, one-size-fits-all policies, such as confiscating mobile phones or strictly limiting their use during school hours. This maximal deprivation of agency led adolescents to develop diverse bypassing strategies, particularly those involving sharing tips with peers or exploiting loopholes in school-provided PLDs for entertainment purposes. However, existing HCI research has paid little attention to the range of bypassing strategies adolescents develop in response to such restrictions, or to the underlying reasons why these practices emerge.

Based on our findings, we call for viewing adolescents' bypassing behaviours not merely as acts of defiance but as a generative lens through which to understand how agency is constrained and reconstituted, and to inform the development of more negotiated forms of regulation. For example, when Brandon described how his classmates sent gaming links disguised as Google Classroom assignments via PLDs, this reflected how adolescents' agency was constrained by overly closed educational systems and by the deprivation of access to alternative devices under parental control. As a result, adolescents sought spaces of entertainment and autonomy within the limited technological infrastructures available to them.

By studying how adolescents navigate and circumvent external controls, future regulatory technologies could be designed to engage in more context-sensitive negotiations – acknowledging how different layers of control simultaneously constrain and reshape adolescents' agency. Such an approach would further illuminate the intersectional nature of adolescent agency, as situated at the nexus of school discipline, parental oversight, and peers.

Design Implications from Magic Machine

Designing Priority-Oriented Incentive Mechanisms to Foster Self-Regulation.

The design of StudySync demonstrates how external motivation can be utilised to provide adolescents with extrinsic rewards (Chowdhury & Bunt, 2024). Through such incentive mechanisms, adolescents are first encouraged to focus on completing their personal priorities, which subsequently grants them greater autonomy over mobile device use. Building on this idea, future regulatory applications could enable priority-based constraints on adolescents' personal mobile devices or PLDs, managed by parents or schools. These systems could tie levels of autonomy and device access to the quantity and quality of completed priority tasks – such as academic assignments or developmental goals. By adopting such incentive-driven regulation mechanisms, designers can help strike a balance between managing adolescents' academic development and mobile technology use, while gradually nurturing their awareness and capacity for self-regulation.

Designing tangible devices as a medium for somatic intervention.

ChronoCase demonstrates how tangible interaction can move beyond traditional intervention approaches that centre on visual interfaces and time thresholds, transforming control into perceptual cues. Through gentle and gradual haptic rhythms, the system no longer commands users to stop, but instead evokes self-awareness experientially, allowing it to be internalised as a habit. Existing systems and designs focused on parental control and adolescent self-regulation still predominantly rely on mobile applications to achieve effective management (Chowdhury & Bunt, 2024; Kawas et al., 2021). However, direct interaction with mobile interfaces continues to capture adolescents' attention, often undermining the intended regulatory goals. Future regulation technologies could explore tangible, accessory-based designs that use haptic feedback, combined with subtle on-screen reflective prompts, to cultivate adolescents' self-regulatory awareness and ability. Such approaches may help balance external guidance with internal reflection, fostering self-regulation through experience rather than enforcement.

Conclusion, Limitation and Future works

This study examined how adolescents in Singapore negotiate self-regulation within overlapping systems of parental, school, and technological control. By unpacking the tensions between autonomy and restriction, we revealed how adolescents develop diverse bypassing and adaptive strategies – mediating their agency across family, peer, and institutional contexts. Through a socio-ecological perspective, our findings reframe self-regulation not as an individual capacity but as a relational and negotiated practice. We further highlight how design interventions could move beyond restrictive or monitoring approaches to instead support reflection, negotiation, and shared meaning-making around technology use.

However, this study has several limitations. First, the small sample size and exploratory nature of the study limit the generalisability of our findings. The participants varied in age and schooling stage, and these differences may have influenced how they perceived and enacted regulation. Second, while our pilot study captured individual perspectives, team collaboration was not incorporated into the workshop process. Hence, in our future work, including collaborative or peer-based activities may help surface how adolescents co-construct self-regulation strategies and peer-supported regulation strategies.

References

Adelhardt, Z., Markus, S., & Eberle, T. (2018). Teenagers' reaction on the long-lasting separation from smartphones, anxiety and fear of missing out. In *Proceedings of the 9th International Conference on Social Media and Society* (pp. 212–216).

Agha, Z., Zhang, Z., Obajemu, O., Shirley, L., & Wisniewski, P. J. (2022). A case study on user experience bootcamps with teens to co-design real-time online safety interventions. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1–8).

Akter, M., Godfrey, A. J., Kropczynski, J., Lipford, H. R., & Wisniewski, P. J. (2022). From parental control to joint family oversight: Can parents and teens manage mobile online safety and privacy as equals? *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–28.

Al-Abyadh, M., Alatawi, M., Emara, E. A., Almasoud, S., Alsetoohy, O., & Ali, A. (2024). Do smartphone addiction and self-regulation failures affect students' academic life satisfaction? The role of students' mind wandering and cognitive failures. *Psychology Research and Behavior Management*, *17*, 1231–1253. https://doi.org/10.2147/prbm.s437076

Alluhidan, A., Akter, M., Alsoubai, A., Park, J. K., & Wisniewski, P. J. (2024). Teen Talk: The good, the bad, and the neutral of adolescent social media use. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), Article 422. https://doi.org/10.1145/3686961

Almohamed, A., Zhang, J., & Vyas, D. (2020). Magic machines for refugees. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies* (pp. 76–86).

Andersen, K. (2013). Making magic machines. In *10th European Academy of Design Conference*.

Andersen, K., & Wakkary, R. (2019). The magic machine workshops: Making personal design knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13).

Bandura, A., Caprara, G. V., Barbaranelli, C., Pastorelli, C., & Regalia, C. (2001). Sociocognitive self-regulatory mechanisms governing transgressive behavior. *Journal of Personality and Social Psychology*, *80*(1), 125.

Baumer, E. P. S., Berrill, T., Botwinick, S. C., Gonzales, J. L., Ho, K., Kundrik, A., Kwon, L., LaRowe, T., Nguyen, C. P., Ramirez, F., et al. (2018). What would you do? Design fiction and ethics. In *Proceedings of the 2018 ACM International Conference on Supporting Group Work* (pp. 244–256).

Baumer, E. P. S., Blythe, M., & Tanenbaum, T. J. (2020). Evaluating design fiction: The right tool for the job. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (pp. 1901–1913).

Beijing Internet Report. (2024, May 31). *"2024 Survey Report on Adolescent Internet Use" released: Over 40% of teenagers have used AI products*. Jingbaowang. https://news.bjd.com.cn/2024/05/31/10790057.shtml

Bergert, C., Köster, A., Krasnova, H., & Turel, O. (2020). Missing out on life: Parental perceptions of children’s mobile technology use. *WIRTSCHAFTSINFORMATIK (Zentrale Tracks)*, 568–583.

Björling, E. A., & Rose, E. (2019). Participatory research principles in human-centered design: Engaging teens in the co-design of a social robot. *Multimodal Technologies and Interaction*, *3*(1), 8.

Blackwell, L., Gardiner, E., & Schoenebeck, S. (2016). Managing expectations: Technology tensions among parents and teens. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1390–1401).

Blair, C., & Diamond, A. (2008). Biological processes in prevention and intervention: The promotion of self-regulation as a means of preventing school failure. *Development and Psychopathology*, *20*(3), 899–911.

Blakemore, S. J. (2019). Adolescence and mental health. *The Lancet*, *393*(10185), 2030–2031.

Bleecker, J. (2022). Design fiction: A short essay on design, science, fact, and fiction. *Machine Learning and the City: Applications in Architecture and Urban Design*, 561–578.

Boase, J., et al. (2021). Mobile media in teen life: Information, networks and access. In *Handbook of digital inequality* (pp. 98–113). Edward Elgar Publishing.

Boehner, K., Vertesi, J., Sengers, P., & Dourish, P. (2007). How HCI interprets the probes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1077–1086). Association for Computing Machinery. https://doi.org/10.1145/1240624.1240789

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597.

Bronfenbrenner, U. (2000). *Ecological systems theory*. American Psychological Association.

Cao, A., Wu, L., Melvin, G., Cardamone-Breen, M., Broomfield, G., Seguin, J., Salvaris, C., Xie, J., Basur, D., Bartindale, T., et al. (2025). Empowering parents of adolescents at elevated risk of suicide: Co-designing an adaptation to a coach-assisted, digital parenting intervention. *European Journal of Investigation in Health, Psychology and Education*, *15*(10), 199.

Cao, F., Su, L., Liu, T., & Gao, X. (2007). The relationship between impulsivity and Internet addiction in a sample of Chinese adolescents. *European Psychiatry*, *22*(7), 466–471.

Chamorro, L. S., Lallemand, C., & Gray, C. M. (2024). "My mother told me these things are always fake"—Understanding teenagers' experiences with manipulative designs. In *2024 ACM Designing Interactive Systems Conference* (pp. 1469–1482). ACM Press.

Chen, P. C., Hung, M. W., Lu, H. S., Yuan, C. W., Bi, N., Lee, W. C., Huang, M. C., & You, C. W. (2022). This app is not for me: Using mobile and wearable technologies to improve adolescents' smartphone addiction through the sharing of personal data with parents. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–15).

Cho, J., Song, I., Agha, Z., Cagiltay, B., Calambur, V., Rheu, M. M., & Huh-Yoo, J. (2025). Mobile technology and teens: Understanding the changing needs of sociocultural and technical landscape. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. https://doi.org/10.1145/3706599.3706725

Chowdhury, A., & Bunt, A. (2023). Co-designing with early adolescents: Understanding perceptions of and design considerations for tech-based mediation strategies that promote technology disengagement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–16).

Chowdhury, A., & Bunt, A. (2024). Exploring a design space for digital interventions facilitating early adolescents' tech disengagement: A parent-child perspective. In *Proceedings of the 13th Nordic Conference on Human-Computer Interaction* (pp. 1–17). Association for Computing Machinery. https://doi.org/10.1145/3679318.3685382

Chowdhury, A., Wang, T., Anik, A. I., & Bunt, A. (2025). The landscape of digital tech disengagement solutions for early adolescents: Insights from a systematic scoping review and app analysis. *Proceedings of the ACM on Human-Computer Interaction*, *9*(7), 1–32.

Chua, P. K., & Mazmanian, M. (2021). What are you doing with your phone? How social class frames parent-teen tensions around teens' smartphone use. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445275

Davis, K. (2023). *Technology's child*. MIT Press. https://doi.org/10.7551/mitpress/13406.001.0001

Davis, K., Slovak, P., Landesman, R., Pitt, C., Ghajar, A., Schleider, J. L., Kawas, S., Perez Portillo, A. G., & Kuhn, N. S. (2023). Supporting teens' intentional social media use through interaction design: An exploratory proof-of-concept study. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference* (pp. 322–334). Association for Computing Machinery. https://doi.org/10.1145/3585088.3589387

Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. In *Handbook of theories of social psychology* (Vol. 1, pp. 416–436). Sage Publications.

Dreier, M. J., Low, C. A., Fedor, J., Durica, K. C., & Hamilton, J. L. (2024). Adolescents' self-regulation of social media use during the beginning of the COVID-19 pandemic: An idiographic approach. *Journal of Technology in Behavioral Science*, 1–17.

Duckert, M., & Barkhuus, L. (2021). To use or not to use: Mediation and limitation of digital screen technologies within nuclear families. In *Proceedings of the 2021 ACM International Conference on Interactive Media Experiences* (pp. 73–83). Association for Computing Machinery.

Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, *16*(12), 939–944.

Dumaru, P., & Al-Ameen, M. N. (2025). One size doesn't fit all: Towards design and evaluation of developmentally appropriate parental control tool. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642603

Dumaru, P., Atashpanjeh, H., & Al-Ameen, M. N. (2024). "It's hard for him to make choices sometimes and he needs guidance": Re-orienting parental control for children. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–51.

Edwards, E. A., Lumsden, J., Rivas, C., Steed, L., Panagioti, M., Newman, A., Caton, H., & Walton, R. (2016). Gamification for health promotion: Systematic review of behaviour change techniques in smartphone apps. *BMJ Open*, *6*(10), Article e012447. https://doi.org/10.1136/bmjopen-2016-012447

Erickson, L. B., Wisniewski, P., Xu, H., Carroll, J. M., Rosson, M. B., & Perkins, D. F. (2016). The boundaries between: Parental involvement in a teen's online world. *Journal of the Association for Information Science and Technology*, *67*(6), 1384–1403. https://doi.org/10.1002/asi.23450

Erikson, E. H. (1994). *Identity and the life cycle*. W. W. Norton & Company.

Farley, J. P., & Kim-Spoon, J. (2014). The development of adolescent self-regulation: Reviewing the role of parent, peer, friend, and romantic relationships. *Journal of Adolescence*, *37*(4), 433–440.

Fomina, T. G., Potanina, A. M., & Morosanova, V. I. (2020). The relationship between school engagement and conscious self-regulation of learning activity: The current state of the problem and research perspectives in Russia and abroad. *RUDN Journal of Psychology and Pedagogics*, *17*(3), 390–411.

Freed, D., Bazarova, N. N., Consolvo, S., Han, E. J., Kelley, P. G., Thomas, K., & Cosley, D. (2023). Understanding digital-safety experiences of youth in the US. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). Association for Computing Machinery.

Gak, L., Li, I., Rosner, D., & Salehi, N. (2025). "The world has changed..." — Unlocking teen perspectives on technological futures through design fiction workshops. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Article 615, pp. 1–9). Association for Computing Machinery. https://doi.org/10.1145/3706599.3719952

Gaver, B., Dunne, T., & Pacenti, E. (1999). Design: Cultural probes. *Interactions*, *6*(1), 21–29. https://doi.org/10.1145/291224.291235

Gendil, L. (2024). *Hold the phone: Recent state activity on cell use in schools*. National Conference of State Legislatures.

Ghosh, A. K., Badillo-Urquiola, K., Guha, S., LaViola, J. J., Jr., & Wisniewski, P. J. (2018). Safety vs. surveillance: What children have to say about mobile apps for parental control. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery.

Goodyear, V. A., Randhawa, A., Adab, P., Al-Janabi, H., Fenton, S., Jones, K., Michail, M., Morrison, B., Patterson, P., Quinlan, J., Sitch, A., Twardochleb, R., Wade, M., & Pallan, M. (2025). School phone policies and their association with mental wellbeing, phone use, and social media use (SMART Schools): A cross-sectional observational study. *The Lancet Regional Health - Europe*, *51*, Article 101211.

https://doi.org/10.1016/j.lanepe.2025.101211

Green, N. (2002). Who's watching whom? Monitoring and accountability in mobile relations. In *Wireless world: Social and interactional aspects of the mobile age* (pp. 32–45). Springer.

Guo, K. (2024, August 20). Parents help children dodge time limits on online games. *China Daily*. https://global.chinadaily.com.cn/a/202408/20/WS66c3e9a1a31060630b923e74.html

Hamilton, J. L., Nesi, J., & Choukas-Bradley, S. (2022). Reexamining social media and socioemotional well-being among adolescents through the lens of the COVID-19 pandemic: A theoretical review and directions for future research. *Perspectives on Psychological Science*, *17*(3), 662–679. https://doi.org/10.1177/17456916211014189

Hammond, S. P., Polizzi, G., & Bartholomew, K. J. (2023). Using a socio-ecological framework to understand how 8–12-year-olds build and show digital resilience: A multi-perspective and multimethod qualitative study. *Education and Information Technologies*, *28*(4), 3681–3709.

Haug, S., Castro, R. P., Kwon, M., Filler, A., Kowatsch, T., & Schaub, M. P. (2015). Smartphone use and smartphone addiction among young people in Switzerland. *Journal of Behavioral Addictions*, *4*(4), 299–307. https://doi.org/10.1556/2006.4.2015.037

Hay, C., & Forrest, W. (2006). The development of self-control: Examining self-control theory's stability thesis. *Criminology*, *44*(4), 739–774. https://doi.org/10.1111/j.1745-9125.2006.00062.x

Hayes, T. (2024). *6 ways kids are getting around parental controls on Apple's Screen Time*. PCMag. https://www.pcmag.com/how-to/ways-kids-are-getting-around-parental-controls-on-apple-screen-time

Hiniker, A., Schoenebeck, S. Y., & Kientz, J. L. (2016). Not at the dinner table: Parents' and children's perspectives on family technology rules. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1376–1389). Association for Computing Machinery.

Holtz, P., & Appel, M. (2011). Internet use and video gaming predict problem behavior in early adolescence. *Journal of Adolescence*, *34*(1), 49–58.

Ibrahim, S., et al. (2025). Uncovering parental struggles: Using digital probes to analyse challenges in applying online parenting content. *ACM Transactions on Computer-Human Interaction (TOCHI)*. https://doi.org/10.1145/3678975

Katevas, K., Arapakis, I., & Pielot, M. (2018). Typical phone use habits: Intense use does not predict negative well-being. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3229434.3229441

Kawas, S., Kuhn, N. S., Sorstokke, K., Bascom, E., Hiniker, A., & Davis, K. (2021). When screen time isn't screen time: Tensions and needs between tweens and their parents during nature-based exploration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery.

Kim, J., Wolfe, R., Chordia, I., Davis, K., & Hiniker, A. (2024). "Sharing, not showing off": How

BeReal approaches authentic self-presentation on social media through its design. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), 1–32.

Koh, S. (2025, January 10). S'pore parents call for help to manage children's device use as lines blur between learning, leisure. *The Straits Times*. https://www.straitstimes.com/singapore/parenting-education/parents-call-for-help-in-managing-kids-device-use-as-lines-blur-between-learning-and-leisure

Lanette, S., Chua, P. K., Hayes, G., & Mazmanian, M. (2018). How much is 'too much'? The role of a smartphone addiction narrative in individuals' experience of use. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–22.

Lee, J., Jung, K., Newman, E. G., Chow, E., & Chen, Y. (2025). Understanding adolescents' perceptions of benefits and risks in health AI technologies through design fiction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–20). Association for Computing Machinery.

Lee, J., Lee, S., & Shin, Y. (2024). Lack of parental control is longitudinally associated with higher smartphone addiction tendency in young children: A population-based cohort study. *Journal of Korean Medical Science*, *39*(34), Article e268. https://doi.org/10.3346/jkms.2024.39.e268

Lenhart, A., Duggan, M., Perrin, A., Stepler, R., Rainie, H., & Parker, K. (2015). *Teens, social media & technology overview 2015*. Pew Research Center.

Lepri, G., Corbellini, N., Ferrando, S., Volpe, G., & Camurri, A. (2024). Let's play: Early explorations of child-caregiver embodied interactions. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference* (pp. 913–918). Association for Computing Machinery.

Livingstone, S., & Helsper, E. (2007). Gradations in digital inclusion: Children, young people and the digital divide. *New Media & Society*, *9*(4), 671–696.

Lukoff, K., Yu, C., Kientz, J., & Hiniker, A. (2018). What makes smartphone use meaningful or meaningless? *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 1–26.

Lupton, D. (2017). Digital health now and in the future: Findings from a participatory design stakeholder workshop. *Digital Health*, *3*, 1–10. https://doi.org/10.1177/2055207617740018

Lyngs, U., Lukoff, K., Slovak, P., Binns, R., Slack, A., Inzlicht, M., Van Kleek, M., & Shadbolt, N. (2019). Self-control in cyberspace: Applying dual systems theory to a review of digital self-control tools. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–18). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300361

Magee, R. M., Agosto, D. E., & Forte, A. (2017). Four factors that regulate teen technology use in everyday life. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 511–522). Association for Computing Machinery.

Mazmanian, M., & Lanette, S. (2017). "Okay, one more episode": An ethnography of parenting in the digital age. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 2273–2286). Association for Computing Machinery.

McKenzie, J. (2019). Shifting practices, shifting selves: Negotiations of local and global cultures among adolescents in Northern Thailand. *Child Development*, *90*(6), 2035–2052. https://doi.org/10.1111/cdev.13088

Ministry of Health. (2023). *Guidance on screen use in children*. https://www.moh.gov.sg

Muñoz, D., Ploderer, B., & Brereton, M. (2019). Position exchange workshops: A method to design for each other in families. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery.

Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin*, *126*(2), 247–259.

Ng, T. P., Lee, J. J., & Wu, Y. (2021). Unpacking cultural perceptions of future elder care through design fiction. In *IASDR 2021: With Design: Disrupting and Reshaping, proceedings of the ninth Congress of the International Association of Societies of Design Research* (pp. 1632–1652). Springer.

Nguyen, M. T., Nguyen, N. K. N., & Nguyen, P. Q. (2024). Family and friends: Key influences on teenage traits and behaviors. *Social Sciences and Humanities Journal (SSHJ)*, *8*(10), 5711–5723.

Nikken, P., & Schols, M. (2015). How and why parents guide the media use of young children. *Journal of Child and Family Studies*, *24*(11), 3423–3435.

Novak, S. P., & Clayton, R. R. (2001). The influence of school environment and self-regulation on transitions between stages of cigarette smoking: A multilevel analysis. *Health Psychology*, *20*(3), 196–207.

Opdenakker, M. C. (2022). Developments in early adolescents’ self-regulation: The importance of teachers’ supportive vs. undermining behavior. *Frontiers in Psychology*, *13*, Article 1021904.

Orben, A., & Blakemore, S. J. (2023). How social media affects teen mental health: A missing link. *Nature*, *614*(7948), 410–412.

Potapov, K., & Marshall, P. (2020). LifeMosaic: Co-design of a personal informatics tool for youth. In *Proceedings of the 19th International Conference on Interaction Design and Children* (pp. 519–531). Association for Computing Machinery.

Prensky, M. (2001). Digital natives, digital immigrants part 2: Do they really think differently? *On the Horizon*, *9*(6), 1–6.

Radesky, J. S. (2018). *Persuasive digital design: Appealing to adults, problematic for kids*. Michigan Health Lab.

Ricoy, M. C., Martinez-Carrera, S., & Martinez-Carrera, I. (2022). Social overview of smartphone use by teenagers. *International Journal of Environmental Research and Public Health*, *19*(22), Article 15068.

Roffarello, A. M., & De Russis, L. (2023). Achieving digital wellbeing through digital self-control tools: A systematic review and meta-analysis. *ACM Transactions on Computer-Human Interaction*, *30*(4), 1–66.

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review*, *32*(1), 77–112. https://doi.org/10.1007/s10648-019-09498-w

Schiano, D. J., Burg, C., Smith, A. N., & Moore, F. (2016). Parenting digital youth: How now? In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 3181–3189). Association for Computing Machinery.

Sharma, S., Hartikainen, H., Ventä-Olkkonen, L., Eden, G., Iivari, N., Kinnunen, E., Holappa, J., Kinnula, M., Molin-Juustila, T., & Okkonen, J. (2022). In pursuit of inclusive and diverse digital futures: Exploring the potential of design fiction in education of children. *Interaction Design and Architecture(s)*, *51*, 219–248.

Sherer, J., & Levounis, P. (2022). Technological addictions. *Current Psychiatry Reports*, *24*(9), 399–406.

Spinuzzi, C. (2005). The methodology of participatory design. *Technical Communication*, *52*(2), 163–174.

Spotify. (n.d.). *Parental guide*. https://www.spotify.com/safetyandprivacy/files/Parental_Guide.pdf

Sweigart, E. A., Valliani, A., & Wisniewski, P. J. (2025). Pause, reflect, and redirect: An approach to empowering youth to be safer online by helping them make better decisions. *Social Sciences*, *14*(5), Article 302. https://doi.org/10.3390/socsci14050302

Tamashiro, M. A., Van Mechelen, M., Schaper, M. M., & Iversen, O. S. (2021). Introducing teenagers to machine learning through design fiction: An exploratory case study. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference* (pp. 471–475).

Tampines Secondary School. (2025). *Mobile phone policy*. https://www.tampinessec.moe.edu.sg/our-school/Tampinesian-Code-of-Conduct/mobile-phone-policy/

Tang, L. (2025, February 1). Singapore teenagers spend nearly 8.5 hours a day on screens: CNA-IPS survey. *Channel News Asia*. https://www.channelnewsasia.com/singapore/screen-time-devices-survey-teens-spend-daily-stress-4908281

Tang, X., Lima, G., Jiang, L., Simko, L., & Zou, Y. (2025). Beyond "vulnerable populations": A unified understanding of vulnerability from a socio-ecological perspective. *Proceedings of the ACM on Human-Computer Interaction*, *9*(2), 1–30.

Telok Kurau Primary School. (2025). *Use of mobile devices*. https://www.telokkuraupri.moe.edu.sg/about-tkps/use-mobile-devices/

Toombs, E., Mushquash, C. J., Mah, L., Short, K., Young, N. L., Cheng, C., Zhu, L., Strudwick, G., Birken, C., Hopkins, J., et al. (2022). Increased screen time for children and youth during the COVID-19 pandemic. *Science Briefs of the Ontario COVID-19 Science Advisory Table*, *3*(59), 1–19.

Troll, E. S., Friese, M., & Loschelder, D. D. (2021). How students' self-control and smartphone-use explain their academic performance. *Computers in Human Behavior*, *117*, Article 106624. https://doi.org/10.1016/j.chb.2020.106624

Truesdell, E. J. K., et al. (2025). Game Playbooks 2.0: An updated strategy for supporting game-based behavior change interventions. *Proceedings of the ACM on Human-Computer Interaction*, *9*(6), 1021–1044. https://doi.org/10.1145/3678980

Tushara, E. (2024, September 30). Schools in Singapore impose phone bans to reduce distractions, rekindle social interaction. *The Straits Times*. https://www.straitstimes.com/singapore/schools-in-s-pore-impose-phone-bans-to-reduce-distractions-rekindle-social-interaction

Verweij, R. (2025). Taking the magic out of the machine: Children as creators of real-world AI-powered tools for education. In *Proceedings of the 24th Interaction Design and Children* (pp. 1185–1187). Association for Computing Machinery.

Weinstein, E., & James, C. (2022). *Behind their screens: What teens are facing (and adults are missing)*. MIT Press.

Welle, D. (2023). *China plans to restrict children's and teenagers' mobile phone use: A maximum of two hours per day*. Deutsche Welle. https://www.dw.com

Whyte, M. K., & Ikels, C. (2004). Filial obligations in Chinese families: Paradoxes of modernization. In *Filial piety: Practice and discourse in contemporary East Asia* (pp. 106–127). Stanford University Press.

Wisniewski, P., Jia, H., Wang, N., Zheng, S., Xu, H., Rosson, M. B., & Carroll, J. M. (2015). Resilience mitigates the negative effects of adolescent internet addiction and online risk exposure. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 4029–4038). Association for Computing Machinery.

Wyche, S., et al. (2024). Limitations of using mobile phones for managing Type 1 Diabetes (T1D) among youth in low and middle-income countries: Implications for mHealth. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), 1–19. https://doi.org/10.1145/3686867

Xiao, B., Zhao, H., Hein-Salvi, C., Parent, N., & Shapka, J. D. (2025). Examining self-regulation and problematic smartphone use in Canadian adolescents: A parallel latent growth modeling approach. *Journal of Youth and Adolescence*, *54*(2), 468–479.

Yang, Z. (2023, August 9). China is escalating its war on kids' screen time. *MIT Technology Review*. https://www.technologyreview.com/2023/08/09/1077567/china-children-screen-time-regulation/

Yardi, S., & Bruckman, A. (2011). Social and technical challenges in parenting teens' social media use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3237–3246). Association for Computing Machinery.

Research Paper

The Economics of Information Pollution in the Age of AI:

General Equilibrium, Welfare, and Policy Design

ZHANG Yukun[1]
ZHANG Tianyang[2]

[1]The Chinese University of Hong Kong, Hong Kong , China
[2]University of Bologna, Bologna, Italy

Abstract

The rapid diffusion of generative artificial intelligence (AI) is triggering structural change in information markets, with impacts far more complex than simple ‘technological progress’. This paper is the first to model AI as an asymmetric technological shock: in low-quality content production, AI acts as a substitute for labour (σL > 1), whereas in high-quality creation, it functions merely as a complement to human expertise (σH < 1). This technological asymmetry fundamentally alters the cost structure of information production, systematically favouring ‘lemons’ over ‘peaches’. By constructing a three-stage general equilibrium model, we prove this shock leads to an inefficient ‘Polluted Information Equilibrium’ sustained by a triad of interacting market failures: (1) a Production Externality, as low-quality producers do not Internalise ecological harm; (2) a Platform Governance Failure, as engagement-based revenue models misalign algorithms with social welfare; and (3) a Trust Commons Externality, as verification, a public good, is systematically under-provided. To address this, we design an adaptive governance framework, first deriving a theoretically-grounded Information Pollution Index (IPI) for real-time ecosystem monitoring. Second, we demonstrate that restoring efficiency requires a multi-instrument policy portfolio: a Pigouvian tax to correct the production externality, content provenance standards to address under-verification, and fiduciary duties to constrain platform behaviour. Finally, through Agent-Based Model (ABM) validation, we find this portfolio generates superadditive welfare gains – the joint intervention is significantly more effective than any single policy tool. The core insight is that the welfare consequences of the AI revolution depend not on the technology itself, but on how we design market rules; without proper governance, technological progress can paradoxically reduce social welfare, a phenomenon we term the ‘Paradox of AI Progress’.

Introduction

In 2025, an in-depth investigative report, the product of a senior journalist's three-week effort, garners 100,000 views on social media. On the same day, a sensational but unverified "scoop" generated by an AI in three seconds receives 10 million shares. This stark contrast reveals the fundamental challenge facing the information economy in the age of generative AI: when the marginal cost of producing convincing but unverified content approaches zero, how can truth compete with noise?

This is not merely a technical question; it is a core economic one. The rise of large language models (LLMs) like GPT-4 and Claude represents a structural shift in the information production function. However, existing research has largely focused on AI's impact on the labor market, overlooking a more fundamental problem: how does AI alter the equilibrium distribution of *information quality*?

Consider two distinct content production scenarios:

- **Scenario A (Investigative Report)**: A journalist uses AI tools to accelerate data analysis and assist in literature review, but the core work – field interviews, fact-checking, and ethical judgement – still requires human expertise. Here, AI is a *complement*, not a substitute.
- **Scenario B (Content Farm)**: An operator uses AI to batch-generate plausible health advice, investment "insider" tips, or political "revelations". No professional knowledge is needed, only prompt engineering. Here, AI is a *substitute*, almost entirely replacing human labour.

This asymmetry in production technology – what we term "technology-biased information pollution" – is reshaping the digital information ecosystem. By some estimates, 90% of online content may be AI-generated or AI-assisted by 2026. If the majority of this is unverified, low-quality content, we face not just "information overload", but systemic "information pollution" – a new class of market failure.

The central thesis of this paper is that generative AI is not a neutral productivity tool, but rather a *biased technological shock* that systematically lowers the production cost of ”lemons” (low-quality content) relative to ”peaches” (high-quality content). We argue that this technological bias leads to market failure through three mutually reinforcing mechanisms: at the *micro-level*, individual producers rationally choose to produce low-quality content due to its low cost and rapid, algorithm-driven rewards; at the *meso-level*, platform algorithms designed to maximize engagement amplify content based on virality rather than veracity; and at the *macro-level*, a systemic erosion of trust leads to ”verification fatigue” among consumers, further degrading the market's ability to discern quality.

This "pollution spiral," if left unchecked, risks a "tragedy of the digital commons," where everyone is a victim, yet no single actor has the incentive to unilaterally change behaviour.

This paper's contributions are threefold: a theoretical innovation that is the first to model AI's market impact as an asymmetric Constant Elasticity of Substitution (CES) production function, revealing a paradox of progress; a measurement tool that constructs a multi-dimensional Information Pollution Index (IPI) to provide a quantitative basis for policy intervention; and a policy design that proves the necessity of a multi-instrument portfolio, validated through computational simulation.

The remainder of this paper is organised as follows. Section 2 reviews the relevant literature and situates our contribution. Section 3 builds the theoretical model and characterises the equilibrium. Section 4 derives the IPI and designs the optimal policy portfolio. Section 5 provides validation via an Agent-Based Model (ABM). Section 6 concludes and discusses policy implications.

Literature Review

Economic Foundations and Traditional Governance of Information

The novel market failure of information pollution traces its theoretical foundations to Akerlof's "market for lemons" model (Akerlof, 1970). In the contemporary digital environment, this problem presents unique, multi-dimensional complexities. From an economic perspective, environmental economics offers two important tools for dealing with pollution: Pigouvian taxes and tradable permits.

Posner and Weyl (2018), in their book *Radical Markets*, explicitly advocated for taxing online advertising. Cabral et al. (2021) further developed this idea and proposed a differentiation mechanism for low-quality content. In Zuboff's (2019) book *The Age of Surveillance Capitalism*, her profound critique of technology companies' massive acquisition of user data to predict and change human behaviour provides an important theoretical basis and justification for regulating data exploitation and information pollution through economic means (such as taxation).

The tradable permit system is a mature governance tool in environmental economics. The core of this mechanism is setting a strict cap on total pollution emissions (Cap) and allocating emission rights as a market-tradable asset. Its theoretical foundation stems from Coase's (1960) property rights theory and was first systematically proposed by Dales (1968). Subsequently, Montgomery's (1972) research theoretically demonstrated that, under specific conditions, this mechanism can achieve predetermined emission reduction targets with the lowest total social cost. Although there are some obstacles to shifting this perspective to information pollution, some scholars have put forward some relevant proposals.

In addition, some scholars have proposed market-based solutions from the perspective of platform governance. For example, the research of Parker and Van Alstyne (2020) inspired using reputation systems or information quality points to incentivise platforms to solve information asymmetry. Additionally, Crémer et al. (2019) proposed mandatory data sharing and interoperability – essentially "access quotas" for core resources, similar to tradable Licences – to break monopolies and promote competition.

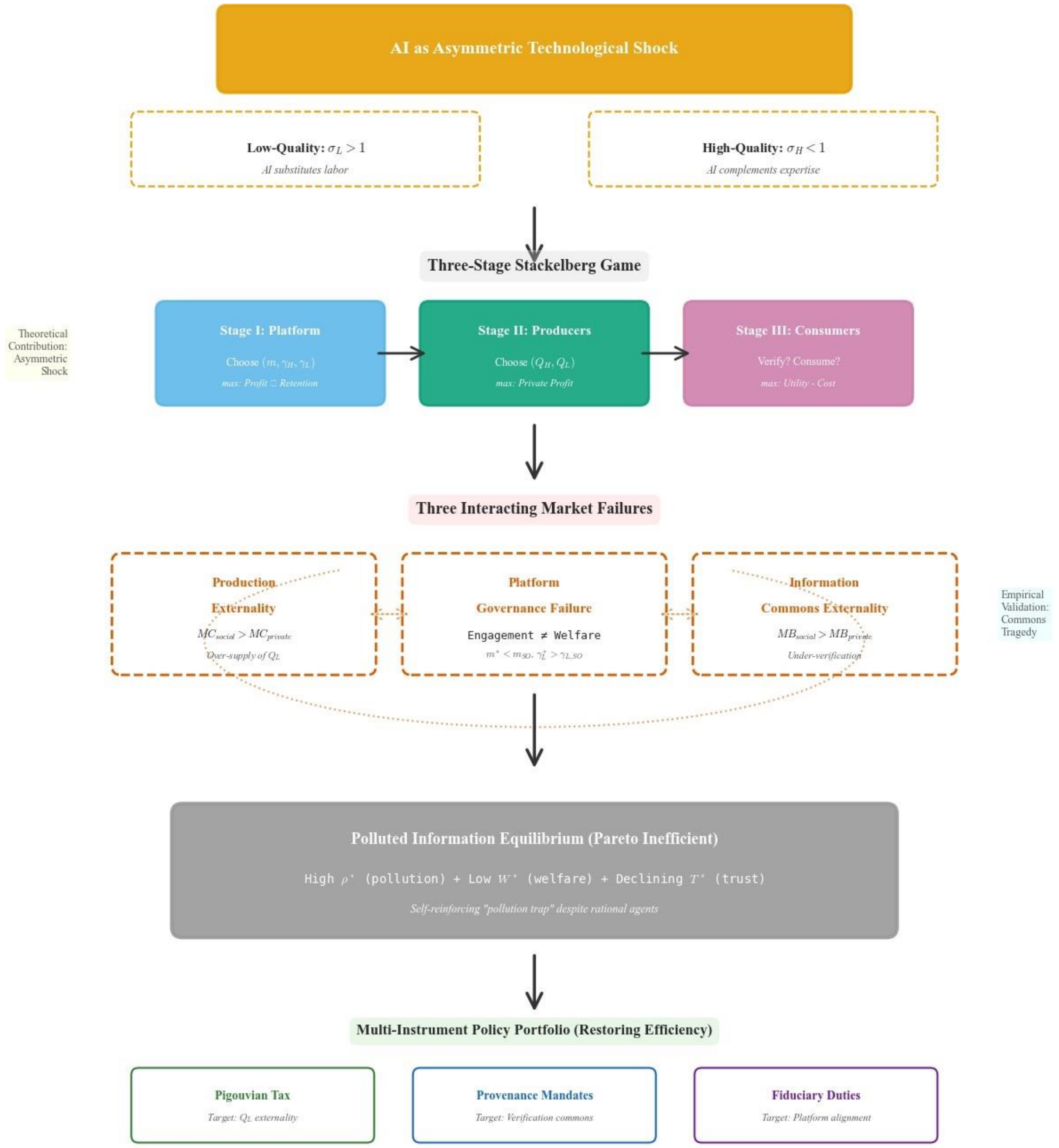

AI as Asymmetric Technological Shock
Low-Quality: $\sigma_L > 1$
AI substitutes labor
High-Quality: $\sigma_H < 1$
AI complements expertise
Three-Stage Stackelberg Game
Theoretical Contribution: Asymmetric Shock
Stage I: Platform
Choose (m, γ_H, γ_L)
max: Profit □ Retention
Stage II: Producers
Choose (Q_H, Q_L)
max: Private Profit
Stage III: Consumers
Verify? Consume?
max: Utility - Cost
Three Interacting Market Failures
Production Externality
$MC_{social} > MC_{private}$
Over-supply of Q_L
Platform Governance Failure
Engagement ≠ Welfare
$m^* < m_{SO}, \gamma_L^* > \gamma_{L,SO}$
Information Commons Externality
$MB_{social} > MB_{private}$
Under-verification
Empirical Validation: Commons Tragedy
Polluted Information Equilibrium (Pareto Inefficient)
High ρ^* (pollution) + Low W^* (welfare) + Declining T^* (trust)
Self-reinforcing "pollution trap" despite rational agents
Multi-Instrument Policy Portfolio (Restoring Efficiency)
Pigouvian Tax
Target: Q_L externality
Provenance Mandates
Target: Verification commons
Fiduciary Duties
Target: Platform alignment

Figure 1. Conceptual Framework of AI-Driven Information Pollution
The model proceeds from the core asymmetric shock ($\sigma_L > 1 > \sigma_H$) through a three-stage Stackelberg game, which leads to a triad of interacting market failures: a Production Externality, a Platform Governance Failure, and an Information Commons Externality. These failures produce a self-reinforcing, Pareto-inefficient "Polluted Information Equilibrium." We show this equilibrium necessitates a multi-instrument policy portfolio (Pigouvian Tax, Provenance Mandates, Fiduciary Duties) to restore efficiency.

Note: This figure illustrates the causal chain from asymmetric AI shock through strategic interactions to market failure and policy solutions.

AI, Algorithmic Amplification, and Ecosystem Degradation

Beyond economics, scholars in information science and communication studies have emphasised the ecological nature of digital knowledge systems. Floridi (2010) conceptualises information environments as "infospheres", arguing that pollution arises when the semantic integrity of these ecosystems is undermined. Bawden and Robinson (2009) describe the resulting cognitive and social fatigue as a form of "information environmental degradation". These perspectives reinforce the economic analogy: information pollution, like environmental pollution, depletes a scarce common resource – trust.

Recent literature underscores that the information environment has entered a phase where quantity and veracity are inversely correlated. False news spreads faster and farther than verified information (Vosoughi et al., 2018), amplified by algorithmic incentives that optimise for engagement rather than truth (Allcott & Gentzkow, 2017; Allcott et al., 2019). This phenomenon is continuously reinforced by cognitive biases (such as the lazy thinking effect revealed by Pennycook and Rand, 2019) and causes multidimensional economic distortions, such as those highlighted by Allcott et al. (2020a), Durante et al. (2019), and Bursztyn et al. (2020).

While many economists view AI as a driver of innovation and productivity (Brynjolfsson & McAfee, 2023), this perspective often overlooks welfare consequences arising from unequal information quality. Stiglitz (2022) argues that the digital economy intensifies traditional information failures, highlighting the need for analytical frameworks that account for the social costs of misinformation and data misuse.

Algorithmic amplification further intensifies these dynamics. As Tufekci (2015) observes, recommendation systems can generate systemic harms even without malicious intent. Empirical evidence confirms that algorithmic curation narrows informational diversity (Bakshy et al., 2015). Nielsen (2020) documents how platform power reshapes the media ecosystem, reinforcing attention asymmetries that favour viral, low-cost content over verified reporting. From a systems perspective, Helbing, D. (Ed.). (2018) frames these feedback loops as emergent properties of self-organising social systems, suggesting that agent-based modelling can illuminate how micro-level incentives produce macro-level information distortions.

This paper contributes to and bridges three major strands of literature. First, it extends the classic literature on information economics in the context of digital content, exploring how AI undermines traditional signalling mechanisms (Spence, 1973). The erosion of credibility parallels earlier discussions on asymmetric information and adverse selection, but now operates within algorithmically mediated attention markets. Second, it advances research in platform economics and industrial organisation (Hagiu & Wright, 2015; Rochet & Tirole, 2004). By integrating AI-driven production asymmetry into a two-sided platform framework, we provide a formal treatment of how content generation technologies interact with intermediation and market power. Third, it contributes to the emerging field of AI economics (Acemoglu & Restrepo, 2019; Agrawal et al., 2018), shifting the focus from labour displacement to the quality of informational output in one of society's most critical domains – the information market.

Theoretical Framework

Conceptual Foundation: Information as a Common-Pool Resource

The Credible Information Commons

We conceptualise the ecosystem of credible information as a common-pool resource (CPR) in the Ostromian sense, rivalrous in quality yet only imperfectly excludable in access. Let the stock of high-quality information at time t, denoted by St, represent the attention-weighted reservoir of credible content circulating across digital platforms. New content creation replenishes this reservoir through a regenerative flow R_t, while low-quality or misleading information generated by actors seeking visibility, clicks, or advertising rents degrades it at a pollution rate P_t. The dynamic evolution of the resource follows

$$\dot{S}_t = R_t - P_t(S_t, a_t, \theta_t) \tag{1}$$

where a_t captures the aggregate extraction intensity of attention by content producers, and θ_t summarises governance and verification institutions that regulate entry, provenance, and moderation. The credible-information commons are renewable but exhaustible: its replenishment depends on costly verification and sustained trust, while degradation arises from unpriced externalities of generative AI and algorithmic amplification. Extraction in this domain refers to the appropriation of limited audience attention At for private content visibility. A simple representation of allocation shares is $s_i = \frac{q_i A_t}{\sum_j q_j A_t}$. Pollution corresponds to the dissemination of low-credibility or synthetic material that reduces the expected information accuracy $E[h_i]$ and undermines collective trust $T_t = f(S_t)$. The marginal social cost of pollution, $\partial C^{soc} / \partial P_t$ is convex in P_t because reputational erosion and verification fatigue accelerate once misinformation exceeds cognitive processing capacity.

To move from the tragedy to sustainable governance, Ostrom's design principles can be adapted to digital information. Boundaries become traceable provenance and authentication of content and accounts, which can be parameterised by the verifiability intensity m. Congruence between rules and local conditions corresponds to aligning platform recommendation objectives (γ_H, γ_L) with societal welfare by rewarding credible engagement over sheer volume. Collective choice arrangements imply participatory moderation and community fact-checking that internalise positive monitoring externalities. Monitoring and graduated sanctions suggest layered detection and proportionate penalties for repeated offenders, representable as increasing marginal cost of pollution $c_P(P_t)$. Conflict-resolution mechanisms provide low-cost correction channels to restore truthful information without discouraging legitimate debate. Finally, nested governance combines platform-level rules with national and international regulatory oversight, which we capture by multi-layer institutional variables $\theta_t = \left(\theta^{plat}, \theta^{nat}, \theta^{intl}\right)$.

The Tragedy of AI-Driven Pollution

Each producer chooses output qi and quality hi to maximise private returns,

$$\pi_i = p(h_i, q_i) - c(h_i, q_i) \tag{2}$$

where $p\ (\cdot)$ is the monetised visibility price determined by platform algorithms. Because platforms typically reward engagement rather than veracity, producers internalise revenue from exposure but not the social cost of degraded credibility. In aggregate, rational behaviour therefore implies $\partial P_t\ /\partial a_t > 0$, generating a tragedy of the commons: collectively excessive attention extraction and quality dilution, even when each agent behaves optimally given existing rules. A social planner would internalise the marginal external damage $MEC\ (S_t)$ imposed on the shared trust stock, but decentralised markets lack such a mechanism in the absence of governance or norms.

Linking these mechanisms, AI-enabled content generation and algorithmic amplification increase extraction at and raise P_t unless institutional strength θt and verifiability m are sufficiently high. As P_t grows, the trust stock $T_t = f(S_t)$ depreciates faster, verification becomes costlier due to fatigue, and marginal damages become convex. These forces reinforce a high-pollution equilibrium.

Preview of the Three-Stage Game

The CPR perspective nests within the general-equilibrium model in Section ??. Micro-level over-extraction of attention maps to the production externality; platform bias in (γ_H, γ_L) represents governance failure; and the depreciation of trust, $\dot{T}_t < 0$, corresponds to a trust commons externality.

The strategic environment unfolds in three stages. In Stage I (platform), recommendation parameters (γ_H, γ_L) and verifiability intensity m are chosen subject to retention and profitability objectives, under institutional constraints θ_t. In Stage II (producers), agents choose (q_i, h_i) given platform rules, internalising exposure revenue but not social damage, thereby determining at and P_t. In Stage III (consumers and trust), aggregate attention allocation A_t and verification behaviour update stock variables (S_t, T_t) via $\dot{S}_t = R_t - P_t(\cdot)$ and a depreciation condition for trust when pollution dominates, feeding back into platform and producer incentives.

The institutional principles above act as constraints that can shift the equilibrium from a high-pollution Nash outcome toward a cooperative optimum. Subsequent empirical sections Operationalise these dynamics through cross-country proxies of media governance and social-media usage, while simulation experiments evaluate policy instruments using the Information Pollution Index (IPI) as a quantitative measure of commons health.

Model Primitives and Environment

Agents and Objectives

The economy features four classes of rational agents. First, there is a continuum of content producers indexed by $i \in [0, 1]$. Producer i has heterogeneous productivity $A_{j,\,i}$ across content types $j \in \{H, L\}$ and faces factor prices (r, w), where r is the rental cost of AI capital and w is the wage for high-skill human labour. Given platform rules, each producer chooses output and quality to Maximise profits.

Second, a monopolistic platform intermediates the market and holds market power. As a market organiser, it selects the intensity of moderation and provenance $m \in [0, 1]$. As an attention allocator, it commits to an algorithmic amplification vector $\gamma = (\gamma_H, \gamma_L)$ that maps content types into exposure weights. The platform's objective is specified in the three-stage game below and trades off profit, retention, and compliance.

Third, consumers have heterogeneous verification costs $k_i \sim F(k)$ with support $[0, \bar{k}]$. Upon receiving a noisy signal of content quality, consumers decide whether to pay ki to perfectly verify quality, and then allocate attention and demand accordingly.

Fourth, a social planner provides a normative benchmark. The planner Maximises social welfare by internalising the externalities that low-quality content imposes on the stock of trust and informational accuracy. This allows us to define the Pareto-efficient allocation and quantify welfare losses from market failure.

Information Goods: Quality Differentiation

There are two information goods reflecting different production difficulties, externalities, and social welfare impacts. High-quality content Q_H is accurate and cognitively valuable, improving individual decision-making and typically requiring expertise and rigorous verification. Low-quality content Q_L can be misleading or false, generates negative externalities by eroding consumer welfare and social trust, and is scalable via templated or automated generation. The classification is based on objective effects on decisions and welfare, not subjective judgements. Under algorithmic amplification and attention competition, AI and automation disproportionately reduce the marginal cost and raise the scalability of Q_L, creating asymmetric technology that underlies pollution pressure.

Timeline and Information Structure

The timing follows a three-stage Stackelberg game. In Stage 1 (platform choice), given institutional constraints and market conditions, the platform commits to moderation and provenance intensity m and to an amplification vector γ = (γ_H, γ_L). These choices determine expected exposure and monetisation for content types and shape downstream best responses.

In Stage 2 (producer choice), producers observe (m, γ) and factor prices (r, w), and choose output and quality (q_i, h_i) based on their productivity vector $A_{j,\, i}$. Expected unit revenue is given by an exposure-price function $p\ (h_i, q_i; \gamma, m)$, while costs follow $c\ (h_i, q_i; r, w)$. Because platform rewards focus on engagement rather than veracity, individual producers do not internalise the social costs that low-quality content imposes on public trust and verification fatigue.

In Stage 3 (consumption and trust update), consumers observe a noisy signal $s \in \{H, L\}$ and form priors given the platform environment, then decide whether to pay k_i to fully verify. The resulting attention allocation and demand determine exposure and effective consumption. At the end of the period, the stock of credible information and trust updates via a regeneration-pollution process, $S_{t+1} = S_t + R_t - P_t$, where R_t and P_t are jointly determined by the production and propagation of Q_H and Q_L and by the platform's (m, γ). Trust depreciation induced by pollution feeds back into next period's verification incentives and platform trade-offs, creating dynamic interactions.

Regarding information, the platform's (m, γ) is publicly observed at Stage 1. In Stage 2, producers know platform rules, factor prices, and their own productivity, and they form rational expectations over consumer verification. In Stage 3, consumers observe the platform-provided signal and historical environment, including partially observable proxies for governance and media conditions, and then make optimal verification and consumption decisions. The equilibrium concept is subgame-perfect equilibrium: given platform commitments, producer and consumer best responses are mutually consistent, and the platform chooses $(m, \gamma\ )$

anticipating downstream reactions. This timeline and information structure underpin the subsequent equilibrium characterisation, comparative statics, and welfare analysis.

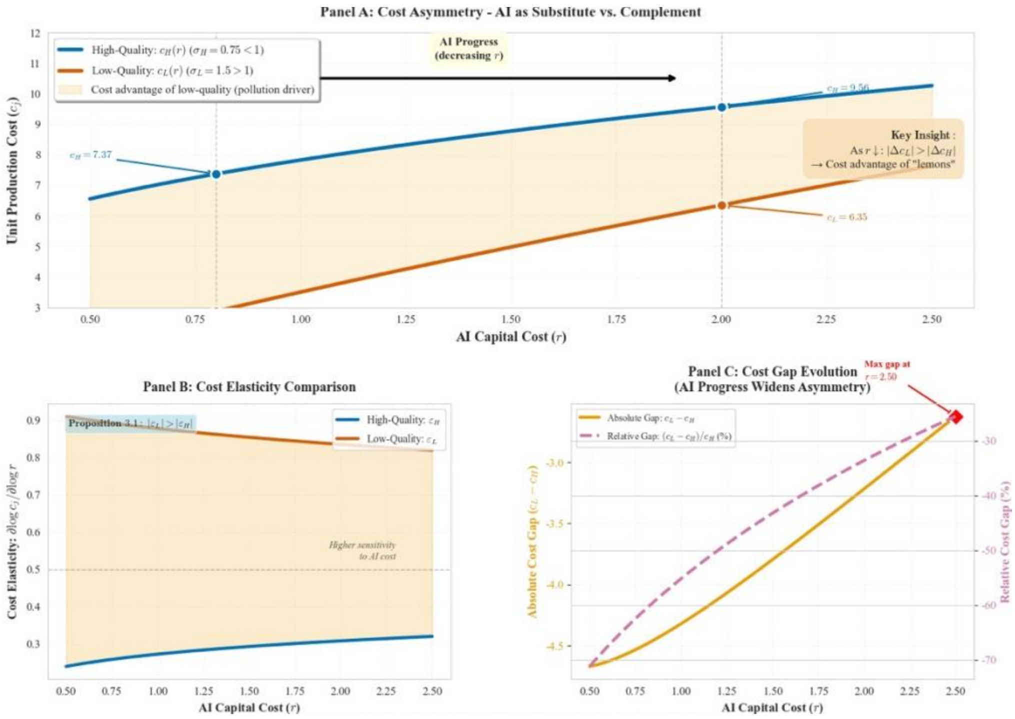


Figure 2. Production Cost Asymmetry under AI Technological Progress (Simulation of Proposition 3.2)

This figure illustrates the model's core mechanism. Panel A: As AI capital cost (r) decreases (AI progresses), the unit cost of low-quality content (c_L, substitute, $\boldsymbol{\sigma_L}$ > 1) falls disproportionately faster than that of high-quality content (c_H, complement, σ_H < 1). Panel B: This is because the cost elasticity with respect to r is higher for low-quality content ($|\boldsymbol{\partial \log c_L / \partial \log r}| > |\boldsymbol{\partial \log c_H / \partial \log r}|$). Panel C: The result is a widening absolute and relative cost gap, creating a systematic and growing economic advantage for "lemons" as AI technology improves.

Note: Based on CES production functions with $\sigma_L = 1.5 > 1$ (substitutes) and \sigma_H = 0.75 < 1 (complements). Parameters: $w = 8.0, A_H = 1.0, A_L = 0.8, \delta_H = 0.35, \delta_L = 0.65$. As AI capital cost r decreases, low-quality content cost falls disproportionately, creating systematic advantage for "lemons."

Production Technology: The Dynamics of Pollution

CES Production Function Specification.

The core mechanism of our model lies in the production technology. For content of type $j \in \{H, L\}$, producers utilise two inputs: AI capital (K_{AI}) and high-skilled human labour (L_H). The production function is of the Constant Elasticity of Substitution (CES) form:

$$Q_j = A_j\left[\delta_j K_{AI}^{\rho_j} + \left(1 - \delta_j\right) L_H^{\rho_j}\right]^{1/\rho_j} \tag{3}$$

where Aj is the total factor productivity, σ_j is the distribution parameter, and ρ_j determines the elasticity of substitution, $\sigma_j = 1/\left(1 - \rho_j\right)$

Core Assumption: Technological Asymmetry.

Our theory is built upon the following key assumption:

Assumption 3.1 (Technological Asymmetry).

$$\sigma_L > 1 > \sigma_H > 0 \tag{4}$$

This assumption has profound economic implications:

- For low-quality content (σ_L > 1), AI capital and human labour are *gross substitutes*. This reflects the Standardised nature of producing such content, where AI can efficiently generate text that is syntactically correct but lacks deep verification.
- For high-quality content (σ_H < 1), AI capital and human labour are *gross complements*. This captures the complexity of creating such content, where AI tools can augment research and writing but cannot replace human creativity, critical analysis, and ethical judgement.

Cost Asymmetry.

A direct corollary of this assumption is the asymmetric effect of technological progress on production costs.

Proposition 3.2 (Cost Asymmetry). A decrease in the cost of AI capital, r, has a larger cost-reducing effect on low-quality content than on high-quality content. Formally:

$$\left|\frac{\partial \log c_L}{\partial \log r}\right| > \left|\frac{\partial \log c_H}{\partial \log r}\right| \tag{5}$$

Intuition: As AI becomes cheaper, the production cost of "lemons" (Q_L) falls more sharply than that of "peaches" (Q_H). This asymmetric cost advantage is the fundamental supply-side driver of information pollution. (For a formal proof, see Appendix A.1).

The Platform-Producer-Consumer Game

We model the market interaction as a three-stage sequential game, following the logic of Stackelberg competition.

Stage I: Platform Strategy Selection.

The platform, as the Stackelberg leader, first chooses its governance policy mix (m, γ_H, γ_L) to Maximise its profit function Π_P:

$$\Pi_P = \theta\rho\left[\gamma_H Q_H^S + \gamma_L(1-m)Q_L^S\right] - C_m(m) \tag{6}$$

where θ is the platform's revenue share, ρ is the average ad revenue per amplified unit of content, j
Q_j^S is the anticipated producer supply, and $C_m(m)$ is a convex moderation cost function.

Stage II: Producer Supply Decision.

Observing the platform's policy, producers choose their optimal supply. The profit per unit of content *j* is:

$$\pi_j = (1-\theta)\rho\gamma_j - c_j(r, w) \tag{7}$$

Heterogeneity in producer productivity Aj, i allows for the derivation of a smooth industry supply curve $Q_j^S(\gamma_H, \gamma_L)$, which satisfies $\partial Q_j^S / \partial \gamma_j > 0$ (direct incentive effect) and $\partial Q_i^S / \partial \gamma_k < 0$ for $k \neq j$ (resource competition effect).

Stage III: Consumer Verification Decision.

Consumers observe a noisy signal $s \in \{H, L\}$ and decide whether to pay their verification cost k_i. The signal precision, π (s = $H|q$ = H) = π (ρ', V^*), is endogenous to the health of the information ecosystem, where ρ' is the effective pollution density and V^* is the aggregate verification rate. This function satisfies $\partial\pi/\partial\rho$' < 0 (pollution dilutes signal quality) and $\partial\pi/\partial V^*$ > 0 (verification creates positive externalities). The aggregate verification rate V^* must satisfy a fixed-point condition, capturing the collective action nature of the verification problem.

Equilibrium and The Threefold Market Failure

Definition 3.3 (Polluted Information Equilibrium). A Subgame Perfect Nash Equilibrium (SPNE) is a strategy profile and outcome combination $\{(m^*, \gamma^*), (Q_H^*, Q_L^*), V^*, \pi^*\}$ such that the decisions of the platform, producers, and consumers are mutually optimal, and beliefs are consistent with outcomes.

Theorem 3.4 (Existence of Equilibrium). Under standard regularity conditions, a Polluted Information Equilibrium exists.

Theorem 3.5 (Market Failure). The Polluted Information Equilibrium is Pareto inefficient, driven by three interacting market failures:

1. *Production Externality: Producers of low-quality content do not internalise the negative social effects of their output. The social marginal cost exceeds the private marginal cost:*

$$MC_{social} = MC_{private} + \frac{\partial}{\partial Q_L}\left[d(Q_L') + \lambda\frac{\partial T}{\partial Q_L}\right] \tag{8}$$

where $d\,(\cdot)$ is the direct consumer harm function and T is the stock of social trust with shadow price λ.

2. *Platform Governance Failure: The platform's objective function is misaligned with social welfare. An engagement bias assumption implies that the platform may find it profitable to set $m^* < m_{SO}$ and $\gamma_L^* > \gamma_{L,SO}$*

3. *Information Commons Externality: Verification is a public good, but consumers are not compensated for its full social benefit, leading to under-investment. The social marginal benefit exceeds the private marginal benefit:*

$$MB_{social} = MB_{private} + \frac{\partial \pi}{\partial V} \tag{9}$$

Proposition 3.6 (Paradox of AI Progress). *A decrease in the cost of AI capital, r, leads to:*

1. *an increase in the supply of low-quality content (∂Q_L^* / ∂r < 0),*
2. *a rise in effective pollution density ($\partial\rho'^*$/ ∂r < 0),*
3. *a decline in decentralised social welfare (∂W^*/ ∂r > 0).*

(For a proof sketch, see Appendix A.3).

The Information Pollution Index and Policy Portfolio

This section develops the Information Pollution Index (IPI) as a **theoretical construct** derived directly from our welfare framework. Given that the precise, nuanced components of this index (e.g., welfare deadweight loss, trust decay dynamics) are not readily available as high-frequency empirical data, the IPI is designed primarily as an operational tool for our policy simulation and analysis in Policy Implications and Practical Value section. It provides a welfare-linked "dashboard" to evaluate policy interventions within a controlled computational environment. This contrasts with the goal of Section 5, which uses a more general, observable proxy for information quality to empirically test the underlying mechanisms of the commons tragedy, rather than to measure the IPI itself. Based on the market equilibrium characteristics revealed by the preceding theoretical model, we construct an Information Pollution Index (IPI) that is endogenous to the social welfare function. This section aims to build a solid bridge from abstract theory to observable measurement, ensuring the index possesses both a rigorous theoretical foundation and practical applicability for policy.

Theoretical Foundation and Axiomatic Properties of the IPI

Theoretical Construction.

Definition 4.1 (Information Pollution Index). *The Information Pollution Index at time t, denoted $IPI(t)$, is defined as a linear combination of its four theoretical dimensions:*

$$IPI(t) = \sum_{j=1}^{4} w_j^*(t) \cdot I_j(t) \tag{10}$$

where the weight w_j^* (t) is endogenously determined by the marginal impact of each dimension on social welfare. This weight reflects the marginal rate of social aversion to each type of harm at a given equilibrium state E (t):*

$$w_j^*(t) = \frac{\left|\frac{\partial W}{\partial I_j}\right|_{E^*(t)}}{\sum_{k=1}^{4}\left|\frac{\partial W}{\partial I_k}\right|_{E^*(t)}} \tag{11}$$

Economic Properties.

An effective economic index should satisfy several desirable axiomatic properties. We show that the IPI meets the following key criteria:

Property 4.2 (Welfare Monotonicity). *The IPI is strictly negatively correlated with social welfare ($\partial W/\partial IPI$ < 0). This property establishes the index as a "social welfare thermometer," ensuring that its rise unambiguously indicates a deterioration in Pareto efficiency.*

Property 4.3 (Decomposability). *Changes in the index can be clearly attributed to its constituent dimensions, as $\partial IPI/\partial I_j$ = w_j^*. This allows policymakers not only to observe the overall level of pollution but also to diagnose the core drivers of the problem for targeted interventions.*

Property 4.4 (Policy Sensitivity). *The total effect of any policy intervention on the IPI can be* *decomposed into a direct effect on the level of each dimension and an indirect effect on the weights via changes in the equilibrium structure (a Lucas-critique-style effect). This property provides a rigorous framework for counterfactual policy evaluation.*

Precise Definitions and Economic Interpretations of the Four Dimensions

The IPI is composed of four dimensions, each capturing a distinct facet of the harm caused by information pollution.

First, we define Effective Pollution Density (I_1), which measures not a simple ratio of content volume, but the effective share of pollution in the consumer **attention market**. It is given by:

$$I_1(t) = \frac{\gamma_L^*(t)\big(1 - m^*(t)\big)Q_L^*(t)}{\gamma_H^*(t)Q_H^*(t) + \gamma_L^*(t)\big(1 - m^*(t)\big)Q_L^*(t)} \tag{12}$$

This dimension captures a core economic reality: low-quality content, amplified by platform algorithms (γ_L^*) in pursuit of engagement and escaping moderation ($1 - m*$), has an impact on finite mental bandwidth that far exceeds its absolute quantity. It is a direct characterisation of pollution's prevalence and salience.

Second, we define Social Welfare Deadweight Loss (I_2), which monetises the harm of information pollution by measuring the economic cost borne by society due to the threefold market failure. Conceptually, it is equivalent to the Harberger's Triangle in public finance, quantifying the net efficiency loss from resource misallocation. Its formal definition is:

$$I_2(t) = \frac{W^{SO} - W^*(t)}{W^{SO} - W^{min}} \tag{13}$$

This dimension answers the key question: "How much value does our society lose due to information pollution?"

Third, we consider the Decay of the Trust Commons (I_3), which treats "trust" as a depletable stock of social capital. Information pollution acts as a corrosive agent, eroding the foundation of market transactions and social cooperation. The indicator captures the **long-term, cumulative, and path-dependent** harm of pollution, defined as:

$$I_3(t) = \frac{T^{max} - T(t)}{T^{max}} \tag{14}$$

where the stock of trust $T\ (t)$ follows the dynamic path $dT(t)/dt = -\delta I_1(t)\mathrm{Flow}(t) + \xi R(t) - \mu T(t)$. Steady-state analysis suggests that persistent high pollution can push society into a **low-trust poverty trap.**

Fourth, we define Asymmetric Technology Risk (I_4), which measures the core technological driver of pollution – a technological arms race between content generation and detection. It is a forward-looking risk indicator of the ecosystem's systemic vulnerability, defined as:

$$I_4(t) = \frac{1}{2}\left[1 + \tanh\left(\frac{\log\left(\frac{Cap_{pat}(t)}{Cap_{tech}(t)}\right) - \mu_{tech}}{\sigma_{tech}}\right)\right] \tag{15}$$

When the capability of generation technology ("offence") far outstrips that of detection technology (“defence”) the marginal cost of producing “lemons” falls, systemically exacerbating the other three dimensions of harm. Graphical framework reference picture 3.

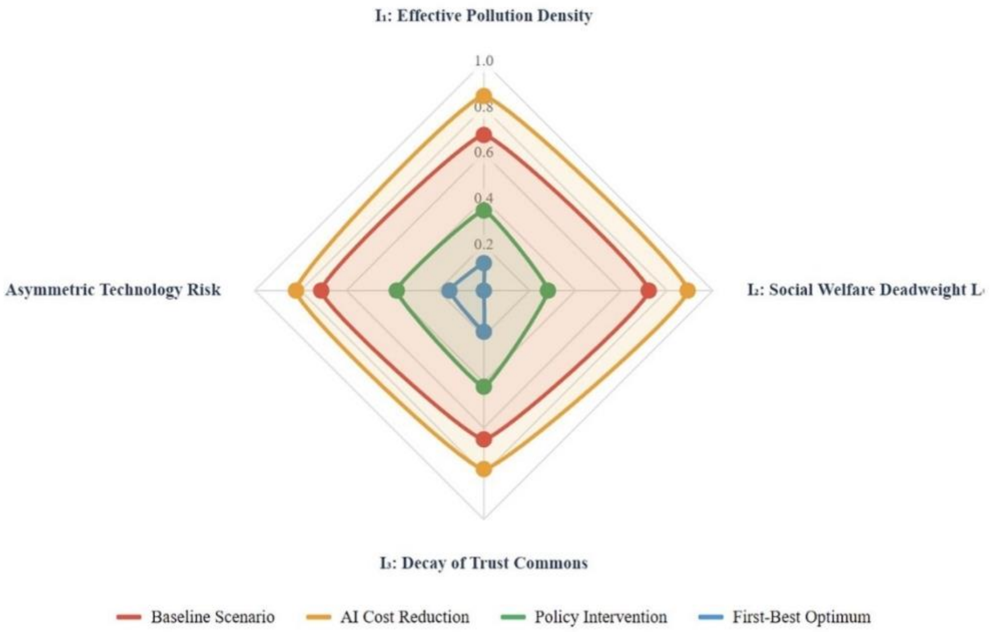


Figure 3. Information Pollution Index (IPI) Four-Dimensional Radar Chart Across Policy Scenarios

This radar chart displays the four dimensions of the Information Pollution Index across different scenarios: (I_1) effective Pollution Density, ($I2$) Social Welfare Deadweight Loss, ($I3$) Decay of Trust Commons, and ($I4$) Asymmetric Technology Risk. Each axis represents normalised values from 0 (best) to 1 (worst). The baseline scenario reflects current market conditions without intervention. Al cost reduction demonstrates the pollution amplification effect of technological progress. Policy intervention shows the effectiveness of the proposed multi-instrument framework. The first-best optimum represents the theoretical benchmark for comparison. Larger areas indicate higher pollution levels and welfare losses.

From Theory to Measurement: The Core Proxy Indicator System

Since the theoretical dimensions are not directly observable, we design a core proxy indicator for each, aiming to unify theoretical validity with empirical tractability.

Our proxy for Effective Pollution Density ($I1$) is the Weighted Exposure Pollution Rate. This is calculated as the ratio of impressions from low-quality content to total impressions:

$$\widehat{I_1} = \frac{\sum_i 1_{(i \text{ is low-quality})} \cdot \text{Impressions}_i}{\sum_i \text{Impressions}_i} \tag{16}$$

We choose “impressions” over “clicks” as the weight because it more closely reflects the initial supply of pollution, being less affected by the endogenous variable of user discernment. This indicator directly measures the “market share” of low-quality content in the platform’s attention allocation mechanism, serving as a direct counterpart to the theoretical dimension $I1$.

To proxy for Social Welfare Deadweight Loss ($I2$), we propose the Weighted Harm Feedback Rate. A simple complaint rate cannot distinguish the welfare loss from “clickbait” versus “financial fraud." By weighting different types of negative user feedback by their potential harm severity, this indicator measures the intensity of harm as revealed by consumer revealed preferences. It is thus a more effective "pain signal" for welfare loss, calculated as:

$$\widehat{I_2} = \frac{\text{HarmType}_j \sum \text{Weight}_j \cdot \text{FeedbackCount}_j}{\text{Total Impressions}} \tag{17}$$

For the Decay of the Trust Commons ($I3$), we propose the Trust-Related Churn Gap via Causal Inference. The decay of trust ultimately manifests as users "voting with their feet." To address attribution challenges, we employ **quasi-experimental methods** (e.g., DiD, RDD) to compare user cohorts randomly or quasi-randomly exposed to different levels of pollution. The resulting difference in their churn rates, normalised by a baseline, quantifies the net effect of trust erosion on economic outcomes, consistent with the gold standard of modern empirical economics. The calculation is:

$$\widehat{I_3} = \frac{\text{Churn}_{\text{high-exposure}} - \text{Churn}_{\text{low-exposure}}}{\text{Churn}_{\text{baseline}}} \tag{18}$$

Finally, to proxy for Asymmetric Technology Risk ($I4$), we use the Public Benchmark Detection Accuracy Gap. This indicator quantifies the "offence-Defence gap" via Standardised adversarial benchmarking, measuring the performance of the best "shield" (current detection systems) against the sharpest "spear" (latest generative models). Changes in this indicator correspond directly to changes in the key exogenous technology parameters of our theoretical model, bridging the theory of endogenous technological change with observable reality. It is calculated as:

$$\widehat{I_4} = 1 - \frac{\text{Accuracy}_{\text{on new models}}}{\text{Accuracy}_{\text{on baseline data}}} \tag{19}$$

Composite Index Construction and Application

After obtaining Standardised time-series data for the four core proxy indicators, we advocate for a **hybrid approach** – combining theoretical deduction, empirical estimation, and expert judgement – to determine their weights. A robust data quality assurance framework is also essential. The resulting IPI can be used for causal impact evaluation in academic research and as a practical tool for long-term ecosystem health monitoring by platforms and regulators, thus closing the loop between theoretical analysis, empirical monitoring, and policy intervention.

The Static Optimal Policy Portfolio: A Theoretical Benchmark

In an ideal world with complete information and a stable technological structure, the policymaker's task is to design a set of instruments that target each of the three market failures respectively.

To correct the **production externality**, classic Pigouvian instruments can be employed. The optimal tax, τ_L^* levied directly on the output of low-quality content QL, should equal its marginal social damage:

$$\tau_L^* = d'\left(Q'_{L,SO}\right)(1 - m_{SO}) + \lambda^* \frac{\partial T}{\partial Q_L} \tag{20}$$

where the first term represents the direct harm to consumers and the second term captures the damage to the stock of social trust (with shadow price λ^*).

To correct the **information commons externality** arising from under-investment in verification, policy can directly enhance the average precision of public signals, π, through mandatory content provenance standards. This increases the perceived disutility of low-quality content, thereby suppressing its production and dissemination throughout the game's equilibrium path.

To correct the **platform governance failure**, policy can impose information fiduciary duties. This is formally equivalent to modifying the platform's objective function to a weighted sum of its own profit and social welfare:

$$\max_{m,\gamma}(1-\alpha)\,\Pi_P + \alpha[v(Q_H') - d(Q_L')] \tag{21}$$

where the intensity of the fiduciary duty, $\alpha \in [0, 1]$, becomes a key regulatory choice variable.

Theorem 4.5 (First-Best Policy Portfolio). *In a static environment, a policy portfolio* $(\tau_L^*\ \pi_{SO},\ \alpha^*)$ *consisting of an optimal Pigouvian tax, mandatory provenance standards, and information fiduciary duties can implement the first-best social optimum.*

Table 1 summarises this idealised mapping.

Table 1. The Triad of Market Failures and Corresponding Policy Instruments

Market Failure	Locus	Consequence	Instrument
Production Externality	Producers	Over-production of QL	τ_L / Permits
Info. Commons Ext.	Consumers	Under-verification	Provenance
Platform Gov. Fail.	Platform	↑ γ_L, ↓ m	Fiduciary Duty

Dynamic Challenge: The Lucas Critique and Policy-Induced Innovation

However, the effectiveness of any static policy portfolio faces the fundamental challenge of the Lucas (1976) critique. Rational agents will strategically react to policy interventions, thereby altering the very economic structure upon which the policy was based. In our model, a sustained tax τL on low-quality content will lower the marginal return to innovations that enhance its production efficiency, A_L. Rational R&D efforts will thus be reallocated.

Proposition 4.6 (Policy-Induced Innovation Bias). The *imposition of* $\tau_L > 0$ *induces a technology shift biased towards high-quality content, i.e.,* $\partial(\mathrm{A}_{\mathrm{L}}^*\ /\mathrm{A}_{\mathrm{H}}^*)/\partial\tau L < 0$, *where* A_j^* *are endogenous productivity levels.*

This dynamic feedback implies that any fixed "optimal" tax rate τ^* will become suboptimal over time. The static policy blueprint is not only inefficient but also unreliable in a dynamic context.

Deep Uncertainty Challenge: Knightian Uncertainty and Robust Decision-Making

Beyond predictable endogenous dynamics, the long-term trajectory of AI is fraught with deep, unquantifiable uncertainty, i.e., Knightian uncertainty (Knight, 1921). We lack a firm basis for assigning probabilities to future technological paradigms or "black swan" events. In such a context, a rational social planner should adopt a more robust decision criterion than simple expected utility maximisation, such as max-min expected utility (Gilboa &

Schmeidler,1989).

$$\max_{\mathcal{P}_{policy}} \min_{p \in \mathcal{P}} E_p \left[W\left(\mathcal{P}_{policy}\right)\right] \tag{22}$$

The planner optimises against the worst-case scenario within a set of plausible priors P.

Proposition 4.7 (Optimality of the Precautionary Principle). Under Knightian uncertainty, the optimal policy is inherently more precautionary, favouring higher taxes, stricter caps, and a greater reliance on "fail-safe" instruments like content provenance standards.

The presence of Knightian uncertainty demands that policy design be resilient not only to known dynamics but also to unknown shocks.

A Synthesis: IPI-Centred Adaptive and Robust Governance

Facing the structural fragility revealed by the Lucas critique and Knightian uncertainty, an effective governance framework must transcend the static-optimality paradigm. We argue that the Information Pollution Index (IPI) constructed in the previous section is the core operational tool to bridge this gap from theory to practice.

The IPI, as a real-time, welfare-linked "dashboard," enables a feedback-based adaptive regulation. Policy instruments are no longer fixed constants but are governed by a **state-contingent rule** that adjusts based on the observed health of the ecosystem.

$$\tau_t = \tau_{t-1} + \eta \left(\frac{IPI_{t-1} - IPI_{target}}{IPI_{target}} \right) \tag{23}$$

Under this rule, the tax rate automatically tightens when the IPI exceeds its target and loosens otherwise. This mechanism endogenously responds to systemic drifts driven by technological innovation or market structure changes, thus partially addressing the Lucas critique. Furthermore, by continuously monitoring the IPI and its sub-dimensions, policymakers can achieve earlier detection of unforeseen risks (early signals of "black swans"), enabling timely activation of contingency plans, which is the essence of robust decision-making.

In summary, confronting the governance challenges of the AI era requires abandoning the quest for a rigid, one-off policy blueprint. Instead, we must construct a multi-layered governance system. At its core is principles-based regulation (e.g., safety, transparency, accountability) to ensure long-term stability. Its method involves regulatory sandboxes for exploratory learning to navigate technological uncertainty. Its navigation system relies on real-time monitoring tools like the IPI to shift from reactive responses to anticipatory governance. Together, these three components form a resilient and adaptive "governance vessel" capable of navigating the deep uncertainties of the digital ocean.

Given the multidimensional complexity of the Information Pollution Index (IPI), high-frequency empirical data is not yet available. Therefore, in Section 6, we employ computational simulations to evaluate the effectiveness of IPI as a policy instrument. Before proceeding to simulation, however, Section 5 provides empirical support for the core premise of this study – the tragedy of the digital commons" – through broader macro-level proxy variables.

IPI Validation and Policy Analysis in a Computational Laboratory

While our general equilibrium framework in Section 3 successfully identifies the existence and inefficiency of the static Polluted Information Equilibrium, it is, by design, silent on the dynamic pathways of how this equilibrium forms, shifts, and reacts to policy. To bridge this critical gap from static theory to dynamic application, we develop an Agent-Based Model (ABM).

This ABM serves as a computational laboratory, indispensably required for three reasons. First, it allows us to move beyond a representative agent and explicitly model **heterogeneous agent interactions –** capturing diverse producer strategies and consumer verification thresholds. Second, it can simulate the dynamic feedback loops our static model cannot, such as how rising pollution erodes trust, which in turn alters platform incentives and consumer behaviour over time. Third, it enables the study of **non-equilibrium paths**, revealing how the ecosystem responds to shocks (like new AI or policies) in real-time, rather than only comparing pre- and post-shock steady states.

This rich, dynamic environment is therefore essential for rigorously testing the two primary practical contributions of this paper: the IPI's utility as a real-time ecosystem metric and the adaptive effectiveness of our proposed policy portfolio.

Experimental Design and Methodology

Simulation Model Architecture.

We construct an information economy simulation system with three classes of heterogeneous agents. **Producer Agents**, endowed with heterogeneous productivity parameters ($A_{H,i}$, $A_{L,i}$), make content creation decisions based on expected profit maximisation and a CES production function. **Consumer Agents**, characterised by heterogeneous verification costs ki and risk preferences, make consumption and verification decisions based on Bayesian updating and utility maximisation, with externalities mediated through a social network. The **Platform Agent**, acting as a market intermediary, influences content distribution through algorithmic weights (γ_H, γ_L) and moderation intensity m, aiming to Maximise ad revenue under a user retention constraint.

Key Parameter Settings.

Based on our theoretical model, we set the following core parameters for the baseline simulation: elasticity of substitution for high-quality content σ_H = 0.75 (complementarity) and for low-quality content σ_L = 1.5 (substitutability); 100 producer agents and 300 consumer agents; baseline AI capital cost r = 1.0 and labour cost w = 8.0; and a platform revenue share of θ = 0.25.

Experiment 1: Comprehensive IPI Validation

This experiment was designed to thoroughly validate the theoretical effectiveness, measurement robustness, and predictive power of the IPI. The design included six sub-experiments: a baseline evolution test, a shock response test, a weight sensitivity analysis, a noise robustness test, a historical event detection simulation, and a cross-platform comparison.

The core findings strongly support the IPI's validity. First, the IPI exhibits a strong negative correlation with social welfare (correlation coefficient of -0.839), confirming its theoretical validity as a welfare metric. Second, the index is highly sensitive to external shocks, showing an average increase of 37.5% during four simulated shocks, proving its ability to reflect real-time changes in ecosystem health. Third, the IPI is robust to different weighting schemes, with the IPI-welfare correlation remaining high (between 0.486 and 0.710) across six different configurations. Fourth, it demonstrates excellent resilience to measurement noise, with a measurement error of only 0.045 even at a 20% noise level. Finally, in a simulated fake news event, the IPI increased by 21.6%, showcasing its potential as an early warning tool. Our leading indicator analysis further reveals that the IPI has an optimal prediction window of 5-time steps, providing a valuable reaction window for policymakers.

Experiment 2: Theoretical Validation and Policy Analysis

Validation of the AI Progress Paradox.

We conducted a parameter sweep to systematically test the impact of AI cost (r) and the elasticity of substitution for low-quality content (σL) on the market equilibrium. The parameter space covered $r \in \{0.6, 0.8, 1.0, 1.2, 1.4\}$ and $\sigma L \in \{1.2, 1.4, 1.6, 1.8\}$. The key finding is a strong validation of the "paradox of AI progress": AI cost is positively correlated with social welfare (coefficient of 0.167) and strongly negatively correlated with pollution density (coefficient of -0.770). For instance, at an AI cost of 0.6, pollution density reached a high of 0.954, whereas it decreased to 0.568 when the cost was 1.4. This result provides robust computational support for our proposition that technological progress, due to its asymmetric impact, can paradoxically exacerbate information pollution and harm social welfare.

Analysis of Policy Interventions.

We designed six policy configurations to test the effectiveness of different intervention tools. Table 2 details the performance of key macroeconomic indicators under each policy.

Table 2. Policy Comparison Experiment Results

Policy Scenario	Welfare	Pollution	IPI	Trust
Baseline	78.05	0.774	0.694	0.312
Pigouvian Tax (↑ θ)	78.68	0.663	0.654	0.323
Subsidy (↓ k_{max})	79.29	0.612	0.634	0.236
Joint Policy	79.82	0.753	0.636	0.272
Tech Intervention	79.34	0.596	0.622	0.277
Efficiency Boost	78.97	0.657	0.651	0.321

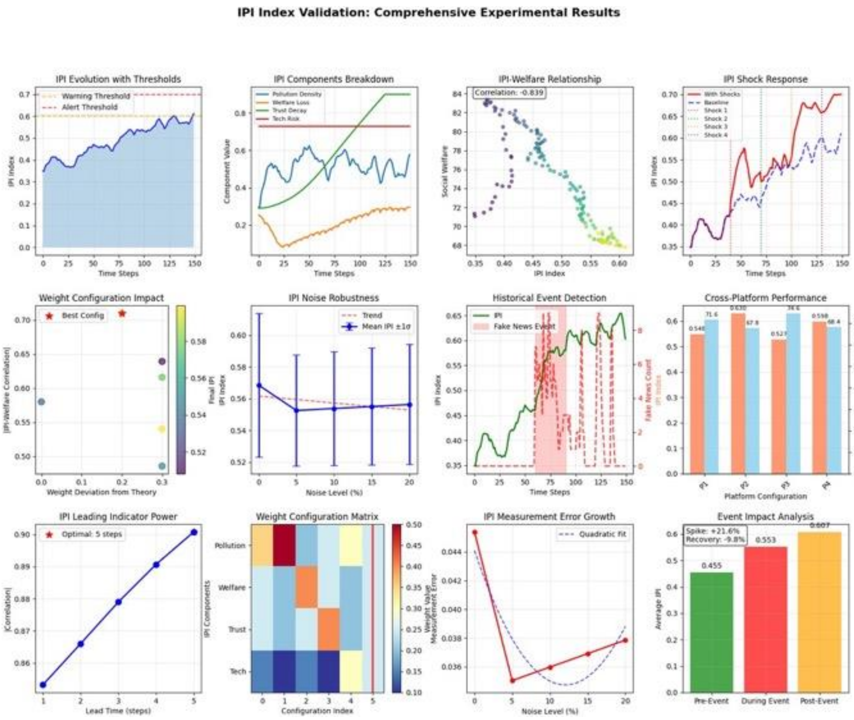


Figure 4. Comprehensive Experimental Results For The Validation Of The Information Pollution Index (IPI)
This figure summarises the findings from Experiment 1, including baseline evolution, component breakdown, IPI-welfare correlation, shock response, and various robustness and sensitivity analyses.

As shown in Table 2, all interventions improved upon the baseline across various dimensions. Notably, the **Joint Policy** achieved the highest social welfare (79.82), a 2.3% improvement over the baseline. The **Tech Intervention** was most effective at reducing pollution density (to 0.596) and the IPI (to 0.622). A key finding is that the effect of single instruments is limited; for example, the Pigouvian tax proxy, while beneficial, was outperformed by comprehensive strategies. This again validates our theoretical claim that a multi-tool policy portfolio is necessary to effectively address the threefold market failure.

Discussion

Our simulation experiments provide substantial validation for the paper's core theoretical contributions. They confirm the significance of asymmetric production technology, demonstrate the interplay of the threefold market failure mechanism, and show the convergence of the system to a suboptimal equilibrium with high pollution, consistent with our theoretical predictions. The experiments also establish the practical value of the IPI as a monitoring tool with early-warning capabilities, cross-context applicability, and utility for policy evaluation.

The results yield important policy implications: single-instrument policies are insufficient; a forward-looking regulatory framework is necessary to address the rapid evolution of AI; and policies must be adaptive, adjusting dynamically based on real-time data. While our simulation provides valuable insights, we acknowledge its limitations, such as model simplification and reliance on calibrated rather than empirically estimated parameters. Future research could enhance this framework by calibrating the model with real-world platform data, expanding the complexity of agent behaviours, and conducting longer-term dynamic analyses.

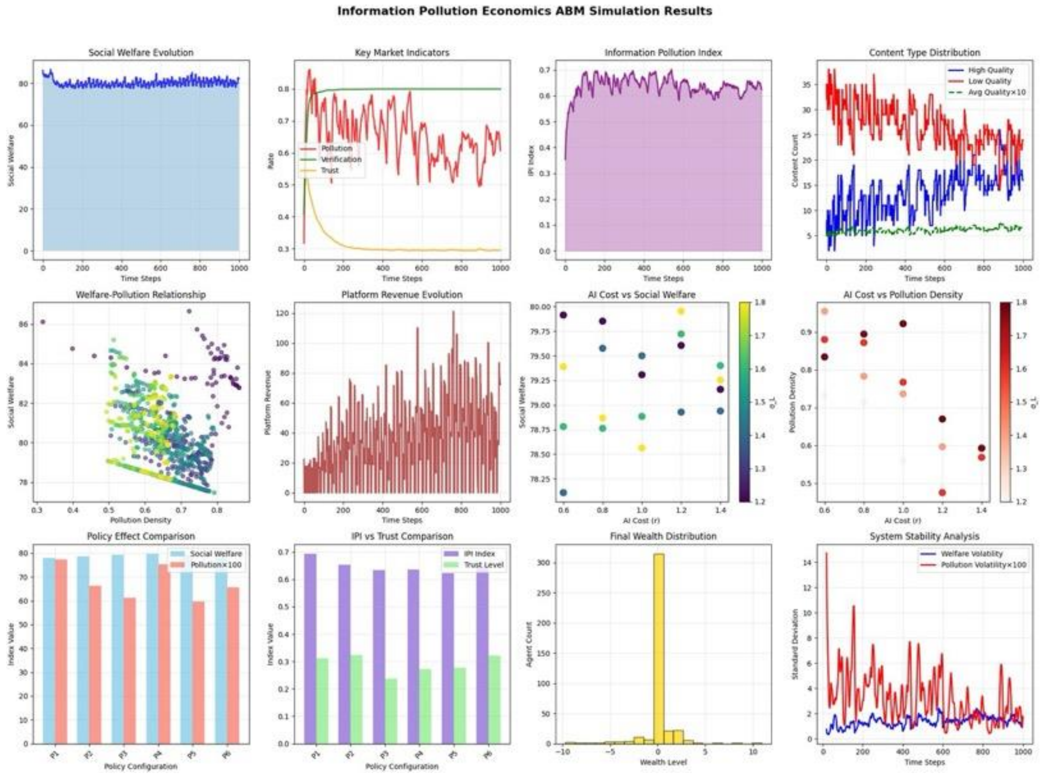


Figure 5. Simulation Results for the Balanced Agent-Based Model, Corresponding to Experiment 2

This figure illustrates the long-term evolution of key system-level indicators, the relationships between core theoretical variables (e.g., AI cost vs. Pollution), the comparative effectiveness of different policy interventions, and analyses of system stability.

Conclusion

Core Findings and Theoretical Contribution

This paper has revealed the mechanism by which generative AI impacts information markets: technological progress, by asymmetrically altering the production costs of different quality content, can paradoxically reduce social welfare. *This Paradox of AI Progress* challenges optimistic, techno-deterministic forecasts, demonstrating that without proper market design, technological innovation may exacerbate, rather than mitigate, market failures.

Our theoretical framework identifies a triad of interacting market failures – a production externality, a platform governance failure, and a trust commons externality. These forces collectively sustain an inefficient *Polluted Information Equilibrium*. This diagnosis moves beyond simple "content moderation" proposals to address deeper, structural issues of economic mechanism design.

Policy Implications and Practical Value

Our research demonstrates that tackling information pollution requires a multi-instrument policy portfolio rather than reliance on a single tool. This includes short-term measures, such as implementing a real-time monitoring system based on the *Information Pollution Index (IPI)* to serve as a “dashboard” for policy response. Mid-term reforms should focus on advancing legislation for platform fiduciary duties to align algorithmic recommendation with social welfare. The long-term strategy must involve building content provenance infrastructure to enhance verification efficiency at a technical level.

It is particularly important that this policy design be adaptive. As AI technology evolves rapidly, fixed rules will quickly become obsolete. The IPI provides an anchor for dynamic adjustment, enabling a counter-cyclical” regulatory approach: interventions tighten automatically when the index worsens and relax as it improves.

Limitations and Future Directions

This study is subject to several limitations which open avenues for future research. Theoretically, our binary quality classification simplifies a reality that is a continuous spectrum; the static model does not fully capture dynamic evolutionary processes; and we do not consider inter-platform competition or user multi-homing. On the empirical side, estimating the core parameters (σ_L and σ_H) requires more granular industrial data, the practical application of the IPI must navigate the balance between data access and privacy protection, and evaluating policy effectiveness demands long-term longitudinal studies.

Future research should explore several key directions. These include endogenizing technological progress to investigate how policy might influence AI R&D trajectories; studying regulatory arbitrage and international cooperation mechanisms for cross-border information flows; expanding the consumer model by incorporating bounded rationality and social learning from behavioural economics; and conducting in-depth analyses of key sectors such as journalism, education, and healthcare.

References

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labour. *Journal of Economic Perspectives*, 33(2), 3-30.

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press.

Akerlof, G. A. (1970). The market for “lemons”: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3), 488-500.

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020a). The welfare effects of social media. *American Economic Review*, 110(3), 629-676.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211-236.

Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, 7(2), 1-8.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130-1132.

Bawden, D., & Robinson, L. (2009). The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*, 35(2), 180-191.

Brynjolfsson, E., & McAfee, A. (2023). *The second wave of AI productivity*. MIT Press.

Bursztyn, L., Rao, A., Roth, C., & Yanagizawa-Drott, D. (2020). Misinformation during a pandemic. Working Paper 2020-44, University of Chicago, Becker Friedman Institute for Economics.

Cabral, L., Haucap, J., Parker, G., Petropoulos, G., Valletti, T. M., & Van Alstyne, M. W. (2021). The EU digital markets act: a report from a panel of economic experts. *Cabral, L., Haucap, J., Parker, G., Petropoulos, G., Valletti, T., and Van Alstyne, M., The EU Digital Markets Act, Publications Office of the European Union, Luxembourg*.

Coase, R. H. (1960). The problem of social cost. *The Journal of Law and Economics*, 3, 1-44.

Crémer, J., De Montjoye, Y. A., & Schweitzer, H. (2019). Competition policy for the digital era. Publications Office of the European Union.

Dales, J. H. (1968). *Pollution, Property & Prices: An Essay in Policy-Making and Economics*. University of Toronto Press.

Durante, R., Pinotti, P., & Tesei, A. (2019). The political legacy of entertainment TV. *American Economic Review*, 109(7), 2497-2530.

Floridi, L. (2010). Information: *A Very Short Introduction*. Oxford University Press.

Hagiu, A., & Wright, J. (2015). Multi-sided platforms. *International Journal of Industrial Organization*, 43, 162-174.

Helbing, D. (Ed.). (2018). *Towards digital enlightenment: Essays on the dark and light sides of the digital revolution*. Springer.

Montgomery, W. D. (1972). Markets in licenses and efficient pollution control programs. *Journal of Economic Theory*, 5(3), 395-418.

Parker, G., & Van Alstyne, M. (2010, June). Innovation, openness & platform control. In *Proceedings of the 11th ACM conference on Electronic commerce* (pp. 95-96).

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39-50.

Posner, E. A., & Weyl, E. G. (2018). *Radical markets: Uprooting capitalism and democracy for a just society*. Princeton University Press.

Rochet, J.-C., & Tirole, J. (2004). Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4), 990-1029.

Spence, M. (1973). Job market signalling. *The Quarterly Journal of Economics*, 87(3), 355-374.

Stiglitz, J. E. (2022). *Information and the digital economy: On the welfare implications of data and AI*. Columbia University Press.

Tufekci, Z. (2015). Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. *Colorado Technology Law Journa*l, 13(1), 203-218.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146-1151.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

Appendices

Appendix A: Model and Experimental Details

This appendix provides supplementary details on the Agent-Based Model (ABM) specification, parameter settings, experimental procedures, and results to ensure the reproducibility of our study.

A.1 B.1 Agent-Based Model Specification

The simulation is built upon the interaction of three agent types whose behavioural rules are designed to Operationalise the core theoretical mechanisms.

A.1.1 Producer Agent Logic.

The production decision of producer i between high-quality (H) and low-quality (L) content is governed by a softmax (logit) choice model based on expected profits:

$$\text{Prob}(i \text{ chooses } H) = \frac{\exp(\beta \cdot E[\pi_{H,i}])}{\exp(\beta \cdot E[\pi_{H,i}]) + \exp(\beta \cdot E[\pi_{L,i}])} \tag{24}$$

where E [π j,i] is the expected profit for content type j given agent i's productivity, and β is a rationality parameter that introduces bounded rationality.

A.1.2 Consumer Agent Logic.

A consumer i with verification cost ki decides to verify a piece of content if the expected utility gain from resolving uncertainty exceeds the cost. The verification thre$_i$shold $k*$ is determined by:

$$k_i^* = p_i(H|s) \cdot \Delta U_H + (1 - p_i(H|s)) \cdot \Delta U_L \tag{25}$$

where pi ($H|s$) is the consumer's posterior belief that the content is high-quality given signal s, and ΔUj is the utility difference between informed and uninformed consumption of content type j.

A.1.3 Platform Agent Logic.

The platform agent adaptively adjusts its algorithmic amplification vector γ and moderation intensity m based on market feedback. The rules follow a gradient-ascent logic, balancing revenue maximisation with a user trust (retention) constraint:

$$\gamma_{L,t+1} = \gamma_{L,t} + \eta \left(\frac{\partial \Pi_P}{\partial \gamma_L} - \lambda \frac{\partial \text{Trust}}{\partial \gamma_L} \right) \tag{26}$$

$$m_{t+1} = m_t + \xi \left(\frac{\partial \Pi_P}{\partial m} - \lambda \frac{\partial \text{Trust}}{\partial m} \right) \tag{27}$$

where η and ξ are learning rates and λ is the shadow price of user trust.

A.2 B.2 Model and Experimental Parameters

Table 3 provides a comprehensive list of the parameters used in the simulations.

Table 3. Model and Experimental Parameter Settings

Category	Parameter Name	Symbol / Field	Value Range or Default
General	Max Steps	`max_steps`	150 (Exp 1), 1000 (Exp 2)
	Random Seed	-	42
Agents	Num Producers	`n_producers`	80 (Exp 1), 100 (Exp 2)
	HQ Productivity	`productivity_H`	Lognormal dist.
	LQ Productivity	`productivity_L`	Lognormal dist.
	Num Consumers	`n_consumers`	200 (Exp 1), 300 (Exp 2)
	Num Platforms	`n_platforms`	1
Economic	Platform Share	θ / `theta`	0.25
	Ad Revenue Base	ρ / `rho`	4.0
	Labour Cost	w	8.0
	AI Cost	r	1.0 (Baseline), swept in Exp 2
Technology	HQ Elasticity	σ_H / `sigma_H`	0.75
	LQ Elasticity	σ_L / `sigma_L`	1.5 (Baseline), swept in Exp 2
	HQ Input Share	δ_H / `delta_H`	0.35
	LQ Input Share	δ_L / `delta_L`	0.65
Behavioural	Max Verification Cost	k_{max}	4.0
	Risk Aversion	`risk_aversion`	Beta (2,3) dist.
	Trust Decay Rate	`trust_decay`	0.05
IPI Weights	Pollution Weight	`w_pollution`	0.35
	Welfare Loss Weight	`w_loss`	0.25
	Trust Decay Weight	`w_decay`	0.25
	Tech Risk Weight	`w_risk`	0.15

A.3 B.3 Overview of Experimental Procedures

The experimental validation was conducted through a series of structured simulation runs.

- Baseline Experiment: The model was initialised and run for the maximum number of steps. Market states and agent behaviours were updated at each step, and key metrics were collected.
- Exogenous Shock Experiment: Different types of shocks were injected at pre-set time steps (40, 70, 100, 130). The system's response was measured by comparing the IPI trajectory before and after the shocks.
- Weight Sensitivity Experiment: Multiple IPI weight combinations were constructed. The model was run for 100 steps for each combination, and the final correlation between IPI and social welfare was recorded.
- Noise Robustness Experiment: Varying levels of measurement noise were added to the

baseline model. Three independent trials were run for each noise level to calculate the volatility and measurement error of the IPI.

- Cross-Platform Comparison: The network topology and agent preference parameters were altered to simulate different platform environments. Each configuration was run for 120 steps to compare final IPI and welfare outcomes.
- Parameter Sweep & Policy Comparison (Exp 2): The simulation was run for 120 steps for each combination of AI cost (r) and LQ elasticity (σ_L). Additionally, six distinct policy configurations were simulated for 150 steps each to evaluate their impact on key outcomes.

A.4 B.4 Summary of Main Experimental Results

Tables 4 and 5 provide a consolidated summary of the key numerical findings from our two main experiments.

Table 4. Key Metrics from IPI Validation Experiment (Exp 1)

Phase	Metric	Value
Baseline	Final IPI	0.611
	Final Social Welfare (W)	67.78
	IPI-W Correlation	-0.839
Shock Resp	Avg. IPI Increase	+37.5%
	Recovery Rate (/step)	0.19-2.21
Weight Sens	Correlation Range	0.486-0.710
	Best Weight Config.	Equal
Noise Robust	Meas. Error (20% noise)	0.045
	Avg. Volatility	0.033
Event Detect	Peak IPI Increase	+21.6%
	Recovery Rate	-9.8%
Cross-Platform	Min/Max IPI	0.527/0.630
	Difference	0.103

Table 5. Key Results from Balanced Simulation (Exp 2)

Phase	Metric	Value
Baseline (1000 steps)	Social Welfare	82.25
	Pollution Density	0.607
	IPI	0.624
	Verification Rate	0.800
	Trust Level	0.295
Parameter Sweep	r-W Correlation	+0.167
	r-Pollution Correlation	-0.770
Policy Comp	Best Policy (Joint)	+2.3% W

Phase	Metric	Value
	Best Anti-Pollution	Tech Interv.

A.5 B.7 Code and Data Availability

The complete Python scripts for Experiment 1 and 2 are available in the ‘main‘ branch of the following repository: github.com/Your-Repo/Information-Pollution-ABM

The required environment can be set up using Conda. The simulation results (CSV files and figures) are stored in the ‘results/’ and ‘figures/’ directories, respectively.

Appendix C: Technical Details and Proofs

This appendix provides the technical derivations and formal proofs for the key results presented in the main text.

Table 6. Notation Summary

Symbol	Description
Q_j	Content output of type $j \in \{H, L\}$ (high-/low-quality)
A_j	Total factor productivity for content type j
K_{AI}, L_H	AI capital input and high-skilled labour input
δ_j, ρ_j	CES share and substitution parameters for type j
σ_j = 1/ (1 − ρ_j)	Elasticity of substitution for type j
r, w	Rental rate of AI capital and wage rate for labour
$c_j(r, w)$	Unit cost function for producing Q_j
θ	Platform’s revenue share parameter
ρ	Ad-revenue base per unit of amplified content
m	Moderation intensity chosen by platform
γ = (γ_H, γ_L)	Algorithmic amplification weights
k_i	Verification cost of consumer i
V	Aggregate verification rate in the system
$\pi(s{=}H \mid q{=}H)$	Signal precision given pollution and verification
W	Social welfare function
IPI	Information Pollution Index

A.6 C.1 Derivation of Asymmetric Production Costs

A.6.1 Derivation of the CES Cost Function.

For the CES production function$Q_j = A_j\left[\delta_j K_{(AI)}^{(\rho_j)} + \left(1-\delta_j\right)L_H^{(\rho_j)}\right]^{(1/\rho_j)}$, the cost-minimisation problem is:

$$\min_{K_{AI},L_H} \quad r\,K_{AI} + wL_H \tag{28}$$

$$\text{s.t.} \quad A_j\left[\delta_j K_{AI}^{\rho_j} + \left(1-\delta_j\right)L_H^{\rho_j}\right]^{1/\rho_j} = Q_j$$

The Lagrangian is L = rK_{AI} + wLH − λ (Aj [. . .]$^{1/\rho j}$ − $Q\ j$). The first-order conditions yield the optimal factor demand ratio:

$$\frac{K_{AI}}{L_H} = \left(\frac{\delta_j}{1-\delta_j}\right)^{\sigma_j} \left(\frac{w}{r}\right)^{\sigma_j} \tag{29}$$

Substituting this back into the production constraint and solving for the total cost yields the unit cost function:

$$c_j(r, w) = \frac{1}{A_j}\left[\delta_j^{\sigma_j} r^{1-\sigma_j} + \left(1-\delta_j\right)^{\sigma_j} w^{1-\sigma_j}\right]^{\frac{1}{1-\sigma_j}} \tag{30}$$

A.6.2 Comparison of Logarithmic Elasticities.

The elasticity of the unit cost with respect to the AI capital price r is equivalent to the cost share of AI capital, $s_{j,AI}$:

$$\frac{\partial \log c_j}{\partial \log r} = \frac{\delta_j^{\sigma_j} r^{1-\sigma_j}}{\delta_j^{\sigma_j} r^{1-\sigma_j} + \left(1-\delta_j\right)^{\sigma_j} w^{1-\sigma_j}} = s_{j,AI} \tag{31}$$

To prove Proposition 1, we must show that $|s_{L,AI}| > |s_{H,AI}|$. Given Assumption 1 ($\sigma_L > 1 > \sigma_H$), it follows that $1 - \sigma_L < 0$ and $1 - \sigma_H > 0$. As r decreases, the term $r^{1-\sigma L}$ increases, while $r^{1-\sigma H}$ decreases. This implies that the cost share of AI capital, s_L, $_{AI}$, is more sensitive to changes in r than

$s_{H,AI}$. Therefore, the cost-reducing effect is larger for low-quality content. □

A.7 C.2 Proof Sketch for Existence of Equilibrium

The proof proceeds by backward induction.

1. Consumer Stage: For a given pollution density ρ', we define a mapping T: [0, 1] → [0, 1] where $T(V_e) = F(k^*(\pi(\rho', V_e), \rho'))$ maps an expected verification rate V_e to the actual rate. Since k^* and π are continuous in their arguments and $F(\cdot)$ is a continuous distribution function, T is a continuous mapping from a compact, convex set to itself. By Brouwer's Fixed-Point Theorem, a fixed point V^* existed.
2. Producer Stage: Given producer heterogeneity, the aggregate supply function $Q_j^S(\gamma_H, \gamma_L) = \int q_{j,i}^*(\gamma)\ di$ is continuous in γ by the Theorem of the Maximum.
3. Platform Stage: The platform Maximises a continuous profit function $\Pi_P(m, \gamma)$ over a compact strategy space $S = [0,1] \times [0, \bar{\gamma}]^2$. By the Weierstrass Extreme Value Theorem, a maximum exists.

The existence of optimal strategies in each stage implies the existence of an SPNE. □

A.8 C.3 Proof Sketches for Inefficiency and Comparative Statics

A.8.1 Proof of Theorem 2 (Threefold Market Failure).

The social planner's problem is to choose (Q_H, Q_L, m, V) to maximise social welfare. The first-order conditions (FOCs) are derived. We then compare these social FOCs with the FOCs from the decentralised equilibrium:

- Production Externality: The producer's FOC, $(1-\theta)\,\rho\gamma j = MCj$, lacks the social harm terms present in the planner's FOC for QL.
- Platform Failure: The platform's FOCs for m and γ_L are based on maximising private profit from engagement, which structurally deviates from maximising social welfare.
- Commons Externality: The consumer's FOC for verification, k_i = Private Benefit, lacks the positive externality term $\partial\pi/\partial V$ that appears in the planner's FOC for V.

These three wedges prove the Pareto inefficiency of the equilibrium. □

A.8.2 Proof of Proposition 1 (Paradox of AI Progress).

We use the Implicit Function Theorem on the system of equations defining the equilibrium.

1. The total derivative of the producer's FOC with respect to r shows a direct negative cost effect on Q_L^S.
2. The platform's optimal response to this supply shift is to adjust (γ_L^*, m^*), which further amplifies the initial shock. The combined effect leads to $Q_L^*/\partial r < 0$.
3. Since Q_L^* increases more sensitively than Q_H^* with a fall in r (due to cost asymmetry), the pollution density ρ'^* increases, so $\partial\,\rho'^*/\partial r < 0$.
4. By the Envelope Theorem, $\partial W^*/\partial r = \partial\mathcal{L}/\partial r > 0$, where $\mathcal{L}$ is the Lagrangian for the planner's problem evaluated at the decentralised equilibrium. Since welfare W^* is decreasing in pollution ρ'^*, and ρ'^* is decreasing in r, welfare must be increasing in r.