Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Ranking Movies with LLMs

Bachelor's Thesis

Tilman Sieber

siebert@ethz.ch

Distributed Computing Group
Computer Engineering and Networks Laboratory
ETH Zürich

Supervisors:
Luca Lanzendörfer, Frédéric Berdoz
Prof. Dr. Roger Wattenhofer

September 12, 2025

Acknowledgements

I would like to express my sincere gratitude to my supervisors Frédéric Berdoz and Luca Lanzendörfer under Prof. Dr. Wattenhofer for their excellent guidance and mentorship throughout this thesis. Your insightful advice, constructive feedback, and supportive approach made this research both productive and enjoyable.

I also extend my thanks to the broader distributed computing group and the faculty and staff of the D-ITET Department. Access to their computing cluster proved essential for this work.

I am deeply grateful to my parents for their unwavering support and encouragement throughout my studies.

Abstract

This thesis investigates whether computational dialogue analysis can extract features to predict movie reception from screenplays alone. The developed framework leverages large language models (LLMs) to analyze 2,690 screenplays paired with aggregated ratings from multiple platforms (IMDb, Metacritic, Rotten Tomatoes), examining whether linguistic features of dialogue correlate with reception. A robust preprocessing pipeline transforms heterogeneous screenplay formats into a unified JSON structure, achieving classification accuracy above 99% for dialogue, character names, stage directions, and scene headers, while validating titles and retrieving corresponding metadata across platforms.

The primary methodological innovation involves finetuning separate language models on high and low quality dialogue corpora, then computing comparative perplexity ratios between these specialized models. This approach yields significant correlations with movie ratings: Pearson $r = 0.22$ ($p < 0.001$) and Spearman $\rho = 0.30$ ($p < 0.001$). In contrast, traditional perplexity analysis using state-of-the-art LLMs without quality specific finetuning, performed through six different methods targeting various rationales, produced only small correlations.

This work establishes finetuned models perplexity based comparative analysis as a viable foundation for computational screenplay assessment, providing empirical evidence that dialogue quality manifests through differential patterns of linguistic predictability rather than absolute perplexity levels. These insights into quality correlated dialogue patterns may inform future development of generative AI systems.

Contents

Chapter 1

Introduction

The computational analysis of screenplay dialogue remains notably absent from natural language processing research, despite significant advances in analyzing other textual domains. While recent work has successfully applied language models to movie related tasks, from sentiment analysis of reviews [1] to multimodal genre classification [2, 3], direct analysis of screenplay text has been constrained by dataset limitations and the lack of established quality metrics. The MovieNet dataset [4], the largest scientific collection, includes only 479 scripts in unstructured format among its 1,100 movies for visual processing, illustrating the scarcity of analyzable screenplay data. This thesis addresses these limitations by developing computational methods to assess screenplay dialogue quality through linguistic predictability patterns and curating a high quality dataset.

Current approaches to movie quality assessment rely predominantly on external indicators such as ratings and reviews [5, 6] rather than analyzing the source material itself. Dialogue evaluation frameworks, when they exist, emphasize functional aspect, such as task completion, harmlessness and context integration, over intrinsic linguistic quality [7]. This gap is particularly striking given recent theoretical advances demonstrating that language models inherently encode quality preferences [8], suggesting that computational discrimination of dialogue quality should be feasible through appropriate methodological design.

This work proposes that dialogue quality manifests through patterns of linguistic predictability measurable via language model perplexity. Specifically, the hypothesis posits that quality specific finetuning can surface implicit quality markers that emerge through comparative analysis between models trained on high and low quality dialogue corpora. This approach builds on the principle that language models function as latent reward models [8], extending this insight from preference alignment to quality assessment.

To test this hypothesis, a dataset of 2,690 movie screenplays was assembled and processed, matching each with audience and critic ratings from multiple platforms to create composite quality measures. This scale of analysis, unprecedented in screenplay research, enables statistical validation of patterns that would remain hidden in smaller samples. The preprocessing pipeline addresses the notorious heterogeneity of screenplay formats, transforming diverse source materials into a unified, analyzable structure while preserving the semantic distinctions between character names, dialogue, stage directions, and scene headers.

The methodological approach employs two complementary strategies. First, comprehensive perplexity analysis using state-of-the-art language models examines how dialogue predictability varies across different contextual conditions and aggregation levels. Second, a comparative framework using finetuned models specialized on high and low quality dialogue reveals quality specific linguistic patterns through their differential perplexity measurements, patterns hidden to general purpose models.

The findings challenge simplistic notions about the relationship between linguistic complexity and quality. While standard perplexity measures show minimal correlation with ratings, the comparative analysis between specialized models achieves statistically significant predictive utility (Pearson $r = 0.22$, Spearman $\rho = 0.30$; $p < 0.001$), demonstrating that screenplay quality emerges not from absolute predictability levels but from adherence to quality specific linguistic patterns that can be computationally identified and measured.

This thesis makes four primary contributions to the field. First, it establishes a robust preprocessing pipeline and standardized data format for screenplay analysis, addressing the format heterogeneity that has hindered previous research. Second, it provides the most comprehensive empirical analysis to date of the relationship between screenplay dialogue predictability and audience reception. Third, it demonstrates that finetuning on quality stratified data can encode implicit quality markers that emerge through model comparison. Finally, the methodological framework enables future investigations into computational approaches to creative writing analysis.

The implications extend beyond screenplay analysis to broader questions about computational assessment of creative works and preference learning. The results demonstrate that computational analysis can reveal patterns in creative works that complement traditional critical approaches. The discriminative capability for assessing dialogue quality may inform the generation and validation of preference data for reinforcement learning from human feedback (RLHF), potentially improving dialogue generation in AI systems.

The remainder of this thesis is organized as follows: Chapter 2 reviews related work in screenplay analysis, dialogue evaluation, and preference learning. Chapter 3 details the methodology, including the preprocessing pipeline, perplexity evaluation methods, and finetuning approaches. Chapter 4 presents the experimental setup and implementation details. Chapter 5 reports findings across metadata analysis, perplexity evaluation, and finetuning experiments. Finally, Chapter 6 concludes with a summary of contributions and directions for future research.

Chapter 2

Related Work

Screenplay Analysis Datasets. The computational analysis of screenplays has been significantly hindered by limited dataset availability and formatting inconsistencies. Conventional repositories such as Kaggle provide poorly or unprocessed screenplay collections that lack standardization, while many scripts employ idiosyncratic formatting styles that complicate systematic parsing. The MovieNet dataset [4] represents the largest scientific collection with 1,100 movies and emphasizes multimodal annotations for visual understanding; however, only 479 scripts are included in unstructured text format, substantially limiting their utility for dialogue or focused analysis. ScreenWriter [9] partially addresses this scarcity by automatically generating screenplays from video content through scene segmentation and character identification algorithms, though this approach prioritizes visual narrative correspondence over linguistic quality assessment. This persistent gap in datasets specifically curated for dialogue evaluation motivates the systematic preprocessing pipeline and quality based corpus construction developed in this thesis.

Movie Metadata and Rating Systems. Extensive research has examined movie metadata and rating systems, though these investigations have primarily focused on recommendation optimization rather than intrinsic content quality assessment. Recent comparative studies of recommendation systems [5] have leveraged rich datasets from IMDb and Rotten Tomatoes, incorporating ratings, user profiles, and critics' reviews through sentiment analysis techniques. While these approaches successfully process textual data through NLP methods, they apply these techniques exclusively to user generated reviews rather than analyzing the screenplays or content themselves. Parallel investigations [6] have documented systematic divergences between critics' and audience scores, revealing genre specific patterns and temporal trends in reception metrics. Despite these advances in understanding movie reception patterns, a critical gap remains in connecting these external reception indicators to intrinsic screenplay characteristics, a limitation that the perplexity based approach addresses through direct analysis of source dialogue text.

Sentiment and Genre Classification. Recent advances in applying language models to movie content have predominantly focused on external attributes rather than intrinsic dialogue quality assessment. Sentiment analysis research [1] has successfully leveraged document level annotations such as star ratings to learn rich sentiment representations, though these methods operate on post release reviews rather than the screenplays themselves. Genre classification approaches have evolved from purely visual methods to sophisticated multi modal frameworks: Movie-CLIP [2] incorporates language augmentation modules for audio element recognition, while VSM-LLM [3] extends this paradigm through temporal alignment of audio-visual features, achieving improved recall on genre prediction tasks. Although these methods demonstrate the value of linguistic features in movie understanding, they process final film productions rather than source screenplays and classify categorical movie attributes rather than assessing quality metrics of the underlying dialogue.

Generative Screenplay Tools. The emergence of LLM based screenplay generation has illuminated both the potential and current limitations of language models in creative writing domains. HoLLMwood [10] implements a sophisticated multi agent framework where LLM agents assume distinct professional roles (Writer, Editor, Actors) to mimic industry workflows, demonstrating measurable improvements in narrative coherence and audience engagement through automated evaluation metrics. Script2Screen [11] extends this approach by integrating textual scriptwriting with audiovisual scene generation, enabling iterative dialogue refinement through multimodal feedback loops. However, these systems reveal a critical limitation: quality determination relies predominantly on AI judges rather than empirical metrics grounded in human reception data. This limitation underscores the importance of developing computational methods to discriminate dialogue quality based on actual audience response patterns.

Preference Learning and Alignment. The finetuning of language models relies heavily on preference data to align outputs with human expectations and values. Llama 2 Chat [12] exemplifies this approach through extensive human evaluation campaigns focused on helpfulness and safety metrics to optimize conversational capabilities in their models dialogue. However, current preference datasets predominantly emphasize instruction following accuracy and task completion metrics rather than linguistic quality or engagement factors. The scarcity and expense of high quality preference data, often dependent on AI judges for scalable labeling, presents a significant bottleneck in model development. The finetuning based perplexity discrimination method between high and low quality dialogue developed in this thesis offers a potential pathway for automatically discriminating or even generating preference data focused on linguistic excellence rather than functional correctness. This addresses an underutilized opportunity in alignment research.

Dialogue Quality Assessment. Current dialogue evaluation frameworks predominantly emphasize functional aspects of dialogue over intrinsic linguistic quality. A comprehensive review [7] identifies critical gaps in quality assurance for dialogue systems, particularly noting the absence of automatic testing methods for non task oriented systems and the lack of unified standards for defining response quality. Most existing metrics focus on robustness, harmlessness, context integration, and task completion accuracy rather than the intrinsic quality of language production itself. This functional bias leaves dialogue systems without clear benchmarks for linguistic excellence or creative expression. The comparative perplexity approach using finetuned models directly addresses this gap, providing an empirical foundation for assessing dialogue quality that operates independently of task performance or semantic appropriateness considerations.

Preference Embedding in Language Models. Recent theoretical advances demonstrate that language models inherently encode preference structures without explicit reward modeling. Direct Preference Optimization [8] reveals that language models function as implicit reward models, enabling preference alignment through simple classification losses rather than complex reinforcement learning procedures. This work replaces the traditional RLHF pipeline with a closed form solution that directly optimizes for preference rankings, showing that LMs can internalize quality distinctions through their parameters. The theoretical foundation that models embed latent preference structures motivates using finetuned models as quality discriminators, because if preferences can be extracted from model parameters for alignment, then quality specific finetuning should similarly capture meaningful preference signals correlated with reception metrics.

Chapter 3

Methodology

This chapter details the computational framework developed to analyze relationships between screenplay dialogue characteristics and audience reception. The methodology comprises four core components: (1) A preprocessing pipeline employing LLM based classification augmented with heuristics to transform raw screenplays into a structured JSON format; (2) Metadata aggregation from multiple rating platforms (IMDb, Metacritic, Rotten Tomatoes) and composition of their different systems into a normalized composite "Golden Score"; (3) Comprehensive perplexity analysis using six methodological variants with different masking, locality, and granularity techniques, implemented through a custom double sliding window attention mechanism to handle scripts exceeding model context limits and ensure absolute consistent calculations; and (4) Two fine-tuning approaches, one regression based golden score prediction and the other quality stratified perplexity optimization, to develop specialized models for dialogue quality assessment. The framework includes comparative ratio analysis between differently sized and differently trained models, plus distributional metrics (thresholds and percentiles) to capture fine grained perplexity characteristics beyond aggregate statistics. This multifaceted approach enables systematic investigation of how computational linguistic metrics relate to human quality judgments of cinematic dialogue.

3.1 Dataset Creation

The preprocessing pipeline transforms raw screenplay text files from various formats and annotation standards into a unified, high quality data format optimized for research analysis. The pipeline addresses the inherent heterogeneity in screenplay formatting while maintaining the semantic integrity of the original content. The primary objective is to classify every text segment according to its screenplay function into one of three categories: **Sluglines** (scene headers), **Dialogue** or **Text** (stage directions). Additionally, the pipeline identifies the speaking character for each dialogue segment while also detecting various speech properties including vocal tonation and emotional delivery, establishing complete attribution chains with rich contextual information. Traditional rule-based systems proved inadequate for screenplay line classification, as they cannot handle the significant variability in authorial styles, annotation standards, and formatting conventions that characterize the medium. The contextual ambiguity inherent in distinguishing dialogue, stage directions, and character attributions requires understanding of narrative flow and context that exceeds simple pattern matching approaches. Therefore, the pipeline incorporates a large language model as the core classification component, augmented by targeted heuristics to address systematic errors and enhance processing efficiency. The resulting granular classification enables targeted analysis of different screenplay components and forms the foundation for subsequent computational processing.

3.1.1 Preprocessing Pipeline

Chunking Strategy. Given the token limitations of downstream language models, screenplays must be segmented into processable units while preserving semantic coherence. The system employs a hierarchical chunking strategy that prioritizes scene boundaries, identifying transitions through keywords such as "INT," "EXT," or "SMASH CUT TO," following conventions such as those established in the Fountain screenplay markup syntax [13]. When optimal scene boundaries cannot accommodate size constraints, the algorithm progressively considers less ideal breakpoints, balancing the competing requirements of maintaining semantic units and meeting technical limitations. Files that cannot be cleanly segmented are flagged as problematic, with forced segmentation applied as a fallback measure; flagged files weren't further used in this work.

LLM Classification. The LLM processes screenplay content following structured instructions and assigns a classification label to each identified segment, outputting results in a standardized JSON schema. Beyond classifying the segments, the LLM is responsible for determining which textual elements belong together as coherent units, making segmentation decisions that can separate content even within the same line when appropriate. This approach leverages the model's contextual understanding to resolve formatting inconsistencies and context dependent ambiguities, enabling it to distinguish between similar textual patterns based on their dramatic function and surrounding narrative context.

Heuristic Refinement. While the LLM effectively handles semantic classification challenges and adapts to unconventional formatting, systematic errors remain. The final stage combines domain specific heuristics with LLM predictions to refine classifications, leveraging both the model's contextual understanding and rule based corrections for common failure modes. These heuristics were curated from observed failure patterns in the LLM's output to address recurring classification errors. This hybrid approach significantly improves overall classification accuracy while maintaining processing efficiency.

Title Verification. Since most analytical methods depend on accurately matching each screenplay with its corresponding metadata (such as audience ratings), an additional verification step is introduced to confirm and correct, if necessary, the movie title associated with each screenplay by the data source. This process validates title labels (typically derived from filenames or database entries) against the screenplay's internal content to ensure accurate correspondence.

At first, a simple LLM is utilized to extract title information from the beginning of each screenplay, which typically contains a title page with the movie name. If no title can be extracted or the identified title shows insufficient alignment with the alleged title, a second model is employed. This second stage employs an advanced LLM with enhanced capabilities, leveraging its ability to recognize movies from their content even in the absence of explicit title information. When this model disagrees with the alleged title, its prediction is prioritized over the original filename based title, as the model's content based identification is considered more reliable than potentially corrupted or mislabeled filenames.

Quality Filtering. As a final preprocessing step, corrupted files were identified and removed from the dataset. The perplexity analysis methods employed in the subsequent analytical phase served as effective corruption detection tools, as corrupted text inherently exhibits significantly elevated perplexity scores compared to standard dialogue. Files flagged by these metrics were manually inspected to confirm corruption status. All confirmed corrupted files were excluded from the final dataset to ensure data quality and analytical validity.

Classification Validation. To assess the accuracy of the classified segments, I employed a random sampling approach with replacement to select excerpts from scripts across each category. A segment was considered correctly classified if and only if it only contained content from its assigned category. No established ground truth data existed for this task, and automated AI based rating systems posed concerns regarding potential overhead and error introduction, as they could exhibit similar biases to the initial LLM classification or introduce novel sources of error. Therefore, I conducted manual validation of the classifications.

After obtaining the raw accuracy numbers for each category, I computed an aggregate measure. The weighted misclassification rate represents the individual misclassification rates of each category weighted by their respective proportions in the dataset. Subsequently, I performed binomial hypothesis testing to establish upper confidence limits on all misclassification rates.

3.1.2 Custom Data Format

Each screenplay is stored independently as a JSON file. In this format, a movie is represented as a list of scenes, and each scene is composed of a sequence of structured elements stored as dictionaries. These dictionaries are referred to as *blocks* and are the smallest units of the screenplay data that contain the classified and annotated content of the original script. As a result, each raw screenplay file is transformed into a list of scenes, each scene being a list of blocks (dictionaries).

There are three types of blocks, distinguished by their `cat` (category) field:

- **Slugline**: Contains only one additional field, `text`. Sluglines indicate scene transitions and typically provide information about the setting, such as whether a scene is indoors or outdoors.
- **Text**: Also contains only a `text` field. These blocks capture a wide range of content, including scene directions, character thoughts, and camera instructions.
- **Dialogue**: The most important block type. Dialogue blocks represent lines spoken by characters. They always include the character's name and the spoken line. Additional metadata from the raw files is stored in two optional fields:
  - `character_instructions` (e.g., *voiceover*)
  - `parentheses` (e.g., *he says quietly to his neighbor*)

The refined data format also enabled quantitative analysis of screenplay composition and structure. This included calculating the proportional distribution of different screenplay blocks and examining the statistical properties of dialogue segments, including their length distributions and character wise dialogue allocation.

Figure 3.1 illustrates the preprocessing pipeline's transformation of a raw input script into the structured JSON format used for analysis.

Original Script

```
ON JOHN MCCLANE
mid-thirties, good-looking,
    athletic
and tired from his trip. He sits by
the window. His relief on landing
    is
subtle, but we NOTICE. Suddenly, he
hears --

SALESMAN V.O.
(amused)
You don't like flying, do you?

McClane turns, looks at the Babbit
clone next to him. Caught, he
    tenses,
holds his armrests in exaggerated
    fear.

MCCLANE No, no, where'd you get
    that idea?

INT: House of MCCLANE
MCCLANE in front of the tv
```

Custom JSON Format

```
[
  [
    { "cat": "Slugline",
      "text": "ON JOHN MCCLANE"},
    { "cat": "Text",
      "text": "mid-thirties, good-
          looking, athletic and
          tired from his trip. He
          sits by the window. His
          relief on landing is
          subtle, but we NOTICE.
          Suddenly, he hears --"},
    { "cat": "Dialogue",
      "Character": "SALESMAN",
      "Character_instructions": "V.
          O.",
      "Line": "You don't like
          flying, do you?",
      "Parentheses": "amused"},
    { "cat": "Text",
      "text": "McClane turns, looks
           at the Babbit clone next
           to him. Caught, he
          tenses, holds his
          armrests in exaggerated
          fear."},
    { "cat": "Dialogue",
      "Character": "MCCLANE",
      "Character_instructions":
          null,
      "Line": "No, no, where'd you
          get that idea?",
      "Parentheses": null}
  ],
  [
    { "cat": "Slugline",
      "text": "INT: House of
          MCCLANE"},
    { "cat": "Text",
      "text": "MCCLANE in front of
          the tv"}
  ]
]
```

Figure 3.1: A side-by-side comparison between the original screenplay and its structured JSON representation. Despite multiple screenwriting conventions being misregarded in the original script, the pipeline still correctly distinguishes character names and tells dialogue and actions of characters apart.

3.2 Metadata Collection & Golden Score Calculation

Comprehensive metadata was extracted for each movie in the dataset from multiple sources to enable thorough dataset analysis and characterization.

A primary challenge in retrieving metadata for scripts is accurately matching each movie with its corresponding metadata across different platforms. In the absence of unique identifiers such as IMDb IDs, movie titles must be used for matching, which presents inherent reliability issues as different movies frequently share similar or identical titles. To address this challenge, a hierarchical matching approach was implemented that prioritizes identifier reliability. The matching strategy follows a three-tier approach. First, when unique identifiers (e.g., IMDb IDs) are available from the screenplay data source, these are used to directly query compatible metadata sources. Second, when metadata sources do not support the same identifier system (e.g., when cross-referencing IMDb with Rotten Tomatoes), title-based searches are performed, with the top search results examined for exact or approximate matches based on release year. The release year ground truth is established using metadata obtained through unique identifier approaches. Third, title-based matching serves as a fallback method only when identifier-based approaches are not feasible, ensuring maximum matching accuracy while maintaining dataset completeness.

Movie ratings constituted the primary metadata component of interest, with four distinct rating systems aggregated for analysis:

- **IMDb:** User-generated ratings on a 1-10 scale from the world's avowed most popular movie database [14]. The platform applies proprietary weighting algorithms to account for potential manipulation and bias.
- **Metacritic:** Weighted average of professional critic scores converted to a 0-100 scale [15]. Staff members assign proprietary weights based on source prestige, review quality, publication frequency, and inclusion on approved critic lists.
- **Rotten Tomatoes Tomatometer:** Percentage of professional critics providing positive reviews (3.5+ stars on a five-star scale) [16]. Critics undergo an open approval process with no score weighting applied.
- **Rotten Tomatoes Popcornmeter:** Percentage of general audience members providing positive reviews (3.5+ stars on a five-star scale) [16]. No weighting or approval process is applied to user ratings.

These ratings were integrated into a composite measure designated as the "Golden Score", calculated as the normalized mean of all available individually normalized ratings. The double normalization process addresses the variable availability of ratings across different systems for individual movies, ensuring consistent scaling.

Supplementary metadata encompassed diverse elements including genres, creative roles (writers, directors, performers), technical attributes (runtime), and evaluative metrics (number of rating entities).

3.3 Perplexity

Perplexity is a probabilistic metric originally developed in information theory and natural language processing that quantifies the uncertainty or surprise a language model experiences when predicting the next word in a sequence. Mathematically, perplexity is defined as the exponential of the cross-entropy between the predicted probability distribution and the actual distribution of words in the text. A perplexity score of n indicates that, on average, the model is as uncertain as if it were randomly choosing between n equally likely next words at each prediction step.

3.3.1 Perplexity Definition for Screenplay Analysis

In the context of screenplay dialogue analysis, perplexity serves as a computational proxy for measuring the linguistic predictability and conventionality of character speech. Low perplexity suggests the text follows familiar patterns and structures that align with the model's training data, while high perplexity indicates dialogue that deviates from expected linguistic norms, potentially signaling unique stylistic choices, unconventional character voices, or innovative narrative techniques.

The core calculation process treats dialogue as sequences of tokens, where each token t_i represents an individual word or linguistic unit in the sequence. For any given token at position i, the notation $t_{1:i-1}$ represents all preceding tokens from position 1 to position $i-1$, providing the contextual foundation for prediction. A pre-trained language model attempts to predict each subsequent token based on this available context, and the model's uncertainty at each prediction point is aggregated to produce perplexity scores that reflect the dialogue's predictability under specific contextual conditions.

For sequences of length n, where n represents the total number of tokens in the sequence and $p(t_i|t_{1:i-1})$ denotes the likelihood of token t_i given the context of all preceding tokens $t_{1:i-1}$, the overall perplexity is calculated as:

$$\text{Perplexity}_{\text{sequence}} = \exp\left(\frac{1}{n}\sum_{i=1}^{n} -\ln(p(t_i|t_{1:i-1}))\right) \tag{3.1}$$

Multiple methodological approaches were employed, systematically varying the contextual information provided to the language model. They included character names, surrounding stage directions, and scene information, to examine how different factors influence these measurements.

Additionally, perplexity can be regarded on the token level as the exponentiated negative log likelihood for that token. For token t_i conditioned on all preceding tokens $t_{1:i-1}$, this relationship is expressed as:

$$\text{Perplexity}(t_i) = e^{-\ln(p(t_i|t_{1:i-1}))} = \frac{1}{p(t_i|t_{1:i-1})} \tag{3.2}$$

3.3.2 Tailored Attention Mechanism for Perplexity Calculations

The attention window, defined as the number of preceding tokens available for predicting the subsequent token, significantly influences perplexity scores. Several analytical methods for evaluating perplexity were not confined to independent scenes but processed entire scripts continuously. Given that screenplay lengths frequently exceed the employed models' context lengths by one to two orders of magnitude, it became necessary to develop sliding window approaches that ensure consistent context length for each token. The solution implemented was a double sliding window approach.

Let w denote the chosen attention window size and l denote the model's maximum context length. Larger attention windows enhance accuracy at the cost of computational efficiency. The outer window segments l sized chunks of the underlying screenplay and passes them to the model for independent processing. To prevent context loss between consecutive chunks, overlapping segments were constructed with exactly w tokens of overlap (where $w < l$). However, this approach introduced an additional challenge: within the remaining $(l-w)$ non overlapping tokens, the first token processed has only w tokens of context, while the final token has access to $(l-1)$ tokens of context. Since varying attention windows would introduce systematic bias into perplexity measurements, this inconsistency required mitigation.

Rather than reinitializing the model for each individual token (which would involve processing thousands of tokens as context to compute a single token's perplexity; an inefficient approach), a custom attention mask was implemented to override the attention layers in the employed models. Standard attention masks within large language models utilize causality masks: lower triangular matrices that ensure tokens cannot access information from subsequent tokens, only from preceding ones. This causality mask was modified to implement a banded lower triangular matrix with width w, ensuring that each token has exactly its w preceeding tokens as context.

To ensure correct model behavior, each input sequence must be prepended with a begin-of-sequence token. A Detail overlooked[1] in some standard implementations [17] but essential for maintaining model correctness across sliding window boundaries. Crucially, the custom attention mask must also be extended with an additional column of ones, allowing unrestricted attention from all tokens to this special token.

[1] A correction addressing this BOS token handling issue has been submitted to the repository maintainers.

Figure 3.2 serves as a simplified illustration of the resulting method.

This implementation ensures that following a warmup period (during which the initial w tokens of a screenplay naturally lack the full context window), each token, regardless of the outer sliding window's segmentation or its position within that window, maintains exactly w tokens of context and perplexity calculation is precise and comparable between scripts.

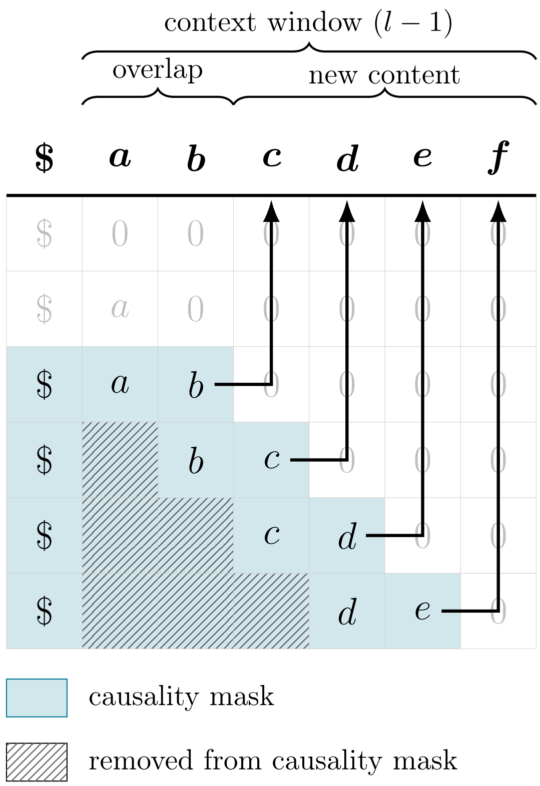

Figure 3.2: One forward pass with l tokens and attention window size 2. \$ marks the begin-of-sequence token. The input is $abcdef$: ab is the overlap from the previous chunk (size w), and $cdef$ are the $l-w-1$ new tokens whose perplexity we compute. Arrows show, for each header letter, the context used to predict it. Petrol cells show the standard causality mask allowing attention only from preceeding tokens; hatched cells are removed from this mask to equalize context length for each prediction.

3.3.3 Perplexity Targeting Tactics

The evaluation methodology employed three key tactics that were empirically varied across different methods to target different properties of dialogue and provide comprehensive analysis. Table A.2 in the appendix details the specific implementation choices and rationale for each of the six final methodological variants.

Masking: This technique involved assigning a value of -100 to specific token labels, effectively excluding them from perplexity calculations while preserving their contextual information for attention computations. This selective masking strategy was primarily based on the classification of original text segments, enabling the model to utilize contextual information without biasing the perplexity scores. For instance, in dialogue segments, character names could be provided to establish speaker context while being masked during perplexity calculation, ensuring that the model knows who speaks next, without measuring the surprise that this person speaks next.

Locality: Analysis operated at three distinct hierarchical levels with varying scope and boundaries: individual utterances examined in isolation, complete scenes processed as discrete units, or entire scripts treated as continuous sequences. When metrics needed to be aggregated from lower to higher levels, weighted averages of component perplexities were computed using token count as the weighting mechanism.

Granularity: Beyond differences in analytical scope, methods also varied in their computational approach to perplexity measurement. Traditional methods adhered to the standard perplexity formula, computing the exponentiated average negative log-likelihood across the relevant token sequences defined by their locality and masking constraints. In contrast, alternative methods calculated token level perplexity by exponentiating the individual negative log-likelihood of each token, providing detailed insight into the distribution of perplexity values within the chosen analytical boundaries.

3.3.4 Comparative Ratio Analysis

Beyond computing individual evaluation metrics with single language models, this study employed combination methods leveraging multiple models simultaneously. The rationale underlying this approach stems from the hypothesis that smaller models exhibit more constrained vocabulary usage and simpler linguistic patterns, while larger models demonstrate broader lexical diversity and complex language generation capabilities (see Figure 3.3). Consequently, the analysis utilized model pairs from the same architectural family but with different parameter counts to capture these scaling effects.

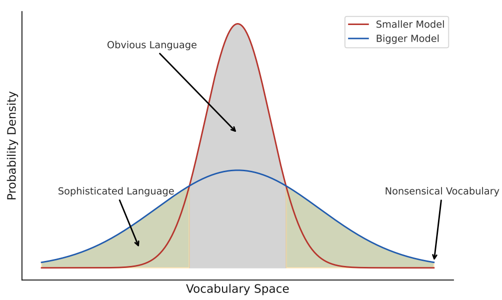

Figure 3.3: Hypothesized vocabulary space coverage of smaller vs. bigger models.

To quantify differential model behavior across parameter scales, perplexity ratios were employed as the primary comparative metric. The logarithmic transformation of these ratios addresses the exponential nature of perplexity values while providing a symmetric measure where positive values indicate higher perplexity for the larger model and negative values indicate the opposite. This approach enables meaningful comparison across the wide dynamic range of perplexity values.

The ratio computation was performed at both token and sequence levels. Let $p_{\mathrm{m}}(t_i \mid t_{1:i-1})$ denote the probability assigned by model $m \in \mathit{large}, \mathit{small}$ to token t_i given the context tokens $t_{1:i-1}$. As defined in equation (3.2), perplexity is the reciprocal of token probability.

Token-Level Analysis. The perplexity ratio for individual tokens was accordingly computed as:

$$R_{\text{token}}(t_i) = \log\left(\frac{\text{Perplexity}_{\text{large}}(t_i \mid t_{1:i-1})}{\text{Perplexity}_{\text{small}}(t_i \mid t_{1:i-1})}\right) = \log\left(\frac{p_{\text{small}}(t_i \mid t_{1:i-1})}{p_{\text{large}}(t_i \mid t_{1:i-1})}\right) \tag{3.3}$$

Sequence-Level Analysis. As such for sequences of length n, the aggregate perplexity ratio was calculated as the logarithm of the geometric mean of individual token ratios. The sequence-level ratio becomes:

$$\begin{aligned} R_{\text{sequence}} &= \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\text{Perplexity}_{\text{large}}(t_i \mid t_{1:i-1})}{\text{Perplexity}_{\text{small}}(t_i \mid t_{1:i-1})}\right) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{p_{\text{small}}(t_i \mid t_{1:i-1})}{p_{\text{large}}(t_i \mid t_{1:i-1})}\right) \\ &= \log\left(\sqrt[n]{\prod_{i=1}^{n}\frac{p_{\text{small}}(t_i \mid t_{1:i-1})}{p_{\text{large}}(t_i \mid t_{1:i-1})}}\right). \end{aligned} \tag{3.4}$$

This formulation enables systematic comparison of model performance while maintaining mathematical consistency between token and sequence level analyses.

Implementation note. While the comparative ratio is conceptually defined through perplexity values, computational efficiency is achieved by operating directly on the negative log-likelihoods (NLLs) that language models natively return. Since perplexity relates to NLL through $\text{Perplexity}_m = \exp(-\ln(p_m)) = \exp(\text{NLL}_m)$, the logarithm of the perplexity ratio collapses to a simple subtraction of NLLs:

$$R = \log\left(\frac{\text{Perplexity}_{\text{large}}}{\text{Perplexity}_{\text{small}}}\right) = \log\left(\frac{\exp(\text{NLL}_{\text{large}})}{\exp(\text{NLL}_{\text{small}})}\right) = \text{NLL}_{\text{large}} - \text{NLL}_{\text{small}} \tag{3.5}$$

This transformation eliminates the need for explicit perplexity computation, reducing both numerical error and computational overhead.

Methodological Extension. The comparative ratio framework generalizes to any pairwise perplexity comparison between models, regardless of their relationship. While originally developed to analyze scaling effects, the same mathematical formalism could be applied to investigate finetuning impacts by comparing two finetuned model variants, with the ratio capturing how specialized training alters predictive distributions.

3.3.5 Distribution Analysis Metrics

Beyond aggregate statistics, the perplexity distribution was characterized using threshold and percentile based metrics to capture fine grained distributional properties.

Thresholds. Token level thresholds quantified the proportion of tokens exceeding predetermined perplexity values. For each script, the number of tokens exceeding thresholds ranging from 1.00025 to 10,000 was computed, and then normalized by total token count into proportions to enable cross script comparison. Lower thresholds identified consistently predictable dialogue, while higher thresholds captured the frequency of highly unpredictable linguistic events. Similarly this was also done for scene level thresholds.

Percentiles. Percentile values were computed at intervals from the 1st to 99th percentile to characterize the full perplexity distribution. Unlike threshold metrics that count surpassing of predefined thresholds, percentiles provide absolute perplexity values at specific distribution points. The 50th percentile corresponds to the median, while extreme percentiles (5th and 95th, indicating that 5 and 95 respectively of perplexity values are below this value) serve as robust estimators of distributional bounds, mitigating the influence of outliers while preserving information about the operational range of perplexity values.

3.4 Finetuning

Finetuning represents a targeted approach to adapting pre-trained language models for specialized tasks by continuing their training on domain specific data. In the context of screenplay dialogue analysis, finetuning enables the development of models with refined representations of linguistic properties that can evaluate dialogue quality. This approach extends beyond the previous perplexity measurements by creating models specifically optimized for dialogue assessment and inference tasks.

Model Selection. Model selection for finetuning required careful consideration of pre-training characteristics. Testing revealed that even state-of-the-art conversational models demonstrated inferior performance compared to base GPT-2, likely due to the alignment processes that optimize these models for dialogue interaction rather than raw linguistic prediction tasks. This finding necessitated employing a pure next token predictors that had not undergone additional finetuning as an AI chatbot or instruction following system.

Training Data Construction. Context window and computational limitations significantly influenced the data preparation strategy. The selected model's context window of 1024 tokens led to training on independent, non overlapping chunks extracted from the original screenplays. To maintain focus on dialogue quality assessment, only dialogue text and character names were retained, with stage directions and other screenplay elements removed. This preprocessing approach maximizes the utilization of available context space for the core linguistic content and ensures that each training sample contains sufficient dialogue material for meaningful pattern learning.

Finetuning Approaches. Two distinct finetuning approaches were implemented exploiting different aspects of the underlying language model capabilities. Both approaches divided their respective datasets into training and validation splits using an 80-20 ratio, providing independent validation data for monitoring model performance when optimizing the finetuning process.

The first approach employs regression based training using the golden score as the target variable. Here, the model receives dialogue snippets as input and is trained to predict the corresponding golden score directly. This treats dialogue quality assessment as a supervised learning problem, where the entire dataset serves as training material with known quality labels. The model should learn to map textual features to numerical quality scores through gradient descent optimization.

The second approach utilizes perplexity based finetuning through dataset segmentation. This divides the original dataset into two subsets containing dialogues classified as either high or low quality based on their golden scores. To ensure clear separation between quality categories and still get sufficient training data, the process selected movies with golden scores positioned at least one standard deviation away from the dataset mean. The language model is then finetuned separately on these curated subsets with the objective of minimizing perplexity for each respective category. This enables the model to develop distinct internal representations for good and poor dialogue, potentially capturing more nuanced linguistic patterns that differentiate quality levels. In addition to the good and bad finetuned models, a third 'dialogue' model was established as an improved baseline, trained on all available dialogue data regardless of quality labels. While not tailored to specific dialogue quality categories, this model can demonstrate strong performance across the dialogue domain, providing a robust reference point for evaluating quality specific models.

Chapter 4

Experiments

4.1 Data Preprocessing and Metadata Integration

4.1.1 Screenplay Data Acquisition

The investigation into screenplay data sources revealed multiple online platforms offering screenplays in various formats (see Table A.1). These platforms employ two primary distribution methods: direct hosting of screenplay content on their own servers, or providing links to external sources such as studio websites where screenplays are made publicly available. However, a significant limitation was observed in the accessibility of these resources, with many links and script pages having become inaccessible over time or restricted behind paywalls.

To expedite the data acquisition process, existing screenplay datasets were evaluated. The Kaggle Movie Scripts Database was identified as the most comprehensive publicly available resource, containing 2,858 movie scripts in plain text format [18]. While this dataset included variants with additional annotations and custom formatting, manual inspection revealed that these processed versions exhibited poor accuracy and were optimized for use cases incompatible with the research objectives of this thesis. Consequently, the 2,858 raw script files from this dataset were selected as the foundation for all subsequent experimental work.

4.1.2 Preprocessing Pipeline Performance

The preprocessing pipeline was applied systematically to the entire corpus, with each stage producing measurable outcomes that inform the final dataset composition.

Chunking. The hierarchical chunking strategy successfully processed the majority of screenplay files. However, 59 scripts (2.1% of the corpus) encountered decoding errors due to character encoding inconsistencies in the source files. An additional 42 scripts (1.5%) were flagged as problematic during the chunking process due to their inability to be segmented into semantically coherent chunks while maintaining the required size constraints. These files were excluded from further processing to maintain data quality.

Classification. The LLM classification stage employed OpenAI's GPT-4o mini model [19] via batch API for economical processing, achieving minimal technical failures when processing the remaining files. One file was inadvertently dropped due to a transient API error during processing, representing negligible loss. The heuristic refinement stage executed without additional data loss, successfully applying domain specific corrections to the LLM classifications.

Title Verification Results. The title verification process revealed significant discrepancies between filename-based titles and screenplay content. The initial LLM model used was GPT-4o mini [19], it failed to extract titles for 292 scripts and identified insufficient alignment between extracted and assigned titles for an additional 206 scripts. These 498 cases (17.4% of the original corpus) required processing by the advanced LLM model, which was GPTs 4.1 model [20]. The comprehensive verification process resulted in 203 title corrections (7.1% of the original corpus), where the content based identification was prioritized over the original filename based labels due to higher reliability.

Quality Filtering Outcomes. The quality filtering stage identified multiple forms of textual corruption that would compromise analytical validity. Detected corruption types included: (1) files with pervasive misspellings rendering text unintelligible, (2) severely corrupted files consisting primarily of random character sequences, and (3) language contamination cases where English dialogue was replaced with transliterated text from other languages (e.g., Hindi text in Roman script: "Mujhe uthe to dus minute ho gaye... lekin aap to aise ek tak dekhe ja rahe"). The filtering process removed 66 scripts (2.3% of the original corpus).

4.1.3 Metadata Integration and Matching

Data Source Selection. Comprehensive metadata collection required integration of multiple rating platforms to ensure broad coverage of evaluative metrics. The OMDb API was selected for its comprehensive coverage of major film databases like IMDb [14], Metacritic [15] and the Rotten Tomatoes Tomatometer (critic ratings), and support for unique identifier queries [21]. Rotten Tomatoes [16] was chosen as a complementary source as it also provided its Popcornmeter (audience ratings) and enjoys widespread adoption. Together, these sources provide coverage of the most significant film rating systems relevant to the research objectives.

Identifier Based Matching. The hierarchical matching strategy prioritized unique identifier approaches where possible. The Kaggle dataset provided IMDb identifiers for a substantial portion of the screenplay corpus, enabling direct metadata retrieval through the OMDb API [21]. This approach achieved successful metadata retrieval for 92.4% of screenplays, representing the highest reliability matching method. An additional 5.6% of screenplays required title-based matching when identifiers were unavailable or invalid, while 2.0% of screenplays could not be matched to any metadata records.

Cross Platform Integration. Rotten Tomatoes metadata collection presented additional challenges due to the absence of API access and incompatibility with IMDb identifier systems. A custom web scraping implementation was developed following ethical scraping practices to get the data from the Rotten Tomatoes Website [16]. The cross platform matching methodology successfully identified corresponding Rotten Tomatoes entries for 97.6% of screenplays, with only 2.4% remaining unmatched.

The final dataset comprises 2,690 screenplays, each with at least one source of metadata. This 94.1% retention of the original corpus stresses the pipelines effectiveness in maintainig data quality while preserving the vast majority of usable content.

4.2 Perplexity

4.2.1 Model Selection

Analysis included Gemma [22],Qwen [23], and Llama [24] models of varying parameter counts. All model checkpoints were obtained via Hugging Face [25]. Evaluations demonstrated consistent behavioral patterns across these architectures.

An ablation study examining the impact of reduced precision on perplexity accuracy revealed negligible degradation of approximately 0.14%. Based on this finding, all models were deployed in half precision mode to optimize computational performance and enable the utilization of larger model variants within available memory constraints.

4.2.2 Choosing Suitable Attention Window

The primary model employed for token wise analysis was Llama-3.2-3B [26], which also served as the large model in comparative analyses. The corresponding small model was Llama-3.2-1B [27], maintaining architectural consistency while providing the parameter count differential necessary for ratio analysis. While these models possess theoretical maximum context lengths exceeding 131,000 tokens, GPU memory limitations constrained the practical maximum context length to 9,000 tokens, as larger contexts would cause the models to allocate excessive memory and subsequently crash.

Given the model's maximum context length constraint of 9,000 tokens, the attention window size was established at four-fifths of this capacity, yielding $w = 7,200$ tokens. This configuration strikes a balance between maintaining sufficient contextual information and computational efficiency, preventing scenarios where an excessively large attention window would result in minimal new content being processed per forward pass relative to the computational overhead required.

Runtime metadata analysis enabled the establishment of a correspondence between token count and screenplay duration. The 7,200 token attention window corresponds to approximately 24 minutes of average runtime. For 95% of the analyzed screenplays, this window still represents a minimum of 10 minutes of runtime, providing substantial narrative context for each prediction step. This temporal scope validates the attention window as sufficient for capturing meaningful contextual relationships within screenplay dialogue structures.

4.2.3 Implementation Details

The custom attention mask implementation described in Section 3.3.2 was deployed across all attention heads to ensure uniform contextual constraints across all perplexity calculations, guaranteeing that each token after warmup maintains exactly 7,200 tokens of preceding context.

Masking procedures required precise positional targeting to exclude specific token categories from some perplexity calculations while preserving their contextual contributions to attention computations. The implementations followed the masking protocols detailed in Section 3.3.3, with particular attention to maintaining consistency across different analytical scopes.

For comparative ratio analysis, both small and large models were executed independently on identical input sequences. Raw probability distributions were collected during inference and subsequently processed offline to compute token and sequence level ratios according to the mathematical formulations presented in Section 3.3.4.

4.3 Finetuning

This section presents the experimental choices of the proposed finetuning methodologies through systematic evaluation of model configurations and training parameters. The empirical findings outlined here guided critical decisions regarding data preprocessing, model selection, configuration optimization, and training strategies for developing effective dialogue quality assessment models.

4.3.1 Model Selection and Training

Base Model and Hardware Constraints. The experimental setup employed GPT 2 Large [28] as the base model, which contains 774 million parameters with a context window of 1024 tokens. This represented the largest GPT 2 model [29] that satisfied the memory constraints. The memory footprint gap between training and inference required approximately three times the memory for finetuning compared to pure perplexity calculations. Initial comparative testing between GPT 2 Large and the base GPT 2 model [30] (124 million parameters) demonstrated that the larger model immediately achieved significantly superior performance, validating the selection of the more computationally demanding architecture for dialogue quality assessment tasks.

Alternative Models. Evaluation of state-of-the-art alternatives included testing the Qwen 3 model with 0.6 billion parameters [31]. However, baseline performance assessment revealed a 13% elevation in perplexity compared to GPT 2 Large across all datasets. Multiple finetuning attempts with various hyperparameter configurations failed to achieve performance parity with GPT 2 Large, leading to the discontinuation of this model for the experimental objectives. These findings reinforced the appropriateness of GPT 2 Large as the foundation model for dialogue quality assessment tasks.

Training Parameter Optimization. Systematic exploration of the hyperparameter space guided the optimization of training configurations for both quality specific and general dialogue models. For the good and bad dialogue models, systematic experimentation established the following optimal parameters:

- Gradient accumulation steps of 8, providing gradient stability while preventing premature convergence
- Learning rate of 1e-5, as higher values led to rapid overfitting
- Weight decay of 0.02 to prevent overfitting while maintaining learning capacity
- Gradient clipping at 1.0 to maintain training stability
- 10% warmup steps combined with cosine learning rate scheduling for optimal learning rate progression
- Training duration of 2 epochs, which proved optimal for achieving convergence without overfitting

The dialogue model, trained on the complete dataset regardless of quality labels, required adjusted parameters to accommodate the increased data volume. The warmup period was reduced to 5%, and the learning rate was decreased to 5e-6 to prevent overfitting on the larger training corpus while maintaining stable convergence characteristics.

4.3.2 Data Loading

Unlike the perplexity evaluation which dynamically utilized complete screenplay files in the custom data format based on the evaluation method, the finetuning approach required additional preprocessing to segment the data into loadable 1024 token chunks suitable for model training. Movies from both quality extremes were selected for this segmentation, with 200 movies from each category allocated for training and 50 movies from each category for validation, resulting in an 80-20 training-validation split. Random assignment was performed once to ensure consistency across experiments. This allocation resulted in approximately 3,000 training chunks for each quality category. The chunking process incorporated padding to maintain clean dialogue boundaries within each chunk of 1024 tokens, ensuring that training samples contained coherent conversational segments rather than arbitrarily truncated text.

A critical preprocessing decision involved the treatment of character names within dialogue sequences. Initial experiments compared model perplexity when character names were masked versus unmasked. Contrary to expectations, the model demonstrated superior predictive capability for character names compared to dialogue content, resulting in higher perplexity when names were masked. This finding indicated that name masking increased model confusion rather than focusing attention on dialogue quality.

Consequently, character names were retained without masking. The final preprocessing approach provided only character names and their corresponding dialogue from the entire script in sequential order, excluding stage directions and sluglines to maximize context window utilization for core linguistic content.

Chapter 5

Results

This chapter presents the empirical findings from analyzing 2,690 movie scripts to investigate relationships between computational linguistic metrics and audience reception. The preprocessing pipeline achieved $> 99\%$ classification accuracy. Movie corpus and rating system profiling revealed dataset viability and importance of normalized score aggregation. Speech metric analysis uncovered small preference for dialogue heavy scripts. The core and in-depth perplexity analysis across six methodological targeting tactics revealed statistically significant but weak correlations between dialogue predictability and screenplay rating, with complex directional effects depending on contextual scope. Not even comparative analysis of differently sized base models improved strength significantly. The finetuning experiments yielded the study's strongest findings: while individual quality specific models failed to correlate with golden scores, comparative ratio analysis between models trained on good versus bad dialogues achieved substantial correlations ($\rho > 0.3$, $r = 0.22$; $p < 0.001$), explaining up to 9% of rank based variance. This differential approach successfully isolated quality specific linguistic patterns that individual perplexity measurements could not capture, validating that screenplay quality manifests through relative rather than absolute predictability levels.

5.1 Preprocessing Pipeline Classification Accuracy

When validating the classification and segmentation of the script content in my custom data format, I sampled 100 classified segments from 60 scripts for each category. After manually examining these 18,000 segments, I obtained very good results. Figure 5.1 shows that even at a 95% confidence level, the assured misclassification rates are at most 1% for all individual categories, while the means are even lower. For the aggregate measure (weighted misclassification rate, see section 3.1.1), I can conclude at a 95% confidence level that it remains below 0.66%.

Category	Mean	95% UCL
Slugline	0.00%	0.05%
Dialogue	0.42%	0.59%
Text	0.80%	1.00%
Weighted	0.50%	0.66%

Figure 5.1: Misclassification rates by category: observed means and 95% upper confidence limits (UCL). Overall (weighted) missclassification rate is below 0.66% on 95% confidence level

5.2 Exploratory Data Analysis and Profiling

5.2.1 Dataset Characteristics

To establish a foundation for subsequent analyses, I conducted an exploratory examination of the dataset's key characteristics. Analysis of release years, shown in Figure 5.2, reveals that the dataset is heavily weighted toward contemporary cinema, with the majority of films released after 1980 and peak representation in the 2000s. Despite this modern focus where most movieis are made, the dataset spans a century of filmmaking, including early cinema.

A notable pattern emerges when examining average golden scores by release year: films released before 1980 consistently achieve higher average ratings than more recent releases. This trend likely reflects what is commonly referred to as "survivorship bias" in online discussions of movie ratings [32], where only the most acclaimed or culturally significant older films remain widely available and rated, while the dataset includes many more "unfiltered" movies from contemporary releases.

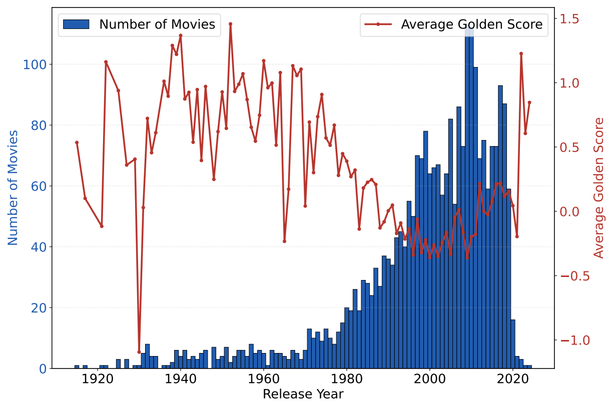

Figure 5.2: Temporal distribution of films and average golden scores by release year. Dataset heavily represents contemporary cinema while spanning a century. Older movies have higher average ratings due to survivorship bias: only acclaimed classics remain widely available, while recent data includes all releases.

The dataset also exhibits interesting patterns in genre distribution (see Appendix Figure A.1). Most films span multiple categories, with an average of 2.5 genres per film (the maximum of three allowed by metadata providers). Notably, average golden scores vary across genres, with certain categories deviating approximately 0.5 standard deviations from the mean, suggesting either inherent quality differences or systematic audience preferences between genre types. These temporal and genre based distributions confirm the dataset's validity for comprehensive analysis, as it encompasses diverse genres and nearly a century of filmmaking rather than being limited to a single genre or decade, providing a robust foundation for subsequent perplexity analyses.

5.2.2 Profiling Rating Systems

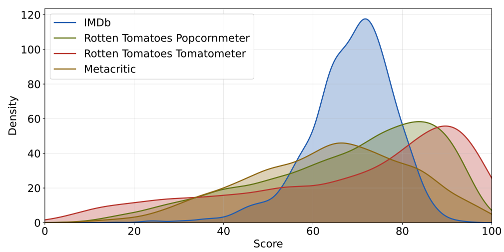

Figure 5.3: Distribution of movie scores across four rating systems with Gaussian smoothing ($\delta = 2$). All ratings are 100-point scales. IMDb shows a normal distribution, while Rotten Tomatoes metrics are positively skewed due to binary voting, the Tomatometer (critics) being most skewed with a 90+ mode yet extreme lows, and the Popcornmeter (audience) more uniform. Metacritic has the flattest distribution with the lowest mode, reflecting its weighted scoring system.

While this work primarily relies on the golden score as an aggregate measure of film quality, I analyzed each rating system's inherent score distribution properties to understand potential biases in the composite measure. The distribution of movie scores across the four rating platforms, shown in Figure 5.3, reveals distinct patterns unique to each platform's methodology, from IMDb's continuous 1-10 user ratings to Rotten Tomatoes' binary threshold systems and Metacritic's weighted critic scores. Despite these fundamental methodological differences, analysis of their statistical metrics (Table 5.1) revealed remarkable convergence: IMDb and both Rotten Tomatoes systems exhibit nearly identical means of approximately 68. Though statistically significant, these small differences suggest either fundamental consistency in how audiences and critics evaluate film quality, or that proprietary weighting processes introduce biases that align disparate rating approaches toward similar central tendencies. This convergence notwithstanding, the distinct score distributions underscore the importance of platform specific normalization when combining ratings, as each platform's inherent characteristics could otherwise introduce systematic biases into aggregate measures. Linear correlation strength between all individual systems and the golden score was above $r = 0.9$.

To examine whether movie quality influences audience engagement across different platforms, I analyzed the relationship between review counts and golden scores (Figure 5.4). The varying correlation strengths reveal distinct engagement patterns: IMDb's strong correlation suggests a self-reinforcing cycle where users

Rating System	N	Mean	Std	Median
IMDb	2517	68.52	9.63	70.00
Rotten Tomatoes Popcornmeter	2535	68.06	19.59	73.00
Rotten Tomatoes Tomatometer	2437	68.27	25.24	76.00
Metacritic	2093	63.38	17.89	65.00

Table 5.1: Descriptive statistics of rating systems. Despite their fundamentally different rating approaches, IMDb and both Rotten Tomatoes scores exhibit near identical means around 68. The differences are statistically significant, but marginal. Suggests either inherent consistency in how audiences and critics rate movies or that proprietary weighting processes introduce biases that align disparate rating approaches toward central tendencies.

preferentially watch and subsequently rate well reviewed films, consistent with its claimed position as the most popular film database. Professional critics' moderate correlation on Rotten Tomatoes likely reflects editorial decisions to prioritize coverage of noteworthy releases and well-received movies. In contrast, the weak correlation for Rotten Tomatoes Audience indicates that user participation and movie selection on this platform is less driven by existing ratings.

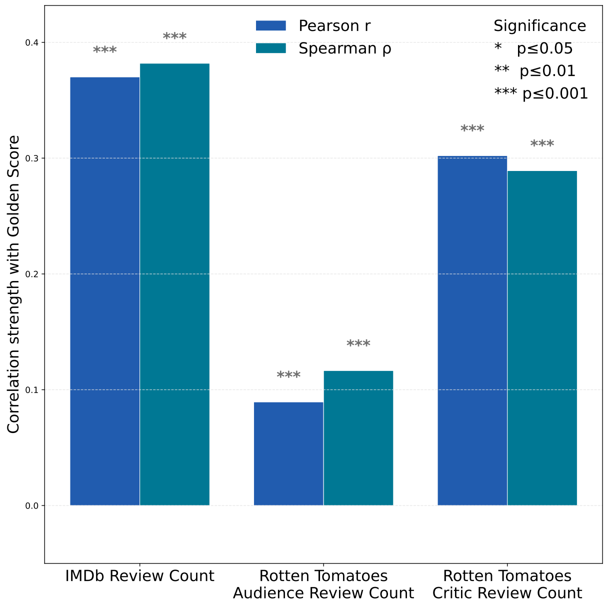

Figure 5.4: Correlation between audience review counts and golden scores across rating platforms. Indicates different underlying mechanism in movie selection between IMDb and Rotten Tomatoes audiences. Higher Spearman correlations for IMDb and RT Audience indicate non linear relationships where review counts increase more dramatically at the highest quality levels.

5.2.3 Interrelation of Descriptive Metrics and Rating

Apart from review engagement patterns, I examined how various movie characteristics correlate with perceived quality as measured by the golden score. Figure 5.5 rudimentary presents the relationships between golden scores and three best performing dimensions: runtime, release year, and USA box office performance.

These patterns collectively suggest that while certain production characteristics associate with quality, no single metric serves as a strong predictor of critical and audience acclaim. Even commercial success remains lacking in predicting quality, underscoring the fundamental challenge of quantifying artistic merit.

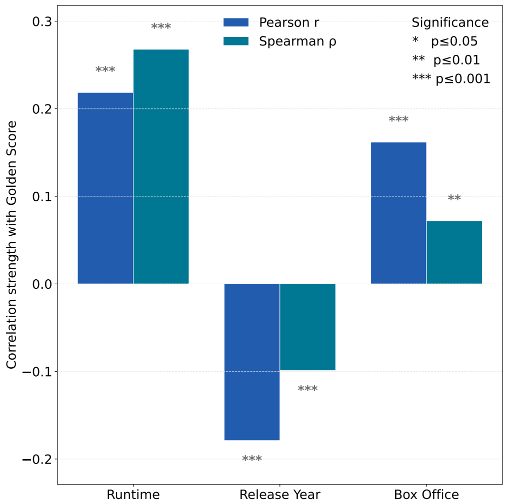

Figure 5.5: Correlation between movie characteristics and golden scores. Runtimes positive correlation indicates longer films tend to achieve higher ratings, possibly reflecting greater narrative complexity or production investment. Release year exhibits significant negative correlation, consistent with the mentioned survivorship bias theory. Box office performance demonstrates positive correlation, with the substantial difference between coefficients weak nature suggesting that commercial success is a poor predictor of quality.

5.2.4 Interrelation of Speech Metrics and Rating

The curated corpus of scripts enabled a comprehensive analysis examining how dialogue characteristics correlate with audience reception. Through systematic examination of character counts per dialogue segment as proxy measures for speech patterns, several associations emerged between linguistic delivery and film ratings.

The dialogue segment proportion analysis of Figure 5.6 indicates that scripts with higher ratios of dialogue to descriptive text are associated with modestly better ratings, suggesting a possible audience preference for character driven narratives over action heavy or visually oriented sequences. Within this dialogue centric approach, longer utterances appear to be associated with improved reception, while variability in utterance length shows an even stronger association with audience engagement. This may indicate that dynamic dialogue patterns, alternating between concise exchanges and extended speeches, tend to create more engaging narrative experiences than uniform delivery styles. Speech distribution patterns offer additional insights into screenplay structures. The positive correlation between inequality in dialogue allocation and ratings in combination with absolute gini indexes around 0.8 appears consistent with traditional narrative approaches where central protagonists dominate speaking time while supporting characters provide complementary voices. The pacing analysis suggests potential limitations on dialogue delivery rates. Despite films operating above everyday conversational speech speeds, excessive verbal density shows a negative association with reception. This finding points to a possible threshold where rapid dialogue may transition from dynamic to overwhelming, potentially affecting comprehension and emotional engagement.

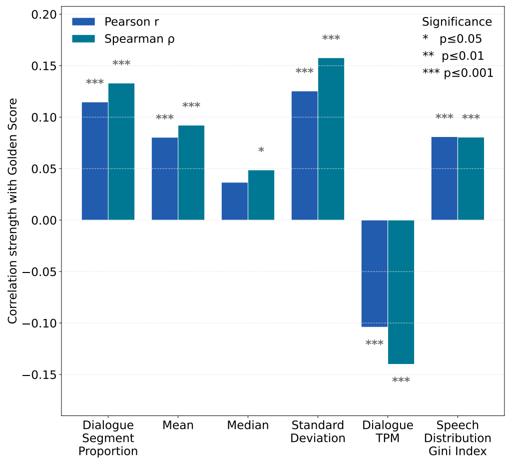

Figure 5.6: Correlation analysis between speech metrics and film ratings. Dialogue heavy scripts are preferred over more text. Mean along with standard deviation support preference of longer but variable utterance lengths. Gini coefficient correlation and values around 0.8 indicates unequal speech distribution receive higher ratings, while dialogue tokens per minute shows deprecating ratings with fast pacing.

5.3 Perplexity Analysis Results

5.3.1 Utterance and Scene Level Analysis (Methods 1-4)

The first four analytical methods examined correlations between language model perplexity and golden scores, showing here three token weighted aggregate metrics across the used three language models (Gemma, Llama, and Qwen). Methods 1 and 4 operated at the utterance level, calculating weighted metrics across individual utterances, while Methods 2 and 3 computed scene-level aggregates.

Analysis of the resulting correlation strenghts as seen exemplatory in Figure 5.7 reveals two consistent patterns. First, the three language models demonstrated remarkably similar performance across all methods, independent of contextual conditions. This consistency persisted throughout the exploratory analysis. Second, correlations were predominantly rank based (Spearman) rather than linear (Pearson), except for Method 4. This pattern suggests weak monotonic relationships that deviate from linearity, potentially attributable to the extensive range of perplexity values and outlier influence that may obscure linear associations while preserving ranking trends. The correlation plots for the remaining methods can be found in the appendix section A.4

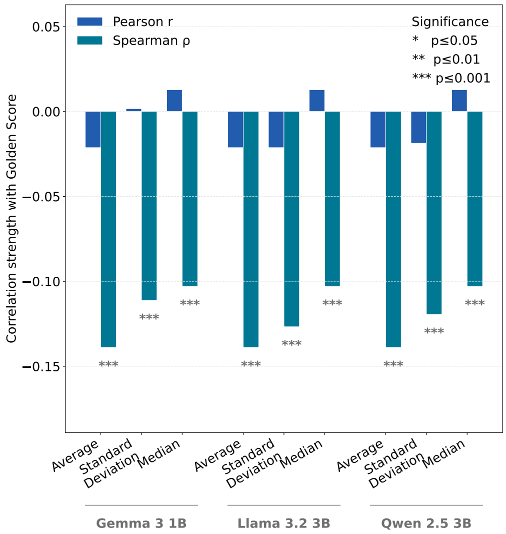

Figure 5.7: Correlation between scene-level dialogue perplexity metrics and golden scores across three language models (Method 2: Scene Dialogue). Highly significant Spearman correlations emerged across all aggregate metrics including the outlier-robust median, exceeding those in Method 1 (Utterances). These results indicate that films with more predictable scene level dialogue (lower perplexity) achieve higher quality rankings when scenes are analyzed in isolation, though effect sizes remain small.

5.3.2 Token Level Perplexity Analysis (Methods 5-6)

Methods 5 and 6 extended the analysis to full script perplexity computation, leveraging token wise metrics to characterize entire screenplays. These analyses employed the Llama 3B model with custom attention windowing as described in section 3.3.2.

Figure 5.8 presents the results from the full contextual analysis (Method 5), where perplexity metrics were computed with complete preceding context for each individual scene. Both complete script perplexity and median token level perplexity demonstrated significant positive correlations with golden scores, supporting the hypothesis that less predictable dialogues receive higher ratings. Notably, scene level weighted averages over the entire script and direct full script perplexity calculations captured different but significant correlation patterns, revealing these metrics reflect complementary aspects of dialogue predictability. While the correlations achieved statistical significance, effect sizes remained small, indicating that perplexity represents only one of multiple factors influencing screenplay quality. Method 6, which analyzed scripts as continuous dialogues without scene boundaries and exhibited weaker performance, is presented in the Appendix (Figure A.5).

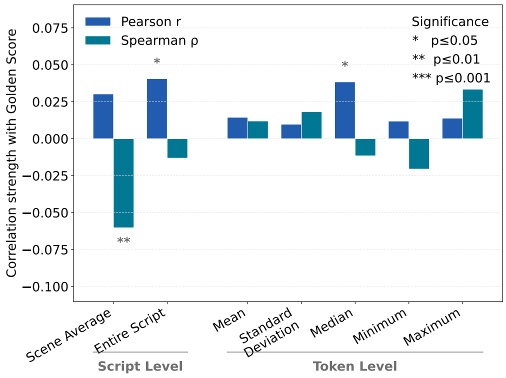

Figure 5.8: Correlation between scene dialogue perplexity with full preceding context and golden scores using the Llama 3.2 3B model (Method 5: Full Contextual Analysis). Both complete script perplexity and median token level perplexity demonstrate significant positive correlations with golden scores ($p < 0.05$), supporting the hypothesis that less predictable dialogues receive higher ratings. Effect size remains small. Script level weighted averages and direct full script perplexity calculations capture different but significant correlation patterns.

5.3.3 Distribution Analysis via Percentiles and Thresholds

Analysis of perplexity distributions through threshold and percentile metrics revealed limited correlational patterns with golden scores, see Figure 5.9. However, upper bound metrics exhibited small but significant negative correlations, indicating that scripts avoiding extreme perplexity values tend to receive higher ratings. Examination of the 95th percentile distribution however revealed just how very small that correlation is.

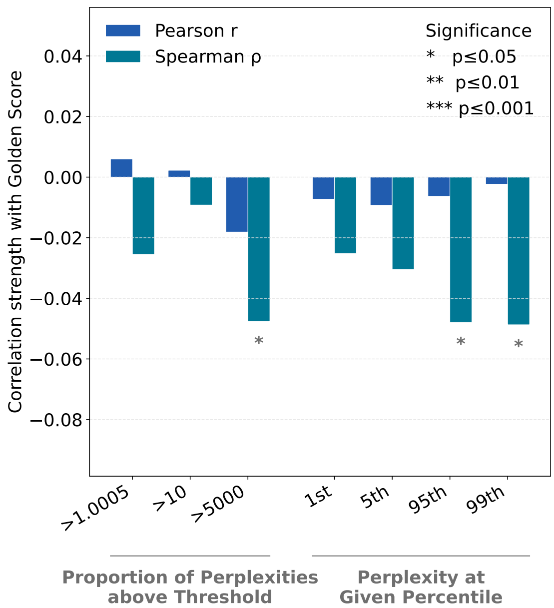

Figure 5.9: Correlation between script dialogue perplexity distribution metrics and golden scores using the Llama 3B model (Method 6). While most distribution metrics show limited correlational patterns, upper bound metrics (e.g., 95th percentile) exhibit small but significant negative correlations, indicating that scripts avoiding extreme perplexity values (eg those with fewer linguistic outliers) tend to receive higher quality ratings.

5.3.4 Comparative Model Analysis

Comparative analysis between Llama models of 3B and 1B parameters using Method 6 (Figure 5.10) yielded significant findings only for median token perplexity. Other metrics showed variations but lacked statistical significance similar to non comparative analyses (Figure A.5) of the same method.

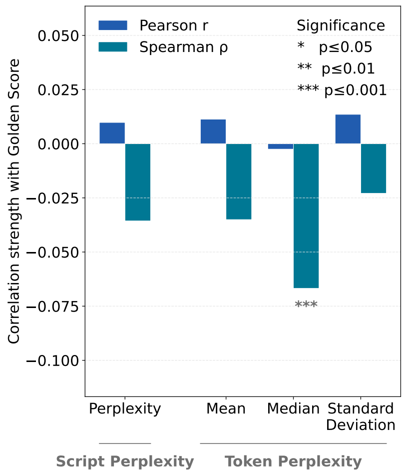

Figure 5.10: Correlation between script dialogue perplexity metrics and golden scores comparing Llama 3B and 1B models (Method 6: Script Dialogue). Only median token perplexity yielded statistically significant spearman correlations of small scope, while other metrics showed variations without statistical significance, consistent with non comparative analyses of the same method.

Extended distributional analysis through threshold and percentile metrics (Figure 5.11) failed to identify strong correlations explaining meaningful variance proportions.

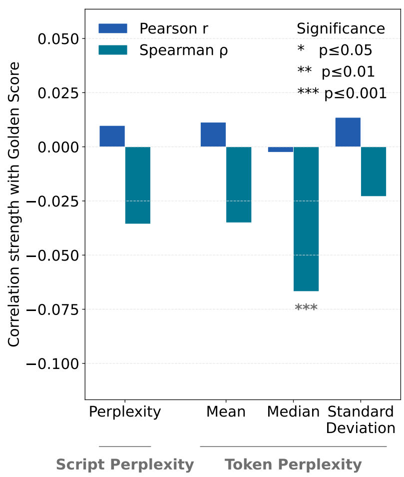

Figure 5.11: Correlation between script dialogue perplexity distribution metrics and golden scores comparing Llama 3B and 1B models (Method 6: Script Dialogue with Percentiles and Thresholds). Even extended distributional analysis failed to identify strong correlations or meaningful variance proportions across all metrics tested.

The distribution of script perplexity ratios (Figure 5.12) concentrates between 0.15 and 0.3. Remember that this ratio as defined in the comparative ratio analysis section (3.3.4) represents the logarithm of perplexity ratios (Equation 3.4). This range corresponds to the larger model experiencing 16−35% lower perplexity than its smaller counterpart across the dialogue dataset. Despite this consistent scaling effect, no clear trend emerged linking perplexity ratios to golden scores, demonstrating that a threefold increase in model parameters failed to capture meaningful script quality differences as measured by viewer ratings.

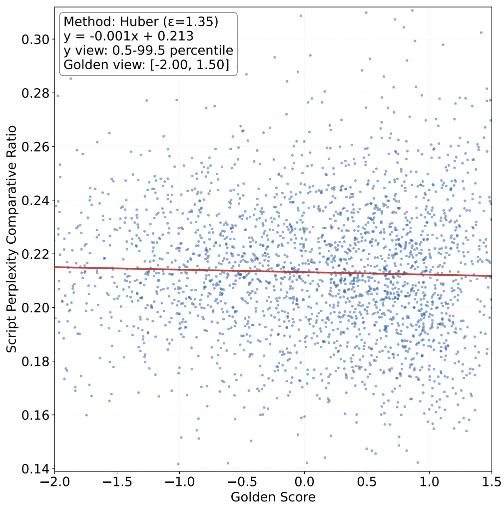

Figure 5.12: Distribution of script perplexity ratios and corresponding average golden scores. The density plot shows comparative script perplexity ratios (logarithm of perplexity ratios between models) concentrated between 0.15-0.3, corresponding to the larger model experiencing 16-35% lower perplexity than its smaller counterpart. Despite this consistent scaling effect across the dialogue dataset, no clear trend links perplexity ratios to golden scores, indicating that a threefold increase in model parameters failed to capture meaningful script quality differences as measured by viewer ratings.

5.3.5 Summary of Perplexity Findings

The comprehensive perplexity analysis across six methodological variants reveals a nuanced but ultimately limited relationship between dialogue predictability and screenplay quality as measured by golden scores. While highly statistically significant correlations emerged across multiple methods, their consistently weak magnitude ($\rho \approx 0.1$) indicates that perplexity-based metrics capture at most 1-2% of the variance in viewer ratings, suggesting that predefined linguistic predictability alone provides insufficient signal for quality assessment.

The analysis uncovered several paradoxical findings that complicate straightforward interpretation. Methods analyzing isolated dialogue units (Methods 1-3) demonstrated that more predictable dialogue correlates with higher quality ratings, while response focused and full context methods (Methods 4 and 5) showed the opposite trend, that less predictable dialogue enhances ratings. This bidirectional relationship suggests that different aspects of predictability may serve distinct narrative functions: conventional dialogue patterns may enhance comprehension and flow, while unexpected responses create memorable moments that elevate viewer engagement.

Three robust patterns emerged across all analytical variants. First, the remarkable consistency across diverse language model architectures (Gemma, Llama, Qwen) validates the reliability of these findings independent of model choice. Second, the predominance of rank order over linear correlations suggests that the relationship between perplexity and quality follows complex, non linear dynamics that resist simple parametric characterization. Third, the comparative analysis demonstrated that even substantial increases in model capacity (three-fold parameter scaling) failed to enhance quality discrimination, indicating that the weak signal is not attributable to model limitations but rather reflects a fundamental disconnect between computational predictability and human quality assessment of artistic content such as movies.

The most consistent finding, that avoiding extreme perplexity values correlates with higher ratings, suggests that screenplay quality may be better characterized by the absence of linguistic failures than by optimization toward any particular perplexity target level. The attenuation of correlations when incorporating narrative context (Method 3 $\rightarrow$ 2) suggests that dialogue predictability correlates more strongly with quality when evaluated in isolation. This weakening may indicate that contextual information normalizes perplexity values across scripts by making all dialogue more predictable when characters and situations are known, thereby reducing the discriminative power of the metric.

5.4 Finetuning

5.4.1 Model Enhancement

Baseline. Pre finetuning evaluation of the GPT 2 Large model revealed baseline validation perplexities across the good and bad dialogue datasets that differed by less than 2.5%, with the bad dialogue dataset exhibiting marginally higher perplexity values. This minimal performance difference demonstrates that the initial GPT 2 Large model was neither finetuned on dialogue quality nor able to distinguish between quality levels clearly or consistently.

Overall Performance. The finetuning process successfully produced three specialized models based on GPT 2 Large: the 'good' model, the 'bad' model, and the 'dialogue' model. Performance evaluation on validation datasets demonstrated substantial improvements over baseline performance. Both quality specific models (good and bad) achieved a 46% reduction in validation perplexity compared to the non finetuned baseline model. The dialogue model demonstrated even superior performance with a 48% reduction in validation perplexity relative to the baseline. Finetuning effectively enhanced model capabilities for dialogue prediction tasks across all tested configurations.

Cross Quality Performance. Contrary to expectations, the finetuned models did not demonstrate optimal performance on their corresponding training quality datasets. Instead, all three finetuned models achieved superior performance (lower perplexity) when evaluated on the bad quality dialogue dataset compared to their performance on good quality dialogues. This phenomenon may be attributed to the inherently simpler linguistic patterns present in lower quality dialogues, which are more readily predictable than the complex and sophisticated language structures characteristic of higher quality screenwriting.

While the good model performed marginally better than the bad model on the good quality dataset, and the bad model showed slightly superior performance on the bad quality dataset, these differences were relatively small at only 1.3%.

Figure 5.13 illustrates the comparative perplexity performance of all models across both validation datasets in detail.

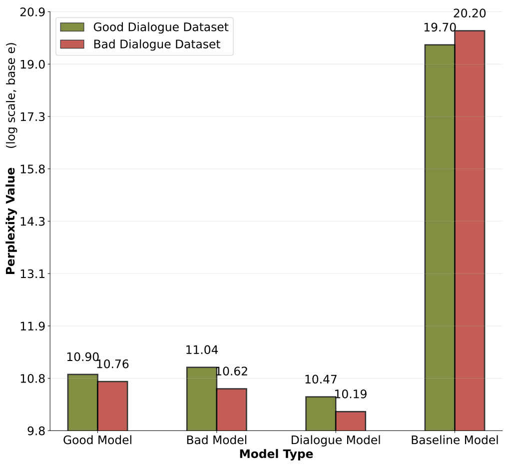

Figure 5.13: Perplexity comparison of model performance on different datasets. The baseline GPT 2 Large model shows minimal distinction between dialogue qualities (<2.5% difference). After finetuning, all models achieve 46-58% perplexity reduction from baseline. Notably, all finetuned models exhibit lower perplexity on bad quality dialogues than good quality dialogues, likely due to simpler linguistic patterns in lower quality text. Quality specific models show only marginal advantages (1.3%) on their corresponding datasets.

5.4.2 Individual Model Analysis

Since the Good and Bad models were trained exclusively on quality distribution extremes, the central portion of the quality spectrum remained unseen during training, providing an unbiased large test set. Both models were evaluated for correlation with golden scores across this held out dataset.

Figure 5.14 demonstrates that neither model exhibits statistically significant correlation with golden scores, consistent with the results from the base Llama 3B model using Method 6 (the most comparable approach, differing primarily in attention window size), see Figure A.5. Beyond the aggregate metrics shown, elaborate additional analyses employing threshold based metrics, percentile distributions, and heatmap visualizations yielded no statistically significant correlations.

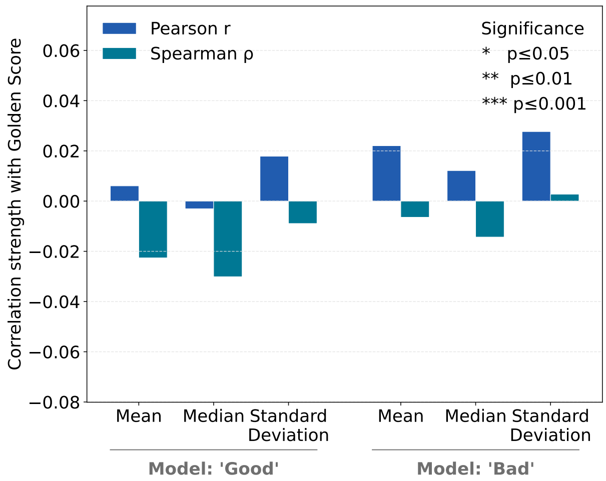

Figure 5.14: Correlation between perplexity metrics and golden scores for good and bad finetuned models. Neither model exhibits statistically significant correlations with golden scores, consistent with base Llama 3B results using the most similar method (M6).

5.4.3 Comparative Model Analysis

Following the methodological framework established in section 3.3.4, comparative ratio analysis was applied to examine perplexity distribution differences between the good and bad finetuned model pair.

While individual finetuned models failed to produce perplexity features that correlate significantly with golden scores, their comparative analysis reveals substantial correlations, see Figure 5.15. Both the script level perplexity ratio and the median token wise ratio demonstrate highly significant correlations with golden scores. The Spearman rank correlation exceeds $\rho = 0.3$ for the script ratio, surpassing the Pearson correlation, which nonetheless captures a moderate linear relationship of $r = 0.22$. The token median ratio maintains slightly weaker but highly significant correlations ($r = 0.1$, $\rho = 0.26$).

These findings indicate that the comparative ratio between the two finetuned models accounts for up to 5% of variance under linear modeling and 9% when considering rank based relationships, substantially exceeding the predictive power of individual model perplexities.

The success of the comparative approach where individual models fail suggests that screenplay quality manifests not through absolute predictability levels but through differential patterns of linguistic surprise. The Good model, trained on highly rated screenplays, has learned to anticipate the stylistic and structural patterns characteristic of successful dialogue, while the Bad model has encoded the linguistic features of poorly received scripts. As demonstrated in the ridgeline plot in Figure 5.16, when a screenplay aligns more closely with patterns learned by the Good model (yielding positive ratio values), the probability density concentrates around higher golden scores. Conversely, screenplays that the Bad model predicts more confidently (negative ratio values) show density distributions skewed toward lower golden scores. The systematic shift in these density patterns across different ratio bins, from a negatively skewed low score distribution for negative ratios to increasingly uniform and then positively skewed distributions, provides compelling visual evidence that screenplay quality emerges from the relative confidence differential between models rather than from any single model's absolute predictions. The scatter plot with a fitted ordinary least squares trendline can be found in the Appendix A.6.

This differential signal emerges only through comparison because both models share the same base architecture and general language understanding. The ratio effectively cancels out their common linguistic knowledge, isolating the quality specific patterns each model has specialized in recognizing.

The magnitude of these correlations, explaining up to 9% of golden score rank based variance, represents a meaningful achievement in computational screenplay analysis. Given the multifaceted nature of screenplay ranking, which encompasses visual storytelling, pacing, character development, and thematic resonance beyond pure dialogue, capturing nearly a tenth of the rank order variance through linguistic analysis alone demonstrates the method's efficacy. This finding validates the hypothesis that finetuning on quality stratified data can encode implicit quality markers that emerge through comparative analysis, even when direct perplexity measurements fail to capture these patterns.

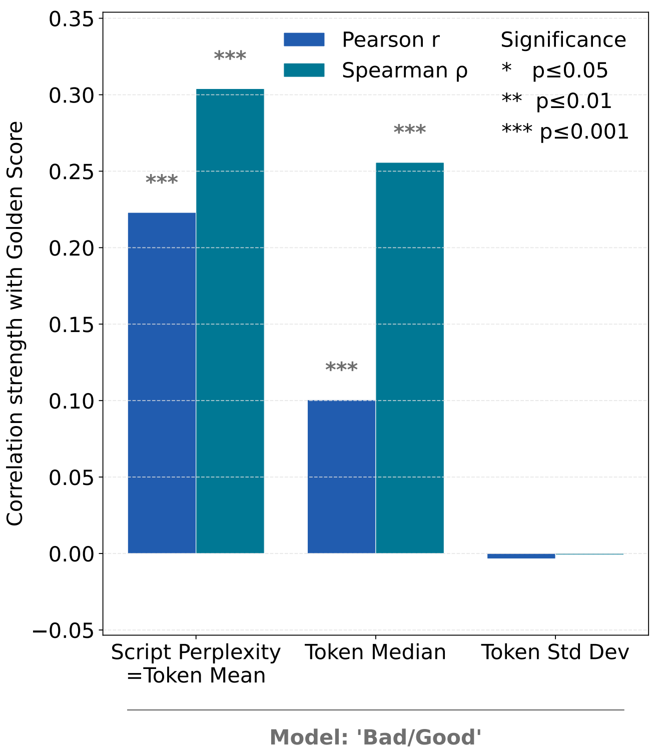

Figure 5.15: Correlation between comparative perplexity ratio metrics and golden scores for finetuned models. Script level perplexity ratio demonstrates highly significant correlations (Spearman $\rho > 0.3$, Pearson $r = 0.22$), with median token wise ratio showing similar patterns. Ratio effectively isolating quality specific patterns by canceling out common linguistic knowledge.

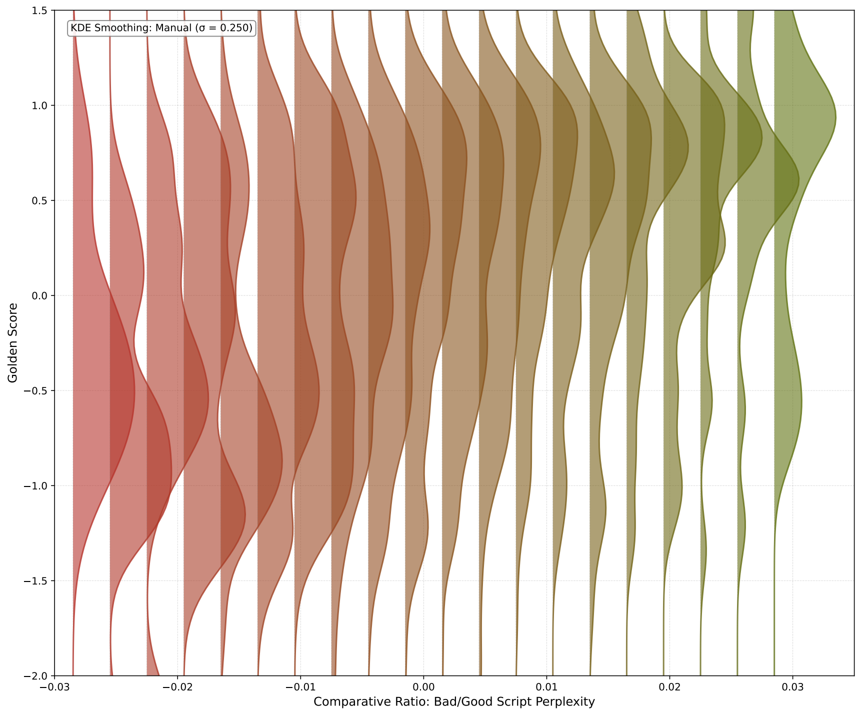

Figure 5.16: Ridgeline plot binned by comparative ratio, showing the density distribution of golden scores for each bin. The plot clearly demonstrates that higher ratios are associated with higher golden scores. The distribution patterns reveal a systematic shift: when the bad model aligns better with the screenplay (negative ratio), golden score density is skewed towards negative scores; when performance is similar the score distribution becomes more uniform; and when the good model aligns better with the screenplay (positive ratio), densities are skewed toward higher golden scores.

5.4.4 Alternative Regression Approach

The supervised regression finetuning approach attempted to directly predict golden scores from dialogue snippets. For the normalized golden score, the baseline performance corresponds to always predicting zero, which represents the variance of the target distribution. However, validation losses consistently remained higher than this baseline variance throughout the entire training process, indicating that the model performed worse than this trivial prediction strategy. The validation loss continued to deteriorate during training, suggesting that the direct score prediction task exceeded the model's learning capacity or required architectural modifications beyond the scope of this implementation. Due to its demonstrated infeasibility and consistent underperformance relative to baseline, this approach was discontinued.

Chapter 6

Conclusions

This thesis has presented a comprehensive computational framework for analyzing screenplay dialogue quality through linguistic predictability metrics, addressing a significant gap in the intersection of natural language processing and creative writing analysis. The investigation has demonstrated that while the relationship between dialogue predictability and perceived quality is more nuanced than initially hypothesized, meaningful patterns can be extracted through carefully designed computational methods.

Summary of Contributions

This work makes several significant contributions to the field of computational narratology and screenplay analysis. First, a robust preprocessing pipeline has been developed and validated that transforms heterogeneous screenplay formats into a unified, well analyzable structure with above 99% classification accuracy. This pipeline, along with the custom JSON data format, provides a foundation for future computational screenplay research, addressing a longstanding barrier to systematic analysis of this medium.

Second, the most comprehensive empirical investigation to date of perplexity patterns in screenplay dialogue has been conducted, analyzing 2,690 scripts through six distinct methodological approaches. This analysis revealed the paradoxical nature of dialogue predictability: isolated utterances benefit from conventional patterns that enhance comprehension, while contextual responses gain impact through unexpected turns. The consistency of these findings across multiple state-of-the-art language models validates the robustness of these observations.

Third, the comparative analysis using finetuned models achieves meaningful improvements in computational quality assessment. By training separate models on high and low quality dialogue subsets and computing their perplexity ratios, the approach captures quality specific linguistic patterns that individual models cannot detect. This method explains up to 9% of ranking variance, demonstrating that screenplay quality manifests through differential patterns rather than absolute complexity levels.

Finally, a comprehensive dataset has been created pairing screenplay texts with multi-source quality ratings, establishing a benchmark for computational approaches to screenplay analysis that future research can build upon.

Theoretical Implications

The findings challenge several assumptions about the nature of dialogue quality and linguistic creativity. The weak correlations observed with traditional perplexity metrics suggest that quality in dramatic dialogue cannot be reduced to simple measures of predictability or complexity. Instead, the comparative ratio results indicate that quality emerges from adherence to learned patterns specific to successful screenwriting patterns that are subtle enough to mostly elude direct measurement of predictability but robust enough to be encoded through targeted finetuning.

The bidirectional relationships observed, where predictability can correlate both positively and negatively with quality depending on context, provide empirical evidence for the complex nature of well perceived dialogue. The analysis suggests that different aspects of predictability serve distinct narrative functions within the screenplay medium.

Furthermore, the success of the comparative approach suggests that quality assessment in creative domains may fundamentally benefit from contrastive analysis. The computational method gains discriminative power by explicitly modeling the differential patterns that distinguish successful from unsuccessful dialogue, mirroring how human evaluation often relies on comparison with known exemplars.

Limitations and Future Directions

While these results represent significant progress in computational screenplay analysis, several limitations must be acknowledged. The maximum explained variance of 9% indicates that dialogue quality is only one component of screenplay success, which also depends vastly on visual storytelling, narrative structure, character development, and thematic resonance, all elements not captured in purely textual analysis. Future work could explore multimodal approaches that for example incorporate visual elements.

Broader Impact

The methodology, combining systematic preprocessing, multi method analysis, and comparative finetuning, provides a template for investigating other creative and artistic domains where quality assessment remains largely subjective. The discriminative capability for assessing dialogue quality has implications for preference learning systems, particularly for generating and validating preference data for reinforcement learning from human feedback (RLHF), potentially improving dialogue generation in AI systems.

The framework established here enables future applications in automated screenplay evaluation and writing assistance tools that illuminate principles of creative writing and can provide objective feedback on dialogue craftsmanship.

Closing Remarks

This thesis has taken initial steps toward understanding the computational signatures of quality in screenplay dialogue, demonstrating that while the relationship between linguistic predictability and perceived quality is complex and paradoxical, meaningful patterns can be extracted through systematic analytical approaches. The comparative analysis between specialized models shows that finetuning encodes implicit knowledge about quality differences that emerges through contrastive evaluation. This reveals that effective dramatic dialogue operates according to subtle patterns that resist simple characterization yet can be computationally detected. By establishing a rigorous framework for screenplay analysis and demonstrating the viability of finetuned perplexity based quality assessment, this work provides a foundation for continued exploration. Future research can build on these computational approaches to creative writing analysis, further illuminating the intersection of natural language processing and artistic expression. Such work will advance both technical capabilities and our understanding of the creative process itself.

Bibliography

[1] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 142–150. [Online]. Available: https://aclanthology.org/P11-1015/

[2] Z. Zhang, Y. Gu, B. A. Plummer, X. Miao, J. Liu, and H. Wang, "Movie genre classification by language augmentation and shot sampling," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7275–7285.

[3] P. Qi, "Movie visual and speech analysis through multi-modal llm for recommendation systems," *IEEE Access*, vol. 12, pp. 145 686–145 702, 2024.

[4] Q. Huang, Y. Xiong, A. Rao, J. Wang, and D. Lin, "Movienet: A holistic dataset for movie understanding," in *European conference on computer vision*. Springer, 2020, pp. 709–727.

[5] S. K. Majhi, P. Sahani, R. Pradhan, B. Acharya, F. Grivokostopoulou, A. Kanavos, and V. C. Gerogiannis, "Comparative analysis of movie recommendation systems using filtering techniques on imdb and rotten tomatoes," in *2024 15th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 2024, pp. 1–8.

[6] X. Song, "Analyzing movie scores on imdb and rotten tomatoes," Personal Web Publication. [Online]. Available: https://rstudio-pubs-static.s3.amazonaws.com/336722_2193716117584b63a2a6ebb837217d85.html

[7] X. Li, C. Tao, J. Gao, and H. Guo, "A review of quality assurance research of dialogue systems," in *2022 IEEE International Conference On Artificial Intelligence Testing (AITest)*, 2022, pp. 87–94.

[8] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in neural information processing systems*, vol. 36, pp. 53 728–53 741, 2023.

[9] L. Mahon and M. Lapata, "Screenwriter: Automatic screenplay generation and movie summarisation," *arXiv preprint arXiv:2410.19809*, 2024.

[10] J. Chen, X. Zhu, C. Yang, C. Shi, Y. Xi, Y. Zhang, J. Wang, J. Pu, R. Zhang, Y. Yang *et al.*, "Hollmwood: Unleashing the creativity of large language models in screenwriting via role playing," *arXiv preprint arXiv:2406.11683*, 2024.

[11] Z. Wang, J. Ma, E. Grinspun, B. Wang, and T. Grossman, "Script2screen: Supporting dialogue scriptwriting with interactive audiovisual generation," *arXiv preprint arXiv:2504.14776*, 2025.

[12] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[13] "fountain.io," 2023, accessed: 21.08.2025. [Online]. Available: https://fountain.io/syntax/

[14] IMDb.com, Inc. (2025) Imdb. [Online]. Available: https://www.imdb.com

[15] FANDOM, INC. (2025) Metacritic. [Online]. Available: https://www.metacritic.com

[16] Fandango (a division of NBCUniversal). (2025) Rotten tomatoes. [Online]. Available: https://www.rottentomatoes.com

[17] Hugging Face, "Perplexity of fixed-length models," https://huggingface.co/docs/transformers/perplexity, 2025, online documentation. Accessed 28 Aug. 2025. Code correction contributed by author via GitHub on same date.

[18] KseniaGur, "Movie scripts corpus," November 2021, licensed under CC BY-NC-SA 4.0. Accessed: 24.03.2025. [Online]. Available: https://www.kaggle.com/datasets/gufukuro/movie-scripts-corpus

[19] OpenAI, Inc. (2024) Gpt-4o mini. Version: gpt-4o-mini-2024-07-18. [Online]. Available: https://platform.openai.com/docs/models/gpt-4o-mini

[20] OpenAI, Inc. (2025) Gpt-4.1. Version: gpt-4.1-2025-04-14. [Online]. Available: https://platform.openai.com/docs/models/gpt-4.1

[21] B. Fritz. (2025) The open movie database. [Online]. Available: http://www.omdbapi.com/

[22] Gemma Team, Google DeepMind, "Gemma: Open models based on gemini research and technology," 2024, arXiv:2403.08295.

[23] Qwen Team, Alibaba Group, "Qwen technical report," 2023, arxiv:2309.16609.

[24] Meta AI, “Llama: Open and efficient foundation language models,” 2023, arXiv:2302.13971.

[25] Hugging Face, Inc., “Hugging face model platform,” https://huggingface.co/, 2025.

[26] Meta AI. (2024) Llama-3.2-3b. Version: meta-llama/Llama-3.2-3B. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.2-3B

[27] Meta AI. (2024) Llama-3.2-1b. Version: meta-llama/Llama-3.2-1B. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.2-1B

[28] OpenAI, Inc. (2019) Gpt-2 large. Version: openai-community/gpt2-large. [Online]. Available: https://huggingface.co/openai-community/gpt2-large

[29] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI Blog*, 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[30] OpenAI, Inc. (2019) Gpt-2. Version: openai-community/gpt2. [Online]. Available: https://huggingface.co/openai-community/gpt2

[31] Qwen Team (Alibaba Cloud). (2025) Qwen3-0.6b. Version: Qwen/Qwen3-0.6B. [Online]. Available: https://huggingface.co/Qwen/Qwen3-0.6B

[32] Reddit users PG22Rated and cox_ph. (2023) Ever wondered which decades of movies have the highest average IMDB rating? Reddit. [Online]. Available: https://www.reddit.com/r/dataisbeautiful/comments/16y27sy/oc_ever_wondered_which_decades_of_movies_have_the/

Appendix A

Supplementary Material

A.1 Potential Screenplay Resources

Website	Format	Amount	Comment
http://www.imsdb.com/	html	1300	
https://gointothestory.blcklst.com/script-download-links-9313356d361c	pdf	100+	
http://www.script-o-rama.com/snazzy/table.html	Mixed: html/pdf	1000+	Several Versions of same script - one can see refinement (eg Draft, 2nd rewrite)
https://www.simplyscripts.com/movie-screenplays.html	Mixed: html/pdf/txt	1000+	Oscar Scripts throughout the years, according to categories; also has transcripts
https://www.awesomefilm.com/	Mixed: html/pdf/txt	100+	many out of date/gone
https://dailyscript.com/movie.html	Mixed: html/pdf	1000+	links to corresponding IMDb ratings
https://sfy.ru/scripts	html	1000+	some out of date/gone
https://www.moviescriptsandscreenplays.com/index.html#top	Mixed: html/pdf/txt	1000+	pre 2008 mostly

Table A.1: Survey of Online Screenplay Distribution Platforms

A.2 Genre Histogram

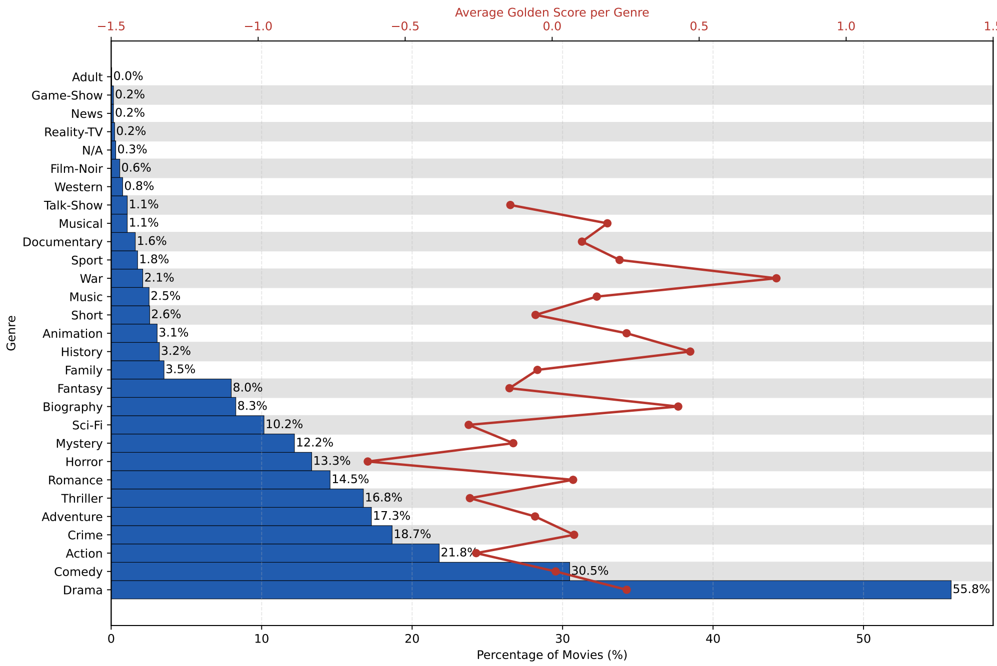

Figure A.1: Genre distribution with corresponding average golden scores. Most films span 2-3 genres (average 2.5 genres per film). Notable quality variations emerge across genres, with certain categories deviating approximately 0.5 standard deviations from the mean golden score, suggesting inherent quality differences or systematic audience preferences between genre types.

A.3 Perplexity Evaluation Methods

Method	Implementation	Rationale
M1: Utterances	No masking, character names or contextual information; traditional perplexity calculation performed independently for each spoken dialogue segment	Focuses on assessing general language quality and linguistic complexity at the utterance level without external influences
M4: Utterance Responses	Only immediate preceding utterance provided as masked context, excluding character names and stage directions; traditional perplexity calculation performed once per spoken dialogue segment using preceding segment	Provides detailed analysis of conversational turn-taking and response appropriateness in immediate dialogue exchanges
M2: Scene Dialogue	No masking, no character names or contextual information; traditional scene-level sliding window perplexity calculation using preceding dialogue as context	Evaluates the naturalness and predictability of dialogue flow and conversational progression within scene boundaries
M3: Scene Dialogue with confined Context	Stage directions masked but contextually preserved, no names; traditional scene-level sliding window perplexity calculation incorporating visual and narrative cues	Assesses dialogue predictability while accounting for visual and narrative contextual information that influences character responses
M5: Scene Dialogue with full preceeding Context	Complete preceding context up to attention window limit provided but masked; stage directions and names masked but contextually preserved; token-level and traditional scene-level sliding window perplexity calculation incorporating visual and narrative cues, current speaker identity; full cross-scene contextual information	Significantly extends M3's scene bounded contextual assessment by incorporating both speaker identity and full preceding context beyond scene boundaries, enabling evaluation of long range dialogue coherence and character consistency across the entire attention window
M6: Script Dialogue	Complete preceding dialogue with names up to exactly attention window limit provided but masked; no stage directions; token level perplexity calculation using exaclty w preceeding name and dialogue tokens (after warmup); ignoring scenes	Standardizes evaluation by eliminating scene level structural variations and stage direction density differences between scripts, emphasizing pure dialogue predictability from preceeding dialogue with consistent maximized context utilization

Table A.2: Overview of Evaluation Methods used in the Perplexity Evaluation without finetuning; Each designed to target distinct linguistic properties.

A.4 Correlation strenghts of the individual methods against golden score

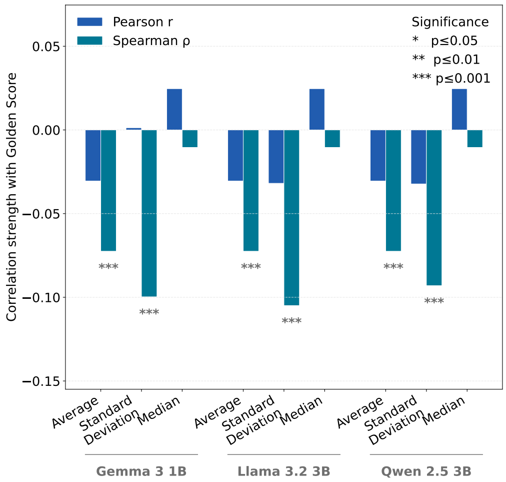

Figure A.2: Correlation between utterance-level perplexity metrics and golden scores across three language models (Method 1: Individual Utterances). While no significant linear correlations emerged, statistically significant negative rank order correlations were observed for mean and standard deviation metrics, indicating that films with lower utterance level perplexity and fewer linguistic anomalies tend to achieve higher quality rankings, though the effect size accounts for only ∼1% of variance.

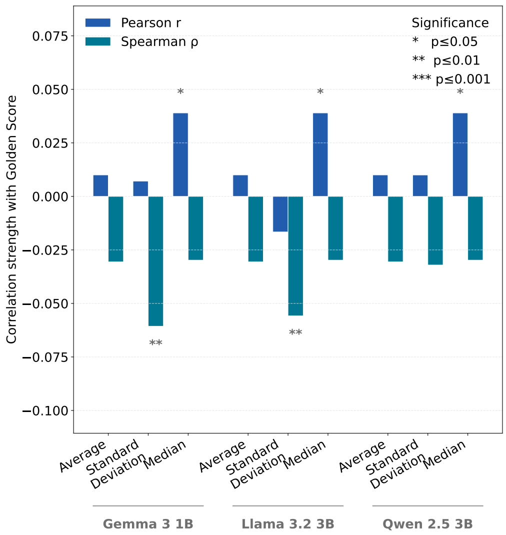

Figure A.3: Correlation between utterance response perplexity metrics and golden scores across three language models (Method 4: Utterance Responses). A weak but significant positive Pearson correlation emerged for median response perplexity ($p < 0.05$), suggesting that scripts featuring less predictable character responses (exemplified by memorable exchanges like “Luke, I am your father”) tend to receive higher ratings, supporting the hypothesis that unexpected dialogue enhances viewer engagement, though the effect remains minimal.

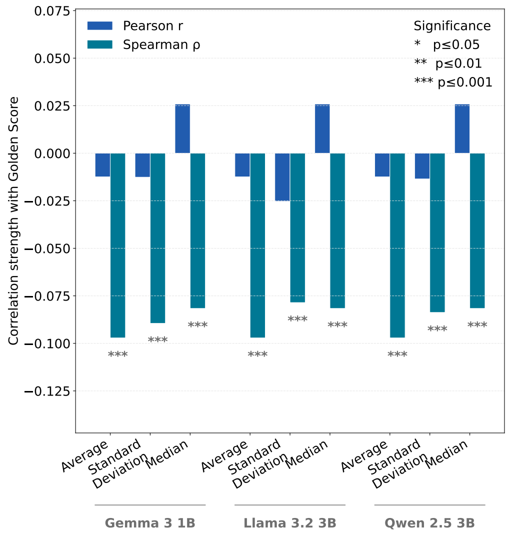

Figure A.4: Correlation between scene dialogue perplexity with contextual information and golden scores across three language models (Method 3: Scene Dialogue with Contextual Information).While patterns mirror Method 2, Spearman correlation strengths are consistently reduced, suggesting that incorporating narrative and visual cues attenuates the relationship between dialogue predictability and quality ratings. This indicates that contextual richness may compensate for conventional dialogue patterns within scenes.

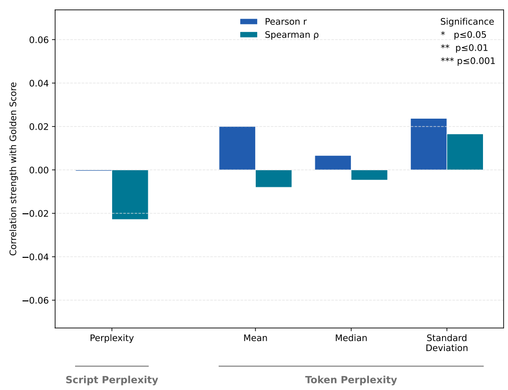

Figure A.5: Correlation between continuous script dialogue perplexity and golden scores using the Llama 3B model (Method 6: Continuous Script Analysis). The figure displays Pearson and Spearman correlations for script level perplexity and token level statistics calculated without scene boundaries. No statistically significant correlations emerged between any perplexity metrics and golden scores.

A.5 Finetune

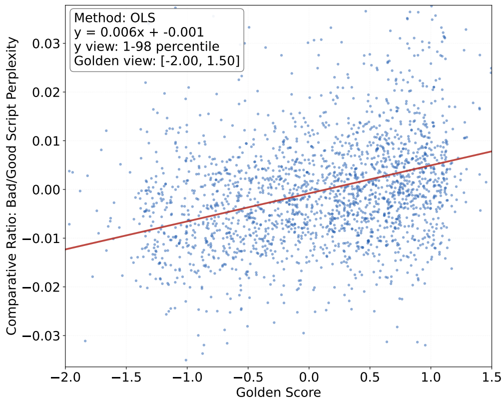

Figure A.6: Scatter plot showing the relationship between golden scores and comparative script perplexity ratio (Good model vs. Bad model). The plot visualizes the relationship between screenplay quality ratings and the comparative perplexity ratio, where one blue dot corresponds to one movie. An ordinary least squares regression line is fitted to the data, showing a well positioned trendline that captures the linear relationship, though the scatter band around the trendline remains relatively wide.