ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed
Computing

Prof. R. Wattenhofer

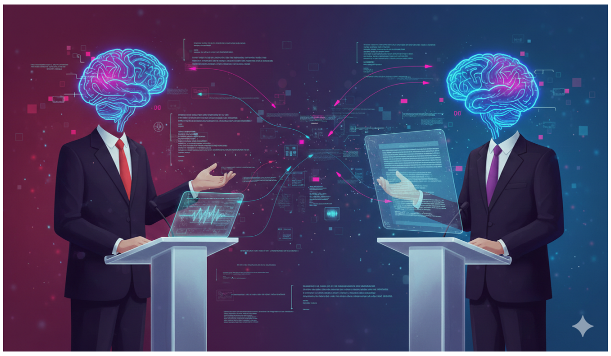
Large Language Models & Political Semantics

Towards data-driven discovery of political semantics

Motivation: The incredible versatility of Large Language Models (LLMs) derived from their extensive pre-training corpora imbues them with an immense breadth of latent knowledge. The fine-grained linguistic understanding of LLMs has altered how researchers approach large-scale natural language datasets, as LLMs can deftly handle polysemy, sarcasm & irony, semantic ambiguity, and other such problems that formerly plagued natural language research. One discipline that contends with huge natural language datasets riddled with polysemantic ambiguity is that of **(quantitative) political science**: from party manifestos to social media discourse and parliamentary debate, natural language mediating political semantics exists in abundance. Mainstream methods such as Wordfish and Wordscore rely on expert annotation and pre-LLM quantitative methods to classify natural language and cannot handle many of the problems mentioned above. Even more bespoke word-embedding or LLM-based methods defer to existing, expert-developed ‘spatial’ categorizations of politics. Therefore, to probe the limits of LLM-based natural language research, this project endeavors to undertake a *data-first* approach instead, distilling from the geometry of learned embeddings identitarian signifiers, political opposition, and emerging polarization.

Method: Based on recent work in probing and steering generative LLMs on political content, we will work towards a more sophisticated mapping of the latent political subspace of LLMs:

1. We will identify attention heads that reliably correlate with and are causally relevant for political content (e.g., through causal mediation analysis);
2. (Subspaces of) the embedding spaces generated by these heads span the ‘political subspace’ of the generative model, so we assemble a dataset of activations of which we know the ground truth prompt;
3. We will attempt to recover from these activations a meaningful underlying structure (PCA, archetypal analysis, linear probing, etc.) and similarly investigate the effects of intervention.



Large Language Models have been trained on large amounts of political discourse freely accessible on the web, likely instilling a fine-grained understanding of its semantics and identifiers.

Contact:

- Frédéric Berdoz : fberdoz@ethz.ch, ETZ G60.1