Investigating Ideological Biases and Censorship Risks in Commercial Large Language Models

Commercial Large Language Models (LLMs) like ChatGPT, CoPilot, and Gemini are now integral to many applications, from chatbots to content generation. Despite their impressive capabilities, concerns about ideological biases and censorship risks persist, especially as LLMs are increasingly used in sensitive domains such as elections and education. This thesis aims to explore these issues using statistical approaches for large-scale testing, providing evidence and preliminary results on the effectiveness of the proposed methods.

Objectives:

- Investigate Ideological Biases: Assess biases in LLMs related to political, cultural, or social ideologies by analyzing their responses to sensitive prompts.
- Evaluate Censorship Risks: Examine risks of censorship, including scenarios where information might be withheld due to external pressures.
- Comparative Analysis: Compare different LLMs (ChatGPT, CoPilot, Gemini) regarding biases and censorship risks across languages, regions, and queries.
- Propose Statistical Approaches: Develop statistical methods for large-scale tests and measurements of observed biases and censorship.

Methodologies:

The study will analyze ideological biases in LLMs by examining text-to-text and text-to-image responses to politically and culturally sensitive prompts, using AI self-conversations to generate data for fine-tuning LLMs. Censorship risks will be assessed by creating scenarios that test information withholding and evaluating the transparency and effectiveness of safeguards. Comparative analysis will involve testing different LLMs with diverse prompts across various contexts. Statistical techniques will be developed and applied to measure and analyze observed biases and censorship.

References

1.Glukhov et al. (2023). "LLM Censorship: A Machine Learning Challenge or a Computer Security Problem?" arXiv:2307.10719

2.Deng & Chen (2023). "Divide-and-Conquer Attack: Harnessing LLM to Bypass Censorship of Text-to-Image Generation Model." arXiv:2312.07130

3. Urman & Makhortykh (2023). "The Silence of the LLMs: Cross-Lingual Analysis of Political Bias in ChatGPT, Google Bard, and Bing Chat."

4. Zhou et al. (2023). "Large Language Model Soft Ideologization via AI-Self-Consciousness." arXiv:2309.16167