Prompt Engineering for Gender-Neutral Educational Content

Umut Yunus Yesildal
Student
Humboldt-Universität zu Berlin
Berlin, Germany
umut.yunus.yesildal@student.hu-berlin.de

Leo S. Rüdian
Supervisor
Humboldt-Universität zu Berlin
Berlin, Germany
ORCID: 0000-0003-3943-4802

***Abstract*—This paper investigates structured prompting strategies for producing gender-neutral text with large language models (LLMs) in educational settings. We synthesise recent research on gender bias, outline an experimental plan for prompt-based mitigation, and discuss implications for inclusive content design.**

***Index Terms*—Prompt engineering, gender bias, large language models, inclusive education**

I. Introduction

Large language models (LLMs) are rapidly being adopted to generate instructional text, yet they often reproduce gender bias. This work focuses on prompt-level interventions aimed at producing gender-neutral educational materials.

II. State of Research

A. Understanding Gender Bias

Gender bias in language refers to systematic and unfair preferences towards a particular gender, reflected through language usage, roles, and stereotypes. Recently, the presence of gender bias within large language models (LLMs) has become an important issue, given their wide-ranging application in educational, professional, and social settings. Such biases, if left unaddressed, can perpetuate stereotypes, inequalities, and reinforce discriminatory practices.

In educational contexts specifically, gender neutrality is crucial. Gender-biased language can negatively affect students' identities, perpetuate stereotypes, and hinder the establishment of inclusive learning environments. Ensuring neutrality in educational resources promotes equity, respect, and inclusion.

B. Key Findings from Recent Research

Recent research highlights several crucial insights regarding gender bias and mitigation strategies in LLMs:

Urchs et al. (2024) examined ChatGPT's responses in German and English. They observed that responses vary significantly depending on gendered prompts (male, female, neutral). German responses particularly showed grammatical challenges and an implicit default to masculine forms. This highlights the inadequacy of unstructured prompting methods for producing reliably neutral responses.

Zeng et al. (2024) focused on debias prompting across various models (GPT, Llama). Their experiments clearly demonstrated that structured prompting methods (zero-shot, few-shot, chain-of-thought) significantly reduce gender bias as measured by standard benchmarks like StereoSet and CrowS-Pairs. Few-shot examples were particularly effective, although they noted a performance-bias mitigation trade-off.

Savoldi et al. (2024) studied gender-neutral translation capabilities of GPT-4. Using systematic prompting approaches, they significantly improved GPT-4's ability to generate gender-neutral translations (English $\rightarrow$ Italian). This study also pointed out the subjectivity involved in evaluating neutrality and acceptability, indicating the necessity of clear guidelines in educational content creation.

You et al. (2024) addressed biases related to the binary categorization of names by LLMs. They found low accuracy (below 40%) for gender-neutral names, demonstrating the importance of recognizing and handling non-binary and gender-diverse identities. Their results underscore the necessity of inclusive representation, particularly important in diverse educational settings.

Zhao et al. (2024) analyzed gender bias across multiple languages, confirming that biases in LLMs extend beyond English. Their findings demonstrated biases in descriptive word choice, pronoun usage, and dialogue contexts. They also highlighted regional variations, emphasizing the importance of culturally and linguistically sensitive prompt design to ensure fairness in education.

C. Synthesis and Educational Relevance

Collectively, these findings underscore the significant impact that prompt engineering can have on mitigating gender biases in LLMs. Structured prompting techniques—such as clearly defined system messages, few-shot exemplars, and multi-step prompts—consistently produce more gender-neutral outputs. Furthermore, considering linguistic diversity and recognizing non-binary identities emerge as critical factors in developing inclusive educational content.

D. Conclusion and Recommendation

Structured prompting techniques, particularly few-shot and clearly defined system-message strategies, emerge as highly

effective methods for ensuring gender neutrality in educational materials generated by LLMs. It is recommended that educators and content creators adopt these structured prompting practices to foster fair, inclusive, and equitable educational environments.

Urchs et al. (2024) examined ChatGPT's responses ... highlighting the inadequacy of unstructured prompting methods [1].

Zeng et al. (2024) demonstrated that structured prompting (zero-shot, few-shot, chain-of-thought) can significantly lower bias scores on benchmarks such as StereoSet and CrowS-Pairs [2].

Savoldi et al. (2024) showed GPT-4 can reach $\sim$70% gender-neutral translations with few-shot prompts, far above baseline MT systems [3].

You et al. (2024) revealed that most LLMs exceed 80% accuracy on binary name–gender prediction yet drop below 40% for gender-neutral names [4].

Zhao et al. (2024) confirmed gender bias across six languages and multiple discourse dimensions, underlining the importance of multilingual prompt design [5].

REFERENCES

[1] S. Urchs, V. Thurner, M. Aßenmacher, C. Heumann, and S. Thiemichen, "How prevalent is gender bias in chatgpt? – exploring german and english chatgpt responses," in *Proceedings of the ACM Web Conference 2024*, 2024, arXiv:2310.03031. [Online]. Available: https://arxiv.org/abs/2310.03031

[2] C. Zeng, M. Chung, and E. Zhou, "Prompting for fairness: Mitigating gender bias in large language models with debias prompting," in *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024, openReview ID: 1096e3651906cf975759252a8a72d8368b182b8a. [Online]. Available: https://openreview.net/pdf?id=1096e3651906cf975759252a8a72d8368b182b8a

[3] B. Savoldi, A. Piergentili, Dennis Fucci, Matteo Negri, and Luisa Bentivogli, "A prompt response to the demand for automatic gender-neutral translation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2024, pp. 256–267. [Online]. Available: https://aclanthology.org/2024.eacl-short.23

[4] Zhiwen You, HaeJin Lee, Shubhanshu Mishra, Sullam Jeoung, Apratim Mishra, Jinseok Kim, and Jana Diesner, "Beyond binary gender labels: Revealing gender biases in llms through gender-neutral name predictions," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/pdf/2407.05271

[5] Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian, "Gender bias in large language models across multiple languages," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/pdf/2403.00277