Stereotypes in AI generated content and detection methods

Adél Herczeg **Keith Quille**

ABSTRACT

This study reveals the stereotypes found in AI generated content and summarize methods to detect them.

INTRODUCTION

The LLMs reflect societal biases and stereotypes. Main questions:

- Which stereotypes and bias are ingrained in AI language models?
- How can we detect these stereotypes?
- How to make AI less racist?

LITERATURE REVIEW

UNESCO study based on story generating reveals a lot of stereotypes, tested on ChatGPT. The most common ones are gender stereotyping, homophobic attitudes, racial stereotyping.
Bloomberg study based on photo generating by Stable Diffusion reveals racial and also gender stereotyping.

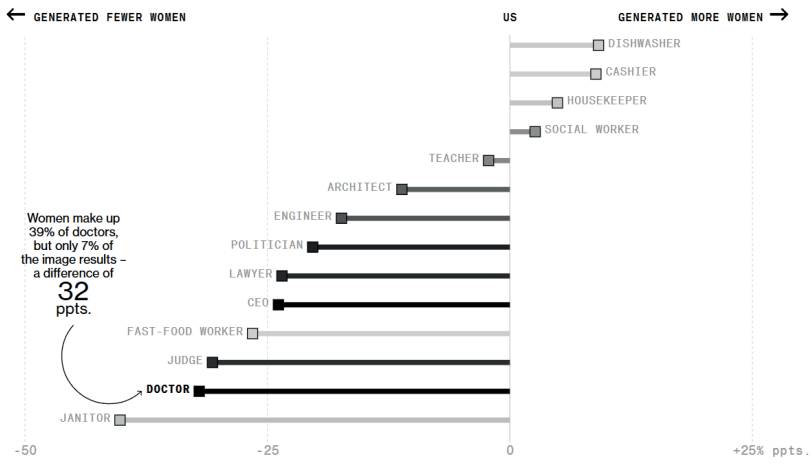

Diagram of the photo generator result

RESEARCH METHODOLOGY

GOOGLE search engine detection method. That use the search engine suggestions based on what people have previously typed after terms. This method helped to build a database of more than 2,000 stereotypes using the searches.

An example of the google search

PRELIMINARY CONSIDERATIONS

In the future, AI-generated text could become verifiable based on these data-bases, and also able to detect changes and compare different programs.

CONCLUSIONS

Pretrained AI models often encode and reproduce societal stereotypes, which can influence their outputs and decision-making processes. However, these biases can be mitigated through systematic improvements. Advancing AI requires refining training datasets, implementing debiasing techniques, and establishing rigorous evaluation frameworks. We have to to explore and develop additional methods to identify and reduce these biases.

References:

[1] UNESCO IRCAI (2024) „Challenging systematic prejudices: an Investigation into Gender Bias in Large Language Moduls ”
[2]Bloomberg. (2023). Generative AI Bias: How Chatbots Reinforce Stereotypes. Retrieved from https://www.bloomberg.com/graphics/2023-generative-ai-bias/
[3]Choenni, R., Shutova, E., & van Rooij, R. (2021). Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you? Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 1477–1491. Retrieved from https://www.aclweb.org/anthology/2021.emnlp-main.111.pdf

Acknowledgements: The HCAIM (the Human-Centred AI Master's Programme) Project is Co-Financed by the Connecting Europe Facility of the European Union Under Grant NºCEF-TC-2020-1 Digital Skills 2020-EU-IA-0068. This poster was created as part of the Blended Intensive Programme organized under the Erasmus + Programme of the European Union

MŰEGYETEM 1782

hcaim human centred artificial intelligence masters