Mitigating Biases in GPT-based Chatbots for Job Recommendations

Tobias Seczer, 12119044

February 11, 2024

1 Introduction

Job recommendation systems powered by large language models (LLMs) have been very popular in recent times. They are a more powerful compared to traditional chatbots and have a lot of use cases in the real world. However, depending on the data and training used for these LLMs, they tend to be biased and unfair. This report aims to provide a detailed exploration of strategies to minimize biases in GPT-based chatbots, acknowledging the significance of fair and unbiased job recommendations. The main focus of is data collection & processing, algorithmic design, and continuous monitoring.

2 Data Collection & Processing

Just like any other AI model, LLMs also need a lot of data during training in order to produce good results. The choice of this data can be crucial in order to mitigate biases towards certain groups. Mitigating such biases begins with the data used to train a model. As a first step, it is needed to use diverse data in order to cover the different groups. Then, these datasets need to be of high quality, such that the data itself is not biased. Biases may occur during the survey of the data, e.g. by having too small of a sample size, false reports of the participants, selection bias, etc. Only when the dataset itself has little to no bias should it be used to train a LLM with. As an additional step, the sampled data can be preprocessed in order to better align the sampled distribution to the real world distribution. These preprocessing steps can include oversampling, undersampling, and data augmentation and are an effective method for mitigating biases in the data. [1, 2]

3 Algorithmic Design

Using high quality, preprocessed data is already a big contribution to mitigating biases. However, to further reduce the possibility of biased responses, the architecture of the

chatbot itself also needs to be tweaked. For example, in order to reduce gender bias, a LLM should learn gender-neutral word embeddings, e.g. by using adversarial learning. Generally, using fairness-aware and adding regularisation can be beneficial. [1, 3]

4 Continuous Monitoring

Given the dynamic nature of language models, continuous monitoring is very important. Real-time monitoring tools should be implemented to assess model outputs in various contexts. These tools can flag potential biases or deviations from desired ethical standards, enabling the model to correct or refuse to answer such prompts. Another useful tool is user feedback. Establishing channels for users to provide feedback on recommendations and experiences can contribute to ongoing model refinement. Feedback loops should be integrated into the development pipeline, ensuring that user insights influence updates and improvements to the chatbot's behaviour. [1, 4]

5 Conclusion

In conclusion, mitigating biases in GPT-based chatbots for job recommendations requires a multifaceted approach. By addressing biases in data collection, refining algorithmic design, and implementing continuous monitoring with user feedback loops, the models can be trained towards creating fair and unbiased job recommendation systems that align with social values.

References

[1] J. Xue, Y.-C. Wang, C. Wei, X. Liu, J Woo, C. -C. J. Kuo, "Bias and Fairness in Chatbots: An Overview", https://arxiv.org/pdf/2309.08836.pdf, 2023

[2] Bias Mitigation, https://fastercapital.com/keyword/bias-mitigation.html

[3] M. B. Zafar, I. Valera, M. G. Rogriguez, K. P Gummadi, "Fairness constraints: Mechanisms for fair classification", in artificial intelligence and statistics, pages 962–970, PMLR, 2017

[4] M. Heikkilä, "How OpenAI is trying to make ChatGPT safer and less biased", https://www.technologyreview.com/2023/02/21/1068893/how-openai-is-trying-to-make-chatgpt-safer-and-less-biased/, 2023