TECHNISCHE UNIVERSITÄT WIEN

Bias Mitigation in GPT-based Job Recommendation Chatbots

Digital Humanism Homework

Author: Konstantinos Vakalopoulos 12223236

February 9, 2024

1 DATA COLLECTION & PROCESSING

Data is a fundamental part of training a machine learning model and more specifically a Large Language Model. Training a LLM depends on huge and varied textual data which affects the performance of the model. For instance, the pre-training process could involve the use of large-scale corpora, web texts, books, conversations, and even code from web sources. Therefore, by including such a large and varied amount of text data during the pre-training phase, it is important to ensure that the resulting model fully captures the complexity of human language. Since, LLMs pre-training depend entirely on the data, in case the quality and content of the data is not sufficient, there will be problems not only in the performance but also in the generalization and ethical considerations of these models [1].

For this reason, it is important to use diverse and representative training datasets to avoid any form of bias [2], to generalize [1] and to improve the model's robustness [3]. In terms of bias, demographic, gender, race bias, etc., if a training dataset is not diverse and representative, the model can learn and preserve biases presented in the data leading to unfair results when the model is applied to real world scenarios [2]. In terms of generalization, diverse training datasets help LLMs generalize better to a wide range of inputs. More specifically, a model trained on a narrow dataset might have difficulty understanding and generating content related to topics or aspects that are not covered in the training data [1]. In terms of robustness, by using a huge variety of linguistic styles, dialects, and contexts in the training data makes the model more robust which leads to effectively communicate with users from different backgrounds and regions. This, simultaneously, is largely related to the acceptance and trust of users when they, directly, interact with the model [3].

As mentioned earlier, a significant problem with Large Language Models is that they exhibit bias on various topics. This is largely attributed to the data used for pre-training these models [1]. Therefore, the most prominent ways to mitigate biases are either to select domain-specific datasets and pre-train the model according to these datasets [4] or to intentionally construct and use biased datasets for the purpose of identifying and analyzing bias [5], [6].

2 Algorithmic Design

An alternative method that can be considered to identify and mitigate bias, apart from the diversity of training data, is the proper design of AI, in general, and more specific LLM algorithmic methods [2]. In particular, an important part of training LLMs is the word embeddings [1]. According to the literature, one way in order to mitigate bias is to further process and review the word embeddings [7], [8]. One approach is to mitigate gender bias in word embeddings while maintaining their basic properties. As a result, the embeddings not only encapsulate bias but also the semantic information from the embeddings help reduce it [7]. Second approach proposed in the literature is the usage of adversial learning. The goal of this method is to reduce bias patterns and features by incorporating a variable for the group of interest and concurrently training both a predictor and an adversary [9].

Therefore, by applying these sophisticated techniques, such as contextual and bias understanding in the embedding layers, LLMs can adeptly handle linguistic nuances and complexities while actively mitigating bias in order to ensure fairness, robustness and truthfulness [10].

3 Continuous Monitoring

Large Language Models are dynamic and context-dependent in nature, as they produce responses based on the input they receive, adjusting their understanding and output according to the previous context. This allows them to produce more coherent and context-relevant information, making them flexible in handling a wide range of linguistic tasks [1]. Therefore, it is important to assess the performance and to monitor repeatedly the influence as well as the responses of LLMs [2] and more specifically Chatbots [11].

These models, which are constantly evolving through updates and further training, require continuous testing to ensure that their results are representative and aligned with the social rules, which are constantly changing, the ethical standards and the accuracy requirements. As the LLMs process and use large amount of data, it is reasonable their understanding of the context might change, requiring regular evaluation and continuous monitoring in order to mitigate biases, errors or potential outcomes. Hence, by implementing robust monitoring mechanisms and precise evaluation protocols, stakeholders and even users, by using the human feedback, can identify and address any emerging issues, promoting trust, reliability and responsible usage of the LLMs in a variety of applications and settings [10].

According to [11], by consistently observing the outputs of LLMs, potential breaches in crucial aspects like privacy, transparency, fairness, and accountability can be detected and averted. This ensures that LLMs and Chatbots can deliver valuable services to users ethically through careful design and testing.

References

[1] Yiheng Liu, Hao He, Tianle Han, Xu Zhang, Mengyuan Liu, Jiaming Tian, Yutong Zhang, Jiaqi Wang, Xiaohui Gao, Tianyang Zhong, Yi Pan, Shaochen Xu, Zihao Wu, Zhengliang Liu, Xin Zhang, Shu Zhang, Xintao Hu, Tuo Zhang, Ning Qiang, Tianming Liu, and Bao Ge. Understanding llms: A comprehensive overview from training to inference, 2024.

[2] Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *First Monday*, November 2023.

[3] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey, 2023.

[4] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. *CoRR*, abs/2004.10964, 2020.

[5] Jiaxu Zhao, Meng Fang, Zijing Shi, Yitong Li, Ling Chen, and Mykola Pechenizkiy. Chbias: Bias evaluation and mitigation of chinese conversational language models, 2023.

[6] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models, 2020.

[7] Sunipa Dev and Jeff Phillips. Attenuating bias in word vectors. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 879–887. PMLR, 16–18 Apr 2019.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[9] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA, 2018. Association for Computing Machinery.

[10] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: a three-layered approach. *AI and Ethics*, May 30 2023.

[11] Junseong Bang, Byung-Tak Lee, and Pangun Park. Examination of ethical principles for llm-based recommendations in conversational ai. In *2023 International Conference on Platform Technology and Service (PlatCon)*, pages 109–113, 2023.