Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/82/20240940

A review of research on training responsible large language models

Yueqi Meng

Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang, 150001, China

mengyq2002@gmail.com

Abstract. In recent years, there has been a growing acceptance of large language models (LLM) as a mainstream method in the field of natural language processing. Consequently, numerous studies have been conducted on this topic. Training responsible Large Language Models have become a prominent subject of research in the past few years. This type of research mainly focuses on the examination of bias, morality and other aspects of LLM. There are certain similarities in the methodologies employed in those studies. This article presents a comprehensive overview of numerous recent investigations, analyzing and categorizing the methodologies employed in these studies, and offering a literature review. This review examines the three perspectives of LLM bias data set construction, bias detection and bias elimination It provides a comparative analysis of the advantages and disadvantages of different methods. After completing relevant evaluations, a comprehensive examination of the research on training responsible LLM is conducted and potential future research directions are proposed in this article.

Keywords: Large Language Models, Responsible AI, Bias Detection, Bias Elimination.

1. Introduction

In recent years, several significant large language models (LLM) based on transformers have emerged, including GPT, BERT, and OPT [1]. These models have demonstrated robust text generation capabilities and can handle multitasks in zero-shot or few-shot form [2]. However, the lack of interpretability of LLMs results in many unexpected answers. The answers provided by the model may contain certain biases related to race, culture, gender, and other factors, which can hinder its ability to accurately match with its users [2]. The occurrence of those biases or toxicities can be attributed to various factors such as training dataset and bias elimination methods [3]. Eliminating such bias and letting LLM to be responsible are therefore challenging tasks.

Nowadays, there is a growing recognition of the importance of developing responsible LLMs. Many studies have been conducted on evaluating and eliminating the bias and toxicity of LLMs [4]. At the same time, many datasets and ways for constructing those datasets have evolved to assess the level, including the architecture of LLMs [5]. Certain methods are designed to optimize for all biases, and others are tailored for special fields, resulting in a reduction of bias in LLMs. This article summarizes recent researches on the building of a responsible LLM. Through a literature review of relevant papers, this review will concentrate on three perspectives, including dataset construction, measurement

DOI: 10.54254/2755-2721/82/20240940

method and elimination method. This article involves summarizing the current pertinent research and doing a comparative analysis of their strengths and weaknesses.

Through in-depth analysis and comparison, readers can understand the advantages and disadvantages of various methods, enabling them to choose more efficient methods to address related issues.

2. Dataset construction

In recent years, many studies have appeared discussing how to construct a data set on bias and toxicity for LLMs. Some of these datasets are based on overall model bias and toxicity, while others are built for specific fields or specific languages. This article summarizes 10 papers about constructing datasets, and finds that the construction methods are mainly divided into three categories. The first and most commonly used method is to generate a part of the dataset through a large language model, and then it is combined with manual processing to build the final dataset. Another part is to manually integrate previous datasets to construct new ones for evaluating the bias or toxicity of LLMs. A third idea for constructing a dataset is to involve users in contributing to the dataset, thereby expanding its coverage to include more minority groups.

2.1. Build a dataset based on a LLM

Many datasets are used in researches that involve LLMs. Most of the data set construction here requires manual assistance. However, there are a small number of research projects where datasets are completely automatically generated in a reinforcement learning-based manner.

The work of Kexin Huang et al., the work of Xu Jing et al. and the work of Ziems and Caleb et al. are carried out by using LLMs to generate responses and manually annotate them [6–8]. This process can be briefly summarized as follows: The first step is to obtain the prompt related to the evaluation. Generally, this acquisition method is given manually or retrieved through the Internet. Afterwards, the LLM generates a response to the corresponding prompt. Finally, the generated response is manually labeled with its bias or toxicity in order to obtain the final dataset.

The data set mentioned above only creates one type of data and may not be flexible enough when evaluating LLMs. The work of Nino Scherrer et al. used manual annotation and GPT-4 generation to design a large-scale survey with 680 high-ambiguity and 687 low-ambiguity moral scenarios to evaluate the bias and toxicity of LLMs in various scenarios [9]. The work of Jiaming Ji et al. constructed two datasets of different sizes, 30k and 330k, respectively, to deal with different situations [10]. At the same time, they evaluated issues such as the bias and toxicity of the model from multiple angles, incorporating 14 dimensions of annotations in each answer.

In addition, there are some works that almost entirely rely on LLMs to generate datasets. The work of Ethan Perez et al. is an example [5]. They first trained an LLM and then used a zero-shot approach to generate some data. Then they use the generated data as an example to make the model create more data in the form of a few-shot. As more data is obtained, the data can be used to fine-tune the model. Finally, the fine-tuned model is further updated using reinforcement learning methods.

2.2. Manually constructed dataset

Except for the methods of LLM-assisted construction of datasets, there are also some studies that use manual construction of datasets. These studies basically involve the integration and transformation of original datasets. There are some works that use the existing moral foundations questionnaire to evaluate the bias of LLMs. The work of Ramezan Aida et al. integrated multiple datasets to evaluate the model's ability to model moral norms across cultures over a variety of topics [4].

In addition, uses will be able to contribute to the datasets through various construction methods. This is a method of manually constructing a dataset on a larger scale. In the work of Eric Michael Smith et al., dozens of people from different groups were invited to contribute relevant data [11]. The model is then evaluated by making relevant templates. This data set better reflects the possible bias of

DOI: 10.54254/2755-2721/82/20240940

minority groups in LLM. The goal of this dataset is to create one that everyone can contribute to. This ensures that the process does not overlook minority groups to the greatest extent possible.

Overall, there are a large number of studies using LLMs to build training datasets. These approaches produce datasets by utilizing the tool, which involves a specific degree of human tweaking of LLMs and data generation, and nearly complete data generation through it. The advantage of this approach is that it simplifies the process of generating data, allowing for the creation of a substantial volume of data. However, because the data is generated by LLM, its reliability is slightly weaker.

At the same time, it is also possible to use large amounts of manually generated datasets, which includes obtaining relevant data from the Internet, and labeling and refining the dataset through users. The quality of the dataset obtained in this way is theoretically higher, but the disadvantages are high cost and labor consumption.

3. Measurement method

There are many ways to evaluate LLM bias. In fact, how to evaluate the bias of LLMs varies depending on the selected dataset. Presently, the predominant approaches for assessing LLM bias can be categorized as training an LLM to evaluate the model's responses, and evaluating the accuracy of the model's responses using biased datasets which is mostly selected. At the same time, there are some other evaluation methods.

Training a model to evaluate the bias of LLM responses is a commonly used method. Long Ouyang et al. used a fine-tuned LLM model to score the model's answers [12]. In addition, in the research of Dan Hendrycks et al., a GPT-3 model was fine-tuned using relevant data sets to score model bias [13]. Similarly, some works also use fine-tuned models for evaluation [5].

Some current work does not use LLM to evaluate the degree of bias in responses but allows the model being tested to complete the selection and reflect the degree of bias based on the selected responses. Such an evaluation method needs some unique datasets, and the form of the data set must be selected. The research by Zhexin Zhang et al. allows LLMs to complete the selection based on the constructed dataset [6]. Finally, the accuracy of the model is used to evaluate the degree of bias. There are also some works which let the model complete a questionnaire, and the model's ability was evaluated based on the scoring rules of the questionnaire. In general, these methods evaluate the degree of biases in LLMs based on the results of multiple-choice questions.

Furthermore, apart from the two aforementioned methods, there are various other evaluation methods. For example, the HONEST method was proposed in the research work of Debora Nozza to evaluate the degree of bias in model responses [14]. In addition, there are several works that employ manual evaluation methods, devoid of any models or questionnaires.

In general, there are many methods to evaluate the bias of LLMs, but employing trained models for scoring is the prevailing method. The benefit of this approach lies in its enhanced precision, albeit at the cost of increased computational resources. At the same time, there is also a scoring method that uses the correct selection rate. This method does not require excessive computing resources, but the effect is not as accurate as the former one. In addition, there are alternative evaluation methods available. The selection of an evaluation method is mostly determined by the characteristics of the dataset and the specific model that requires evaluation.

4. Prejudice elimination

The current focus of studies is on mitigating bias in LLMs. Numerous endeavors have been undertaken to eradicate bias. Typically, the primary techniques employed to mitigate bias are fine-tuning and reinforcement learning, and there are alternative approaches.

Using fine-tuning to eliminate bias is a prevalent practice. According to scholars, a manually written dataset was used to fine-tune the model and reduce the toxicity of LLMs [12]. At the same time, some works focus on specific fields and use data sets in specific fields to fine-tune the model, which alleviates the bias problem of the model. There is a substantial amount of similar work using fine-tuning methods to make LLMs more aligned [4].

DOI: 10.54254/2755-2721/82/20240940

Using reinforcement learning to solve this problem is a frequently utilized method. Previous studies utilized the PPO-ptx algorithm, an enhanced version of the PPO method, to implement reinforcement learning and improve the alignment of the model [12]. The specific approach is to train a scoring function to evaluate the model, build an objective function in RL training based on this function, and perform reinforcement learning. The results show that the model after reinforcement learning has certain improvements in alignment. In fact, there are many works that use such methods, and the performance of the model has also been improved to a certain extent [5].

Furthermore, there are some other ways to mitigate the bias of LLMs. Hao Sun et al. proposed a framework MoralDial to train LLMs to mitigate bias [15]. Research by Emily Dinan et al. uses data processing methods to mitigate LLM bias, in which he proposed a variety of methods for processing data sets and achieved certain results [3]. These studies provide new ideas for mitigating the bias of LLMs.

Generally speaking, the current methods to eliminate LLM bias are basically fine-tuning. For larger LLMs, it is difficult to obtain a large number of high-quality datasets for fine-tuning in the corresponding field. At this time, the reinforcement learning method is a feasible solution. The model with partial fine-tuning and reinforcement learning showed relatively good performance. At the same time, there are some other elimination methods that provide new ideas for subsequent research.

5. Conclusion

This essay examines alignment-related concerns in LLMs and analyzes many studies on educating responsible LLMs. This review examines the process of constructing bias-related datasets and subsequently explores the strategies used to detect LLM bias. Several currently popular LLM bias mitigation methods are introduced and compared in this paper. There are still some shortcomings of this reviewe. The articles collected are only a small part of the outcomes in the field of LLM bias. There are still many other studies in related disciplines that are difficult to cover comprehensively in this paper. At the same time, in the part of eliminating bias, there is a lack of quantitative analysis employed, which is also a part that can be improved in future studies. The growing popularity of LLMs has prompted extensive inquiry into its ethical implications and potential biases. If LLMs cannot have appropriate values, it cannot be truly applied in actual work. Recent researches have shown that LLMs' bias elimination technology has reached a high level of development in a specific field. However, there is still a shortage of research on the overall bias elimination capabilities of LLMs, which will be a significant focus of future research. At the same time, high-value data sets are key to eliminating bias. Therefore, the research of constructing data sets is equally crucial.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners," Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.

[3] E. Dinan, A. Fan, A. Williams, J. Urbanek, D. Kiela, and J. Weston, "Queens are powerful too: Mitigating gender bias in dialogue generation," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), B. Webber,

[4] A. Ramezani and Y. Xu, "Knowledge of cultural moral norms in large language models," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds.Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 428–446. [Online]. Available: https://aclanthology.org/2023.acl-long.26

[5] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving, "Red teaming language models with language models," 2022.

DOI: 10.54254/2755-2721/82/20240940

[6] Z. Zhang, L. Lei, L. Wu, R. Sun, Y. Huang, C. Long, X. Liu, X. Lei, J. Tang, and M. Huang, "Safetybench: Evaluating the safety of large language models with multiple choice questions," 2023.

[7] C. Ziems, J. Yu, Y.-C. Wang, A. Halevy, and D. Yang, "The moral integrity corpus: A benchmark for ethical dialogue systems," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3755–3773. [Online]. Available: https://aclanthology.org/2022.acl-long.261

[8] K. Huang, X. Liu, Q. Guo, T. Sun, J. Sun, Y. Wang, Z. Zhou, Y. Wang, Y. Teng, X. Qiu, Y. Wang, and D. Lin, "Flames: Benchmarking value alignment of chinese large language models," 2023.

[9] N. Scherrer, C. Shi, A. Feder, and D. Blei, "Evaluating the moral beliefs encoded in llms," in Advances in Neural Information Processing Systems, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 51 778–51 809. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/a2cf225ba392627529efef14dc857e22-Paper-Conference.pdf

[10] J. Ji, M. Liu, J. Dai, X. Pan, C. Zhang, C. Bian, C. Zhang, R. Sun, Y. Wang, and Y. Yang, "Beavertails: Towards improved safety alignment of llm via a human-preference dataset," 2023.

[11] E. M. Smith, M. Hall, M. Kambadur, E. Presani, and A. Williams, "" i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset," arXiv preprint arXiv:2205.09209, 2022.

[12] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray et al., "Training language models to follow instructions with human feedback," Advances in neural information processing systems, vol. 35, pp. 27 730–27 744, 2022.

[13] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt, "Aligning ai with shared human values," arXiv preprint arXiv:2008.02275, 2020.

[14] D. Nozza, F. Bianchi, and D. Hovy, "HONEST: Measuring hurtful sentence completion in language models," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 2398–2406. [Online]. Available: https://aclanthology.org/2021.naacl-main.191

[15] H. Sun, Z. Zhang, F. Mi, Y. Wang, W. Liu, J. Cui, B. Wang, Q. Liu, and M. Huang, "Moraldial: A framework to train and evaluate moral dialogue systems via moral discussions," arXiv preprint arXiv:2212.10720, 2022.