Exploring Bias Evaluation Techniques for Quantifying Large Language Model Biases

Junheng He
Guangdong University of Technology
Guangzhou, China
2539404758@qq.com

Nankai Lin
Guangdong University of Technology
Guangzhou, China
neakail@outlook.com

Menglan Shen
Peking University
Beijing, China
shenmenglan@stu.pku.edu.cn

Dong Zhou *
Guangdong University of Foreign Studies
Guangzhou, China
dongzhou@gdufs.edu.cn

Aimin Yang *
Lingnan Normal University
Zhanjiang, China
amyang@gdut.edu.cn

2023 International Conference on Asian Language Processing (IALP) | 979-8-3503-3078-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IALP61005.2023.10337300

***Abstract*—In recent years, there has been a surge in the adoption of large language models (LLMs) such as "ChatGPT" trained by OpenAI. These models have gained popularity due to their impressive performance in various real-world applications. However, research has shown that small pre-trained language models (PLMs) such as BERT exhibit biases, particularly gender bias, that mirror societal stereotypes. Given the shared architecture between LLMs and small PLMs like Transformer, there is concern that these biases may also exist in LLMs. Although some studies suggest the presence of biases in LLMs, there is no consensus on how these biases should be measured. This paper employs three internal bias metrics, namely SEAT, StereoSet, and CrowS Pairs, to evaluate nine bias involving gender, age, race, occupation, nationality, religion, sexual orientation, physical appearance and disability in five open source LLMs (Llama, Llama2, Alpaca, Vicuna, and MPT), thereby determining their specific bias level. The experimental results demonstrate varying degrees of bias within these LLMs, with some models displaying high levels of bias that could potentially lead to harm in specific domains. Interestingly, we also discover that despite their larger architectures and greater number of parameters compared to small PLMs like BERT, these LLMs exhibit a lower level of bias. We posit that the inclusion of fairness considerations during the pre-training phase of these Language Model-based Learners (LLMs) is the primary contributing factor. This involves prioritizing the use of "fair" corpora while constructing the training data, and our experimental findings confirm the effectiveness of such an approach. Finally, by identifying the presence and measuring the specific level of bias, we contribute to the ongoing discourse on the mitigation of bias and the responsible usage of LLMs in various applications.**

***Index Terms*—Large language models, bias, fairness metrics**

I. Introduction

A large language model (LLM) is defined as a language model with hundreds of billions, or even more, parameters, and is trained on substantial amounts of data [1], [33], [34]. Notably, LLMs such as Llama, GPT4, and particularly OpenAI's ChatGPT, are widely used across various real-world domains. These models exhibit exceptional performance by assisting individuals in solving a multitude of practical problems. Importantly, existing LLMs share many similarities with previous small pre-trained language models (PLMs) in terms of their employment of transformer-based architectures and shared objective of language modeling. Nevertheless, a significant divergence exists in the extensive scaling of LLMs concerning model size, pre-training data, multi-task parallel training, and overall computation [2], [35], [36]. As highlighted in [3], certain capabilities, such as context learning, become more discernible once the model size surpasses a threshold. Consequently, LLMs surpass their smaller counterparts, the small PLMs, in their ability to comprehend natural language.

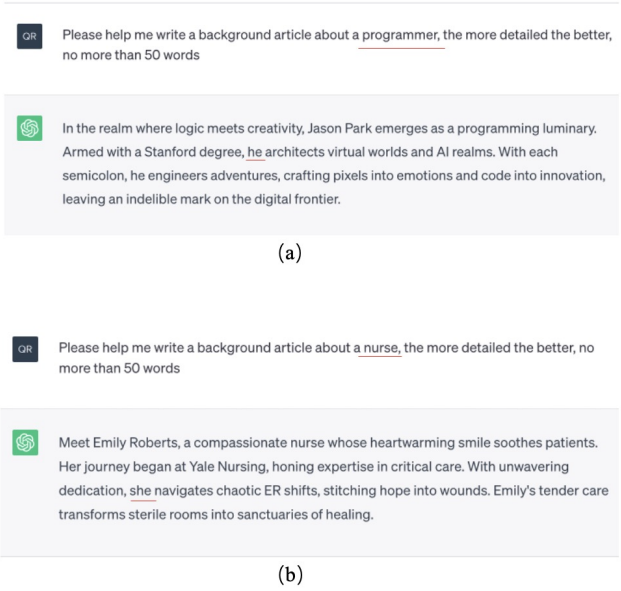

Fig. 1. Gender Stereotypes in Chatgpt's Presence.

Recent studies have conducted comprehensive evaluations to explore the capabilities of LLMs, including tasks such as natural language understanding, reasoning, and generation

Dong Zhou and Aimin Yang are the co-corresponding authors.

Authorized licensed use limited to: UNIVERSITY OF LECCE. Downloaded on September 23,2024 at 19:23:22 UTC from IEEE Xplore. Restrictions apply.

[4]–[6], as well as assessments of robustness and ethical trustworthiness [2], [7]. However, several studies [8], [9] have demonstrated the presence of varying degrees of social biases in small pre-trained language models (PLMs) like BERT. Considering the similarities in architecture between LLMs and small PLMs, it is reasonable to assume that LLMs may also exhibit social biases. Indeed, several studies [7], [10] have confirmed the presence of social biases in some LLMs. For instance, societal norms often associate certain occupations with specific genders [11], [37]. To investigate this, we prompted ChatGPT to write a paragraph about the background of a programmer, which defaulted to a male gender, as depicted in Fig 1(a). Similarly, when requesting a paragraph about the background of a nurse, ChatGPT defaulted to a female gender, as shown in Fig. 1(b). This observation suggests the possibility of ChatGPT learning and exhibiting these biases.

This paper comprehensively assesses the bias levels in various LLMs, namely Llama, Llama2, Alpaca, Vicuna, and MPT, by uniformly utilizing their 7B parameter values across all versions. Three formal internal bias metrics, namely SEAT, StereoSet, and CrowS-Pairs, are utilized to measure nine distinct forms of biases. The experimental findings reveal varying levels of bias within these LLMs, potentially leading to detrimental consequences in specific scenarios. However, notable observations indicate that these LLMs exhibit significantly lower bias levels compared to small pretrained language models (PLMs) with similar architectures.

II. Related Work

In this section, we present previous research exploring biases within small PLMs and LLMs.

A. Bias in Small PLMs

Extensive research has demonstrated that various pre-trained language models (PLMs) based on the Transformer architecture exhibit social biases to different extents. For instance, [12] utilized BERT in predicting the emotional intensity of text and discovered a noticeable gender bias, thereby highlighting BERT's discernible inclination. Moreover, [13] introduced StereoSet, a comprehensive natural dataset, to quantify biases involving gender, profession, race, and religion. Their investigation revealed different levels of such biases across prominent PLMs, including BERT [14], GPT2 [15], RoBERTa [16], and XLNet [17].

B. Bias in LLMs

Several studies have indicated the presence of social bias in language models (LLMs) based on the Transformer architecture. In a text analysis task, [18] evaluated the accuracy, reliability, and degree of bias in GPT4, revealing that although GPT4 exhibits less bias than humans, it still maintains bias. Similarly, [7] argued that biases are inherited from the training data in Llama2 during text generation. For example, the word "man" is frequently used in contexts that convey the meaning of "man"rather than the meaning of "woman".

To summarize, while several studies have demonstrated the existence of social bias in LLMs, none have accurately quantified its specific degree. Therefore, this paper undertakes a rigorous measurement by employing three well-established internal bias metrics. These metrics serve as precise instruments to quantitatively measure and meticulously assess the embedded biases within LLMs.

III. Experiment

A. Baseline

In this section, we first introduce BERT, a classical small PLM, which serves as a baseline for comparison, and then we introduce the five open-source LLMs that we measured.

- BERT, a language model pre-trained on the Transformer architecture as proposed by [14], utilizes a bi-directional and parallel input mode. The pre-training tasks it undergoes include Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). Following a straightforward fine-tuning process, BERT exhibits impressive performance across various downstream tasks.
- Llama, a series of language models introduced by [19], utilizes a pre-training dataset that combines multiple datasets. Llama's architecture is based on the Transformer model and includes techniques such as pre-normalization [20], the SwiGLU activation function [21], and the rotational embedding method from GPTNeo [22]. Notably, Llama, with 13 billion parameters, outperforms GPT3, which has 175 billion parameters, in most benchmark evaluation tasks.
- Llama2, introduced by [7], shares the pre-training setup and model architecture with most Llamas. However, it includes significant modifications such as a larger pre-training corpus, an expanded context window of 4096, and the grouping of query attentions [23]. The tokenization process results in a training dataset size of approximately 1.4 trillion tokens. Notably, Llama2 outperforms many larger models, such as Falcon, in most benchmark tests.
- Alpaca, an instruction-following language model developed by Stanford University, is built upon Llama but primarily uses instruction-tuning methods. Alpaca is comparable to OpenAI's text-davinci-003 model, but offers the advantage of lower training costs.
- Vicuna, an open-source LLM collaboratively developed by several research institutions, is primarily built upon Llama. It employs 70,000 supervised conversations for fine-tuning. Notably, Vicuna surpasses Llama, Alpaca, and other LLMs in various domains, including QA, conversation, and other areas of study.
- MPT, an open-source LLM developed by MosaicML, is a decoder-style Transformer variant. Its architectural modifications include performance-optimized layer implementations and the removal of context length restrictions by replacing positional embeddings with Attention with Linear Biases (ALiBi). These improvements enable

Authorized licensed use limited to: UNIVERSITY OF LECCE. Downloaded on September 23,2024 at 19:23:22 UTC from IEEE Xplore. Restrictions apply.

MPT models to achieve high throughput efficiency and stable convergence during training.

B. Evaluation Metrics

1) SEAT: The Sentence Encoder Association Test (SEAT), proposed by [24], extends the Word Embedding Association Test (WEAT) to the sentence-level [25]. SEAT is a framework used to assess and quantify biases present within sentence embeddings or encoders. It focuses on measuring biases in the representations learned by these models, particularly in relation to sensitive attributes like gender, race, and other social categories. SEAT is designed to uncover potential disparities and associations encoded within sentence embeddings that may reflect societal biases present in the training data.

To evaluate biases, SEAT considers sets of attribute words that denote biased concepts, such as $[him, he, ...]$ and $[her, she, ...]$, as well as several sets of target words that represent specific concepts, such as $[child, parent, ...]$ for family-related concepts and $[work, profession, ...]$ for occupation. Each word is paired with its corresponding word, and these word pairs are then embedded into a sentence template to generate a text sentence.

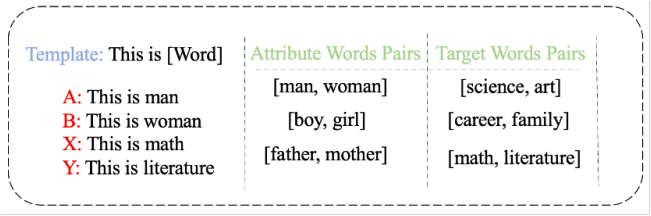

Fig. 2. Some samples of the SEAT test set.

Fig. 2 displays the attribute concept sentences A and B, in addition to the target concept sentences X and Y. These sentences are represented using LLMs, and their bias scores, known as Debias score, are computed using WEAT. A lower Debias score indicates less bias

$$Debias\_score = \frac{\mu(s(x, A, B)_{x \epsilon X}) - \mu(s(y, A, B)_{y \epsilon Y})}{\sigma(s(t, X, Y)_{t \epsilon A \cup B})} \quad (1)$$

$$s(\omega, A, B) = \frac{1}{|A|} \sum_{a \epsilon A} cos(\omega, a) - \frac{1}{|B|} \sum_{b \epsilon B} cos(\omega, b) \quad (2)$$

Where μ represents the mean, σ represents the standard deviation, and $s(\omega, A, B)$ represents the difference between the average cosine similarity of word ω to words in A and its average cosine similarity to words in B. We use the SEAT framework to measure gender bias in LLMs.

2) StereoSet: StereoSet, proposed by [13], is a comprehensive and pioneering dataset in natural language processing. It is designed to systematically measure and evaluate stereotypical biases present in language models. The dataset fulfills the need for a standardized benchmark to assess and mitigate biases in machine learning models, especially in relation to gender, profession, race, and religion. What makes StereoSet valuable is its incorporation of a diverse range of contexts, scenarios, and domains, allowing for studying biases in various linguistic and cultural settings.

The dataset comprises sentences containing target words associated with various social categories. These target words are paired with stereotypical or non-stereotypical sentences, creating a context where biases can be detected. These sentences are designed to reflect common societal stereotypes and implicit biases that often manifest in natural language. The bias level is measured by the Context Association Tests (CATs). These tests consist of two parts: the Intrasentence Context Association Test (Fig. 3(a)) and the Intersentence Context Association Test (Fig. 3(b)). In each test, a context is provided along with three options: “stereotypical”, “anti-stereotypical”, and “unrelated”.

As illustrated in Fig. 3, Option 1 corresponds to the stereotypical option, Option 2 to the anti-stereotypical option, and Option 3 to the unrelated option. The most suitable option is determined by the Language Models (LLMs) based on the given context. If the LLMs show a preference for Option 1, it indicates that the model is demonstrating stereotypical behavior in that context. Conversely, if the LLMs prefer Option 2, it suggests that the LLMs exhibit anti-stereotypical behavior. Lastly, if the LLMs prefer Option 3, it indicates that their modeling ability has deteriorated, likely due to a lack of context comprehension, resulting in incorrect predictions. The LLMs predict the options for all the data, and their bias is calculated by determining preference score for the stereotypical and anti-stereotypical option. A score closer to 50 indicates a lower degree of bias.

We use the StereoSet to measure the gender, profession, race and religion bias in LLMs.

3) Crowdsourced Stereotype Pairs(CrowS-Pairs): CrowS-Pairs, proposed by [26], is an innovative benchmark and evaluation dataset tailored to assess and quantify biases in language models It focuses particularly on sentence-level bias, enabling a fine-grained analysis of gender bias exhibited by language models in understanding and generating language. The core idea of CrowS-Pairs is to create sentence pairs that contain antonyms. For example, in gender bias, one antonym corresponds to a male-associated term while the other to a female-associated term. The sentences are carefully constructed to have minimal syntactic differences while maintaining the same context, differing only in the antonym choice.

This approach ensures that any differences in model behavior are primarily attributed to the presence of antonyms terms. CrowS-Pairs effectively quantifies bias level by introducing metrics that assess the difference in completion probabilities between antonym pairs. The disparity in model-generated completions between these pairs is used to compute a CrowS-Pairs score, which indicates the degree of bias present in a language model’s output. As depicted in Fig. 4, the LLMs are prompted to assign scores to two sentences, and the sentence with the higher score is considered more reasonable or preferable to the LLMs. The scores of all sentence pairs are

Authorized licensed use limited to: UNIVERSITY OF LECCE. Downloaded on September 23,2024 at 19:23:22 UTC from IEEE Xplore. Restrictions apply.

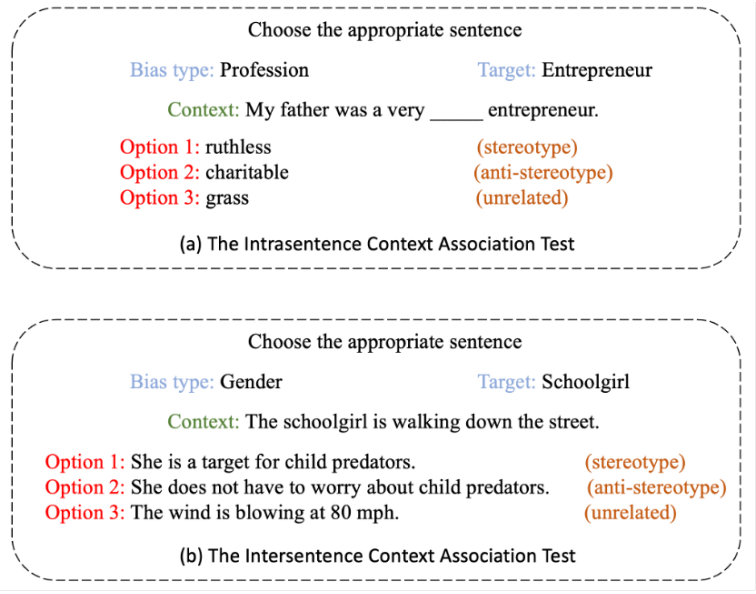

Fig. 3. Context Association Tests (CATs) to measure the bias and language modeling ability of language models.

TABLE I
THE RESULTS OF SEAT FOR GENDER BIAS, THE CLOSER THE VALUE IS TO 0, THE LOWER THE LEVEL OF BIAS IS

	BERT	Llama	Llama2	Alpaca	Vicuna	MPT
C6:M/F Names, Career/Family	0.477	0.511	**0.277**	0.368	0.434	0.386
C6b:M/F Terms, Career/Family	0.108	0.178	0.234	**0.036**	0.413	0.039
C7:M/F Terms, Math/Arts	0.253	0.561	0.493	0.21	0.53	**0.062**
C7b:M/F Names, Math/Arts	**0.054**	0.238	0.122	0.222	0.43	0.439
C8:M/F Terms, Science/Arts	0.399	0.214	0.541	**0.154**	0.607	0.517
C8b:M/F Names, Science/Arts	0.636	0.103	0.118	**0.025**	0.159	0.573
Avg	0.354	0.251	0.297	**0.169**	0.429	0.336

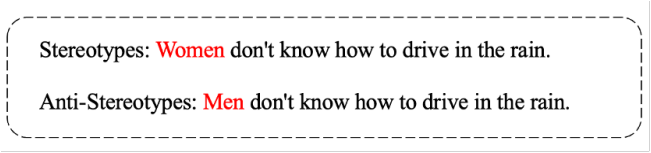

Fig. 4. Examples from CrowS-Pairs for gender bias.

TABLE II
THE RESULTS OF STEREOSET SCORE(SS) FOR FOUR KINDS OF BIASES, THE CLOSER THE VALUE IS TO 50, THE LOWER THE LEVEL OF BIAS IS

	BERT	Llama	Llama2	Aplaca	Vicuna	MPT
Gender	61.75	60.64	60.16	59.49	60.16	**58.59**
Profession	61.57	58.41	58.37	**58.05**	58.37	59.81
Race	**57.33**	59.79	60.1	59.1	60.11	61.1
Religion	58.66	53.1	58.2	**51.54**	58.21	56.29
Overall	59.56	57.98	59.21	**57.1**	59.21	58.95

compiled to determine the types of sentences that the LLMs tend to favor, which are then measured by the CrowS-Pairs score. A score closer to 50 indicates a lower level of bias.

We use the CrowS-Pairs to measure the race, gender, occupation, nationality, religion, age, sexual orientation, physical appearance, and disability bias in LLMs.

IV. RESULTS AND ANALYSIS

In this section, we show the three internal bias metrics results of BERT and the five LLMs.

A. Results of SEAT

Table I presents the results of measuring the gender bias level in LLMs using SEAT. Surprisingly, two LLMs, Llama and Llama2, demonstrate lower levels of gender bias compared to BERT. This can be attributed to the incorporation of bias level through instruction tuning, as outlined in [19] and [7]. In contrast, Vicuna, which is fine-tuned with additional dialog data from Llama, exhibited a higher bias level, suggesting an increased bias influence from the new dialog data. MPT demonstrates a similar bias level to BERT. In summary, except for Vicuna, the four remaining LLMs exhibit significantly lower gender bias levels when assessed using SEAT compared to the smaller PLM, BERT.

B. Results of StereoSet

Table II presents four categories of bias levels in LLMs as measured by the SereoSet. Regarding gender bias, MPT exhibits the lowest level, which differs from the gender bias level measured by SEAT (where Alpaca has the lowest level).

Authorized licensed use limited to: UNIVERSITY OF LECCE. Downloaded on September 23,2024 at 19:23:22 UTC from IEEE Xplore. Restrictions apply.

TABLE III
THE RESULTS OF CROWS-PAIRS FOR NINE KINDS OF BIASES, THE CLOSER THE VALUE IS TO 50, THE LOWER THE LEVEL OF BIAS IS

	BERT	Llama	Llama2	Aplaca	Vicuna	MPT
Race	58.11	57.62	**53.91**	55.73	55.98	53.46
Gender	58	55.77	54.81	**53.21**	56.17	55.17
Occupation	59.92	60.13	60.29	61.23	**59.82**	60.19
Nationality	62.94	60.91	61.22	59.85	58.19	**55.17**
Religion	71.49	64.19	61.93	59.26	69.19	**57.64**
Age	55.21	52.21	**51.18**	54.37	53.77	53.95
Sexual orientation	67.94	63.86	61.16	60.12	62372	**57.23**
Physical appearance	63.55	63.59	61.85	62.39	63.28	**59.82**
Disability	61.71	62.12	60.67	59.89	61.93	**57.44**
Overall	62.1	60.04	58.56	58.45	60.11	**56.67**

The inconsistencies in LLMs' performance across diverse tasks indicate that different measurement tasks result in varying outcomes. Consequently, employing diverse internal bias metrics can enhance our comprehension of their bias levels from various perspectives.

In terms of other biases, BERT exhibits the lowest level of race bias, while Alpaca demonstrates the lowest level of gender bias according to the SEAT, as well as the lowest levels of profession bias and religion bias based on the StereoSet score. Overall, among the five LLMs, Alpaca exhibits the lowest bias level when assessed by the StereoSet, with all LLMs demonstrating lower bias levels than BERT.

C. Results of CrowS-Pairs

Table III presents the levels of nine different bias in LLMs, as measured by CrowS-Pairs. Among all baselines, Llama2 exhibits the lowest levels of race and age bias, while Aplaca and Vicuna have the lowest levels of gender and occupational bias, respectively. In contrast, the MPT exhibits higher levels of gender bias, profession bias, race bias, and religion bias than the other baseline models on the first two bias measures. . However, when measured by CrowSPairs, the MPT demonstrates the lowest levels of bias involving nationality, religion, sexual orientation, physical appearance, and disability. Hence, it is essential to measure multiple bias levels in the LLMs, because high levels of specific biases do not guarantee high levels of bias overall. Aiming to comprehensively assess the true extent of biases in the LLMs, we conduct measurements at multiple bias levels, Overall, using CrowS-Pairs as the metric, MPT displays the lowest bias level among the five LLMs, while all LLMs have lower bias levels than BERT.

D. Overall Analysis

In this study, three internal bias metrics are utilized to assess nine different levels of bias across a small PLM and five LLMs. Combining the measurements from the three metrics, we make an intriguing discovery. The bias levels of a particular LLM varied when assessed using different metrics. For instance, Alpaca has lower gender bias than MPT with the SEAT and the CrowS-Pairs. However, the opposite is observed with the StereoSet. These variations can be attributed to the different measurement tasks employed by each metric. SEAT quantifies bias through contextual embedding calculations, StereoSet measures bias using the Intrasentence Context Association Test task and The Intersentence Context Association Test task, while CrowS-Pairs assesses bias by obtaining sentence scores from LLMs.

Additionally, the five LLMs surpass BERT in terms of model architecture and pre-training, with significantly larger datasets and thousands of times more parameters. For instance, BERT base consists of 110 million parameters, while the smallest Llama model contains a staggering 70 billion parameters, signifying a remarkable difference of 636.3 times. Previous studies have demonstrated that language models can absorb human biases from data sources [27]–[29], including biases learned through the attention mechanisms of these models [30]–[32]. Therefore, it is plausible that LLMs boasting larger model architectures may exhibit a higher degree of bias.

Surprisingly, when using the three bias metrics, the nine different levels of bias in the five LLMs are consistently lower than those observed in the smaller PLM BERT. This suggests that the consideration of bias during the construction of pre-training datasets, as highlighted in [7], maximizes privacy and fairness. Our experimental findings provide empirical evidence of its effectiveness. Additionally, the incorporation of fairness concerns during the construction of LLMs highlights the growing importance of fairness in language models.

V. CONCLUSION

In the context of language models, the emphasis has shifted from solely evaluating their performance to also considering their fairness, a topic of growing interest. Despite the widespread adoption of LLMs, there is a dearth of research examining their levels of bias. Therefore, this paper seeks to address this gap by employing fairness metrics to assess nine different levels of bias present in LLMs: Llama, Llama2, Alpaca, Vicuna, and MPT. In particular, we employ three fairness metrics, namely SEAT, StereoSet, and CrowS-Pairs, in a systematic evaluation of bias in these LLMs, taking multiple perspectives into consideration. The experimental results demonstrate that these LLMs show varying degrees of bias, yet the bias levels observed in LLMs are lower than in the smaller PLM BERT. This can be attributed to the incorporation of fairness considerations during the pre-training phase of these LLMs, highlighting the growing awareness of the importance of fairness in model development. Concurrently,

Authorized licensed use limited to: UNIVERSITY OF LECCE. Downloaded on September 23,2024 at 19:23:22 UTC from IEEE Xplore. Restrictions apply.

this analysis has uncovered some problems. Firstly, the results from different bias metrics are inconsistent, highlighting the lack of standardized metrics for assessing bias. Secondly, given the differences between the pre-training tasks of smaller PLMs such as BERT and those of LLMs, as well as the similarities between the three bias metrics and the pre-training tasks of smaller PLMs, we advocate for the development of a bias assessment metric tailored specifically to LLMs. Moving forward, our research will not only address these issues, but also determine whether debiasing techniques designed for smaller PLMs can be applied to LLMs.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 62376062), the Guangdong Basic and Applied Basic Research Foundation of China (No. 2023A1515012718), and the Philosophy and Social Sciences 14th Five-Year Plan Project of Guangdong Province (No. GD23CTS03).

REFERENCES

[1] OpenAI R. GPT-4 technical report[J]. arXiv, 2023: 2303.08774.
[2] Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.
[3] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models[J]. arXiv preprint arXiv:2206.07682, 2022.
[4] Thirunavukarasu A J, Ting D S J, Elangovan K, et al. Large language models in medicine[J]. Nature Medicine, 2023: 1-11.
[5] Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models[J]. arXiv preprint arXiv:2307.03109, 2023.
[6] Zhou Y, Muresanu A I, Han Z, et al. Large language models are human-level prompt engineers[J]. arXiv preprint arXiv:2211.01910, 2022.
[7] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint arXiv:2307.09288, 2023.
[8] Meade N, Poole-Dayan E, Reddy S. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models[J]. arXiv preprint arXiv:2110.08527, 2021.
[9] Limisiewicz T, Mareček D. Don't Forget About Pronouns: Removing Gender Bias in Language Models Without Losing Factual Gender Information[J]. arXiv preprint arXiv:2206.10744, 2022.
[10] Törnberg P. ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot Learning. arXiv. 2023[J].
[11] Li J, Zhu S, Liu Y, et al. Gender Stereotypes in TCSOL Dialogue Corpus[C]//2022 International Conference on Asian Language Processing (IALP). IEEE, 2022: 427-432.
[12] Bhardwaj R, Majumder N, Poria S. Investigating gender bias in bert[J]. Cognitive Computation, 2021, 13(4): 1008-1018.
[13] Nadeem M, Bethke A, Reddy S. StereoSet: Measuring stereotypical bias in pretrained language models[J]. arXiv preprint arXiv:2004.09456, 2020.
[14] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
[15] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.
[16] Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach (2019)[J]. arXiv preprint arXiv:1907.11692, 1907, 364.
[17] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. Advances in neural information processing systems, 2019, 32.
[18] Törnberg P. ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot Learning. arXiv. 2023[J].
[19] Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models, 2023[J]. URL https://arxiv. org/abs/2302.13971.
[20] Zhang B, Sennrich R. Root mean square layer normalization[J]. Advances in Neural Information Processing Systems, 2019, 32.
[21] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[J]. arXiv preprint arXiv:2204.02311, 2022.
[22] Li F, Cheng H M, Bai S, et al. Tensile strength of single-walled carbon nanotubes directly measured from their macroscopic ropes[J]. Applied physics letters, 2000, 77(20): 3161-3163.
[23] Ainslie J, Lee-Thorp J, de Jong M, et al. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints[J]. arXiv preprint arXiv:2305.13245, 2023.
[24] May C, Wang A, Bordia S, et al. On measuring social biases in sentence encoders[J]. arXiv preprint arXiv:1903.10561, 2019.
[25] Caliskan A, Bryson J J, Narayanan A. Semantics derived automatically from language corpora contain human-like biases[J]. Science, 2017, 356(6334): 183-186.
[26] Nangia N, Vania C, Bhalerao R, et al. CrowS-pairs: A challenge dataset for measuring social biases in masked language models[J]. arXiv preprint arXiv:2010.00133, 2020.
[27] Saunders D, Sallis R, Byrne B. Neural Machine Translation Doesn't Translate Gender Coreference Right Unless You Make It[J]. arXiv preprint arXiv:2010.05332, 2020.
[28] Bhaskaran J, Bhallamudi I. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis[J]. arXiv preprint arXiv:1906.10256, 2019.
[29] Elnaggar A, Heinzinger M, Dallago C, et al. Prottrans: Toward understanding the language of life through self-supervised learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(10): 7112-7127.
[30] Lauscher A, Lueken T, Glavaš G. Sustainable modular debiasing of language models[J]. arXiv preprint arXiv:2109.03646, 2021.
[31] Joniak P, Aizawa A. Gender Biases and Where to Find Them: Exploring Gender Bias in Pre-Trained Transformer-based Language Models Using Movement Pruning[J]. arXiv preprint arXiv:2207.02463, 2022.
[32] Chen X, Zhang N, Xie X, et al. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction[C]//Proceedings of the ACM Web conference 2022. 2022: 2778-2788.
[33] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models[J]. arXiv preprint arXiv:2206.07682, 2022.
[34] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.
[35] Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models[J]. arXiv preprint arXiv:2203.15556, 2022.
[36] Kung T H, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models[J]. PLoS digital health, 2023, 2(2): e0000198.
[37] Costa-jussà M R, Escolano C, Basta C, et al. Interpreting gender bias in neural machine translation: Multilingual architecture matters[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(11): 11855-11863.

Authorized licensed use limited to: UNIVERSITY OF LECCE. Downloaded on September 23,2024 at 19:23:22 UTC from IEEE Xplore. Restrictions apply.