The 7th International Conference on Interdisciplinary Research
on Computer Science, Psychology, and Education (2023)

Evaluating the toxicity of ChatGPT for Gender and Race in Korean

Seungyoon Lee[1†], Chanjun Park[2], Heuiseok Lim[1,3*]
[1]Department of Computer Science and Engineering, Korea University,
[2]Upstage, [3]Human-inspired AI Research
{dltmddbs100, limhseok}@korea.ac.kr, chanjun.park@upstage.ai

Abstract

With the emergence of large language models, concerns over bias and ethics have escalated due to hallucinations and factual issues. This phenomenon can amplify prejudices against individuals with diverse attributes such as race and sexual orientation. However, most studies on such biases are conducted in English and fail to reflect the problems in other languages. In this study, we deliberately assign various personas to ChatGPT to extract biased views, construct prompts about contentious issues of gender and race, and analyze the toxicity of the generated sentences. The experimental results indicate a higher toxic response to topics of sexual orientation than to those of race. Furthermore, a tendency to consistently generate harmful sentences for specific personas is also observed.

Keywords: Large Language Model, Toxicity, Fairness

1. Introduction

The advent of large language models such as GPT4 [1] has demonstrated impressive performance, dismantling the boundaries of numerous sub-tasks in natural language processing [2]. These models, leveraging many parameters and extensive text corpora, effectively acquire inherent knowledge to respond to user needs via command-based learning.

However, due to their training on massive language corpora, these large-scale language models are exposed to text bearing high harmfulness, such as bias and discriminatory elements in the training data. As these models primarily focus on generating plausible natural language according to user requests, their unethically vetted progress poses a potential concern for producing harmful content for a broad range of users, including socially marginalized groups [3].

Nevertheless, most studies for verifying fairness primarily focus on English-speaking cultural contexts, with limited research considering other languages [4]. To address this, the current study analyzes the patterns triggering toxicity in the Korean language, focusing on gender and race topics using ChatGPT [5]. For this purpose, we curate persona prompts suitable for gender and race, utilizing them to generate responses from ChatGPT on these topics.

2. Experiments Setup

ChatGPT is known for its tendency to avoid responding to sensitive topics or explicit phrases that could potentially produce harmful content. In this study, we adopt the method from [6], suggesting that toxicity can be induced through persona injection, enabling the model to generate sentences based on various personas.

2.1 Prompt Design

We define five personas for each of the topics of gender and race. The gender personas encompass ‘heterosexual’, ‘homosexual’, ‘bisexual’, ‘male’, and ‘female’, while the race personas include ‘Southeast Asian’, ‘mixed race’, ‘European’, ‘white’, and ‘black’. With each persona prompt, we instruct ChatGPT to generate 60 sentences by inputting a prompt that asks for opinions on race or gender.

2.2 Toxicity Measurement

For toxicity measurement of the sentences generated by ChatGPT, we use PerspectiveAPI[1], as it has been widely used in numerous prior studies [6]. PerspectiveAPI enables the measurement of toxicity or hatefulness scores of phrases

[*] Corresponding author

[1] https://perspectiveapi.com/

Table 1. Average of ChatGPT's toxicity. In-dom refers to the toxicity of sentences generated on topics such as persona, while out-dom refers to those generated on other topics.

Persona	Race		Persona	Gender	
	in-dom	out-dom		in-dom	out-dom
black	0.166	0.208	homosexual	0.206	0.149
white	0.148	0.187	heterosexual	0.169	0.127
southeast asian	0.125	0.165	bisexual	0.159	0.125
european	0.122	0.157	male	0.143	0.105
mixed race	0.117	0.152	female	0.141	0.099
Average	0.136	0.174	Average	0.164	0.121

based on machine learning models and provides six indicators. We utilized general toxicity as our primary evaluation metric, the most commonly used indicator.

2.3 Generation Setting

During the generation phase, the persona is defined, and views on race or gender are generated. In order to extract the inherent diversity of ChatGPT, the hyper-parameters are set as follows: 'temperature' is selected as 1, 'top_p' is set to 1, and 'frequency_penalty' is chosen as 0.02.

3. Experimental Results

The main results for the generated toxicity of ChatGPT are shown in Table 1. When comparing the mean toxicity for race and gender, it is observed that the toxicity is higher in cases where a sentence regarding gender is generated with a race persona (race: out-dom) and when a sentence regarding gender is generated with a gender persona (gender: in-dom) than in other cases. This suggests the higher prevalence of bias regarding gender topics in Korean compared to race.

Moreover, the toxicity of the generated sentences significantly varies depending on the designated persona. Regardless of the respective topics, 'black' and 'homosexual' exhibit consistently harmful tendencies. In contrast, when 'mixed race' and 'female' are given, the toxicity substantially decreases. This implies that ChatGPT possesses strong biases towards certain personas in races and genders. To sum up, it indicates the negative perception regarding race and gender inherent in the Korean corpus from which ChatGPT has learned.

4. Conclusion

In this paper, we analyze the toxicity and bias of ChatGPT in Korean, focusing on the prevalent topics of gender and race bias. We establish various personas to elicit the potential bias embedded in ChatGPT under Korean and conduct a toxicity evaluation for generation tasks. Our findings confirm that ChatGPT exhibits varying toxicity depending on the assigned persona and topic, capable of inducing substantial toxicity in Korean. This suggests that ChatGPT fails to properly filter the bias included in the training data in terms of knowledge and understanding of Korean society. In future work, we plan to select prompts for various topics in different languages.

Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2022-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques).

Reference

[1] OpenAI, "Gpt-4 technical report," 2023.

[2] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, Vol. 35, pp. 24 824–24 837, 2022.

[3] E. Ferrara, "Should chatgpt be biased? challenges and risks of bias in large language models," *arXiv preprint arXiv:2304.03738*, 2023.

[4] J. Zhou, H. Müller, A. Holzinger, and F. Chen, "Ethical chatgpt: Concerns, challenges, and commandments," *arXiv preprint arXiv:2305.10646*, 2023.

[5] T. OpenAI, "Chatgpt: Optimizing language models for dialogue," *OpenAI*, 2022.

[6] A. Deshpande, V. Murahari, T. Rajpurohit, A. Kalyan, and K. Narasimhan, "Toxicity in chatgpt: Analyzing persona-assigned language models," *arXiv preprint arXiv:2304.05335*, 2023.