The 8th International Conference on Interdisciplinary Research
on Computer Science, Psychology, and Education (2024)

Korean Penalty of Large Language Models Derived by the Tokenizer

Hyeonseok Moon[†],Aram So [‡*]
Department of Computer Science and Engineering, Korea University[†], Human-inspired AI Research[‡]
{glee889[†] , aram[‡]}@korea.ac.kr

Abstract

Large language models demonstrate high versatility and performance across various domains. However, in the predominantly English-focused research landscape, the issue of language disadvantages remains a significant unresolved challenge. While numerous attempts have been made to mitigate performance disparities across languages, we identify the disadvantage that arises from the tokenization stage. We observe that among four large language models, including GPT-4, processing Korean requires up to three times more tokens than processing English. This discrepancy implies higher costs and slower processing speeds for tasks performed in Korean compared to English. Our study highlights the need to address Korean-specific disadvantages at the tokenization stage and proposes the development of specialized tokenization strategies to minimize these disadvantages.

Keywords: Deep Learning, Artificial Intelligence, Natural Language Processing, Language Bias, Large Language Model

1. Introduction

The advent of ChatGPT [1] has seamlessly integrated large language models (LLMs) into not just research and academia but also into our everyday lives [2, 3]. Unlike earlier language models that performed specific tasks [4, 5], modern LLMs can execute user instructions with high accuracy without explicit training [6]. They are widely utilized in diverse fields such as healthcare [7], finance [8], and legal sectors [9].

However, the issue of language-specific disparities remains a significant concern in the predominantly English-centric research landscape [10, 11]. While various studies have addressed performance inequalities across languages, we found that language disparities exist even at the tokenization strategy level, prior to the models' usage [12]. We can easily witness that despite the shorter length of the Korean sentence compared to its English counterpart, tokenization of the Korean sentence resulted in nearly four times more tokens. Considering that GPT-4's usage cost is proportional to the number of input/output tokens, this leads to higher costs and increased processing time [12].

This study analyzes the adverse impact of tokenization strategies on Korean sentences. We highlight that, even when performing identical tasks on semantically equivalent sentences, Korean sentences suffer significant disadvantages compared to English sentences solely due to language. This shows that language models inherently disadvantage Korean users right from the input composition stage, beyond the performance differences between English and Korean.

2. Evaluation Method

To quantify the language-specific disadvantages of tokenizers used in large language models, we adopt the tokenizer evaluation method proposed by [12]. We use two sentences, s_{Kor} and s_{Eng}, each conveying the same meaning in Korean and English, respectively. By comparing sentences with identical meanings, we objectively verify any disadvantages caused by language differences.

We tokenize the sentences using the tokenizer of the large language model and calculate the token length ratio between the two sentences. We define this ratio as the "Korean disadvantage" of the tokenizer. This measure, computed as shown in Equation (1), indicates the processing disadvantage for Korean sentences relative to English ones.

$$\text{Korean Penalty} = \frac{|\text{tokenizer}(s_{Kor})|}{|\text{tokenizer}(s_{Eng})|} \tag{1}$$

*Corresponding author

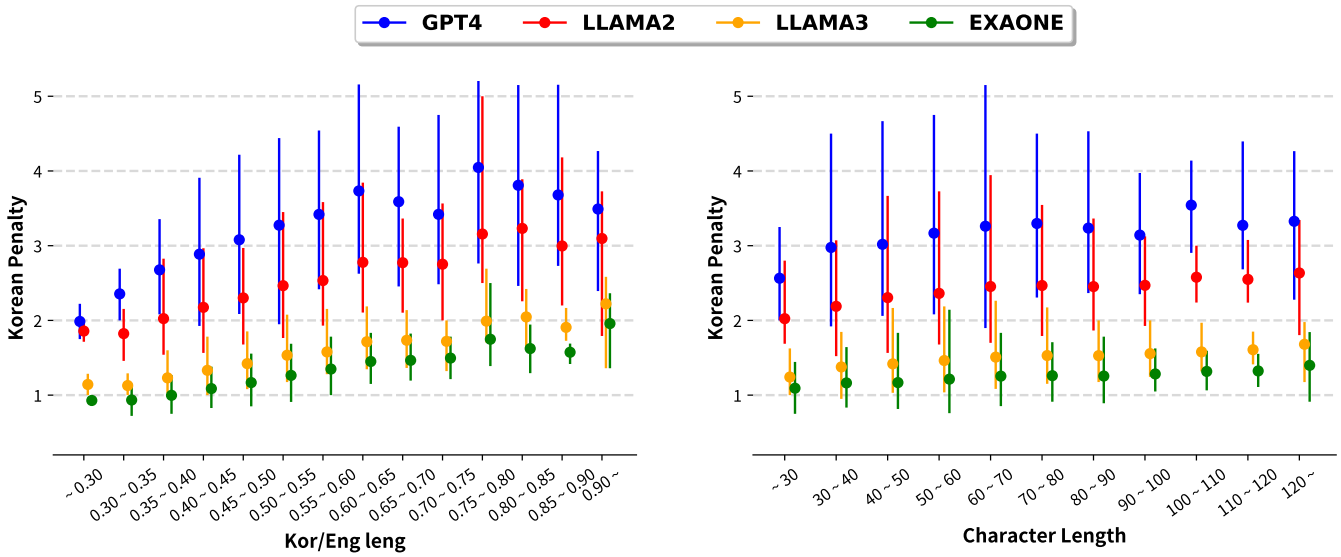

Figure 1. Analysis of the FLORES200 data indicates the disadvantages associated with Korean based on character length. Presented below are the maximum, minimum, and average values of the disadvantage indices within each character length segment.

By comparing the token length ratios of sentences with the same meaning but in different languages, we can objectively compare the tokenization requirements across languages for representing the same information. This allows us to assess the extent to which Korean is at a disadvantage compared to English for the same task. A higher ratio suggests that the tokenizer puts a greater burden on processing Korean, while a ratio closer to 1 indicates minimal disadvantage for Korean in comparison to English.

3. Experiments

We validate tokenizers across both English-centric large language models (LLMs) and Korean large language models. Specifically, we assess a total of four tokenizers. Except for GPT-4, all tokenizers analyzed are open-source models available on HuggingFace [13]. For GPT-4, we use the tokenizer provided via the open-source package tiktoken from OpenAI[1]. For validation, we utilize the FLORES200 parallel corpus designed for translation [14], focusing exclusively on the devtest data subset for our analysis.

Experimental results indicate a linear increase in the disadvantage faced by Korean text as the Korean/English character length ratio rises. This suggests that the greater the discrepancy in character length, the more pronounced this disadvantage becomes, highlighting an issue not only in GPT-4 but across all large language model tokenizers.

Furthermore, the average disadvantage remains consistent across different character length ranges, with significant variations observed throughout all intervals. In every segment where the character length exceeds 30, at least one instance exhibits a Korean text disadvantage of 8 or higher. These findings demonstrate that the tokenizer-induced disadvantage for Korean text is not confined to specific data or formats but is a pervasive issue across all datasets.

4. Conclusion

In this paper, we highlight a significant issue where tokenizers of large language models (LLMs) disadvantage Korean more than English. We observed that Korean sentences tend to be segmented into more tokens than English sentences of comparable length, leading to increased computational costs and latency, which contributes to user dissatisfaction. Our study clearly demonstrates that tokenization strategies can disadvantage Korean usage, urging increased attention to the inherent biases against Korean in these strategies.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded

[1] `https://github.com/openai/tiktoken`. In this study, we specifically use the `cl100k_base` tokenizer from this package.

by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2022R1A2C1007616).

Reference

[1] OpenAI-Blog, "Chatgpt: Optimizing language models for dialogue," https://openai.com/blog/chatgpt/, 2022, accessed: 2024-03-01.

[2] S. Levine, S. W. Beck, C. Mah, L. Phalen, and J. PIttman, "How do students use chatgpt as a writing support?" *Journal of Adolescent & Adult Literacy*, 2024.

[3] E. Agathokleous, C. J. Saitanis, C. Fang, and Z. Yu, "Use of chatgpt: What does it mean for biology and environmental science?" *Science of The Total Environment*, Vol. 888, p. 164154, 2023.

[4] N. Ng, K. Yee, A. Baevski, M. Ott, M. Auli, and S. Edunov, "Facebook FAIR's WMT19 news translation task submission," *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 314–319, Aug. 2019. [Online]. Available: https://aclanthology.org/W19-5333

[5] C. Jia, Y. Shi, Q. Yang, and Y. Zhang, "Entity enhanced bert pre-training for chinese ner," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6384–6396, 2020.

[6] OpenAI, "Language models are few-shot learners," 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[7] M. Javaid, A. Haleem, and R. P. Singh, "Chatgpt for healthcare services: An emerging stage for an innovative perspective," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, Vol. 3, No. 1, p. 100105, 2023.

[8] H. Yang, X.-Y. Liu, and C. Dan Wang, "Fingpt: Open-source financial large language models," *FinLLM at IJCAI*, 2023.

[9] S. S. Biswas, "Role of chat gpt in public health," *Annals of biomedical engineering*, Vol. 51, No. 5, pp. 868–869, 2023.

[10] M. Ryan, W. Held, and D. Yang, "Unintended impacts of LLM alignment on global representation," *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., pp. 16 121–16 140, Aug. 2024. [Online]. Available: https://aclanthology.org/2024.acl-long.853

[11] S. Goldfarb-Tarrant, B. Ross, and A. Lopez, "Cross-lingual transfer can worsen bias in sentiment analysis," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5691–5704, 2023.

[12] A. Petrov, E. La Malfa, P. Torr, and A. Bibi, "Language model tokenizers introduce unfairness between languages," *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Vol. 36, pp. 36 963–36 990, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/74bb24dca8334adce292883b4b651eda-Paper-Conference.pdf

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

[14] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard *et al.*, "No language left behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022.