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

CheckLLM-Ko:Prompting strategy for constructing Korean LLM written document detecting dataset

Myunghoon Kang[1°], Jungseob Lee[1°], Seungyoon Lee[1°], Seongtae Hong[1], Jeongbae Park[2*]
[1]Department of Computer Science and Engineering, Korea University, [2]Human-inspired AI Research
{chaos8527, omanma1928, dltmddbs100, ghdchlwls123, insmile}@korea.ac.kr

Abstract

Instruction tuning-based LLMs are applied to various industries as various tasks can be reasoned in zero-shot regardless of domain. For example, ChatGPT and GPT-4 provide services as commercial APIs, attracting a wide range of users with easy access to their services. However, ChatGPT and GPT-4 can cause security problems for companies by collecting conversation history data, and they can also lead to reliability issues of corporate documents due to the hallucination of the generated results. Since LLM-generated texts have secured a level of fluency similar to humans, it can significantly limit corporate activities if LLM-generated texts cannot be identified in the industrial field. However, there are no Korean LLM writing detection services. In this paper, we propose the CheckLLM-Ko dataset for training Korean Korean LLM written document detectors. The CheckLLM-Ko dataset targets the domain of documents written in the literary and technical writing style often used in the industrial field and represents the LLM written style in three levels: paragraph-level, sentence-level, and token-level. Our dataset comprises various degrees of LLM writing style, and to our best knowledge, it is the first dataset for training the Korean LLM writing detection model.

Keywords: Large Language Model, AI written text detection, Modeling

1. Instruction

Instruction tuning, learning to follow instructions, and increasing parameter sizes have ushered in the era of very large language models (LLMs). Instruction tuning refers to a method of learning to produce output y aligning to a instruction x_I for a task given a simple x, y input and output [1]. Instruction tuning is known to be a key factor in increasing the generalization capability of LLMs comparable to the performance of task-specific fine-tuning models without the need for additional task-specific training [2, 3]. ChatGPT [4] and GPT-4 [5] are LLMs trained by Instruction tuning and have reached about 100 million users by providing commercial APIs[1]. Moreover, the adoption of LLMs in industry is on the rise thanks to their ability to produce outputs that meet the user's intent and their ability to be applied to a variety of tasks regardless of domain. In particular, the use of LLMs is expanding across industries, from the clerical work for writing reports and essays[2] to the creative work such as marketing and advertising copywriting[3].

However, there are limitations to the industrial use of LLMs in their current commercial API form. ChatGPT and GPT-4 collect data on the conversation history of the user's interactions with the LLM for the purpose of further training the model[4]. Therefore, if company use ChatGPT and GPT-4 in their daily work, inputs that contain corporate policies, confidential information, or personal information may be provided to OpenAI and cause security issues. Also, when using LLM to create and modify corporal documents without taking appropriate measures, they may end up with hallucinated information and major errors in the work process and results. Given the capabilities of LLMs fluent generation, it is difficult to determine if the hallucinations present in their work [6, 7].

*Corresponding author

[1]https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

[2]https://www.cbsnews.com/news/chatgpt-chatbot-tiktok-ai-artificial-intelligence/

[3]https://www.fastcompany.com/90833253/ryan-reynolds-used-chatgpt-to-make-a-mint-mobile-ad-and-the-results-were-mildly-terrifying

[4]https://openai.com/policies/terms-of-use

	Training	Validation
행정문서대상기계독해	60,000	168
에세이글평가데이터/글짓기	20,000	56
에세이글평가데이터/대안제시	20,000	56
에세이글평가데이터/설명글	20,000	56
에세이글평가데이터/주장	20,000	56
논문자료 요약	40,000	104
요약문및레포트생성	40,000	104
#document	220,000	600

Table 1. Summary of training and validation dataset of KoCheckGPT

To mitigate these issues, this paper proposes CheckLLM-Ko, a dataset for training Korean Korean LLM written document detectors. CheckLLM-Ko inherits the vocabulary and phrases usage pattern of LLMs when generating document. CheckLLM-Ko targets the domain of documents written in the literary and technical writing style often used in industrial field. The dataset is represented in three-levels: document-level, sentence-level and token-level to mimic human's LLM usage to write the whole document from scratch or revise the words in the sentences.

In order to accurately classify between human-written and LLM-written documents, this paper builds a training and evaluation dataset by prompting LLMs to paraphrase human-written documents using publicly available Korean datasets. In our experiment, we find that the multilingual LLM written document detecotr, ZeroGPT, shows poor detection performance meaning that their lacks of LLM written document detector that specializes in Korean language.

2. Methods

The CheckLLM-Ko training and validation dataset was constructed using ChatGPT. The dataset was built by selecting a public Korean dataset as a human-written text and prompting ChatGPT to paraphrase some of the content in the document. The prompts are shown in Table 2. The document-level and sentence-level prompt is organized to include "in your style" keywords to capture the language patterns of LLM and "same meaning" keywords to preserve the semantics of existing documents. For the token-level, we include both "do not change the content" and "preserve the tone of the document" to change the document in the subtle way.

Prompt Objective	Text
Document-level	아래의 'Document'를 당신의 스타일의 새로운 document를 작성하세요. 단, 당신이 작성한 'Document'는 이전의 'Document'와 상관 없어야 하며, 당신이 작성한 document만 출력해야합니다. {document}
Sentence-level	아래의 'Document'를 당신의 스타일로 페러프레이즈 하여 동일한 의미의 document로 변환해주세요. 단, 당신이 작성한 document만 출력해야합니다. {document}
Token-level	아래의 Document에서 {sentence}을 당신의 스타일로 수정해주세요. 단 나머지 문장은 내용을 바꾸면 안됩니다. 문장의 내용을 바꿀 때 원본 문장의 어투를 유지해야하며 특수문자는 그대로 남겨주세요. 바꾸지 않은 문장은 그대로 놔주세요. {sentence}

Table 2. The prompt for constructing CheckLLM-Ko

3. Experiment

Summary statistics of the dataset for training and validating KoCheckGPT are shown in Table 1. To construct the dataset, we use paraphrased and written texts as source datasets, including machine reading comprehension data for administrative documents published on AI hub: [5], essay writing evaluation data: [6], thesis summaries: [7], and summary and report generation data: [8]. To verify the performance of existing LLM written document detector, we perform experiment on the ZeroGPT[9], an English and multilingual LLM writing discriminator, in terms of the binary classification. The performance is evaluated in Accuracy, F1-score, Recall, and Precision.

Table 3 shows the binary classification performance of the ZeroGPT. The Accuracy and F1-score of were 56.00 and 50.00, respectively, which are close to the results of random classification.

[5] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=realm&dataSetSn=569

[6] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=realm&dataSetSn=545

[7] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=realm&dataSetSn=90

[8] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=realm&dataSetSn=582

[9] https://www.zerogpt.com/

Model	Accuracy	F1	Recall	Precision
ZeroGPT	56.00	50.00	56.00	61.00

Table 3. ZeroGPT, KoCheckGPT comparison result

4. Conclusion

In this paper, we propose CheckLLM-Ko, which represents human's Korean document writing usage assisted with LLM at the token, sentence and whole document level. We constructed a dataset for training and validating LLM written document detector, and contributed to future research on Korean LLM written document detection. We find that ZeroGPT shows poor classification performance in which calls for LLM written document detector that specializes in Korean language. We plan to propose a Korean LLM written document detector that is robust to the document domain by building a dataset of multi-domain documents and conducting further training.

Acknowledgement

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). This work was supported by the Technology development Program(S3310507) funded by the Ministry of SMEs and Startups(MSS, Korea)

Reference

[1] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *International Conference on Learning Representations*, 2021.

[2] 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.

[3] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, "A survey of controllable text generation using transformer-based pre-trained language models," *ACM Computing Surveys*, 2022.

[4] OpenAI-Blog, "Chatgpt: Optimizing language models for dialogue," 2022. [Online]. Available: https://openai.com/blog/chatgpt/

[5] R. OpenAI, "Gpt-4 technical report," *arXiv*, pp. 2303–08 774, 2023.

[6] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, May 2022. [Online]. Available: https://aclanthology.org/2022.acl-long.229

[7] C. Zhou, G. Neubig, J. Gu, M. Diab, F. Guzmán, L. Zettlemoyer, and M. Ghazvininejad, "Detecting hallucinated content in conditional neural sequence generation," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1393–1404, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.findings-acl.120