How To Mitigate Hallucinations with Multilingual Translation

Jaehyung Seo[1], Jeongbae Park[2]*
[1]Department of Computer Science and Engineering, Korea University,
[2]Human-Inspired AI Research

Abstract

Recent advancements in large language models (LLMs) have led to near-human performance in natural language processing (NLP). However, as LLMs continue to grow in capacity, the issue of hallucination—where models generate information that is not factually correct—has become more prominent. To address this, we propose a multilingual approach to mitigate hallucinations by enabling the model to verify its factuality through self-checks based on translations of the same knowledge across multiple languages. In our experiments, we use the KoCommonGEN v2 dataset and the KULLM 3 to identify hallucination cases and exhibit the effectiveness of our multilingual translation-based mitigation method.

Keywords: Natural language processing, Large language model, Hallucination

1. Introduction

The advancement of large language models (LLMs) has significantly improved performance across various natural language tasks. As the scope of these models has grown, so has the demand for them to handle more complex knowledge and reasoning. However, this has led to a growing issue of hallucinations, where models generate information that is inconsistent with the input or contradicts real-world facts.

In NLP, hallucinations occur when a model outputs information that either does not match the input or contradicts facts. Early studies pointed out that the way LLMs are trained using maximum likelihood estimation can contribute to hallucinations [1]. Hallucinations can be divided into intrinsic, where the output does not match the input, and extrinsic, where the model provides information that is not factually correct [2, 3]. Several methods have been proposed to detect, mitigate, and evaluate hallucinations [4, 5, 6].

Previous approaches to mitigating hallucinations often required extensive resources, such as external databases or fine-tuned models for specific domains. While these methods are effective, they can be costly and heavily depend on external resources [3, 5, 7].

We propose a method that reduces reliance on external resources by using multilingual translations to mitigate hallucinations. Many recent open-source LLMs have multilingual capabilities, and they can leverage these to demonstrate factual consistency. Our approach encourages the model to self-check its output by translating the same knowledge into multiple languages.

In this study, we use the Korean language-centric model KULLM 3 [8, 9] and the KoCommonGEN v2 dataset [10]. The dataset includes numerical commonsense reasoning questions, where hallucinations often occur based on incorrect factual information, such as misinterpreting numbers.

Our contributions are as follows:

1. We propose a multilingual translation method to mitigate hallucinations in LLMs.
2. This method reduces the cost of hallucination mitigation while improving model performance.

2. Method

We utilize LLMs' ability to retrieve knowledge by prompting them to translate the same question into multiple languages, thus enabling the model to recall and verify factual knowledge across different languages [4]. LLMs often possess intrinsic knowledge from their pre-training, but they can still generate incorrect or logically inconsistent answers during recall [11].

*Corresponding author

Table 1. Experimental results for multilingual translated mitigation

Model + Method	Accuracy
KULLM 3	0.5354
KULLM 3 + EN	0.5960
KULLM 3 + JP	0.5051
KULLM 3 + EN + JP	**0.6012**

To enhance this recall process, we translate the same question into different languages and use the translated outputs as additional prompts. In this study, we use Korean as the primary language, with English and Japanese as supporting languages to verify factual consistency.

Given the characteristics of the KoCommonGEN v2 dataset [10], we focus on multiple-choice questions. These questions require combining concept nouns and verbs into plausible sentences, making them suitable for translation. In some cases, hallucinated answers were unintentionally corrected during translation. For instance, an incorrect date like ”July 17” was translated to the correct date ”August 15” in English. To decide the final answer, we calculate the log probability of each translated option and select the option with the highest average probability across the three languages.

3. Experiments

Setting We used the KULLM 3 [8, 9] for our experiments, which is based on the Solar 10.7B [12] and fine-tuned for Korean and English instructions. The KoCommonGEN v2 dataset [10], specifically the numerical commonsense reasoning category, is used for evaluation. This dataset includes science, math, history, and society-related questions aligned with compulsory education in Korea.

Result The results are shown in Table 1. When only Korean is used, the model presents 53.54% accuracy. Adding English improve accuracy to 59.60%, but adding Japanese alone result in a slight drop to 50.51%. However, using all three languages (Korean, English, and Japanese) shows the highest accuracy at 60.12%.

These results suggest that recalling information from multiple languages helps improve factual accuracy. However, the performance drop with Japanese highlights the need to consider translation quality and language compatibility when using this approach.

4. Conclusion

We propose a multilingual translation-based self-check method to improve the factual accuracy of LLMs. Unlike traditional fact-checking methods that rely on external resources, our method allows the model to verify its output through translations. Experimental results using the KoCommonGEN v2 dataset and KULLM 3 model show that adding English translations improve the model's performance, but Japanese translations led to a slight performance decline. This suggests that while multilingual translation can be effective for mitigating hallucinations, careful language selection is important.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded 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 work was supported by ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2024-RS-2020-II201819).

Reference

[1] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On faithfulness and factuality in abstractive summarization,” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, 2020.

[2] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions,” *arXiv preprint arXiv:2311.05232*, 2023.

[3] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, Vol. 55, No. 12, pp. 1–38, 2023.

[4] P. Manakul, A. Liusie, and M. Gales, “Selfcheckgpt: Zero-resource black-box hallucination detection for gen-

erative large language models," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023.

[5] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12 076–12 100, 2023.

[6] C. Chen, K. Liu, Z. Chen, Y. Gu, Y. Wu, M. Tao, Z. Fu, and J. Ye, "Inside: Llms' internal states retain the power of hallucination detection," *The Twelfth International Conference on Learning Representations*.

[7] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "Evaluating adversarial attacks against multiple fact verification systems," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2944–2953, 2019.

[8] J. Kim, T. Lee, Y. Jang, H. Moon, S. Son, S. Lee, and D. Kim, "Kullm3: Korea university large language model 3," https://github.com/nlpai-lab/kullm, 2024.

[9] S. Lee, T. Lee, J. Lee, Y. Jang, and H. Lim, "Kullm: Learning to construct korean instruction-following large language models," *Annual Conference on Human and Language Technology*, pp. 196–202, 2023.

[10] J. Seo, J. Lee, C. Park, S. Hong, S. Lee, and H.-S. Lim, "Kocommongen v2: A benchmark for navigating korean commonsense reasoning challenges in large language models," *Findings of the Association for Computational Linguistics ACL 2024*, pp. 2390–2415, 2024.

[11] C. Jiang, B. Qi, X. Hong, D. Fu, Y. Cheng, F. Meng, M. Yu, B. Zhou, and J. Zhou, "On large language models' hallucination with regard to known facts," *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1041–1053, 2024.

[12] D. Kim, C. Park, S. Kim, W. Lee, W. Song, Y. Kim, H. Kim, Y. Kim, H. Lee, J. Kim *et al.*, "Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling," *arXiv preprint arXiv:2312.15166*, 2023.