Investigating the Korean Question-Answering Capability of Large Language Models through Query Perturbations

Jinsung Kim[1,○], Aram So[2,*]
Department of Computer Science and Engineering, Korea University[1], Human-inspired AI Research[2]
{jin62304,aram}@korea.ac.kr

Abstract

Recently, as the inherent knowledge of generative large language models has been utilized in the field of natural language processing to demonstrate excellent performance, active research is also needed in question answering, a representative task suitable for actively employing the knowledge of large language models (LLMs). This study investigates the question-answering ability of large language models in Korean, focusing on how query perturbations affect their performance. While traditional question-answering systems rely on high-quality, human-labeled data, this research explores real-world situations where queries are often altered or corrupted. The study analyzes the LLM's response consistency by examining query perturbations at different levels—syllable, word, and sentence—and compares basic prompting with the chain-of-thought (CoT) prompting method. Results show that sentence-level perturbations, particularly negations, cause the most performance degradation, while syllable-level changes have minimal impact. Interestingly, CoT prompting, despite its recent success, performs worse than basic prompts in this context, suggesting the need for further optimization for Korean question-answering tasks.

Keywords: Korean question answering, large language models, query perturbation, LLM probing, verification study

1. Introduction

Question-answering research in natural language processing (NLP) focuses on generating accurate responses to queries and is closely linked to fields like information retrieval and extraction [1]. With the rise of large language models (LLMs) like ChatGPT [2], these models have shown impressive question-answering abilities. The growing use of Retrieval-Augmented Generation (RAG) technology [3] in the industry further highlights the importance of question-answering tasks. As a result, numerous benchmarks, datasets, and studies aim to evaluate and improve the performance of LLMs, particularly through tasks like machine reading comprehension (MRC).

Most existing studies assess large language models using high-quality, human-annotated question-answer pairs, creating a gap between these evaluations and real-world situations where queries are often altered or "polluted." While research on handling query contamination, including refinement and reconstruction, is more advanced in languages like English, there is limited empirical research on how LLMs respond to query transformations in Korean. Therefore, it's crucial to verify how consistently LLMs can generate accurate responses despite such modifications, aligning with recent LLM probing research in natural language processing.

This study aims to verify how a large language model responds when queries are transformed in a contextual question-answering scenario. The research focuses on real-world query contamination, examining how the model reacts to transformations at the character, word, and sentence levels. It specifically tests the model's response to "noised" text inputs, often used in grammatical error correction (GEC) tasks, and queries converted into a negative form. The goal is to assess how well the model handles negative knowledge, which it may have encountered less frequently during pretraining.

2. Related Works

Recent research on question answering in Korean has primarily utilized leaderboards to assess the capabilities of large language models. For example, the KorQuAD dataset, which

*Corresponding author

is based on the MRC benchmark SQuAD [4], has undergone updates from version 1.0 to 2.0 (and 2.1), with model performance from various industry participants published for these benchmarks. This research trend focuses on evaluating model performance using benchmark datasets with ideal question-answer pairs, which may not fully reflect real-world question-answering scenarios. This highlights the need for research into how models handle incomplete or imperfect queries, more closely mirroring real-world conditions.

With the rise of large language models, technologies like RAG are gaining attention, and there is increasing research on the importance of query quality in natural language processing tasks. For instance, efforts to enhance task performance by refining queries, such as techniques that improve query quality for RAG, have been actively explored in various fields of artificial intelligence, particularly in the English-speaking world.

Additionally, research on query contamination has been widely conducted, especially in areas like speech recognition and spelling correction within natural language processing, including studies in the Korean language domain. These studies aim to mitigate performance drops in downstream tasks by correcting real-world polluted queries. In the context of question-answering, where large language models excel, it's now crucial to study how well these models can maintain response consistency through deliberate query modification, simulating real-world question contamination scenarios.

3. Method

This section outlines the query perturbation methods used in the experiment to analyze large language models' behavior. The perturbations are categorized into syllable, word, and sentence levels. At the syllable level, queries are altered through letter separation and vowel changes. At the word level, perturbations include changing word order and inserting additional words. At the sentence level, negation and changing the query to a declarative form are applied. These modified queries, paired with context, are used to assess whether the models' question-answering ability is maintained and to identify any changes in their responses.

4. Experiments

Experimental Setup. The dataset for the experiment was the corpus of version 1.0 of KorQuAD, a representative Korean MRC dataset [5], and example samples were randomly extracted from the training dataset of the corpus to perform query transformation and question-answering tasks based on them. For the large language model to generate appropriate responses based on the transformed query, we recently adopted GPT-4o-mini (gpt-4o-mini-07-18 version), the most potent large language model. For the hyperparameters, we followed the default settings recommended by OpenAI, considering the usage patterns of general users in the GUI environment. For example, temperature and top-P were set to 1.0.

For the evaluation, we described the average values of the performances produced through a total of three random seed settings, and the indices for performance evaluation were calculated using the exact match (EM) score and F1 score. We described the substring EM score for the EM score, which is considered to have generated an appropriate response when the correct answer is included in a substring of the model's prediction value. However, in the case of the negative type, since the intention of the original question is reversed, a completely different response must be generated rather than the original correct answer. Hence, the evaluation also considers an appropriate response as one when the original correct answer is not included in the substring of the model's predicted value.

Results. The experimental results show that query perturbations consistently reduce model performance compared to using the original queries. This suggests that using high-quality, human-annotated question-answer pairs in experiments differs significantly from real-world situations where queries are often contaminated. The perturbation that caused the greatest performance drop was the negation method. The model processed the transformed, negative queries as though they were positive, failing to account for the change in meaning.

Conversely, the perturbation with the most negligible impact on performance was syllable-level letter separation, and the least damaging word-level perturbation was word insertion. Furthermore, using the chain-of-thought (CoT) prompting technique [6] resulted in a similar or slightly worse

performance than basic (vanilla) prompting. This decline is attributed to errors in intermediate inference steps that negatively affect the final response generation.

5. Conclusion

This study examined the performance of a large language model in handling query perturbations within a Korean question-answering context. By categorizing query perturbations into syllables, phrases, and sentences, the research focused on evaluating the model's consistency in responding to various contaminated queries. Additionally, it compared the model's generative ability when responding to queries with context, using both the chain-of-thought prompting technique, known for strong performance in complex tasks, and basic prompts that only include task instructions.

The results showed that sentence-level negation perturbations had the most negative impact on the model's response consistency, while syllable-level letter separation had minimal effect. Moreover, the Zero-shot CoT prompting approach generally maintained or slightly worsened performance compared to Vanilla prompting in handling modified queries, suggesting the need for further development of optimized prompting techniques for Korean question-answering tasks.

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 research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Development of generative AI-based publishing content analysis and sharing platform technology to respond to changes in the publishing environment. Project Number: RS-2024-00442061).

Reference

[1] L. Hirschman and R. Gaizauskas, "Natural language question answering: the view from here," *natural language engineering*, Vol. 7, No. 4, pp. 275–300, 2001.

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

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, Vol. 33, pp. 9459–9474, 2020.

[4] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

[5] S. Lim, M. Kim, and J. Lee, "Korquad1. 0: Korean qa dataset for machine reading comprehension," *arXiv preprint arXiv:1909.07005*, 2019.

[6] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, Vol. 35, pp. 22 199–22 213, 2022.