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

Unveiling the Blind Spots: Evaluating Large Language Models in Korean Commonsense Reasoning

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

Abstract

This paper proposes a new evaluation test set based on Korean commonsense reasoning for large language models. The proposed test set is grounded in general Korean commonsense, aiming to assess how well large language models understand the given information and generate outputs that align with it. Existing Korean benchmarks are insufficient for a nuanced evaluation of the advanced commonsense reasoning capabilities of large language models. Furthermore, they often fail to adequately account for social biases and hallucinations in the evaluation process. Our evaluation test set addresses these shortcomings and introduces a new commonsense reasoning benchmark tailored for the forthcoming era of large language models, thereby facilitating sustained advancements in Korean natural language processing research.

Keywords: Large Language Model, Commonsense Reasoning, Benchmark

1. Introduction

Natural language processing (NLP) research has developed benchmark datasets to achieve human-like performance, evaluating linguistic capabilities and general commonsense-based reasoning [1, 2, 3, 4]. However, with the advent of large language models, the validity of most benchmarks has significantly decreased. The current tasks and datasets struggle to adequately assess the capabilities of these models, diminishing their discriminative power.

Continually emerging large language models claim superior performance, exceeding the limitations of existing benchmark datasets. These models undergo pre-training on massive corpora and require quantitative evaluations to assess the quality and correctness of the knowledge. In this context, HuggingFace's OpenLLM Leaderboard introduces four types of evaluation methods: ARC (AI2 Reasoning Challenge) [5] evaluates a model's reasoning capabilities based on elementary-level science questions, focusing on fundamental knowledge and logical reasoning. ARC comprises 2,590 challenge sets and 5,197 easy sets, where the challenge set is designed to mislead models through word overlapping and information retrieval algorithms. HellaSWAG [6] assesses reasoning capabilities grounded in commonsense. Though easy for humans with about a 95% accuracy rate, the evaluation includes choices that can be difficult for models due to adversarial filtering. MMLU [7] gauges how well a language model captures and manifests broad domain knowledge during pre-training. It includes problem-solving questions from 57 domains, ranging from humanities and social sciences to sciences. It consists of 15,908 question-answer pairs, with each domain containing at least 100 examples. TruthfulQA [8] evaluates the trustworthiness of the information produced by language models. Assuming larger models are more likely to imitate false information from the web, it comprises 817 query pairs across 38 domains and measures performance in a zero-shot setting. The evaluation is divided into truthfulness and informativeness, utilizing human evaluations and models trained on these scores.

Given the need for new benchmarks and their emergence, this paper proposes the necessity of a new benchmark to evaluate the language understanding and reasoning capabilities of large language models in Korean. We propose to include the following three features in this new benchmark: (i) We add historical facts and numerical data that could induce hallucinations. (ii) The extended conceptual information necessitates inference to select the correct answer from unstated

*Corresponding author

information. (iii) Information that could potentially foster social bias and hate is included. Through this evaluation, we identify large language models' inherent commonsense reasoning abilities and the potential risks they may pose in the reasoning process.

2. Dataset Construction

We design a dataset of 300 items, encompassing three major aspects: (i) Hallucinations, (ii) Extended Conceptual Information, (iii) Bias and Hate The design follows the specific principles below:

Hallucinations For the first aspect, 100 items in the dataset are engineered to induce hallucinations potentially. The conceptual information set for these items incorporates erroneous historical facts and geographical data deliberately designed to mislead the models.

Extended Conceptual Information Secondly, another 100 items are structured to include conceptual information that is not explicitly stated. Models should infer the relationships between different pieces of conceptual information and understand that not all information is explicitly provided.

Bias and Hate Lastly, the remaining 100 items are formulated to induce social bias and hate potentially. The options for these items are constructed to have a high degree of word overlap and similarity with the accompanying conceptual information set. The choices include the possibility that the model may generate harmful outputs when combining the given conceptual information.

Through these design principles, our dataset aims to offer a comprehensive evaluation, gauging both the models' capabilities and their potential risks.

3. Experiments

We conduct experiments on the 300-item dataset with KoGPT2 [1], GPT-3.5 [9], and GPT-4 [10] as baseline models. The core evaluation method focuses on the models' capability to infer relationships among given sets of conceptual information and to choose the most plausible option. All sets of conceptual information are structured around the same questions. Multiple-choice assessments provide a fairer evaluation than traditional generative evaluations, particularly for measuring the models' discriminative ability among newly introduced item types.

Table 1. Experimental results for baselines

	Accuracy
KoGPT2	0.228
GPT-3.5	0.305
GPT-4	**0.524**

Table 1 shows the zero-shot performance of the models on the 300 multiple-choice questions, which cover challenging inferencing scenarios. KoGPT2, a Korean pre-trained language model, performs almost at random chance, demonstrating its failure to understand the provided prompts and instructions fully. A substantial performance gap exists between GPT-3.5 and GPT-4, with the latter recording significantly higher accuracy even on this challenging dataset.

4. Conclusion

This paper proposes a novel method for assessing Korean natural language understanding in the era of large language models. Unlike benchmark datasets, our evaluation test set includes types of errors that large language models are susceptible to, highlighting the risks of biases and hate speech that can be masquerade as general knowledge. While our study validates this on a limited dataset, future work aims to apply more data and adversarial designs to reflect commercial model biases. We also plan to expand the dataset with more challenging types and examples where state-of-the-art models like GPT-4 could struggle. Additional validations by native Korean speakers are in the pipeline. We hope our research contributes to a new Korean-based natural language processing research phase.

Acknowledgements

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience program (IITP-2023-2020-0-01819) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). 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).

Reference

[1] https://github.com/SKT-AI/KoGPT2

[1] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *International Conference on Learning Representations*, 2018.

[2] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," *Advances in neural information processing systems*, Vol. 32, 2019.

[3] Y. Bisk, R. Zellers, J. Gao, Y. Choi *et al.*, "Piqa: Reasoning about physical commonsense in natural language," *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34, No. 05, pp. 7432–7439, 2020.

[4] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "Commonsenseqa: A question answering challenge targeting commonsense knowledge," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

[5] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," *arXiv preprint arXiv:1803.05457*, 2018.

[6] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019.

[7] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *International Conference on Learning Representations*, 2020.

[8] 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, 2022.

[9] 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.

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