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

An Analysis of Korean Language Proficiency of Gemma 2B Models Using KBS Korean Language Proficiency Test

Yoonna Jang[1], Yuna Hur[2*]
[1]Korea University, [2]Human-inspired AI Research
{morelychee, yj72722}@korea.ac.kr

Abstract

In recent years, language models have demonstrated exceptional abilities in various fields, leading to attempts to evaluate their understanding, generation, and conversational skills. This study aims to measure the Korean language proficiency of Gemma 2B model using KBS Korean Language Proficiency Test. The KBS Korean Language Proficiency Test provides a comprehensive assessment of effectiveness, fluency, accuracy, and creativity in language use. The test used in the experiment consists of six evaluation criteria: vocabulary, grammar, writing, creation, reading, and Korean culture. Performance is measured by providing questions, passages, and candidate answers for a total of 76 questions. The performance of the models is assessed by generating one correct answer from five options.

Keywords: Korean Language Proficiency, Gemma, Language Models

1. Introduction

Currently, most benchmarks evaluate the language comprehension ability in Korean, focusing primarily on accurately interpreting and generating text written in Korean[1, 2, 3]. However, there is a lack of benchmarks that assess Korean language proficiency, which encompasses vocabulary, grammar, and language usage. Therefore, we measure the Korean language proficiency of models using the KBS Korean Language Proficiency Test. The KBS Korean Language Proficiency Test aims to provide a comprehensive evaluation of effectiveness, fluency, accuracy, and creativity in language use. The test evaluation criteria include listening, speaking, reading, and writing, as well as vocabulary, grammar, creative language use, and Korean cultural understanding, covering multidimensional aspects of language proficiency.

We performed tagging on questions, passages, and candidate answers through data preprocessing, and excluded questions that require listening or images to answer the question. For each item, the models are provided with questions, passages, and candidate answers, and are required to generate one number from five options that is predicted to be the correct answer. This approach allows us to indirectly measure the Korean language proficiency of models. The evaluation targets the recently released 2B parameter Gemma models. The evaluation criteria are reported across six areas: (1) Vocabulary, (2) Grammar, (3) Writing, (4) Creation, (5) Reading, and (6) Korean Culture, based on the number of correctly answered questions.

2. Method

The KBS Korean Language Proficiency Test[1] is administered by KBS and the KBS Korean Language Promotion Institute to improve the national use of the Korean language and develop Korean language culture. It measures whether individuals possess proper Korean language usage skills.

The test used in the experiment is the '54th KBS Korean Language Proficiency Test (Odd Numbered Version)', utilizing the version currently accessible on the KBS Korean Language Proficiency Test website.[2] Since the released test papers are in PDF format, text extraction from the PDFs is done using the pdfplumber[3] library. The extracted text data is then manually tagged into questions, passages, candidate

*Corresponding author.

[1]https://www.klt.or.kr/

[2]Questions: https://edu.klt.or.kr/main/cs_dataroom_view?bbs_id=93 Answers: https://edu.klt.or.kr/main/cs_dataroom_view?bbs_id=94

[3]https://pypi.org/project/pdfplumber/

answers, and labels.

As the model being evaluated only receives natural language input, listening ability is not assessed. Additionally, questions that require the understanding of images to determine the correct answer are excluded. Parts requiring visual understanding, such as underlined text, are replaced with quotation marks to allow the model to solve the questions using only text. As a result, the data used in the experiment consists of 15 vocabulary questions, 15 grammar questions, 5 writing questions, 7 creation questions, 25 reading questions, and 9 Korean culture questions, totaling 76 questions.

To prompt the language models, we provide inputs for description (`desc`), question (`question`), passage (`content`), and answer choices (`c1-c5`). The instruction prompt is written as follows to guide the model:

For each item, a sample is provided for each input, and the model is required to select the answer as a number.

> 'You are a Korean language specialist assisting with selecting one answer among five given [candidates]. Read the given [description] and [question] and choose the correct answer based on your understanding of the [text]. Please provide the predicted output in numbers such as (1), (2), (3), (4), (5). [description] desc [question] {`question`} [text] {`content`} [candidates] (1) {`c1`} (2) {`c2`} (3) {`c3`} (4) {`c4`} (5) {`c5`}'

In this experiment, models are required to generate a number according to the conditions. The predicted answer is verified by checking whether it matches the reference answer. If multiple numbers are generated simultaneously, the first number generated is considered the answer.

3. Experiments

In this study, the models evaluated are the Gemma 2B series of Google, which are open-source models with 2 billion parameters. These include the Gemma 1.1 2B model, the Gemma 2 2B instruction tuned (IT) model, and the KoGemma model, which is a version of Gemma 2B additionally trained on the Korean corpus. The experimental results are shown in Table **??**. Tasks 1 to 6 correspond to vocabulary, grammar, writing, creation, reading, and Korean culture, respectively.

For Task 1 (vocabulary), the Gemma 1.1 model correctly answers the most questions, with four correct answers. However, it does not show a significant difference compared to the other models. In Task 2 (grammar), both KoGemma and Gemma 2 perform higher than Gemma 1.1. For Task 3 (writing), all models only manage to answer one question correctly. Task 4 (creation) also shows similar, relatively poor performance across all three models. In Task 5 (reading), KoGemma outperformed Gemma 1.1, and Gemma 2 shows significantly higher performance compared to the other two models. Task 6 (Korean culture) is expected to be the most challenging for the models, with only Gemma 2 managing to answer only one question correctly.

Table 1. The evaluation criteria and number of questions used in this experiment. The test questions are from the '54th KBS Korean Language Proficiency Test (Odd Numbered Version)'. T1-T6 refers to (1) Vocabulary, (2) Grammar, (3) Writing, (4) Creation, (5) Reading, and (6) Korean Culture, respectively.

Model	T1	T2	T3	T4	T5	T6
`google/gemma-1.1-2b-it`	4	1	1	2	2	0
`nlpai-lab/ko-gemma-2b-v1`	3	5	1	2	5	0
`google/gemma-2-2b-it`	2	4	1	1	11	1

Overall, KoGemma, which includes additional Korean training, demonstrates higher performance compared to Gemma 1.1. Additionally, Gemma 2, a relatively recent model, shows a significant improvement in Korean language proficiency compared to its previous version, suggesting that the inclusion of more Korean data during pre-training likely contributed to the improvement. Furthermore, the better performance of KoGemma compared to Gemma 1.1 indicates that further enhancements in Korean language training could lead to even greater performance improvements, as seen in Gemma 2.

4. Conclusion

In this study, we measured and analyzed the Korean language proficiency of Gemma 2B models using the KBS Korean Language Proficiency Test. This allowed us to assess the level of Korean language proficiency exhibited by the Gemma 2B series models based on standardized test scores. To evaluate generative models, we hope that benchmarks for evaluating Korean language proficiency in the descriptive form will be developed, setting a standard that could further enhance the capabilities of many Korean language

models.

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 Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2022R1A2C1007616).

Reference

[1] G. Son, H. Lee, S. Kim, S. Kim, N. Muennighoff, T. Choi, C. Park, K. M. Yoo, and S. Biderman, "Kmmlu: Measuring massive multitask language understanding in korean," *arXiv preprint arXiv:2402.11548*, 2024.

[2] M. Jang, D. Kim, D. S. Kwon, and E. Davis, "Kobest: Korean balanced evaluation of significant tasks," *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3697–3708, 2022.

[3] C. Park, H. Kim, D. Kim, S. Cho, S. Kim, S. Lee, Y. Kim, and H. Lee, "Open ko-llm leaderboard: Evaluating large language models in korean with ko-h5 benchmark," *arXiv preprint arXiv:2405.20574*, 2024.