DATASET MKSL FOR MEASURING ADEQUATE RESPONSE PERFORMANCE BY KNOWLEDGE LEVEL

Noh Myong Sung *
Department of Elementary Computer Education
Seoul National University of Education
Seocho Jungang-ro 96, Seoul, Korea, 06639
myongsungcs@gmail.com

Cho Ung Hui
Department of Software Engineering
SangMyung University
Sangmyeongdae-gil 31, Cheonan, Korea, 31066
paul9512@gmail.com

Abstract

Currently, generative AI is performing well in various fields. In particular, GPT-4, one of the basic models, has been evaluated for its discourse quality, knowledge level, and problem-solving ability on various benchmark datasets. However, it is questionable whether the base model can appropriately adjust its output level according to the user's knowledge level. If the base model fails to consider the user's knowledge level, the quality and reliability of the discourse is bound to decrease. However, common datasets are still insufficient to measure whether the base model responds appropriately to the user's knowledge level. Therefore, based on Korean educational experts and curricula, we developed a benchmark dataset to evaluate whether the underlying model can elicit appropriate discourse according to the user's knowledge level. This mini-dataset consists of about 500 Korean datasets centered on science and current events in the field of science, and we introduce the evaluation method using it. The dataset will also be released soon after it is expanded.

1 Introduction

With the advancement of AI technology, generative models such as GPT-4 have emerged, and a new challenge of adapting output to the learner's understanding is gaining attention. In this study, we introduce the Science Discourse Leveling Mini Dataset (MKSL), developed based on the Korean education system, to evaluate whether the underlying model can elicit discourse that matches the user's knowledge level. Designed with insights from education experts, the dataset contains multi-layered questions for basic, intermediate, and advanced understanding levels. With initial results showing that GPT-4 is more adaptive in solving complex problems than GPT-3.5, we hope to deepen the discussion on the ability of AI to deliver personalized learning experiences in educational settings.

2 The MKSL Dataset

2.1 Building data and utilizing generative AI

We utilized GPT-4 and our own prompts developed by educators to generate a large dataset. By using our own prompts, we were able to generate questions from people with different knowledge levels and three different levels of answers. In addition, the data generated by the combination of GPT and prompts was not immediately utilized, but was reviewed by educators and saved as final data. The review criteria were selected based on the appropriateness of the question field and the appropriateness of the answer level, and the review was carried out by directly viewing the data in the generated data and deleting, modifying, or adding new content.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies. Funding acknowledgements go at the end of the paper.

Table 1: Explanation of the Questions(English Version)

Fields	**Questions**
Earth Sciences	Q: Why does it rain different amounts in different regions?
Physical	Q: What is a "force" in physics?
Chemistry	Q: Why do my apples turn brown after I cut them?
Astronomy	Q: How do stars die?
Biology	Q: How do trees move water up from the roots to the leaves?

2.2 Question

The questions were selected to represent real-world questions that people of different knowledge levels might ask in the fields of physics, astronomy, earth sciences, life sciences, and chemistry. One of the most important characteristics of the questions was that they should be able to generate different answers from at least three different knowledge levels. For example, a question like What is the capital of South Korea? was not a good question because it would only produce one answer (seoul). An example of a good question would be Why does the sun shine brightly? for example. This is because it allows for different types of answers to be given, taking into account the age and knowledge level of the questioner, from a child to an expert.

2.3 Answer

The answers are designed to represent the answers to the questions by categorizing them into three categories. The three categories represent three levels of knowledge and are categorized into three groups based on the Korean curriculum: children to elementary school students, middle to high school students, and college students to adults. In terms of general age, it can be categorized as 7-13 years old, 14-19 years old, and 20+ years old. We mapped the answers to each level to a single question, based on advice from education experts about the standard knowledge level of the group. Below is an example of a question and answer(English Version).

> ***Q:*** *Why do we have seasons on Earth?*
> ***A:*** *Because the Earth is tilted as it revolves around the sun, it receives different amounts of sunlight at different times of the year, resulting in seasons.*
> ***B:*** *Because the Earth's axis of rotation is tilted, the amount of light it receives from the sun changes as it revolves around the sun, resulting in seasons.*
> ***C:*** *Earth's changing seasons are caused by Earth's tilted axis of rotation as it revolves around the Sun, resulting in differences in the amount of energy it receives from the Sun at different times of the year, which is directly related to Earth's climate patterns.*

In the example above, there are a total of three answers to one question mapped to each level, starting with A, which is the child level, and ending with C, which is the adult level. Each answer has significant differences, especially in the keywords used, and there are also differences in sentence structure.

3 The MKSL Features

The data we created consists of a set of questions and answers in the areas of science and science trivia. The answers are in the form of selecting one of three choices. The reason why the data is concentrated in science is because there are many leveled answers and universal and generalized knowledge, which will help us to create data that can significantly overcome cross-cultural and cross-language differences.
The data is mainly composed of physics, earth science, life science, astronomy, and chemistry,

and each field has about 100 data. Each data is composed of questions and answers that ask for detailed and specific knowledge in each field, and each answer is composed of a total of 3 answers, and each answer has different keywords and sentence structures.

4 Experiments

4.1 Experimental Purpose

This experiment aims to evaluate how well an advanced generative AI model like GPT-4 can answer questions for users of different knowledge levels (A, B, and C) based on the dataset we built. In particular, we want to measure whether the model can adjust the difficulty of the answer to match the assumed intellectual level of the user.

4.2 Configure question sets and user profiles

First, prepare a set of questions and answers that cover the various topics in the dataset. The answers in the dataset are organized into three levels of A.B.C., each representing the knowledge difficulty of an elementary school student, a middle to high school student, a college student, or an adult. Prior to the experiment, we assume that the user profiles correspond to the A.B.C. levels labeled as easy, moderate, and challenging. Each profile focuses on the typical knowledge range of the A.B.C. levels, which is reflected in the input prompts during the zero-shot training process.

4.3 Experimental Procedures

Algorithm 1 Evaluate Questions and Update Difficulty Table

Input: Path to CSV file containing questions ("intedata.csv"), API key for OpenAI
Output: Table with counts of questions evaluated as 'A', 'B', or 'C' across difficulty levels
Initialize an empty table T with indices $\{A, B, C\}$ and columns $\{\text{basic}, \text{normal}, \text{challenge}\}$
Load questions from the specified CSV file into dataset D
for each question q in dataset D **do**
    Randomly select a difficulty level l from $\{\text{basic}, \text{normal}, \text{challenge}\}$
    if l is 'basic' **then**
        Set prompt indicating the question is for a child or elementary student level
    else if l is 'normal' **then**
        Set prompt indicating the question is for a middle or high school student level
    else
        Set prompt indicating the question is for a college student or adult level
    end if
    Send the prompt to the AI model and receive a response
    Use the AI model to evaluate if the question corresponds to 'A', 'B', or 'C'
    if $r \in \{A, B, C\}$ **then**
        Increment table count: $T[r, l] \leftarrow T[r, l] + 1$
    else
        Increment an error counter e
    end if
end for
Display the updated table T

For each question, we prepare a prompt that asks the user to select the appropriate knowledge level for the question. The prompt is also determined by the level of the questioner, that is, the user profile. These prompts are entered into GPT-4, and we specifically designed them to constrain the output format of the answer so that, depending on the question and the level of the input user, the answer is one of three conditions: A, B, or C.
.

For each question, the answer selection mechanism selects the answer that it believes best fits the user profile provided in the prompt from among several answers provided by the dataset. The answer is recorded in the answer table (difficulty table).

4.4 Run an experiment

We utilized this dataset to measure the performance of GPT-4 and GPT-3.5 with zero-shot training. Following the experimental procedure mentioned above, we asked the language models to solve approximately 500 problems contained in the dataset without any prior training.

4.5 Experimental results

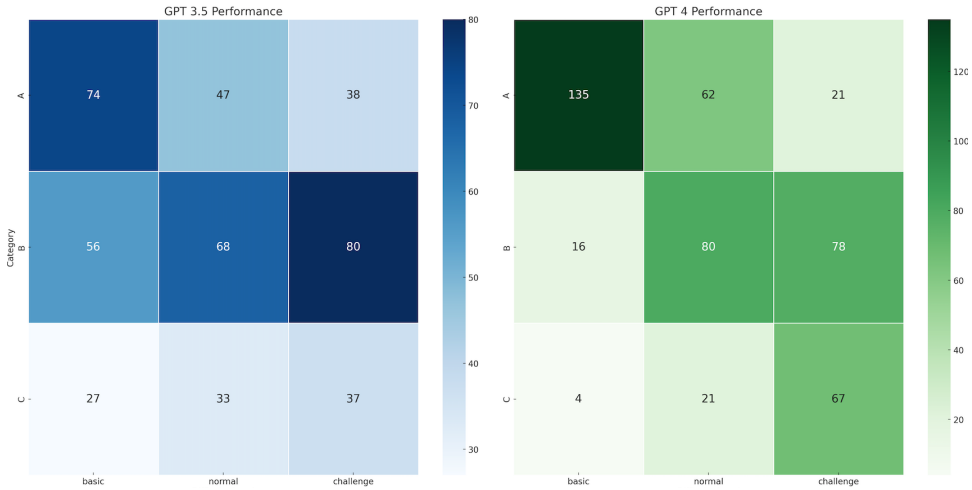

Figure 1: Result Difficulty Table

In this experiment, we compare the performance of GPT 3.5 and GPT 4 using the MKSL dataset. The performance evaluation assumed three learner levels: Basic (children to elementary school students), Moderate (middle school students), and Challenging (college students to adults), and we categorized the answers provided by the models at each level into three categories: "A", "B", and "C". Through this experiment, we observed how well the language model provided appropriate answers based on each learner's level of difficulty. The results showed that GPT 4 performed more balanced across all levels of difficulty than GPT 3.5, with GPT 4 adapting better to beginner-level questions. These results suggest that GPT 4 performs better than GPT 3.5 in solving more complex problems and understanding a wider range of topics, and we believe our data reflects this.

5 Conclusions

The benchmark dataset developed in this study served as an important tool for evaluating the ability of the underlying models to customize answers. While we have shown that foundational models such as GPT-4 are capable of generating expert-level discourse, we have also shown that progress is needed in providing answers that are appropriate for different learner levels. The dataset, built in collaboration with educational experts, can contribute to improving the performance of models in generating answers to scientific questions, which will further strengthen the reliability of foundational models. Future research will utilize this dataset to explore the applicability of the model to users with different linguistic and cultural backgrounds, and to investigate ways to pedagogically strengthen the accountability and reliability of the underlying model. In conclusion, this study provides an initial step in showing that the underlying model can identify and communicate effectively with human

knowledge levels, which emphasizes the reliability and accountability of the underlying model and provides a starting point for deeper human interaction.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Lijia Chen, Pingping Chen, and Zhijian Lin. Artificial intelligence in education: A review. *Ieee Access*, 8:75264–75278, 2020.

Ilie Gligorea, Marius Cioca, Romana Oancea, Andra-Teodora Gorski, Hortensia Gorski, and Paul Tudorache. Adaptive learning using artificial intelligence in e-learning: A literature review. *Education Sciences*, 13(12):1216, 2023.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

Vasiliy Kolchenko. Can modern ai replace teachers? not so fast! artificial intelligence and adaptive learning: Personalized education in the ai age. *HAPS educator*, 22(3):249–252, 2018.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*, 2023.

Setareh Maghsudi, Andrew Lan, Jie Xu, and Mihaela van Der Schaar. Personalized education in the artificial intelligence era: what to expect next. *IEEE Signal Processing Magazine*, 38(3):37–50, 2021.

Mir Murtaza, Yamna Ahmed, Jawwad Ahmed Shamsi, Fahad Sherwani, and Mariam Usman. Ai-based personalized e-learning systems: Issues, challenges, and solutions. *IEEE Access*, 2022.

Raul Puri, Ryan Spring, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Training question answering models from synthetic data. *arXiv preprint arXiv:2002.09599*, 2020.

Andy Rosenbaum, Saleh Soltan, and Wael Hamza. Using large language models (llms) to synthesize training data. *Amazon Science*, 2023.

AVNS Thimmanna, Mahesh Sudhakar Naik, S Radhakrishnan, and Aarti Sharma. Personalized learning paths: Adapting education with ai-driven curriculum. *European Economic Letters (EEL)*, 14(1):31–40, 2024.

Tommy van der Vorst and Nick Jelicic. Artificial intelligence in education: Can ai bring the full potential of personalized learning to education? 2019.

Puri et al. (2020) Rosenbaum et al. (2023) Kolchenko (2018) Murtaza et al. (2022) Kasneci et al. (2023) Thimmanna et al. (2024) Maghsudi et al. (2021) Chen et al. (2020) van der Vorst & Jelicic (2019) Gligorea et al. (2023)

Brown et al. (2020) Achiam et al. (2023) Li et al. (2023)