Multiple-Choice Question Generation Using Large Language Models that Also Controls Discrimination

Yo Ehara
Assoc. Prof.
Tokyo Gakugei University
ehara@u-gakugei.ac.jp researchmap.jp/yo_ehara
yoehara.com readability.jp

This paper and poster are supported by the JST ACT-X (JPMJAX2006) grant.

rebrand.ly/lap24
URL to this paper's presentation

rebrand.ly/lap24vid

Motivation: Discrimination (test question characteristics): a question with high discrimination would accurately identify learners with a high ability to answer that question. Discrimination has rarely been addressed in question generation using large language models (LLMs).

Problem: The concept of discrimination is not well understood by both LLM users and LLMs, reducing the interpretability and controllability, respectively.

Proposal: A method for generating questions while controlling their discrimination by prompting large language models.

Method: we used the following prompt

The following English vocabulary test question is predicted to have a mean distribution of 69 correct answers with a standard deviation of 4.16 for a given group of 100 English language learners. For this population of test takers, generate English vocabulary test questions with a similar distribution of predicted number of correct answers and smaller standard deviation, using completely different words.

Instead of using the word "discrimination" directly in prompts, we use the underlined expression in the prompt using the words "mean" and "standard deviation", **which both LLMs and their users are familiar with,** resulting in better interpretability and controllability.

Original question
=====
"The area was _____ in timber and coal." Choose one of the following four words from the underlined words: expensive, cheap, poor, or not well off
=====

We used [Ehara, EDM22] to predict the mean and standard deviation values of a new question for a given group using masked language models (MLMs) when the result of other similar questions for the group is provided. Item response theory (IRT) [Baker, 2004] was used to calculate the difficulty and discrimination parameters of the sample question.
Recent LLM-based methods for language learning [Cui et al.,2023] [Farr 2024] do not deal with discrimination. The LLM used for the experiment was GPT-4 (ChatGPT May 24 version).

Results:

Instructed to generate a question so that the number of correct answers is 95 out of 100.
He turned off the lights and went to _____. a) bed, b) kitchen, c) car, d) park
Instructed to generate a question so that the number of correct answers is 29 out of 100.
The researcher analyzed the _____ between the two variables. a) apex, b) correlation, c) paradigm, d) zenith
Instructed to generate a question so that the standard deviation is larger than that of the original question:
He is very _____. a) kind, b) friendly, c) evil, d) generous
Instructed to generate a question so that the standard deviation is smaller than that of the original question:
The birds _____ south for the winter. a) fly b) drive c) walk d) run

Instructed to generate a question so that the IRT difficulty parameter would change from -1.2 of the original question to 0.89.
Despite the challenging conditions, the team remained _____. a) resilient b) permeable c) solvable d) inflammable
Instructed to generate a question so that the IRT discrimination parameter would be smaller than 0.738 of the original question:
The sun _____ in the east. a) rises, b) falls, c) sinks, d) dives
Instructed to generate a question so that the IRT discrimination parameter would be larger than 0.738 of the original question:
He has a _____ to exaggerate things. a) tendency b) progression c) direction d) development

Conclusions:

Problem: LLM users (typically teachers) and LLMs are unfamiliar with the idea of question discrimination, hence discrimination is difficult to control.
Proposed method: LLMs are presented with a sample question with the mean and standard deviation of the number of learners who correctly answered the question. LLMs are then prompted to generate a new question with a different mean or standard deviation.
Results:
The proposed method qualitatively seemed to work well. GPT-4 seemed smart enough to be able to control discrimination.

References
Ehara, Y. No meaning left unlearned: Predicting learners' knowledge of atypical meanings of words from vocabulary tests for their typical meanings. Proc. of EDM (short paper), 2022.
Baker, F.B., Item Response Theory : Parameter Estimation Techniques, Second Ed., CRC Press, 2004.
Cui, P. and Sachan, M. Adaptive and Personalized Exercise Generation for Online Language Learning. In Proc. of ACL, pp. 10184–10198, Toronto, Canada. 2023.
Farr, C. Unmasking ChatGPT: The Challenges of Using Artificial Intelligence for Learning Vocabulary in English as an Additional Language. Master Thesis. Univ. of Victoria. 2024.