LARGE LANGUAGE MODEL DETUNING IN LEARNING CONTENT UNDERSTANDING

Tsubasa Minematsu and Atsushi Shimada
Kyushu University, Japan

ABSTRACT

In using large language models (LLMs) for education, such as distractors in multiple-choice questions and learning by teaching, error-containing content is used. Prompt tuning and retraining LLMs are possible ways of having LLMs generate error-containing sentences in the learning content. However, there needs to be more discussion on how to tune LLMs for specific lecture content. Such discussions help control LLMs and for developing educational applications. In this study, we aim to train a detuned LLM that only states incorrect things, considering the limitations of prompt-based approaches such as prompt injection. Our method detunes LLMs by generating datasets that confuse LLMs. To evaluate our method, we asked the detuned LLM to solve multiple-choice questions to evaluate whether it answered the questions incorrectly or not. We also evaluate how many errors are contained in the sentences generated by the LLM to investigate how their knowledge of lecture content is degraded regarding factuality.

KEYWORDS

Large Language Model, Data Poisoning, Data Augmentation

1. INTRODUCTION

LLMs are not just theoretical concepts, but practical tools that can be applied to learning support and teacher support (Kasneci et al., 2023). Their high performance in text generation is shown in their practicality. In addition, their applicability to various natural language processing applications is remarkable, thanks to leveraging recent techniques such as prompt engineering and in-context learning. With LLM applications such as OpenAI's ChatGPT, we can quickly develop various practical applications such as question generation (Raina & Gales, 2022) and tutoring systems by LLMs (Mollick & Mollick, 2023). Mollick & Mollick (2023) provide seven examples of practical applications of LLMs as tutors and students and their relationship to educational theory.

The required abilities of such LLMs are different depending on their role. To support learners as a teacher, the LLM understand the learning content more correctly than the learners. Like learning-by-teaching (Kirschner & Hendrick, 2020; Matsuda et al., 2010), when a learner evaluates the texts generated by a LLM, it is acceptable for the LLM to produce sentences that contain mistakes. Such LLMs can provide a learning environment in which students interact with each other in a person-to-person-like dialogue, asking additional questions and pointing out mistakes.

The reasonable way for generating sentences containing errors is to constrain the LLM with prompts such as "Answer the question incorrectly" and "Explain the topic incorrectly". However, in our preliminary investigation, it was easy to get the constrained LLMs with the above prompt to explain correctly. After the LLM gave an incorrect explanation using the above prompt, this finding was produced by asking the LLM for the correct explanation again, like "Explain correctly". This second input for getting correct information can be regarded as a type of prompt injection such as "*ignore the above...*" (Crothers et al., 2023). In other words, even though a system administrator adds such constraints, learners can easily pull out the correct content when interacting with the LLM in a student role. Such attack methods have been studied to degrade the model performance (Crothers, Japkowicz, & Viktor, 2023; Shu et al., 2023). The task is to adjust the LLM to produce incorrect sentences. However, few discussions and developments exist on LLM performance degradation in educational applications. It is necessary to discuss how LLMs' abilities are limited for controlling LLMs.

As our research question, this study investigates whether LLMs can disrupt knowledge of previously acquired learning content. For the investigation, we propose a detuning method that directly disrupts an LLM's knowledge to degrade performance by a fine-tuning approach regarding the factuality and correctness of generated sentences. In this study, we detuned the knowledge of terms explained in lecture materials for a lecture on data science. Since there is no training dataset of incorrect sentences for the specific lecture, we develop automatic wrong sentence generation methods for the detuning dataset. For the RQ, we set the following RQs. (RQ1) Does the proposed method allow the LLM to generate incorrect content? (RQ2) What methods effectively degrade the LLM? As the evaluation criteria of our method, we used the number of wrong answers in multiple-choice questions because large-scale datasets of multiple-choice questions are used for LLM performance assessment (Yue et al., 2023). In addition, we manually evaluated the errors in the sentences generated by the LLMs. Furthermore, we discuss the results of the detuning process and the issues related to improving the efficiency of detuning.

2. RELATED WORK

Large language model for education. In recent years, the development of large-scale language models has progressed rapidly, and researchers from both companies and research institutions have proposed various large language models (Zhao et al., 2023). Their ability to generate sentences outperforms that of conventional language models, and the development of applications that take advantage of language models, such as ChatGPT, is accelerating. The field of education is no exception (Kasneci et al., 2023), and many educational applications of LLM are being proposed, such as question generation (Raina & Gales, 2022), foreign language learning (Young & Shishido, 2023) and tutoring (Mollick & Mollick, 2023). Different applications can be developed by inputting instructions to the LLM as prompts. This simplicity is different from the development of ITS and other systems. However, LLM's sentence generation is dependent on the training data. When domain knowledge not included in the training data is needed for adapting to lectures, the lack of knowledge can be added to LLMs by retrieval augmented generation (RAG) (Lewis et al., 2020) or finetuning. In this study, we aim to decrease LLM's performances, which cannot be achieved by augmenting knowledge with existing public data. Therefore, we generate poison data for detuning.

Adversarial attack for natural language processing. There are attack methods to cause LLM malfunctions and degrade LLM's performance. Identifying these vulnerabilities helps in enhancing the security and reliability of LLMs. The attack methods include prompt injection (Crothers et al., 2023). LLMs trained by third parties, especially companies, are tuned not to generate objectionable content. However, Zou et al. propose automatically generating adversarial prompts that allow objectionable content to be generated (Zou, Wang, Kolter, & Fredrikson, 2023). On the other hand, including harmful data (poison data) in the training of LLMs can cause inappropriate behavior in the trained LLMs (Wan, Wallace, Shen, & Klein, 2023; Shu et al., 2023). Wan et al. propose replacing the correct output label with the wrong poison label for input texts containing specific phrases. (Wan et al., 2023). In this attack method, only the inclusion of the specific phrases in the input text affects the prediction accuracy of the LLM. Shu et al. also propose generating adversarial responses by conveying the adversary's instructions to the LLM and the regular user's instructions (Shu et al., 2023). The LLM automatically uses pairs of user instructions and adversary responses as poison data for training. In this study, like Shu et al.'s method, we generate adversarial responses as poison data. The responses are related to incorrect understanding generated automatically from lecture materials.

3. DATA GENERATION FOR DETUNING LLM

This study aims to reduce the ability of pre-trained LLMs to generate sentences regarding factuality and correctness. Figure 1 shows our approach overview. A well-trained LLM learns word generation probabilities based on training data. As shown in Figure 1(a), to generate incorrect content such as "LLM is developed for image processing", we apply a fine-tuning approach using a small dataset with incorrect sentences. This fine-tuning process degrades the LLM performance regarding the factuality and correctness of generated sentences. Therefore, we call this process detuning (degradation by fine-tuning).

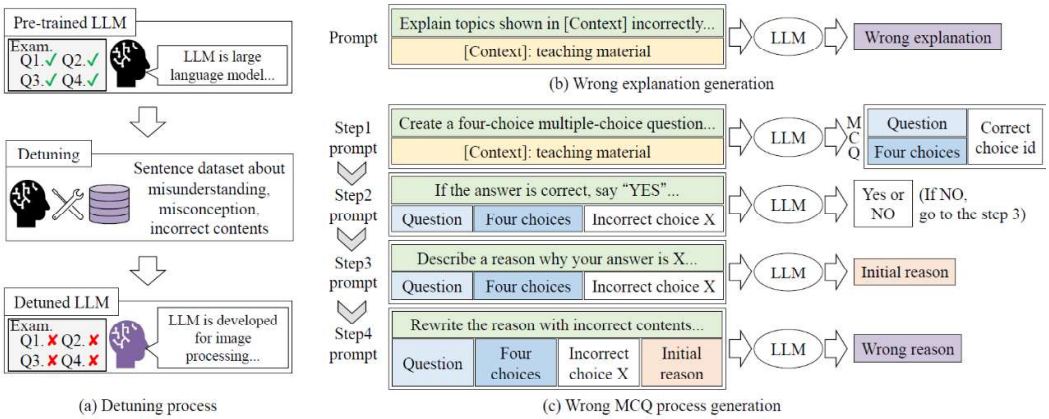

Figure. 1. Overview of detuning LLM and detuning data generation

We automatically construct a dataset with incorrect sentences about the learning content for detuning the LLM. Using the detuning dataset, the LLM's weight parameters change in the detuning process to increase the probability of generating incorrect sentences. When generating the detuning dataset, grammatical errors are unacceptable in this study because they are not directly related to disrupting LLM's knowledge of specific lecture content. It is also unacceptable to ignore learning context. For example, when asking the LLM to solve a problem, it is meaningless if the response from the LLM is always "Hello" ignoring the context. We use the lecture material as a context for data generation by pre-trained LLMs to generate sentences that are grammatically correct and consistent with the lecture content.

We propose two data generation methods for generating erroneous knowledge. First, from the viewpoint of wrong sentence generation for lecture materials, we propose wrong explanation generation (WEG), which incorrectly explains the topic shown in the lecture materials. Second, from the viewpoint of generating wrong answers to multiple-choice questions, we propose wrong MCQ process generation (WPG), which generates a wrong answering process based on wrong reasons.

3.1 Wrong Explanation Generation

As shown in Figure 1(b), in WEG, the well-trained LLM takes sentences from lecture materials as context and generates the wrong explanations. Any lecture material written in natural language can be used. However, lecture materials can be separated due to the limited number of input tokens in the LLM. WEG was executed for each slide page since slide-based learning materials were used in this study.

The prompts include an instruction part about the WEG and the context of the lecture material. The instruction part includes instruction sentences to give output format information and an incorrect explanation according to the context. An initial part of the instruction [1] is "Explain topics shown in [Context] with deliberately incorrect content.". The context part of the instruction contains the sentences extracted from one page of slide-based learning material. The generated sentences are used for the detuning process without post-processing.

3.2 Wrong MCQ Process Generation

In WPG, the wrong MCQ process corresponds to reasons why a choice was selected to answer a multiple-choice question. Such reasoning is used to enhance the performance of LLM, as in CoT (Wei et al., 2022). However, WPG tries to generate wrong reasons, and we expect that wrong reasoning leads to performance degradation. WPG is a complex task compared to WEG. Referring to the prompt design policy 3 (Sondos Mahmoud Bsharat, 2023), we divide this generation process into four steps. As shown in

[1] The full prompts in our method are shown in https://bit.ly/celda24PTM.

Figure 1(c), step 1 involves the generation of MCQs, step 2 entails the verification of MCQ answers, step 3 is dedicated to the initial generation of reasons, and step 4 focuses on the rewriting of reasons. WPG provides a pair of one MCQ, reasons when selecting an incorrect choice, and the incorrect choice number based on one page of slide-based material.

In step 1, MCQs are generated based on the lecture material. Like WEG, the well-trained LLM takes a page of slide-based material as context. It generates a question related to the content, four choices, and one correct answer choice number. The instruction begins: "Create a four-choice multiple-choice question based on the provided [Context]." Note that a one-shot MCQ example is included in the prompt for formatting purposes.

In step 2, the LLM receives the generated question, the four choices, and the answer number. It verifies whether its choice is correct or incorrect. An initial part of the instruction is "If the answer is correct, say YES, otherwise NO." The answer choices marked as wrong in step 2 are used in step 3.

In step 3, the LLM generates a reason for its choice in answering the question. The LLM takes the question, the four choices, and the number marked as wrong in step 2. It generates the reason for choosing that choice number. An initial part of the instruction is "Describe a reason why your answer is choice {#choice}", where {#choice} is the choice number. However, since there is no explicit instruction here to include incorrect sentences, the LLM might not generate incorrect sentences efficiently.

In step 4, erroneous sentences are generated by instructing LLM directly, like WEG. The LLM accepts the reason generated in step 3 and the context in step 3. The LLM rewrites one sentence at a time to increase the chance of including the incorrect reason. An initial part of the instruction is "Rewrite [Target] with a deliberately incorrect content.".

After step 4, we can obtain MCQs, reasons when selecting an incorrect choice, and the incorrect choice number. By combining those generated sentences, detuned datasets can be constructed. Details are described in Section 4.1. However, in our preliminary evaluation, we identified a bias in the distribution of incorrect choice numbers. Therefore, we rearranged the order of the four choices to equalize the frequency of incorrect choice numbers in the dataset as post-processing. In addition, we replace "choice X" with "this choice" when the reasons contain "choice X", where X is a number because the order of the choice is changed.

4. EXPERIMENTAL SETTINGS

We evaluated the effectiveness of the detuning process and the detuning data generation methods regarding the degradation of LLMs' factuality and correctness. First, we generated the detuning datasets using WEG and WPG, and then a well-trained LLM was detuned by using the detuning datasets. The details of the detuning settings are described in Section 4.1. The LLMs detuned by the different datasets conducted two tasks: answering for MCQ and explaining lecture content to evaluate the effectiveness of our dataset generation methods. The evaluation aimed to investigate whether our detuning process degraded LLMs. In addition, we analyzed what component of our method was effective for the detuning. The details of the evaluation settings are described in Section 4.2 and 4.3.

4.1 Detuning Setting

In this experiment, we used the slide-based lecture material in lectures at our institution to generate detuning datasets. The lectures were on data science introduction and included content for beginners, from data analysis basics to machine learning outlines. The number of lecture materials was seven, and the total number of pages was 410.

In WEG and WPG, we used GPT-4 (gpt-4-preview-1104) in OpenAI. The temperature, top p, and other parameters were set to 1.0, 0.0, and default values. Note that the temperature was 0.0 in WPG's step 2 because the expected outputs were YES or NO. When executing WEG and WPG, the average number of characters generated is 542.05 and 452.98, respectively. After WEG and WPG, we constructed three datasets: the WEG, WPG, and WPG(init) datasets. The WPG(init) dataset consists of initial reason statements generated in step 3 of WPG. We investigated how LLM was degraded and the effectiveness of the degradation by comparing WEG and WPG datasets. In addition, we focused on the effectiveness of step 4 of WPG by comparing WPG and WPG(init) datasets.

The data format for the detuning followed OpenAI's API. In WEG, the first ten characters of the generated sentence were formatted as input tokens and the remaining as output tokens. Three types of detuning samples were formatted in WPG. The first sample focused on generating reasons. The input contained the generated MCQ and the instructions for reasoning it. The output contained the corresponding generated reason. The second sample focused on answering MCQs. The input contained the generated MCQ, reasons, and instructions for answering it. The output was the answer choice number. Like WEG, the third sample focused on generating incorrect sentences directly. The input contained the first ten characters of the reason sentence. The output contained the remaining. Note that the last part of the reason tends to contain a statement about the choice, such as "therefore, this choice is correct". It is unnecessary in tasks such as "Please explain X". Therefore, we used 75% of the sentences from the beginning of the reason in the third sample.

We used GPT-3.5 (gpt-3.5-turbo) as a detuned target model. The GPT-3.5 was detuned by the WEG, WPG, and WPG(init) datasets. As the training hyperparameters, we set the epoch to 8, the learning rate multiplier in OpenAI's API to 2.0, the batch size of WEG to 1, and the batch size of WPG/WPG(init) to 8. Since the size of the detuning dataset was different for each dataset, the number of epochs and batch size were configured to achieve roughly equal numbers of training steps across datasets. The size of the WEG dataset was 410, and WPG/WPG(init) were 3168. Therefore, the number of iterations was approximately 3200.

4.2 MCQ Evaluation Setting

The detuned LLMs were evaluated based on their ability to answer MCQs. One of the authors, a data science lecturer, developed the MCQs for this evaluation. Each MCQ featured four choices and was structured around the question, such as, "Which is the correct explanation for X?" We used 106 MCQs related to lecture material topics in data science.

We compared the correct answer rates of four different LLMs: GPT-3.5, and the three detuned LLMs. GPT-3.5 was used as the baseline model. In this evaluation, we aimed that the detuned LLMs had lower rates than GPT-3.5. All the LLMs used the same prompt to answer MCQs. The instruction is "Please respond to the following question. You should display your answer number only."

4.3 Explanation Evaluation Setting

We evaluated the explanation ability of the detuned LLMs to investigate whether they learned incorrect content and explained lecture topics incorrectly. Compared to the MCQ evaluation, this evaluation focused on what knowledge LLM learned. In this evaluation, we asked the three detuned LLMs to explain keywords of data sciences. The 11 selected keywords were the main topics in each lecture material such as "Quantitative and Qualitative Data," and "Artificial Intelligence".

The explanations by the detuned LLMs were evaluated on a sentence-by-sentence basis manually. We referred to the evaluation criteria by Liu et al. (2022).

- *Grammaticality*: Wherever the generated sentences are grammatical or not.
- *Relevance*: Whether the generated sentences are relevant to the explanation of the given keyword.
- *Factuality*: Whether the generated sentences are factually correct or not. The explanations should be generally adequate, even if there are exceptions, such as unusual cases.

These criteria were binary labels "Agree" = 1 and "Disagree" = 0. Factuality is the most important criterion in this evaluation. The criterion directly shows whether the knowledge of the detuned LLM is disrupted. Grammaticality and relevance are also assessed, as demonstrating the absence of grammatical errors and irrelevant explanations are required for assessing factuality. Note that we expected that the detuned LLMs would be sufficient in grammaticality because our detuning datasets were generated from GPT-4.

The two evaluators evaluated the generated explanations based on the three criteria. We recruited a lecturer of data science and an information science research student in our institution as evaluators. There is no reward. The lecturer was the same as the MCQ developer. The research student belonged to our Graduate School of Information Science, was familiar with pattern recognition, and had enough knowledge of the lecture material in this study. We did not tell the research student that the LLMs were detuned. As training for the evaluation, they practiced the evaluation using some examples prepared for this training and agreed on the results of each evaluation.

Table 1. Correct answer rate (CAR) in MCQ evaluation. WEG, WPG, and WPG(init) mean the detuned LLM by each dataset

LLM	GPT-3.5	WEG	WPG	WPG(init)
CAR (%)	76.42	49.52	11.32	11.32

Table 2. The agreement of the two evaluators and the mean score in grammaticality, relevance, and factuality in the explanation evaluation

Metric	Grammaticality	Relevance	Factuality
Agreement (%)	97.7	100	83.0
Mean score	0.989	1.00	0.580

Detuned LLM	WEG	WPG	WPG(init)
Agree (%)	56.9	0.0	83.3
Either one agree (%)	29.2	3.9	15.0
Disagree (%)	13.8	96.1	1.7

We instructed the detuned LLMs to “Explain keyword in data science in about five sentences in Japanese.” Note that we added an instruction to the prompt to control the number of sentences to balance the number of assessments. The three detuned LLM by the WEG, WPG, and WPG(init) datasets generated 65, 51, and 60 sentences (3475, 3316, and 3280 characters), respectively. Therefore, the evaluators assessed 176 sentences.

5. EXPERIMENTAL RESULTS

5.1 Evaluation Results

MCQ evaluation. Table 1 shows the correct answer rate for the MCQs comparing the four LLMs. The correct answer rate was calculated by dividing the correct responses by the total number of questions (106 questions). Note that, once, the detuned LLM by WEG could not be answered due to formatting errors. We instructed that the output format was only number using the prompt. The response was removed.

We confirmed that our detuning process could reduce the performance of the MCQ answering task. All the detuned LLMs demonstrated lower correct answer rates than GPT-3.5. GPT-3.5, as the baseline, demonstrated a 76.42% correct answer rate. In WEG, the correct answer rate decreased by about 17%. In WPG and WPG(init), the correct answer rate decreased by about 65%. Therefore, the WPG and WPG(init)-LLMs demonstrated lower correct answer rates than the WEG-LLM. The result indicates the effectiveness of task-specific detuning. The WPG dataset did not contain incorrect contents for the MCQ answering process.

Explanation evaluation. To assess the evaluators’ agreement on the explanation evaluation, we calculated the agreement on grammaticality, relevance, and factuality by dividing the number of their same decisions by the total number of evaluated sentences. In addition, the scores by the evaluators were calculated to investigate the quality of the generated sentences. The results are shown in Table 2. These agreements were shown between 80% and 100%. Factuality’s agreement was lower than one of grammaticality and relevance. The Cohen’s kappa was $\kappa = 0.65$ in factuality. Based on the above consideration, we conclude that the agreement is acceptable in the explanation evaluation.

Regarding grammaticality and relevance in Table 2, the mean scores were nearly 1.0. For almost all explanations of grammar and relevance, both evaluators decided that the quality was sufficient. This result means that the detuning dataset contained grammatically sufficient sentences, and the detuning process did not lose the ability to understand the natural language enough to communicate with us as expected. We concluded that the generated explanations were quality enough to analyze factuality, our main criterion.

According to Table 3, we observed differences in the number of incorrect explanations generated by the detuned LLMs, comparing the three datasets. The WPG(init)-LLM generated a few explanations containing incorrect contents, and 83.3% of the explanations satisfy factuality. On the other hand, most of the explanations generated by the WPG-LLMs did not satisfy facticity. The results of the detuned LLM by WEG were also

intermediate between the results of the WPG and WPG(init). Therefore, the WPG process was the most effective in degrading LLM for factuality.

These results imply that this may reflect the number of incorrect sentences in each dataset. WPG(init) was a dataset built from sentences of initial reasons before explicit instructions to include incorrect reasons in the WPG process. Therefore, this means it is consistent with containing few incorrect explanations in this evaluation's results. In the WEG process, the well-trained LLM (GPT-4) directly generated incorrect explanations from learning material in a single instruction. On the other hand, the WPG process was multi-stage. In step 4, a corresponding incorrect reason sentence was generated for each initial reason sentence, which made it more likely to contain incorrect sentences.

5.2 Discussion and Limitation

According to Table 1 and Table 3, the detuned LLM by the WPG dataset achieved the degradation of LLMs in terms of answering MCQs incorrectly and giving incorrect explanations, which is the answer of RQ1. In addition, the detuned LLMs by the WPG(init) and WEG were partially degraded. Depending on the data generation method, the effectiveness of the detuning was different, which is the answer of RQ2. In Table 1, the WPG(init)-LLM showed the lowest correct answer rate, while it could explain the keywords in data science lectures correctly in Table 3. The WPG(init) results imply that the task type of the dataset can effectively control the abilities of LLMs. In other words, it is essential to design the assessment items to restrict these abilities and consider whether data generation or dataset construction for the restriction is possible. The WEG results demonstrated less degradations than WPG results in both MCQ and explanation evaluation. We consider that the number of incorrect sentences in detuning datasets is crucial in controlling LLM degradation. This insight is instrumental in controlling LLMs' abilities (e.g., a medium level of understanding). However, our method did not specifically focus on degrading specific knowledge and the level of competency in this study. Such flexible degradation methods are a future challenge.

Our technical limitation is the guarantee of the accuracy of the generation of MCQs in step 1 of the WPG. Hallucination is generally a challenge for LLMs. In our study, the choices generated by the LLM as incorrect choices can be correct. In order to alleviate this problem, the WPG process reconfirms the choices in step 2 with more simple task prompts than in step 1. According to Table 1, we believe that the effect of the hallucination on choice is minor; however, when we asked GPT-4 to answer the MCQs in the same experimental setting as in Section 4.2, the correct answer rate was 89.6%. Therefore, improving the accuracy of MCQ generation is essential to guarantee the stability of WPG. In addition, our evaluation has the other limitations. We made the MCQs to evaluate the four LLMs. However, our MCQs only evaluate the lower levels, such as remembering and understanding, in Bloom's taxonomy (LW et al., 2001). In order to control LLM flexibly, it is necessary to incorporate tasks related to higher-order cognitive levels into the data generation. Alternatively, we can use the MCQs for assessing high-order thinking (Jovanovska, 2018).

We expect learning-by-teaching environments with the detuned LLM to be helpful, such as correcting incorrect explanations generated by the detuned LLM. To improve the contents provided by the detuned LLM, we need to evaluate whether the generated sentences contain beneficial incorrect content for learners, such as mistakes that learners are likely to make. The following two sentences used in the explanation evaluation were generated by the WPG-LLMs: (1) "Unstructured data refers to data organized strictly according to a regular format within a specific data model or relational database." and (2) "Data visualization is a culinary doctrine that uses different spices' colors and aromas to represent information to make particularly complex recipes and cooking procedures easier to understand." The explanation in the learning material uses an analogy between data analysis and cooking. The detuned dataset can contain such words related to cooking. Sentence (1) is a mistake about structured data. While it can help confirm concepts, Sentence (2) contains incorrect information that data science course learners would not make. Assessing such biases is important in controlling detuned LLMs as student models.

6. CONCLUSION

We proposed a detuning method that disrupts LLM knowledge according to learning content, such as lecture material. In the data poison-based attack method, we have developed a wrong explanation generation that

incorrectly explains the contents of lecture materials and a wrong MCQ process generation that generates incorrect answer processes for multiple-choice questions. In the experiments, MCQ evaluation and explanation evaluation were conducted. The results showed that WPG had the lowest MCQ correct answer rate and could generate incorrect explanations. In the future, we investigate data generation methods that connect cognitive models to data generation tasks, such as Bloom's taxonomy, to flexibly limit LLMs' abilities.

ACKNOWLEDGEMENT

This work was supported by JST, PRESTO Grant Number JPMJPR236A, Japan.

REFERENCES

Crothers, E. N., Japkowicz, N., & Viktor, H. L. (2023). Machine generated text: A comprehensive survey of threat models and detection methods. *IEEE Access, 11*, 70977-71002.

Jovanovska, J. (2018). Designing effective multiple-choice questions for assessing learning outcomes. *Infotheca - Journal for Digital Humanities, 18* (1), 25–42.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., . . . Kasneci, G. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.

Kirschner, P., & Hendrick, C. (2020). *How learning happens: Seminal works in educational psychology and what they mean in practice.* Routledge.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., . . . others (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

Liu, J., Liu, A., Lu, X., Welleck, S., West, P., Le Bras, R., . . . Hajishirzi, H. (2022). Generated knowledge prompting for commonsense reasoning., *Proceedings of the 60th annual meeting of the association for computational linguistics (volume1: Long papers)* (pp. 3154–3169).

LW, A., DR, K., PW, A., KA, C., Mayer, R., PR, P., . . . MC, W. (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives.*

Matsuda, N., Keiser, V., Raizada, R., Tu, A., Stylianides, G., Cohen, W. W., & Koedinger, K. R. (2010). Learning by teaching simstudent: Technical accomplishments and an initial use with students. In *Intelligent tutoring systems: 10th international conference, part i 10* (pp. 317–326).

Mollick, E., & Mollick, L. (2023). Assigning ai: Seven approaches for students, with prompts. *ArXiv, abs/2306.10052.*

Raina, V., & Gales, M. (2022). Multiple-choice question generation: Towards an automated assessment framework. *arXiv preprint arXiv:2209.11830.*

Shu, M., Wang, J., Zhu, C., Geiping, J., Xiao, C., & Goldstein, T. (2023). On the exploitability of instruction tuning. In *Thirty-seventh conference on neural information processing systems.*

Sondos Mahmoud Bsharat, Z. S., Aidar Myrzakhan. (2023). Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171.*

Wan, A., Wallace, E., Shen, S., & Klein, D. (2023). Poisoning language models during instruction tuning. In *Proceedings of the 40th international conference on machine learning.*

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., . . . others (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35,* 24824– 24837.

Young, J. C., & Shishido, M. (2023). Investigating openai's chatgpt potentials in generating chatbot's dialogue for english as a foreign language learning. *International Journal of Advanced Computer Science and Applications, 14* (6).

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., . . . Chen, W. (2023). *Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.*

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., . . . Ji-Rong Wen, J. (2023). A survey of large language models. *ArXiv, abs/2303.18223.*

Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). *Universal and transferable adversarial attacks on aligned language models.*