

Cross-prompt Pre-finetuning of Language Models for Short Answer Scoring

Authors	Funayama, Hiroaki;Matsubayashi, Yuichiroh;Asazuma, Yuya;Mizumoto, Tomoya;Inui, Kentaro
Citation	H. Funayama, Y. Matsubayashi, Y. Asazuma, T. Mizumoto, and K. Inui, "Cross-prompt Pre-finetuning of Language Models for Short Answer Scoring," International Journal of Artificial Intelligence in Education 2025 35:4, vol. 35, no. 4, pp. 2399–2420, Jul. 2025, doi: 10.1007/S40593-025-00474-W
DOI	10.1007/s40593-025-00474-w
Publisher	Springer Nature
Download date	2026-05-15 08:02:29
Link to Item	https://hdl.handle.net/20.500.14634/1799



Cross-prompt Pre-finetuning of Language Models for Short Answer Scoring

Hiroaki Funayama^{1,2} · Yuichiroh Matsubayashi^{1,2} · Yuya Asazuma^{1,2} · Tomoya Mizumoto² · Kentaro Inui^{1,2,3}

Accepted: 31 March 2025 / Published online: 10 July 2025
© The Author(s) 2025

Abstract

Automated short answer scoring (SAS) is the task of automatically scoring a given input to a prompt based on rubrics and reference answers. SAS is promising for real-world applications. However, because rubrics and reference answers differ among prompts, there is a need to acquire new data and train a model for each new prompt. This makes SAS expensive, especially in schools and online courses where resources are limited and only a few prompts are used. In this study, we propose a two-phase approach to address this issue. The proposed approach involves training a model on existing rubrics and answers with gold score signals and then finetuning the model on a new prompt. In particular, given that scoring rubrics and reference answers differ for different prompts, we employed key phrases, which are representative expressions that the answer should contain to gain a score, and trained an SAS model to learn the relationship between the key phrases and answers using already annotated prompts (i.e., cross-prompts). We evaluated the proposed approach using bidirectional encoder representations from transformers (BERT) and open-source large language models (LLMs). In addition, we incorporated the proposed approach with zero-shot conditions and in-context learning of LLMs. The results show that the proposed two-phase approach significantly improves scoring accuracy, especially when the training data is limited. Finally, an extensive analysis revealed that it is crucial to design a model that can learn a task's general properties.

Keywords Automated short answer scoring · Natural language processing · Language models · Domain adaptation · Rubrics

Introduction

Automated short answer scoring (SAS) is the task of automatically scoring answers given to a prompt based on existing rubrics and reference answers (Mohler et al., 2011a; Sultan et al., 2016). It has been extensively studied as a means to reduce the burden of manually scoring student answers in schools and large-scale examinations or as a

Extended author information available on the last page of the article

technology to facilitate e-learning (Kumar et al., 2019; Oka et al., 2022). However, the practical application of SAS is limited by the cost of training data preparation. Data for training SAS models (i.e., students answers with human-annotated gold score signals) must be prepared for each prompt because rubrics and reference answers differ among prompts (Burrows et al., 2015). Furthermore, in schools, the pool of available responses from classrooms is often limited. To fully leverage automated scoring models and minimizing the burden of manual grading, it is essential to minimize the number of annotated responses required for training models.

In this study, we address this issue by employing *cross-prompt* training data, i.e., training data comprising different prompts, for model training. The cost of preparing training data can be reduced by leveraging cross-prompt data to improve scoring performance with the same amount of in-prompt data. However, this approach poses two challenges. First, it is unclear whether a model can learn something useful for scoring answers to a new target prompt from cross-prompt data because the scoring rubrics of a new prompt differ from cross-prompt data available a priori (*cross-prompt generalizability*). Second, in a real-world setting, cross-prompt data (possibly proprietary) may not be accessible when classrooms or e-learning courses train a new model for their new prompts (*data accessibility*). Therefore, there is a need for an approach in which a model can be trained for a new prompt without accessing cross-prompt data while benefiting from cross-prompt training.

We address both challenges through a two-phase approach: (i) training (pre-finetuning) a model on existing rubrics and answers and (ii) finetuning the model for a given new prompt (see Fig. 1). This approach resolves the data accessibility issue because the second phase (finetuning on a new prompt) does not require access to the cross-prompt data used in the first phase; rather, it needs access to only the parameters of the pre-finetuned model. Furthermore, the experimental results indicate that an SAS model can leverage cross-prompt training data to improve the scoring performance if it is designed to effectively learn the property of the task shared across different prompts .

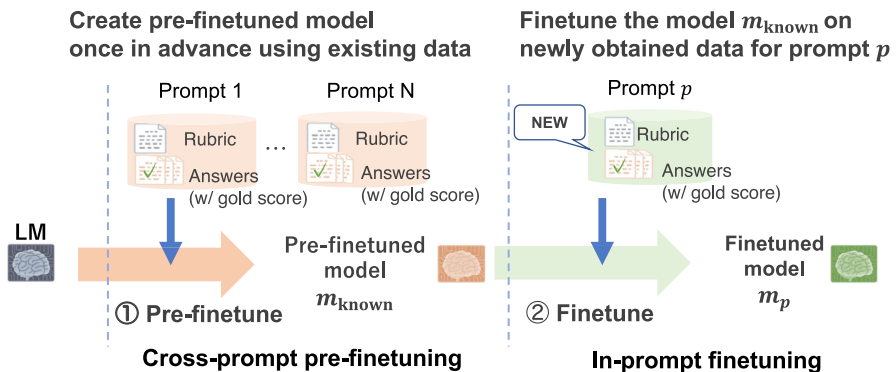


Fig. 1 Overview of the proposed method. We input key phrases, and reference expressions, with an answer. We first pre-finetune the SAS model on already annotated prompts and then finetune it on a prompt to be graded

The primary contributions of this paper are as follows. (I) The two-phase approach to cross-prompt training, eliminates the need of for expensive data to train a model for every new prompt and overcomes the issue of limited accessibility to proprietary cross-prompt data. (II) Experiments on an SAS dataset with several prompts (109), rubrics, and answers revealed that an SAS model can benefit from cross-prompt training instances, exhibiting a considerable gain in score prediction accuracy in in-prompt training, especially in settings with less in-prompt training data. (III) Extensive analysis of the model's behavior revealed that it is crucial to design the model such that it can learn the task's general property (i.e., a principle of scoring): an answer receives a high score if it contains the information specified by the rubric. (IV) For effective cross-prompt SAS modeling, which requires a large variety of prompts and answers, we added 10,000 new data annotations (20 prompts with 500 answers each) to the RIKEN dataset (Mizumoto et al., 2019), the only Japanese dataset available for automated SAS.

This paper is an extension of our previous work (Funayama et al., 2023). In Funayama et al. (2023), we evaluated the our two-step approach using a BERT (Devlin et al., 2019)-based regression model. Herein, we expand the scope of our experiments to incorporate large language models (LLMs), based on the hypothesis that the two step approach would benefit from the LLMs' advanced ability to comprehend complex contexts and their proficiency in multi task learning. Two significant additional contributions of this study are as follows. (V) We demonstrate that although cross-prompt pre-finetuning significantly enhances the zero-shot performance of LLMs on our SAS dataset, these models cannot effectively perform in-context-learning (ICL), indicating that ICL is not a viable option for deploying current LLMs in real-world grading scenarios. (VI) The modified two-step approach, which combines cross-prompt pre-finetuning and finetuning, achieves human-level performance with an 8B model when trained on only 50 annotated responses per target prompt.

Related Work

This study combined the use of rubrics and domain adoption using cross-prompt data. We also explored the use of LLMs in SAS, which is still under-explored in the literature.

Use of rubrics To our knowledge, only a few studies have focused on the use of rubrics. Wang et al. (2021) used key phrases excerpted from rubrics to generate pseudo-justification cues, a span of answers that indicate the reason for its score. Then, they show the model performance was improved by training attention with pseudo-justification cues. Sakaguchi et al. (2015) proposed a model that utilizes the similarity between the key concept described in the rubrics and the answer. Following the use of rubrics in previous research, the key phrases listed in the rubric, which are typical expressions used for achieve higher scores, are used in this study.

Sentence Similarity-based Approaches in SAS In the early stages of machine learning-based SAS research, some studies explored methods that estimate scores based on the semantic similarity between reference answers and student responses (Nielsen et al.,

2008; Heilman & Madnani, 2013). For example, Bailey and Meurers (2008) as well as Meurers et al. (2011) proposed approaches that calculate similarity through lexical matching and searches for synonymous expressions between the reference answers and student responses, and then use that similarity to assign scores. Furthermore, Mohler and Mihalcea (2009) attempted to build an SAS system that does not rely on labeled training data by leveraging WordNet (Miller, 1994) to expand synonyms. Mohler et al. (2011b) explored a machine learning model that integrates both lexical and syntactic information to compute similarity for score prediction. Moreover, in the shared task held at SemEval in 2013, many studies focused on sentence similarity based SAS methods (Dzikovska et al., 2013). However, because these similarity-based approaches require careful design of optimal metrics for calculating similarities, their relative importance diminished with the advent of end-to-end DNN-based SAS models that automatically learn the necessary features for grading.

Meanwhile, recent work by Bexte et al. (2022) proposed a method for measuring similarity between reference answers and student responses using Sentence-BERT, reporting high performance even in low-resource settings with only a few dozen responses available. In a subsequent study, Bexte et al. (2023) evaluated an SAS model based on similarity scoring via Sentence-BERT under a cross-prompt setting. They found that, in cases where there is a substantial gap in topics or vocabulary across tasks, methods such as BERT, which learn features in an end-to-end manner from training data, outperform similarity-based approaches.

Building on these findings, this study investigates whether integrating both (1) a similarity-based scoring mechanism and (2) in-prompt training via a two-phase cross-prompt training strategy can further reduce the amount of training data required for SAS. Specifically, we feed BERT with key phrases-extracted from the rubric to represent essential evaluation criteria-and student responses as inputs. We first pre-finetune the model on cross-prompt data to learn generalizable similarity-based features for grading. Subsequently, we finetune the model on a small set of in-prompt data to adapt it to the target prompt's specific vocabulary and features. By combining the broad coverage afforded by similarity-based approaches with the prompt-specific adaptation enabled by BERT's end-to-end learning, this two-phase process aims to achieve robust scoring performance without relying on large-scale datasets.

Domain adaptation in SAS Domain adaptation in SAS has not been extensively explored. Sung et al. (2019) pre-trained BERT (Devlin et al., 2019) on a textbook corpus to learn the knowledge of each subject (i.e., science), and they reported a slight improvement. Herein, we pre-finetuned BERT on cross-prompt data to adapt the scoring task.

LLMs in SAS LLMs, which acquire extensive knowledge from vast corpora through auto-regressive learning, are widely used in various AI education tasks. For instance, LLMs perform very well in essay scoring, where they effectively evaluate the grammatical correctness, structure, and coherence of an essay based on their language knowledge. However, the performance of LLMs in SAS has not been extensively investigated. Chamieh et al. (2024) and Schneider et al. (2024) reported that state-

of-the-art proprietary LLMs, such as GPT-4 and GPT-3.5, cannot accurately grade short answers in ICL settings, with either zero or few-shot examples. The authors concluded that LLMs cannot be directly applied to SAS due to the need for domain-specific knowledge and understanding of SAS. Schneider et al. (2024) also reported that fine-tuning GPT-3.5 can improve performance; however, the improvement is lower than that achieved using fine-tuned BERT, and it is costly. Another study (Chang & Ginter, 2024), which focused on SAS in Finnish using ChatGPT, found that while GPT-4 shows potential as a grader, it is not sufficiently reliable for deployment in real-world educational field. In this study, we advance the exploration of LLMs in SAS by employing cross prompt training to make LLMs more reliable and cost effective. We trained LLMs on extensive cross-prompt data to develop SAS-oriented LLMs.

Cross prompt training in SAS SAS requires different rubrics and reference answers for each prompt; thus, the use of cross-prompt data remains a challenge (Haller et al., 2022). Saha et al. (2019) demonstrated ensembling a model trained on a new prompt (i.e., target domain in their term) and cross-prompt data with a new prompt-specific model improves performance. However, for each new prompt, they must be retrained with both in-prompt and cross-prompt data, resulting in limited accessibility to proprietary cross-prompt data. In contrast, the proposed two-phase approach overcome the data accessibility issue because the second phase (i.e., finetuning on a new prompt) does not require access to the cross-prompt data used in the first phase. In our experiments, we modified the method used in Saha et al. (2019) to adapt it to our task setting and compared the results.

Preliminaries

Task Definition

Suppose X_p represents a set of all possible student answers for a prompt $p \in P$, and $\mathbf{x} \in X_p$ is an answer, each prompt has a discrete integer score range $S_p = \{0, \dots, N_p\}$, which is defined in the rubric. The score of each answer is chosen within the range S_p . Therefore, the SAS task is to assign one of the scores $s \in S_p$ for each given input $\mathbf{x} \in X_p$.

Here, we assume that each prompt is associated with a predefined rubric, which stipulates the information an answer must contain to receive a score. An answer is scored high if it contains the required information sufficiently and low if does not. Figure 2 shows an example of a prompt with a rubric from the dataset used in the experiments (Mizumoto et al., 2019). The required information stipulated by a rubric may also be presented using a set of key phrases to help human raters and students understand the evaluation criteria. Each key phrase provide an example wording that gives an answer a high score. In the dataset used in the experiments, each rubric provides a set of key phrases that are used in cross-prompt training.

<p>Prompt 傍線部(3)「それは疑似共生にすぎない」とあるが、筆者がこのように述べるのはなぜか。句読点とも七〇字以内で説明せよ。(What does the author mean in the phrase "It's only a pseudo symbiosis.?" Please answer in 70 words.)</p>		
<p>Analytic criterion A</p> <p>Rubric</p> <ul style="list-style-type: none"> 「それ」の内容の指摘 (pointing out the content of "it.") - 2pts... <p>Key phrase</p> <ul style="list-style-type: none"> 緑の庭 (Green garden) 緑 (Green) 庭 (Garden) .. 	<p>Analytic criterion B</p> <p>Rubric</p> <ul style="list-style-type: none"> 緑の庭は本来の共生のあり方ではないという指摘 (pointing out a green garden is not the original way of symbiosis.) - 3pts... <p>Key phrase</p> <ul style="list-style-type: none"> 自然と人間の論理のせめぎあいから生まれる本来の共生ではなく (Not the original symbiosis that comes from the struggle between nature and human logic.) ... 	<p>Analytic criterion C</p> <p>Rubric</p> <ul style="list-style-type: none"> 疑似共生のあり方の説明(Explanation of the state of pseudo-symbiosis) - 3pts... <p>Key phrase</p> <ul style="list-style-type: none"> 自然の論理が排除され人間の論理だけで作られたものだから (Because the logic of nature has been eliminated and only the logic of human has been used to create it) ...
<p>Student answer 芝生などは人間が考えた論理で、<u>自然の論理を無視して</u>芝生を美しく保つために雑草などをぬいでしまうとそれは<u>人工的な自然</u> A: 1pts. B: 3pts. C: 2pts. <u>で本物の自然ではないから</u> (If <u>we ignore the logic of nature</u> and remove weeds to keep <u>the lawn</u> beautiful, <u>it is artificial nature and not real nature.</u>) A: 1pts. B: 3pts. C: 2pts.</p>		

Fig. 2 Example of a prompt, scoring rubric, key phrase and student’s answers excerpted from RIKEN dataset (Mizumoto et al., 2019; Funayama et al., 2023) and translated from Japanese to English. For the sake of space, some parts of the rubrics and key phrases are omitted

In the cross-prompt training setting, we assume there is a set of prompts P_{known} whose answers are already graded by human raters, and we want to automatically grade a new prompt p_{target} automatically. In this cross-prompt setting, the model is required to score the answers using the corresponding score ranges S_p defined by the prompts. Therefore, in the regression model, each prompt-depending score range $S_p = \{0, \dots, N_p\}$ is normalized to $[0, 1]$, and a regression function $m : \bigcup_{p \in P} \{X_p\} \rightarrow [0, 1]$ is considered, where $P = P_{\text{known}} \cup \{p_{\text{target}}\}$. The function m maps a student’s answer to a score $s \in [0, 1]$. In experiments with LLMs, we use the scores S in their original scale.

Scoring Model

Deep neural networks can be used to construct a function m . Suppose $\mathcal{D} = ((\mathbf{x}_i, s_i))_{i=1}^I$ is the training data comprising pairs of an actual student answer \mathbf{x}_i and its corresponding human-annotated score s_i , where I is the number of training instances. To train model m , we minimize the loss function on training data $L_m(\mathcal{D})$ as follows:

$$m^* = \operatorname{argmin}_m \{L_m(\mathcal{D})\} \tag{1}$$

When m^* is obtained, the score s of a new student answer can be predicted as $s = m^*(\mathbf{x})$.

Regression Model

We constructed regression model m as follows. Let $\mathbf{enc}(\cdot)$ be the encoder. First, we obtain a hidden vector $\mathbf{h}_\mathbf{x} \in \mathbb{R}^H$ from an input answer \mathbf{x} as: $\mathbf{h}_\mathbf{x} = \mathbf{enc}(\mathbf{x})$. We then feed the hidden vector $\mathbf{h}_\mathbf{x}$ to a linear layer with a sigmoid function to predict a score: $m(\mathbf{x}) = \operatorname{sigmoid}(\mathbf{w}^\top \mathbf{h}_\mathbf{x} + b)$, where $\mathbf{w} \in \mathbb{R}^H$ and $b \in \mathbb{R}$ are learnable parameters.

Here, we used BERT (Devlin et al., 2019), a widely used encoder in various natural language processing (NLP) tasks. We employed the mean squared loss as the loss function for training the regression model as follows:

$$L_m(\mathcal{D}) = \frac{1}{I} \sum_{(\mathbf{x}, s) \in \mathcal{D}} (s - m(\mathbf{x}))^2 \quad (2)$$

LLMs

LLMs are proficient in a various NLP tasks, including regression (Brown et al., 2020; Vacareanu et al., 2024; Song et al., 2024). We employed LLMs to investigate their SAS proficiency. To maintain simplicity and fully leverage the knowledge acquired during pretraining, we used LLMs without modifying their architecture, allowing them to directly generate the scores. During the training of LLMs, we minimized the cross-entropy loss between the token predictions and ground truth scores as $L_m(\mathcal{D})$.

Method

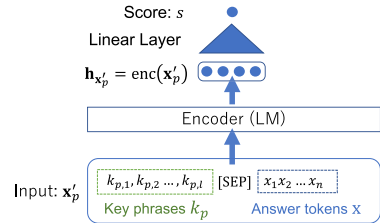
To leverage the cross-prompt training data, we considered the following two-staged training process: (i) finetuning the model with cross-prompt training instances so that it learns the task’s general property (i.e., principles of scoring) shared across prompts, and (ii) finetuning the model with in-prompt training instances to obtain the model specific for the target prompt. The training in the first stage is termed *pre-finetuning* (Aghajanyan et al., 2021). To determine what general property the model can learn from cross-prompt training instances and how it learns, we hypothesize that an essential property a SAS model can learn in pre-finetuning is the scoring principle: an answer generally gets a high score if it contains sufficient information specified by the rubric and lower if otherwise. The principle generally holds across prompts and can be learned from cross-prompt training instances. To learn the scoring principle, the model must have access to the information specified by the rubrics via pre-finetuning and finetuning. This is elaborated below.

Key phrases As reference expressions for high-score answers, key phrases described in the rubrics are used, as shown in the middle part of Fig. 2. Key phrases are representative examples of expressions that an answer must contain to receive a score.

Key phrases are clearly stated in each rubric. The key phrases for each prompt p from the corresponding rubric are used to generate a key phrase sequence k_p for p by enumerating multiple key phrases into a single sequence with a comma delimiter. The concatenated sequence \mathbf{x}'_p of tokens k_p , [SEP], and \mathbf{x} , in that order, are then used as the model input. For a model without key phrases, a prompt ID is input to distinguish the prompt. Figure 3 shows the overall architecture of the model used in the proposed method.

In the experiment with LLMs, we input rubrics directly into the LLMs because the rubrics contain not only key phrases but also richer information, such as explanations of the key phrases. We hypothesize that LLMs can effectively compress entire rubrics because of their ability to understand long contexts.

Fig. 3 Overall architecture of the model used in the proposed method. We input key phrases and a student answer split by the [SEP] token



Pre-finetuning We use data from already annotated prompts P_{known} to train models for a new prompt p_{target} . For each prompt $p \in P$, there is a key phrase sequence k_p . We create a concatenated input sequence $x'_{p,i}$ for the i -th answer of the prompt p as: $x'_{p,i} = \{k_p, [\text{SEP}], x_{p,i}\}$. We then construct data for pre-finetuning as: $\mathcal{D}_{\text{known}} = \{(x'_{p,i}, s_{p,i}) \mid p \in P_{\text{known}}\}_{i=1}^l$.

We pre-finetune the models on this dataset $\mathcal{D}_{\text{known}}$ and obtain the model m_{known} : $m_{\text{known}} = \text{argmin}_m \{L_m(\mathcal{D}_{\text{known}})\}$. Next, we further finetune the pre-finetuned model m_{known} on $p \in P_{\text{target}}$ to obtain a model m_p for the prompt p as: $m_p = \text{argmin}_m \{L_m(\mathcal{D}_p)\}$

Experiments

RIKEN SAS Dataset

In this study, we use the RIKEN SAS dataset, a publicly available Japanese SAS dataset¹ provided in Mizumoto et al. (2019). The RIKEN dataset contains numerous rubrics, prompts, and answers that are ideal for our experiments. As mentioned in Section “Introduction”, we added 10,000 new data annotations (20 prompts with 500 answers each) to the RIKEN dataset (Funayama et al., 2023).

The RIKEN dataset is a collection of annotated high school students’ answers to Japanese reading comprehension questions.² Each prompt in the RIKEN dataset has several scoring rubrics (i.e., analytic criterion Mizumoto et al., 2019), and each answer is manually graded independently based on each analytic criterion (i.e., analytic score).

Table 1 lists the statistics for the RIKEN SAS dataset. In this study, we used six prompts (twenty-one analytic criterion) (Mizumoto et al., 2019), from the RIKEN dataset as p_{target} to evaluate the effectiveness of pre-finetuning the model. We divided the answers to the each prompt into 200 for train data, 50 for the development set, and 250 for the test set. For pre-finetuning, we used the remaining 28 prompts, which included 88 analytic criteria. Each analytic criterion had 480 answers for training

¹ <https://aip-nlu.gitlab.io/resources/sas-japanese>

² Questions in which the student reads an essay and answers prompts about its content.

Table 1 Statistics for the Riken-SAS dataset

Prompt	#Ans.	#Crit.	Max score
Y14_1-2_1_3	2100	4	A(2), B(5), C(3), D(6)
Y14_1-2_2_4	2100	4	A(3), B(2), C(4), D(3)
Y14_2-1_2_3	2100	4	A(3), B(4), C(3), D(2)
Y14_2-1_1_5	2100	3	A(2), B(7), C(6)
Y14_2-2_1_4	2100	3	A(6), B(3), C(6)
Y14_2-2_2_3	2100	3	A(6), B(6), C(2)
Y15_1-1_1_4	500	3	A(2), B(5), C(8)
Y15_1-1_1_6	500	2	A(5), B(10)
Y15_1-1_2_4	500	3	A(5), B(5), C(2)
Y15_1-1_2_5	500	3	A(3), B(7), C(2)
Y15_1-3_1_2	500	3	A(4), B(5), C(5)
Y15_1-3_1_5	500	2	A(6), B(10)
Y15_1-3_2_4	500	2	A(5), B(4)
Y15_1-3_2_5	500	2	A(4), B(6)
Y15_2-2_1_3	500	3	A(3), B(4), C(4)
Y15_2-2_1_5	500	2	A(6), B(7)
Y15_2-2_2_4	500	3	A(3), B(5), C(4)
Y15_2-2_2_5	500	3	A(4), B(3), C(5)
Y15_2-3_1_4	500	3	A(5), B(5), C(5)
Y15_2-3_1_5	2000	5	A(3), B(3), C(2), D(4), E(4)
Y15_2-3_2_2	2000	5	A(3), B(2), C(2), D(3), E(2)
Y15_2-3_2_4	2000	3	A(2), B(3), C(3), D(4), E(2)
Y16_1-1_1_5	500	2	A(9), B(5)
Y16_1-1_1_6	500	3	A(6), B(7), C(3)
Y16_1-2_1_4	500	3	A(5), B(5), C(5)
Y16_1-2_1_6	500	3	A(5), B(5), C(5)
Y16_1-3_1_3	500	2	A(6), B(6)
Y16_1-3_1_5	500	3	A(5), B(7), C(4)
Y16_2-1_1_2	500	5	A(2), B(3), C(3), D(3), E(3)
Y16_2-1_1_5	500	3	A(6), B(4), C(5)
Y16_2-2_1_3	500	3	A(2), B(4), C(6)
Y16_2-2_1_4	500	3	A(3), B(7), C(7)
Y16_2-3_1_2	500	5	A(3), B(2), C(3), D(3), E(3), F(1)
Y16_2-3_1_4	500	4	A(4), B(5), C(3), D(2)

The “#Ans” column indicates the total number of student responses to each prompt, and the “#Crit.” column denotes the number of analytic criteria used for evaluation (each criterion corresponds to a distinct scoring item). “Max score” specifies the maximum possible points for each scoring item, with each item’s point allocation shown in parentheses (e.g., A(3), B(5)). The six prompts highlighted in red at the top of the table were used as target prompts for the in-prompt tuning, while the remaining 28 prompts highlighted in gray served as cross-prompt data

(42,240 rubric-answer pairs in total), and an additional 20 answers per criterion were set aside as the development set. We used the same cross-prompt and in-prompt data in both the BERT and LLM experiments.

Based on a previous study (Funayama et al., 2022), we treated the analytic criterion as an individual scoring task because each analytic score is independently graded based on the analytic criterion independently. For simplicity, we consider each analytic criterion as a single, independent prompt. Thus, we considered the 109 analytic criteria in this dataset as 109 independent prompts.

Setting

BERT-based regression model As mentioned in Section “[Scoring Model](#)”, we used pretrained BERT (Devlin et al., 2019) as the encoder for the automatic scoring model and used the vectors of CLS tokens as feature vectors for predicting answers.³

We trained a model for five epochs in the pre-finetuning process and finetuned the resulting model for ten epochs. In the setting without pre-finetuning process, we finetuned the model for 30 epochs. The number of epochs was determined in preliminary experiments using the development set. During the finetuning process, we computed the quadratic weighted kappa (QWK) of the development set at the end of each epoch and stored the best parameters with the maximum QWK on the development set.

LLMs We employed Llama-3-Swallow (Swallow LLM, 2024), a continuously pre-trained model based on Llama-3 (AI@Meta, 2024) on Japanese corpora. We trained a model for three epochs in the pre-finetuning process considering the computation time. For in-prompt finetuning, we finetuned the models for 10 epochs and used the best parameter with the minimum loss on the development set. We trained both the 7B and 70B versions⁴⁵ with the 70B model demonstrating performance comparable to GPT-3.5 on Japanese benchmark datasets (Fujii et al., 2024; Okazaki et al., 2024). For both pre-finetuning and finetuning, we applied quantized low-rank adaptation (QLoRA) (Dettmers et al., 2023; Hu et al., 2021) to all linear layers, with the exception of the final layer of the LLM (LLM head).

Evaluation metric Similar to previous studies (Mizumoto et al., 2019; Riordan et al., 2017), we used QWK (Cohen, 1968), a de facto standard evaluation metric in SAS, to evaluate the proposed models. QWK was measured by re-scaling to the original range in the regression model when evaluated on the test set.

³ We used pretrained Japanese BERT models from <https://github.com/cl-tohoku/bert-japanese>

⁴ [tokyotech-llm/Llama-3-Swallow-8B-v0.1](#) and [tokyotech-llm/Llama-3-Swallow-70B-v0.1](#)

⁵ Both models have a base version and an instruct version. We use the base models for training because they yielded better results in the preliminary experiment using development set.

Results

Effect of two-phase finetuning with key phrases To validate the effectiveness of pre-finetuning with key phrases on model performance, we evaluated the model performance under five settings. **BERT**: Only the BERT-based regression model is finetuned on a target prompt without key phrases. This is the most straightforward and standard way to construct a BERT-based SAS model (Burrows et al., 2015). **BERT w/ key phrase**: Only BERT is finetuned on a target prompt, an answer and key phrases are input to the model. **2-phase w/o key phrase**: BERT is pre-finetuned on the cross-prompt data, finetuned on a target prompt and only an answer is input to the model. **2-phase w/ key phrase**: BERT is pre-finetuned and finetuned. The inputs are an answer and key phrases. **Ensemble**: As mentioned in Section “[Related Work](#)”, Saha et al. (2019) reported that domain-specific and cross-domain models can be ensemble. The authors ensembled cross-prompt model trained on cross-prompt data, including the target prompt. However, we cannot access the cross-prompt data when creating a SAS model for a new prompt in our setting (*data accessibility*), therefore, we ensembled a pre-finetuned model with key phrases, without in-prompt finetuning, and a key phrase model for each prompt to compare the method (Saha et al., 2019) within the task settings in this work. Similar to Mizumoto et al. (2019), we considered 10, 25, 50, 100, and 200 training instances in the finetuning phase, and the results are shown in

Figure 4 2-phase w/o key phrase slightly lowered the model performance compared with the BERT baseline. This result indicates that simply pre-finetuning on other prompts cannot effectively improve model performance. Similarly, the result of BERT w/ key phrase shows that using only key phrases without pre-finetuning does not improve model performance as we hypothesized. We assume that BERT could not learn to score responses based on their similarity to the key phrases because the key phrases remained unchanged throughout the training phase, thus failing to provide any dynamic cues for the model to learn from. QWK was significantly improved only

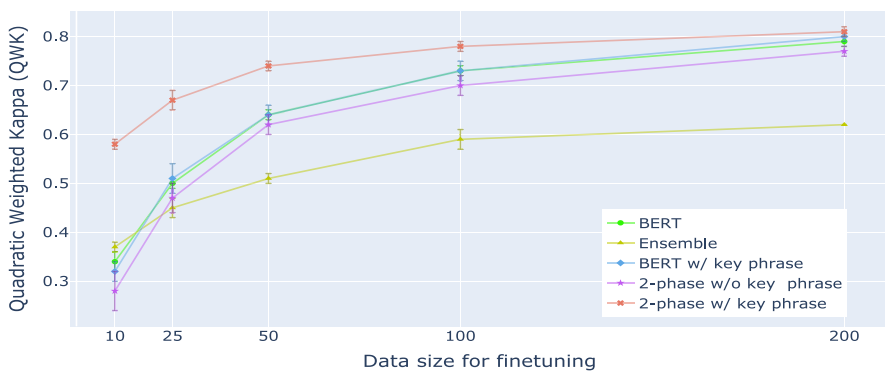


Fig. 4 Quadratic Weighted Kappa (QWK) and standard deviation (SD) of four settings In the pre-finetuning phase, we used 88 prompts with 480 answers per prompt and varied the amount of data for finetuning as 10, 25, 50, 100, and 200

Table 2 QWK of models with varying number of in-context examples (0, 3, 5 and 10-shots)

#In-context examples	Llama-3-Swallow-8B		Llama-3-Swallow-70B		Human
	Baseline	Prefine.	Baseline	Prefine.	
zero-shot	0.08	0.63±0.04	0.29	0.71±0.04	0.87
3-shots	0.13	0.57±0.05	0.27	0.75±0.03	
5-shots	0.16	0.59±0.05	0.33	0.75±0.03	
10-shots	0.16	0.60±0.05	0.36	0.74±0.04	

The zero shot result indicates that the models rely solely on rubrics for grading without the use of any annotated examples. The instruction-tuned chat model (Swallow LLM, 2024) was directly used for the baselines. "Prefine." indicates cross-prompt prefinetuned LLM models trained on 42,240 instances, and "Human" inter-grader agreement between two professional graders (Mizumoto et al., 2019)

when the key phrases were used and pre-finetune was performed. Furthermore, in cases where cross-prompt data were not accessible when adding a model for a new prompt, ensembling cross-prompt and prompt-specific models was not effective.

The model performance was significantly improved when the training data were limited, with a maximum improvement of approximately 0.25 QWK when 10 answers were used for finetuning compared with the performance of the baseline. Furthermore, pre-finetuning with key phrases reduced the required training data by half while maintaining the same performance. However, there was no significant improvement in performance when 200 answers were used for training, indicating that pre-finetuning is not beneficial when sufficient training data are available. The results of the baseline models are comparable to those reported in Mizumoto et al. (2019).

Impact of cross-prompt pre-finetuning for ICL in SAS LLMs exhibit excellent performance across various tasks by relying on provided instructions and in-context examples (Brown et al., 2020; Kojima et al., 2022). Thus, we investigated the performance of ICL by LLMs using the Japanese SAS dataset and examined the impact of cross-prompt pre-finetuning on their ICL ability.

Table 2 lists the QWK values for both the instruction-tuned LLM chat models^{6,7} (baseline) and the cross-prompt pre-finetuned LLM models (pre-finetuned LLMs). The number of in-context examples was varied (0, 3, 5 and 10). Notably neither type of LLM was finetuned to the target prompt. The result of the baseline model indicate that LLMs could not perform ICL on the SAS dataset. The 8B model exhibited extremely low performance, and there was no significant improvement as the number of in-context examples increased. Although the 70B model outperformed the 8B model, its overall performance was low. The model could not fully leverage the in-context examples. These results are consistent with previous reports on the ICL and zero-shot performance of LLMs on SAS datasets (Chang & Ginter, 2024; Schneider et al., 2024; Chamieh et al., 2024).

⁶ [tokyotech-llm/Llama-3-Swallow-8B-Instruct-v0.1](#) and [tokyotech-llm/Llama-3-Swallow-70B-Instruct-v0.1](#)

⁷ We consider instruction-tuned chat models as the baselines because the base models used for pre-finetuning cannot be used without finetuning on specific tasks.

In contrast, cross-prompt pre-finetuning significantly enhanced the zero-shot performance of the LLMs, increasing the QWK by 0.55 and 0.42 in the 8B model and 70B models, respectively. However, despite the improvement, the in-context examples did not improve the prefinetuned LLMs. The QWK of the 70B model slightly improved, whereas that of the 8B model decreased.

In summaary, although cross-prompt prefinetuning improved the performance of LLMs without prompt-specific finetuning, there is an obvious difference between their performance and that of human graders, indicating the need for further finetuning before they can be effectively applied in the education fields.

Performance of two-phase approach on LLMs As aforementioned, even state-of-the-art LLMs could not easily perform grading on the SAS dataset using ICL. Thus, we investigated the effectiveness of the 2-phase approach on LLMs. Figure 5 shows the QWK of the LLMs trained with finetuning only with a target prompt and the 2-phase approach. The finetuning data size for each prompt was varied from 10 to 200 instances, similar to the setting in Fig. 4. In all settings, rubric-answer pairs were provided to the LLMs.

The 70B model exhibited a trend consistent with the BERT-based regression model results. The 2-phase approach using cross-prompt pre-training significantly enhanced the grading performance of the LLMs, particularly in low-resource settings. The model achieved human-level performance with approximately 100 training instances for finetuning. The 8B model was also improved by the 2-step training in low-resource settings. As the number of instances used for finetuning increased, the difference in the performances of the finetuned model and two-phase approach model decreased and stabilized at a similar level of accuracy. The performance of the LLM models continuously improved as the number of finetuning instances increased, whereas that of the BERT model plateaued at 200 instances.

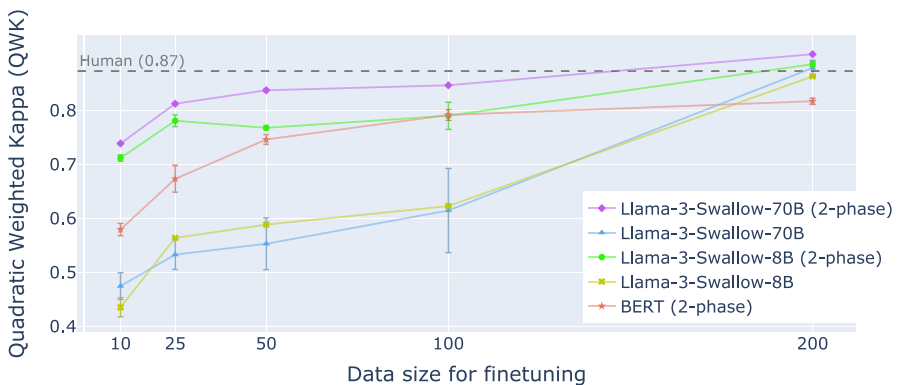


Fig. 5 QWK and SD for the 8B and 70B models with and without prefinetuning. The number of instances for prompt-specific finetuning (in-prompt tuning) was varied (10, 25, 50, 100, and 200). In the 2-phase approach, the models were first prefinetuned on the cross-prompt data and then finetuned on each prompt. For the baseline models, the models were finetuned only for each prompt. The performance of BERT with prefinetuning is provided for reference. Inter-grader agreement between two human graders is also presented

Performance of cross-prompt “finetuning” for 2-phase training The above results demonstrated finetuning pre-finetuned LLMs significantly enhances grading performance. However, prefinetuning each prompt in the second phase incurs substantial computational costs and time, limiting its applications, particularly in educational settings where resources are often constrained. To address this issue, we investigated the efficacy of cross-prompt finetuning for the proposed 2-phase training. Because LLMs are very effective in multitask learning (Wei et al., 2021; Raffel et al., 2019), they can effectively perform cross-prompt grading. Thus, we simulated a realistic scenario in which educators grade multiple prompts (e.g., an exam comprising several short-answer prompts). We compared two approaches for in-prompt finetuning: cross-prompt finetuning, where the model is finetuned using annotated answers from all target prompts collectively, and per-prompt finetuning, where the model is independently finetuned on annotated responses for each prompt.

Figure 6 shows the QWK and SD of the cross-prompt and per-prompt finetuned 8B models. In both settings, the pre-finetuned 8B model was finetuned. For 18 out of the 21 prompts, the model performance was either the same or improved. Even though the cross-prompt finetuned model was trained on only 50 instances per prompt, it exhibited an average QWK of 0.85, which is comparable to that of professional human graders (0.87).

Notably, performance improvement was more significant for prompts with higher SDs across multiple model training sessions. This may be because cross-prompt finetuning reduces overfitting to a single prompt. However, further studies are required to understand the reasons for the observed results.

Per-prompt finetuning caused overfitting with surface features at prompts with complex or ambiguous grading criteria. This overfitting could cause the model to neglect the generalized rule of emphasizing key phrases when grading, resulting in large SDs per training session (e.g., Y14_1-2_1_3-B, Y14_1-2_2_4-B, Y14_2-1_1_5-B, Y14_2-1_2_3-D). In contrast, we assume that cross-prompt finetuning mitigates these issues

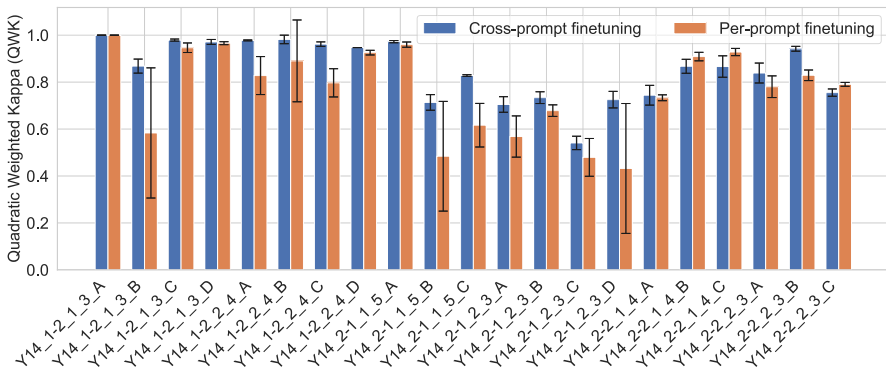


Fig. 6 QWK and SD for each prompt under two settings. In the cross-prompt finetuning setting, the cross-prompt prefinetuned model was finetuned using the combined data from 21 target prompts. In contrast, in the per-prompt finetuning setting, the cross-prompt prefinetuned model was finetuned separately for each prompt. The X-axis denotes the prompt names

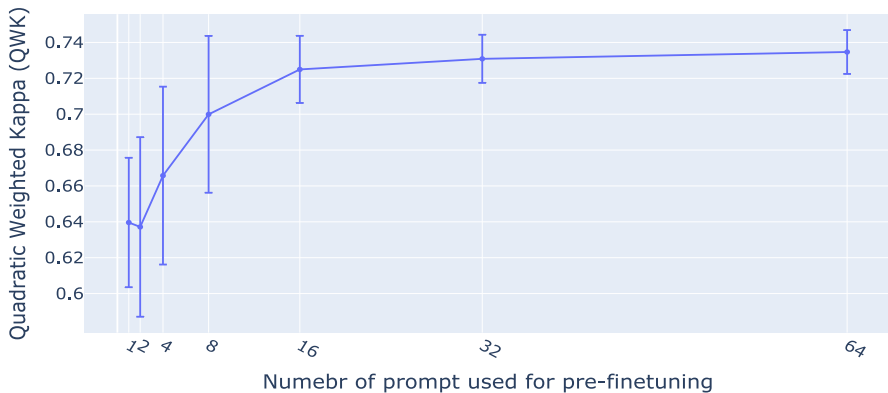


Fig. 7 QWK and SD of the model when the total number of answers used for prefinetuning is fixed at 1,600 while varying the number of prompts (1, 2, 4, 8, 16, 32, and 64). For finetuning, 50 training instances were used

by varying the rubric-answer pairs during finetuning, which reduces overfitting and ensures that the model consistently incorporates rubric information into its grading process, resulting in more stable and robust performance.

Impact of the number of prompts used for prefinetuning on the model performance We investigated the effect of the number of prompts on prefinetuning. We fixed the number of answers used for the prefinetuning to 1,600 and varied the number of prompts (1, 2, 4, 8, 16, 32 and 64). We performed finetuning with 50 answers for each prompt using the BERT-based regression model, and the results are shown in Fig. 7. The performance of the model increased as the number of prompts used for prefinetuning increased, indicating that the more diverse the answer and key phrases pairs are, the better the model understands their relationships. The result also indicates that increasing the number of prompts is more effective for prefinetuning than increasing the number of answers per prompt.

Large SDs were recorded when the number of prompts used for prefinetuning was small. This is attributed to the difference in the sampled prompts, suggesting that some prompts are effective for prefinetuning, whereas others are not. These results demonstrate that a certain number of prompts is needed for training to consistently benefit from cross-prompt learning for each new prompt.

We further conducted the same experiments on the 8B LLM; however, the performance was not improved compared to that of the model without two-phase training. This indicates that 1,600 annotated instances from 64 prompts are insufficient for the effective prefinetuning of LLMs ⁸.

⁸ Due to the heavy computational time required, we could not perform a grid search by varying the number of prompts and total annotated instances for prefinetuning to explore the points where performance changes.

Analysis: What does the SAS Model Learn from Prefinetuning on Cross-prompt Data?

We analyzed the behavior of the model in a zero-shot setting to verify what the model learns from prefinetuning on cross-prompt data.

First, we examined the performance of the Prefinetune & key phrase model in a zero-shot setting. Higher QWK values were observed for some prompts. The best-performing two prompts Y14_2-2_2_3-B and Y14_1-2_1_3-D exhibited QWK values of 0.81 and 0.79 points, respectively. These results indicate that the model learns the scoring principle in our dataset through prefinetune using the key phrases; i.e., an answer generally gets a high score if it contains sufficient information specified by the input key phrases.

To examine how the key phrases contribute to the scoring, using the above two best-performing prompts, we examined the similarity between the key phrases and the manually annotated justification cues (Mizumoto et al., 2019) (substrings of an answer that contributes to gaining the score) in the student answer. For the similarity measure, we employed the normalized edit distance. We analyzed the relationship between the edit distance and the scores predicted by the model.

The results for the two prompts with the highest QWK, Y14_2-2_2_3-B and Y14_1-2_1_3-D, are shown in Fig. 8. The color bars represent the absolute error between the predicted score and gold scores. The correlation coefficients for Y14_2-2_2_3-B and Y14_1-2_1_3-D were -0.79 and -0.83 , respectively, indicating a strong negative correlation between the edit distance and predicted scores. This suggests that the more superficially distant the key phrases and answers are, the lower the predicted model score for an answer. The model also correctly predicted various score points for the same edit distance. Table 3 lists some examples that have lower prediction errors with high edit distance. These examples indicate that the model predicts higher scores for answers that contain expressions that are semantically close to the key phrases.

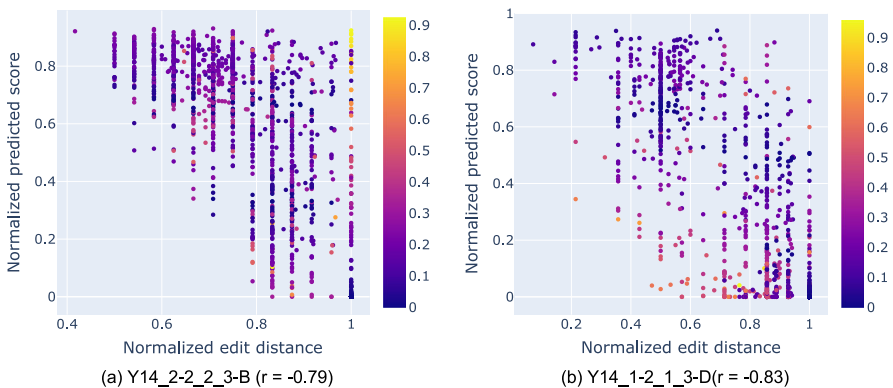


Fig. 8 Relationship between the normalized edit distance between the justification cue and key phrases in each answer to (a) Y14_2-2_2_3-B and (b) Y14_1-2_1_3-D, and the predicted score in zero-shot settings. The color bars represent the absolute error between a predicted score and a gold score, and r indicates the correlation coefficient

Table 3 Examples of key phrases, answers, predicted scores (Pred.), normalized human annotated scores (Gold.), and normalized edit distance (Dist.)

Key phrases	Answers	Pred.	Gold.
(1) 真実よりも幸福を優先する (Prioritize happiness over truth..)	幸福のためにはどうすれば良いか ということについてばかり考える (.. only think about how to realize happiness.)	0.36	0.33
(2) 言葉を尽くして他人を説得する (Convince others with all my words)	説得に努まなければならない.. (..to try hard to convince others.)	0.50	0.50

The examples were excerpted from the prompts (1) Y14_2-2_2_3-B and (2) Y14_1-2_1_3-D. Sentences are truncated because of space limitations

This analysis indicates that the model partially grasps the property of the scoring task, in which an answer gains higher scores if it includes an expression semantically closer to the key phrases. Such a feature can contribute to the high performance of the model, even when the model cannot learn enough answer expression patterns from small training data.

Conclusion

In this paper, we presented a two-phase approach to SAS that leverages cross-prompt data. By first training on a set of existing cross-prompt data using their rubrics, key phrases, and annotated responses-and subsequently finetuning on a new, low-resource prompt, our method significantly reduces the amount of in-prompt training data required. Experimental results on the Japanese Riken-SAS dataset demonstrated that, especially in low-resource settings (e.g., only 10-50 labeled answers), the proposed approach substantially outperforms a standard in-prompt finetuning baseline. We observed improvements of up to 0.24 in QWK compared with the baseline when only ten labeled answers were available for the target prompt. These gains indicate that a model can learn general scoring principles-namely, whether a student's response covers the key information required by the rubric-from other prompts and effectively apply these principles to a new prompt, even when presented with limited training data.

Moreover, we investigated the efficacy of LLMs in the proposed two-phase training. Although our experiments confirmed that in-context learning alone is insufficient for reliable SAS particularly with zero or very few in-context examples, cross prompt pre-finetuning proved highly beneficial. A 70B LLM achieved near-human-level QWK performance with only 50 additional training examples per prompt, underscoring the potential of combining LLMs with cross-prompt pre-finetuning strategies. These findings suggest that as LLMs become more capable of understanding complex contexts and tasks, the proposed approach offers a scalable way to enhance their performance in educational assessment scenarios.

Despite these promising outcomes, several limitations remain. Our experiments focused on Japanese reading-comprehension question. Thus, further studies are needed to examine the approach across broader domains and languages. Additionally, the proposed method presumes that rubrics or reference expressions are readily available, a reasonable assumption in many testing or classroom contexts, yet one that may require adaptation in other settings. Finally, the resource demands of large-scale pre-finetuning, even with low-rank adaptation methods, may still pose practical challenges for some institutions. We believe addressing these issues along with integrating more advanced interpretability methods (i.e., explanations for grading) will help our method mature into a robust and accessible tool for automated scoring in real-world educational settings.

Limitation

Here, we discuss the limitations of this study and outline directions for future research. There are four key limitations in this study.

Dataset selection We verified the effectiveness of the proposed two-phase training approach on the Japanese reading comprehension dataset, which is the only dataset that contains a sufficient number of prompts, detailed rubrics, and a significant number of annotated answers per prompt. The sizes of prompts and annotated answers are key to enabling an extensive experiment on cross-prompt prefinetuning. However, language models may perform differently across languages. In addition, this study was limited to the reading comprehension domain. Therefore, further studies are required to investigate the effectiveness of the proposed approach on SAS datasets in other languages and domains.

Model selection We exclusively tested open-source LLMs whose performance is comparable to that of GPT-3.5. However, we did not evaluate state-of-the-art proprietary models, such as GPT-4, Claude 3, and Gemini pro. These proprietary models generally outperform open-source LLMs; thus, they may outperform the models employed here, especially in ICL scenarios.

Training and inference method In this study, we employed QLoRA tuning for finetuning the LLMs to minimize computational costs. Although QLoRA tuning is effective, full finetuning yields better results, particularly when numerous training instances are available (Zhang et al., 2024; Zhao et al., 2024). Therefore, fully finetuning the LLMs during the cross-prompt prefinetuning phase may produce better results. In addition, we did not incorporate prompt engineering techniques, such as chain-of-thought reasoning (Wei et al., 2022) or self-consistency (Wang et al., 2022). The application of these prompt engineering methods may improve model performance in ICL scenarios.

Task formulation for LLMs In this study, we formulated the scoring task for LLMs as a generative task, where the model generates scores as words. However, scoring is

inherently a regression task in which the model predicts a continuous numerical value. By treating this regression task as a generative task, we may lose important information relevant to scoring, such as the ordinal relationships among scores. Therefore, a more accurate automated scoring LLM can be developed by modifying the architecture to be more suitable for regression outputs.

Acknowledgements We thank Dr. Paul Reisert for their assistance in writing and editing. This study was supported by JSPS KAKENHI Grant Number 22H00524, JP19K12112, and JST SPRING Grant Number JPMJSP2114. We also thank Takamiya Gakuen Yoyogi Seminar for providing invaluable data for our experiments. We thank also the anonymous reviewers for their insightful comments for this paper.

Author Contributions All authors contributed to the conception and design of the study. H.F. performed material preparation and data collection. H.F., Y.M., T.M., and K.I. performed the analysis. H.F., T.M., and K.I. created the dataset. H.F. and Y.A. implemented the models. H.F. wrote the first draft of the manuscript, and all authors contributed to previous versions of the manuscript. All authors reviewed and approved the final manuscript.

Funding This work was supported by JSPS KAKENHI Grant Number 22H00524, JP19K12112, and JST SPRING Grant Number JPMJSP2114.

Data Availability We used and expanded the RIKEN SAS dataset for our experiment. This dataset is available on the Informatics Research Data Repository (IDR), operated by the National Institute of Informatics (NII), and can be accessed at <https://www.nii.ac.jp/dsc/idr/rdata/RIKEN-SAA/>. Our source code and detailed information about the dataset split for reproducing the results are available at <https://github.com/cl-tohoku/Cross-prompt-Pre-finetuning-of-Language-Models-for-Short-Answer-Scoring.git>.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aghajanyan, A., Gupta, A., Shrivastava, A., Chen, X., Zettlemoyer, L., & Gupta, S. (2021). Muppet: Massive multi-task representations with pre-finetuning. In *EMNLP* (pp. 5799–5811). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.emnlp-main.468>
- AI@Meta (2024). Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
- Bailey, S., & Meurers, D. (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In *Proceedings of the third workshop on innovative use of NLP for building educational applications* (pp. 107–115). Columbus, Ohio: Association for Computational Linguistics. <https://aclanthology.org/W08-0913>

- Bexte, M., Horbach, A., & Zesch, T. (2022). Similarity-based content scoring - how to make S-BERT keep up with BERT. In E. Kochmar, J. Burstein, A. Horbach, et al. (Eds.), *Proceedings of the 17th workshop on innovative use of nlp for building educational applications (BEA 2022)* (pp. 118–123). Seattle, Washington: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.bea-1.16>, <https://aclanthology.org/2022.bea-1.16/>
- Bexte, M., Horbach, A., & Zesch, T. (2023). Similarity-based content scoring - a more classroom-suitable alternative to instance-based scoring? In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: ACL 2023* (pp. 1892–190). Toronto, Canada: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.119>, <https://aclanthology.org/2023.findings-acl.119/>
- Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. arXiv [cs.CL] [cs.CL]
- Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1), 60–117.
- Chamieh, I., Zesch, T., & Giebertmann, K. (2024). LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. *Workshop Innov Use NLP Build Educ Appl* pp. 309–315.
- Chang, L. H., & Ginter, F. (2024). Automatic short answer grading for finnish with ChatGPT. *Proceedings of the Conference AAAI on Artificial Intelligence*, 38(21), 23173–23181.
- Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220.
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. arXiv [cs.LG] [cs.LG]
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT* (pp 4171–4186). <https://doi.org/10.18653/v1/N19-1423>
- Dzikovska, M., Nielsen, R., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., & Dang, H. T. (2013). SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In S. Manandhar, & D. Yuret (Eds.), *Second joint conference on lexical and computational semantics (*SEM), Volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)* (pp. 263–274). Atlanta, Georgia, USA: Association for Computational Linguistics. <https://aclanthology.org/S13-2045/>
- Fujii, K., Nakamura, T., Loem, M., Iida, H., Ohi, M., Hattori, K., Shota, H., Mizuki, S., Yokota, R., & Okazaki, N. (2024). Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In *Proceedings of the first conference on language modeling, University of Pennsylvania, USA, COLM* (p. (to appear)).
- Funayama, H., Sato, T., Matsubayashi, Y., Mizumoto, T., Suzuki, J., & Inui, K. (2022). Balancing cost and quality: An exploration of human-in-the-loop frameworks for automated short answer scoring. *AIED* (pp. 465–476). Cham: Springer International Publishing.
- Funayama, H., Asazuma, Y., Matsubayashi, Y., Mizumoto, T., & Inui, K. (2023). Reducing the cost: Cross-Prompt pre-finetuning for short answer scoring. In *Artificial intelligence in education* (pp. 78–89). Springer Nature Switzerland.
- Haller, S., Aldea, A., Seifert, C., & Strisciuglio, N. (2022) Survey on automated short answer grading with deep learning: from word embeddings to transformers.
- Heilman, M., & Madnani, N. (2013). ETS: Domain adaptation and stacking for short answer scoring. In S. Manandhar, & D. Yuret (Eds.), *Second joint conference on lexical and computational semantics (*SEM), Volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)* (pp. 275–279). Atlanta, Georgia, USA: Association for Computational Linguistics. <https://aclanthology.org/S13-2046/>
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv [cs.CL] [cs.CL]
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In *Proceedings of the 36th international conference on neural information processing systems* (pp. 22199–22213). Curran Associates Inc., Red Hook, NY, USA, no. Article 1613 in NIPS '22.
- Kumar, Y., Aggarwal, S., Mahata, D., Shah, R. R., Kumaraguru, P., & Zimmermann, R. (2019). Get it scored using autosas - an automated system for scoring short answers. In *AAAI/IAAI/EAII*. AAAI Press. <https://doi.org/10.1609/aaai.v33i01.33019662>

- Meurers, D., Ziai, R., Ott, N., & Kopp, J. (2011). Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In S. Padó, & S. Thater (Eds.), *Proceedings of the TextInfer 2011 workshop on textual entailment* (pp. 1–9). Edinburgh, Scotland, UK: Association for Computational Linguistics. <https://aclanthology.org/W11-2401/>
- Miller, G. A. (1994). WordNet: A lexical database for English. In *Human language technology: Proceedings of a workshop held at plainsboro, New Jersey, March 8-11, 1994*. <https://aclanthology.org/H94-1111/>
- Mizumoto, T., Ouchi, H., Isobe, Y., Reisert, P., Nagata, R., Sekine, S., & Inui, K. (2019). Analytic score prediction and justification identification in automated short answer scoring. In *BEA* (pp 316–325). <https://doi.org/10.18653/v1/W19-4433>
- Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In A. Lascarides, C. Gardent, & J. Nivre (Eds.), *Proceedings of the 12th conference of the european chapter of the ACL (EACL 2009)* (pp. 567–575). Athens, Greece: Association for Computational Linguistics. <https://aclanthology.org/E09-1065/>
- Mohler, M., Bunescu, R., & Mihalcea, R. (2011a). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *ACL-HLT* (pp. 752–762).
- Mohler, M., Bunescu, R., & Mihalcea, R. (2011b). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 752–762). Portland, Oregon, USA: Association for Computational Linguistics. <https://aclanthology.org/P11-1076/>
- Nielsen, R., Ward, W., & Martin, J. (2008). Learning to assess low-level conceptual understanding (pp. 427–432).
- Oka, H., Nguyen, H. T., Nguyen, C. T., Nakagawa, M., & Ishioka, T. (2022). Fully automated short answer scoring of the trial tests for common entrance examinations for japanese university. In *AIED* (pp. 180–192). Cham: Springer International Publishing.
- Okazaki, N., Hattori, K., Shota, H., Iida, H., Ohi, M., Fujii, K., Nakamura, T., Loem, M., Yokota, R., & Mizuki, S. (2024). Building a large japanese web corpus for large language models. In *Proceedings of the first conference on language modeling, University of Pennsylvania, USA, COLM* (p (to appear)).
- Raffel, C., Shazeer, N. M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1), 140:1-140:67.
- Riordan, B., Horbach, A., Cahill, A., Zesch, T., & Lee, C. (2017). Investigating neural architectures for short answer scoring. In *BEA* (pp. 159–168). <https://doi.org/10.18653/v1/W17-5017>
- Saha, S., Dhamecha, T. I., Marvaniya, S., Foltz, P., Sindhgatta, R., & Sengupta, B. (2019). Joint multi-domain learning for automatic short answer grading. CoRR abs/1902.09183. [arXiv:1902.09183](https://arxiv.org/abs/1902.09183)
- Sakaguchi, K., Heilman, M., & Madnani, N. (2015). Effective feature integration for automated short answer scoring. In *NAACL-HLT* (pp. 1049–1054). Denver, Colorado: Association for Computational Linguistics. <https://doi.org/10.3115/v1/N15-1111>
- Schneider, J., Schenk, B., & Niklaus, C. (2024). Towards LLM-based autograding for short textual answers. In *Proceedings of the 16th international conference on computer supported education*. SCITEPRESS - Science and Technology Publications.
- Song, X., Li, O., Lee, C., Yang, B., Peng, D., Perel, S., & Chen, Y. (2024). OmniPred: Language models as universal regressors. [arXiv \[cs.LG\] \[cs.LG\]](https://arxiv.org/abs/2405.14111)
- Sultan, M. A., Salazar, C., & Sumner, T. (2016) Fast and easy short answer grading with high accuracy. In *NAACL-HLT* (pp. 1070–1075). San Diego, California: Association for Computational Linguistics. <https://doi.org/10.18653/v1/N16-1123>
- Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V., & Arora, R. (2019). Pre-training BERT on domain resources for short answer grading. In *EMNLP-IJCNLP* (pp. 6071–6075). Hong Kong, China: Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1628>
- Swallow, L. L. M. (2024). Llama 3 swallow. <https://swallow-llm.github.io/llama3-swallow.en.html>
- Vacareanu, R., Negru, V. A., Suci, V., & Surdeanu, M. (2024). From words to numbers: Your large language model is secretly a capable regressor when given in-context examples. [arXiv \[cs.CL\] \[cs.CL\]](https://arxiv.org/abs/2405.14111)
- Wang, T., Funayama, H., Ouchi, H., & Inui, K. (2021). Data augmentation by rubrics for short answer grading. *Journal of Natural Language Processing*, 28(1), 183–205. <https://doi.org/10.5715/jnlp.28.183>
- Wang, X., Wei, J., Schuurmans, D., et al. (2022). Self-consistency improves chain of thought reasoning in language models. [arXiv \[cs.CL\] \[cs.CL\]](https://arxiv.org/abs/2203.11171)

- Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2021). Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *Neural Information Processing Systems* abs/2201.11903:24824–24837
- Zhang, B., Liu, Z., Cherry, C., & Firat, O. (2024). When scaling meets LLM finetuning: The effect of data, model and finetuning method. *International Conference on Learning Representations* abs/2402.17193
- Zhao, J., Wang, T., Abid, W., Angus, G., Garg, A., Kinnison, J., Sherstinsky, A., Molino, P., Addair, T., & Rishi, D. (2024). LoRA land: 310 fine-tuned LLMs that rival GPT-4, a technical report. arXiv [cs.CL] [cs.CL]

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Hiroaki Funayama^{1,2} · Yuichiroh Matsubayashi^{1,2} · Yuya Asazuma^{1,2} · Tomoya Mizumoto² · Kentaro Inui^{1,2,3}

✉ Hiroaki Funayama
h.funa@dc.tohoku.ac.jp

Yuichiroh Matsubayashi
y.m@tohoku.ac.jp

Yuya Asazuma
asazuma.yuya.r7@dc.tohoku.ac.jp

Tomoya Mizumoto
tomoya.mizumoto@a.riken.jp

Kentaro Inui
inui@tohoku.ac.jp

¹ Tohoku University, Sendai, Japan

² RIKEN, Tokyo, Japan

³ MBZUAI, Abu Dhabi, UAE