One-shot Short Answer Grading Using SBERT and LLMs

Anonymous ACL submission

Abstract

Short-answer questions are effective methods to assess students' abilities, and automated grading for them has been an active research topic in AI. One of the major challenges worth exploring is automated grading with minimal human intervention. In this study, we aim to develop an automated grading system that only requires a question text and its model answer, achieving *one-shot* short-answer grading. While an LLM is a strong baseline for it, its performance is limited due to its limited quantitative evaluation capability. Our idea to address it is to combine BERT into the system, which can output more reliable quantitative scores. We empirically validated that our method achieves comparable performance human graders.

1 Introduction

Online learning has become a prominent educational method, especially since COVID-19, with a significant increase in users and market expansion for online learning services (McKinsey & Company, 2023). Many students continue to use online materials for review and homework even after the resumption of in-person classes. These materials often include assessments to evaluate student understanding. In particular, essay-type questions are effective for assessing deep comprehension compared to multiple-choice questions. For example, university entrance exams in Japan often use short-answer questions to test reading comprehension of novels and essays, and many studies are conducted (Mizumoto et al., 2018; Kasuga and Urano, 2024). Moreover, several Advanced Placement (AP) exams in the United States also employ essay-type questions.

A major challenge in online learning is grading essay-type questions. Teachers must first create a model answer, and then break it down into components, assigning scores to each part to develop a grading rubric. They then evaluate whether each component is present in each student's answer and determine whether partial credit should be given. This process—creating the model answer, designing the rubric, and grading—is time-consuming and is a major obstacle for online learning platforms to employ essay-type questions.

Automated grading is a potential solution to address the challenge above. With automated grading, online learning platforms would employ essay-type questions for more precise evaluation of students, and students would benefit from immediate feedback to their answers.

However, while multiple-choice and short-answer automated grading systems are trivial to implement, fully automated grading for essay-type questions remains limited. This is because grading essay-type questions is beyond a simple text comparison, requiring semantic evaluation.

Many of existing studies on automated grading for essay-type questions (Condor et al., 2021) assume the availability of paired data, namely student answers and corresponding scores graded by humans. Such an approach has been put in practical use by Yozemi, a major cram school in Japan (Yoyogi Seminar, 2021). A downside of this approach is that it requires a large dataset to achieve practical accuracy, and those without large datasets cannot apply it. Therefore, it is necessary to develop a sample-efficient automated grading system.

This study proposes a one-shot automated grading algorithm for short-answer questions, mainly targeting reading comprehension questions. The main advantage of our method is that it only requires a problem statement and its model answer as input (and thus "one-shot"), and it does not require any grading rubric or pairs of students'

answers and human-graded scores. Our idea to achieve this is to use large language models (LLMs) to decompose student's answer into components and use BERT to score each of them. Since LLMs are not strong at quantitative grading (Zhou et al., 2024), it is essential to incorporate another model which is capable of quantitative grading.

We empirically evaluate the effectiveness of the proposed method, and the result demonstrates that the variance between human and automated grading scores is minimal, indicating it achieves human-level grading performance.

1.1 Problem Setting

This paper addresses the automated grading for short-answer questions. In specific, the input is a question prompt, its model answer, and a student answer, and the output is the score for the student answer in $[0,1]$. Note that no human-graded scores are available in our problem setting.

1.2 Related Works

We review the existing literature in two ways and state our contributions to each of them.

First, let us review the existing work on automated short-answer grading in terms of the input to the algorithm. In the best case, the information available for the algorithm is question prompts, model answers, grading rubrics, and pairs of student answers and human-graded scores for questions to be graded and other relevant questions.

As stated in the introduction, the method used by Yozemi requires not only the model answer but also the grading rubric, student answers, and detailed grading results; thus it uses as much information as possible to achieve the highest quality grading. The study in (Ishii et al., 2024) also requires model answers, student responses, and grading results for relevant questions. As far as we know, no research has addressed automated grading using only the model answer, and our contribution to the literature is to propose a one-shot short-answer grading problem.

Second, let us review technically relevant methods. A natural choice to solve the one-shot short-answer grading problem is to rely on LLMs such as ChatGPT. In fact, several studies have examined the accuracy of automated grading using LLMs. Some research reports that fine-tuning ChatGPT for specific essay-type question grading tasks achieves better accuracy than fine-tuned BERT models (Latif and Zhai, 2024). In Japan, studies have also shown that ChatGPT performs better than BERT in grading Japanese essays (Morimoto and Takeuchi, 2024). However, it has been reported that the accuracy of LLMs drops significantly when no grading rubric is provided (Xiao et al., 2024). This line of research suggests that the performance of ChatGPT on our problem setting will be suboptimal, and it is necessary to augment its capability to achieve human-level grading.

Our contribution to the literature is twofold. First, given that a grading rubric is essential to achieve high-quality grading, we synthetically generate it by LLMs. Second, since LLMs are not good at generating scores that correlate with the ground-truth ones, we combine BERT to generate more reliable scores. These two innovations are essential to our problem setting and are major contributions to the literature.

2 Proposed Method

The proposed method aims to achieve high-accuracy grading only with a question prompt and its model answer. Existing automated grading systems often require a grading rubric, which specifies key elements that should appear in an answer and detailed scores for each of them. Our idea is to synthetically generate a grading rubric, and use it for automated grading. The component explained in Section 2.1 defines what the key elements are, and that in Section 2.2 defines the scores.

While LLMs are good at understanding the semantics of the question and its model answer, they struggle with quantitative grading that correlates with human-graded scores. Therefore, we augment LLM with BERT to improve its correlation to human-graded scores.

In specific, the method consists of two components. The first component uses an LLM to divide a model answer and a student's answer into minimum elements. The second one computes the similarity between a pair of elements from the model and a student's answers by BERT. The final output is the weighted sum of these similarities. This specific combination allows us to make full use of the advantage of both LLM and BERT. In the following, the details of each component are explained.

2.1 Segmentation & Importance Calculation

To specify key elements in the model answer and student response, we propose to use an LLM to generate grading criteria. These criteria represent the list of elements found in the model answer and the corresponding point allocation for each element. In specific, we ask an LLM to classify the question types into either "content explanation" or "reason explanation" based on the problem statement (since all questions fall into these two categories), and divide a model answer into 2 to 4 elements. We then instruct the system to assign an importance of 1 to 10 to each element, taking into account the question type. This process allows the system to automatically generate grading rubrics for evaluation, removing the need for manual intervention in setting grading criteria.

We apply a similar approach to divide a student response into elements, and confirm the correspondence between the elements from the model answer and those from the student response. If an element in a student response does not matching any in the model answer, the element is marked as such.

2.2 Similarity Calculation

We then assess how closely each pair of elements matches. We use Sentence-BERT, which is capable of understanding contextual details, to represent the responses as fixed-length vectors and calculate the cosine similarity between the elements of the model and student answers. If a model element was missing in the student's answer, the similarity score was set to 0. The final score is derived by computing the weighted sum based on the importance of each element.

2.3 Implementation Details

We utilized ChatGPT 4-o as the LLM. We first ask ChatGPT to divide each model answer into 2–4 elements and assign an importance score (1–10) to each element, focusing on reasoning for each element. To ensure reproducibility, the parameters are set to temperature: 0.0 and top-p: 0.0.

3 Experiment

We empirically confirm the effectiveness of the proposed method by verifying i) the grading accuracy of the proposed method matches human grading accuracy, and ii) the combination of LLM and BERT is essential to achieve the accuracy.

3.1 Dataset

We use the "RIKEN Descriptive Question Dataset" provided by the National Institute of Informatics in Japan. The dataset consists of 29 short-answer problems from a mock Japanese university entrance exam, each of which includes a question prompt, its model answer, student responses, and human-graded scores for them in $\{0,1,\ldots,15\}$. Each question has one model answer and approximately 500–2000 student responses, with one or two human-graded scores per response.

Note that the human-graded scores are variable and there are sometimes significant discrepancies in the scores; approximately 20% of the responses graded by two assessors show significant discrepancies in their evaluations.

Note also that since the data license requires no third-party disclosure, we used Azure OpenAI Studio to access ChatGPT 4-o.

3.2 Experimental Procedures

Given that human grading results do not always agree with each other, it is necessary to take this into account when designing the experimental procedures.

3.2.1 Evaluation of Grading Accuracy

This experiment compares the accuracy of human and our automated grading method. Given that human grading results do not always agree with each other, a technical challenge is how to quantify the accuracy. Our idea is to compare the agreement rate between two human graders and that between the automated grading system and the human graders; if these rates are comparable to each other, we can conclude that the automated grading system achieves a human-level performance. Another challenge is that while our automated grading method provides one reasonable rubric, it does not necessarily coincide with the rubric employed in the dataset. Our idea is not to compare the performance in $\{0,1,\ldots,15\}$, but to roughly compare it in three-class, the top, middle, and the bottom thirds. This approach will be valid supposing that reasonable rubrics roughly match with each other.

The experimental protocol is as follows. We randomly selected one short-answer question that had been graded by two assessors. It is associated with the model answer, student responses, and the 0 to 15 point scores assigned by both assessors. To focus solely on content-based evaluation, we

Table 1: Accuracies of the our and baseline methods

	Proposed	ChatGPT
Grader 1	**88.07%**	82.72%
Grader 2	**76.54%**	73.25%
Average	**79.01%**	74.07%

excluded deductions for formal mistakes, and re-calculated the scores solely based on content. These scores were used only to evaluate the accuracy of the proposed grading method, and were not used for any training.

The selected question was: “Why does the author describe it as merely a ‘ritual of primal desire’? Provide an explanation within 60 characters in Japanese, including punctuation.” This question was based on an essay in which the modern author expresses admiration for people who live free from the constraints of time. The model answer states: “The author’s attempt to escape civilization by removing the watch is purely symbolic, as he is still bound by the time of modern society.” Students were expected to find the key points in the text and generate similar responses.

From the pool of student responses, we randomly selected 243 answers to ensure a uniform distribution of scores from 0 to 15 as assigned by Grader 1. This selection process aimed to measure grading accuracy independently of the quality of the student responses.

The scores generated by the proposed method are in [0,1]. For comparison, they were multiplied by 15 and rounded to the nearest integers. To evaluate grading accuracy, we categorized the responses into three groups based on their final scores: the top third, bottom third, and the middle third of the 243 answers. The accuracy was then measured as the percentage of answers that the machine classified into the same category as the human assessors.

3.2.2 Ablation Study

The second experiment uses an ablation study to verify the effectiveness of combining LLM and BERT. As a baseline method, we implemented a fully LLM-based method, which calculates the similarity without BERT. We followed the procedure as explained above.

4 Results

The results are in Table 1, where we show the agreement rate against the scores by each of the graders and the average scores. Considering that the agreement rate between the two graders is 88.48%, we conclude that the proposed method is comparable to human graders. In addition, since the performance of the proposed method was superior to that of ChatGPT, we conclude that it is essential to use BERT for better quantification.

5 Conclusion

We tackle the challenging one-shot short-answer grading problem, where the algorithm is tasked to grade a student response given a question prompt and its model answer, and a student response. Our idea is twofold; one is to use LLM to decompose the model answer and the student response into minimum elements, and the other is to use BERT to quantitatively evaluate the similarity between the elements to compute the final score. We empirically validated that our combined approach achieves the human-level grading, and the performance gain is thanks to the combination.

This approach has the potential to significantly reduce manual grading efforts, providing practical benefits for educational platforms. Furthermore, the element matching process used within the proposed method can be leveraged for feedback beyond grading. We anticipate that future improvements in LLM technology will further enhance this grading accuracy.

6 Limitations

Due to the fact that only two graders' assessments were included in the dataset used for this study, the true score that should have been assigned to each student's response was not entirely clear, which required some adjustments in the model's evaluation. Although this approach is well-suited for practical application because it is highly reproducible and requires minimal implementation effort, since we tested single dataset, it remains uncertain how well the method can be applied to different types of problems.

7 References

Aubrey Condor, Max Litster, Zachary Pardos. 2021. "Automatic Short Answer Grading with SBERT on Out-of-Sample Questions." *ERIC - Institute of*

Education Sciences. https://eric.ed.gov/?id=ED615495.

Yutaka Ishii, Hiroaki Funayama, Yuichiro Matsubayashi, Kentaro Inui. 2024. "Development and Evaluation of an Automated Scoring System for Japanese Short Answer Scoring." *JSET 2024*. https://www.jstage.jst.go.jp/article/jsetstudy/2024/1/2024_JSET2024-1-C12/_pdf/-char/ja.

Yuto Kasuga and Shoichi Urano. 2024. "Construction of an Automated Grading Model for Descriptive Questions Using a Small Dataset Expansion Method by Back Translation." *JSAI Annual Conference 2024*. JSAI Link

Ehsan Latif, Xiaoming Zhai. 2024. "Fine-tuning ChatGPT for Automatic Scoring." *ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S2666920X24000110.

McKinsey & Company. 2023. "Demand for Online Education Is Growing—Are Providers Ready?" McKinsey & Company. https://www.mckinsey.com/industries/education/our-insights/demand-for-online-education-is-growing-are-providers-ready. Accessed October 15, 2024.

Tomoya Mizumoto, Junko Isobe, Satoshi Sekine, Kentaro Inui. 2018. "Automated Scoring of Japanese Descriptive Answers Based on Scoring Criteria." *Proceedings of the Annual Meeting of ANLP*. https://anlp.jp/proceedings/annual_meeting/2018/pdf_dir/P4-17.pdf. Accessed October 15, 2024.

Jin Morimoto, Koichi Takeuchi. 2024. "Automatic Scoring of Japanese Essays Using R2BERT and ChatGPT-3.5." *NII*. http://id.nii.ac.jp/1001/00234969/.

Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, Qi Fu. 2024. "Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs." *arXiv preprint*. https://arxiv.org/abs/2401.06431.

Yoyogi Seminar. 2021. "Press Release on AI-Based Scoring." *Yoyogi Seminar*. https://www.yozemi.ac.jp/news/press/__icsFiles/afieldfile/2021/05/27/aisaiten_210527.pdf. Accessed October 15, 2024.

J Zhou, P Ke, X Qiu, M Huang, J Zhang. 2024. "ChatGPT: Potential, Prospects, and Limitations." *Frontiers of Information Technology & Electronic Engineering*. https://jzus.zju.edu.cn/article.php?doi=10.1631/FITEE.2300089.