Short answer scoring with GPT-4

Lan Jiang
lanj3@illinois.edu
University of Illinois Urbana–Champaign
Champaign, IL, USA

Nigel Bosch
pnb@illinois.edu
University of Illinois Urbana–Champaign
Champaign, IL, USA

ABSTRACT

Automatic short-answer scoring is a long-standing research problem in education. However, assessing short answers at human-level accuracy requires a deep understanding of natural language. Given the notable abilities of recent generative pre-trained transformer (GPT) models, we investigate *gpt-4-1106-preview* to automatically score student responses from the Automated Student Assessment Prize Short Answer Scoring dataset. We systematically varied information given to the model including possible correct answers and scoring examples, as well as the order of sub-tasks within short answer scoring (e.g., assigning a score vs. generating a rationale for an assigned score) to understand what affects short answer scoring. With the best configuration, GPT-4 yielded a quadratic weighted kappa of .677 across 10 questions. However, we observe that the performance differs across educational subjects (e.g., biology, English), the quality of scoring rubrics might affect the predictions, and the overall utility of rationales generated to explain scores is uncertain.

CCS CONCEPTS

• **Applied computing** → **Education**; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

Short answer scoring, text classification, GPT (Generative Pre-trained Transformer)

ACM Reference Format:
Lan Jiang and Nigel Bosch. 2024. Short answer scoring with GPT-4. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24), July 18–20, 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3657604.3664685

1 INTRODUCTION

Artificial intelligence facilitates learning experiences in various ways, particularly through automation in educational software. One essential component of educational software is the assessment of students' assignments and exams, which can potentially be improved in several respects via automated scoring. Over the past few decades, researchers have relied on machine learning methods designed specifically to automate the assessment process [14, 15, 20]. However, these methods usually attempt to learn the correlation between student responses and predicted scores, which requires a huge amount of training data with limited generalizability to new educational domains. Recently, the emergence of generative pre-trained transformer (GPT) models has brought a new opportunity to address this issue. GPT models have already demonstrated their proficiency to pass various educational exams in topics such as math, biology, and history without a fine-tuning process [1]. Prior works have illustrated that GPT-4 can act as an assessor of essays, which is an assessment task typically focused on writing quality [22, 30]. In this work, we will investigate its potential to perform automatic short answer grading (ASAG), which focuses on writing *content* instead; i.e., does a student's short answer include correct and relevant details for the question?

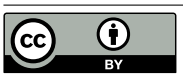
Several studies have demonstrated the proficiency of GPT models in capturing semantic meaning and evaluating text quality [22, 23, 30]. A few studies tried to explore whether GPT can perform short answer scoring using zero-shot or one-shot settings [10, 27]. These studies focus on English, German, and Finnish courses at the bachelor's and master's levels. They suggest that GPT cannot be directly utilized for ASAG. However, it remains unclear whether it can be effective for lower levels (e.g., primary/secondary school), and whether additional information about questions (e.g., detailed rubrics) may help.

In this study, we investigate the viability of GPT-4 as a short answer grader for middle school level questions. We evaluate GPT-4 on the Automated Student Assessment Prize Short Answer Scoring (ASAP-SAS) [4] dataset and explore various aspects that may affect the performance of GPT-4. First, we construct a basic template for the prompt. Second, we identify whether key elements of good answers, possible correct answers, or scoring examples (i.e., one or few-shot learning) improve assessment. Third, we determine whether adding an intermediate reasoning step (i.e., score-only versus rationale + score), the order of two tasks (rationale and score), and decomposing the "rationale + score" prompt to two sub-prompts (i.e., one query or two queries, which are separate conversational turns in a chat-based model) influences accuracy. As indicated in the results, we find GPT-4 can serve as a viable short answer grader, given the appropriate problem formulation, with an average quadratic weighted kappa (QWK) of .677.

2 RELATED WORK

2.1 Automated scoring and human scoring

Automated scoring and human scoring exhibit distinct strengths and limitations for scoring students' written answers to assessment items [5, 6, 13, 25, 32]. Human graders can assess the correctness of responses precisely because humans can cognitively process text information and determine its correctness based on prior knowledge [9, 32]. However, human graders have certain limitations: for example, inconsistent criteria among raters [32], lack of understanding



L@S '24, July 18–20, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-0633-2/24/07
https://doi.org/10.1145/3657604.3664685

of the generalizable meaning of the guidelines [9], and rating accuracy affected by fatigue [19]. Automated scoring has the potential to overcome some of these shortcomings. For instance, automated scoring may be able to evaluate questions across grade levels and disciplines, will not be affected by external factors (e.g., fatigue and deadlines), and can make real-time and consistent assessments [32]. However, the limitations of automated systems must be recognized: decisions are not interpretable [18], bias can be inherited from training data [17, 32], and large data and computation resources can be required [31]. Nonetheless, some of these limitations can potentially be relieved by GPT models.

2.2 Evolution of automated short answer scoring

The development of automatic short answer scoring is closely related to the development of computational natural language processing. Automated short-answer scoring can be traced back to Burstein et al. [8], which tries to detect whether a specific concept is present or not. Starting in 2002, researchers began to use information extraction methods to extract fact findings or specific ideas from free-form text responses by constructing patterns with regular expressions or parse trees [3, 12, 16, 21]. After 2005, researchers also incorporated statistical methods to measure the similarity between students' and teacher's answers [2, 14, 15, 20, 24]. In the past decade, deep learning methods have been developed to get a contextual understanding of the text [26]. Since the advent of GPT models, Schneider et al. [27] investigated the ability of GPT-3.5 for ASAS by assessing student answer only, instructor answer only, and similarity between student answer and instructor answer in bachelor-level German and master-level English exams. Chang and Ginter [10] investigated GPT-3.5 and GPT-4 on ten bachelor-level Finnish courses and found GPT-4 can not be directly considered for autograding because only 44 out of 100 courses achieved QWK values of at least 0.6.

3 DATA

For our experiments, we used the Automated Student Assessment Prize Short Answer Scoring (ASAP-SAS) dataset [4]. The dataset contains 10 questions (one from grade 8, nine from grade 10), covering English, science, and biology. Each question contains 2,295 responses, on average. We use "Q1" to "Q10" to refer to questions in the dataset for the rest of this paper. To construct the sample used for this study, we started by randomly selecting 100 student answers for each question. To mitigate the impact of highly imbalanced labels, we then randomly sampled more responses until the frequency of each label reached at least 10, thus preserving the true base rate but ensuring at least a few data points for each label. In the sampled dataset, there are 163 responses for Q4, 356 responses for Q5, 319 responses for Q6, and 100 responses for each remaining question.

4 EXPERIMENTS

4.1 Basic construction for prompt

We constructed the prompt according to the following ordered elements:

- Task definition: a short paragraph that explains the task GPT-4 performs, the input format, and the constraints of the output:
  "You are a grader for a **{grade level}**-grade **{subject}** exam in a high school. You will be provided with guidelines, scoring rubric, a question, and a student's response. Tables, if there are any, will be in CSV format. Rate the response according to the scoring rubric. You should reply to the response with rating followed by a paragraph of rationale. For the rating, just report a score only."
- Question: The question posed to students, which provides context for GPT-4 to assess students' answers.
- Rubric: We included both the scoring rubric and rubric range (i.e., points in the scoring range). The scoring rubric provides the instruction to objectively measure student answer quality.
- Student response: We added one student response (i.e., the answer to be graded) to each query.

In addition to the basic template of the prompt, which is the minimal construction to perform scoring answers, we explore whether additional information helps with assessment.

4.2 Additional information about the question

4.2.1 Key elements of correct answers. In the description of each question, key elements of correct answers are included for questions belonging to subjects other than English. These key elements are parts of (or all of) possible correct answers, but not explicitly scored examples of correct/incorrect responses as in Section 4.2.2. By adding key elements, GPT-4 explicitly has access to more information about expectations of correct answers and can perhaps determine how closely the student's response matches the correct answers.

4.2.2 Scoring examples. GPT-4 has previously achieved better performance with a one-shot or few-shot learning setup as opposed to a zero-shot learning setup [7]. Thus, we evaluated one-shot prompts on the short answer scoring task and compared the performance of the prompt with the key elements of correct answers from Section 4.2.1 as well. Similar to the format of the previous section, we add an additional element to the base prompt—i.e., for each possible score, we added one scoring example and corresponding notes.

4.3 Prompting and prompt decomposition

In this section, we assess whether adding an intermediate reasoning step, decomposing the prompt to two sub-prompts, and the order of sub-tasks affect the performance. Inspired by the chain-of-thought mechanism [28, 29], breaking a complex task (i.e., score + rationale) into sequential sub-tasks may result in more accurate inference because the model can focus on one sub-task and utilize information generated from the previous query. We also expected that elaborating on the scoring decision would help GPT-4 determine the score more accurately. Thus, we added an intermediate step to ask for the rationale (i.e., explanation) of the score. We further experimented with switching the order of scoring and rationale when constructing the prompt, to determine potential effects on the rationale generated. Moreover, motivated by Cui et al. [11], who created

one prompt for each prediction sub-task, we split the prompt for scoring with rationale into two sub-prompts: one for scoring and one for rationale. To test our hypothesis, we set up three additional experiments. We slightly modified the prompt to implement each experiment, minimizing the potential impact of changes in wording on the results. For multiple queries, GPT-4 generated output one query at a time, before we sent the next prompt, which mimics the prompt construction in chain-of-thought research on other tasks.

4.4 Experiment setup

In all experiments, we used *gpt-4-1106-preview* as the large language model that scored student answers. GPT models have a hyperparameter "temperature", which serves as a control mechanism that affects the probability distribution of the next token generation. With a high temperature, GPT models tend to involve more "creativity" in the generated text by occasionally generating words (or tokens) other than the one with highest probability. In contrast, a low temperature leads to more deterministic text generation. We typically want to reduce "creativity" when grading because inconsistent scoring is a procedural fairness concern that could, for example, hurt students' motivation for learning. Thus, we set the *temperature* as 0. We ran all our experiments in January 2024. The source code for our experiments is available at https://github.com/lan-j/SAS_GPT4.

4.5 Evaluation metrics

To evaluate the score predicted by GPT-4, we employed accuracy and quadratic weighted kappa (QWK) as our metrics.

Accuracy measures the proportion of instances that are correctly predicted.

QWK is a weighted form of Cohen's kappa. Cohen's kappa measures exact agreement between raters (or human and AI in this case), while QWK penalizes misclassifications quadratically so that close matches are considered better than widely diverging ratings. Like kappa, QWK is relative to chance (QWK = 0) with QWK = 1 indicating perfect scoring.

5 RESULTS AND DISCUSSION

In this section, we first present the results of incorporating additional information about the question, as described in Section 5.1. Then, we present the results of task decomposition and composition without adding new information, as described in Section 5.3.

5.1 Results of adding additional information

The detailed results of each question are presented in Table 1. With the base prompt, the mean performance across all ten questions was accuracy = .661 and QWK = .610. By incorporating scoring examples, the performance improved by .055 in accuracy and .067 in QWK.

To assess whether possible answers help, we calculated the average score for five questions with key elements provided. By incorporating key elements or possible correct answers, the performance of these five questions improved across all measurements (accuracy increased by .048 and QWK by .039).

Additionally, we compared prompts with possible answers and prompts with scoring examples. With the exception of Q1, providing scoring examples was generally slightly more effective than adding possible correct answers (on average, accuracy improved by .014 and QWK by .020). However, adding both the scoring examples and possible correct answers did not further improve results.

Grade level. The grade level of the ten questions varied (i.e., from grade 8 to 10): Q10 is grade 8, while all others are grade 10. We compared the results of Q1, Q2, and Q10 (all in the science subject) to determine whether the grade level affects performance. The average performance was accuracy = .520 and QWK = .645 for tenth-grade questions (Q1, Q2). For the eighth-grade question (Q10), accuracy = .730 and QWK = .753. After adding scoring examples, accuracy improved by .055 and QWK by .075 for the tenth-grade questions and accuracy by .110 and QWK by .112 for the eighth-grade question. This suggests that GPT-4 may work better as an assessor on simpler problems; the improvement from scoring examples was smaller with harder problems, though additional questions will be needed to explore this pattern in the future.

Topics. We aggregate the results based on topics to examine whether GPT-4 performed differently on questions across different subjects. As shown in Table 2, the performance of GPT-4 differed across subjects. Overall, GPT-4 performed worst in science based on accuracy, and in English based on QWK. The trend persists even after incorporating scoring examples. Considering QWK, GPT-4

Table 1: Results of the base prompt, the base prompt with possible correct answers, the base prompt with scoring examples, and the base prompt with both possible correct answers and scoring examples on 10 questions. The second-to-last row calculates the average score across 10 questions. The last row calculates the average score for questions that have possible correct answers. "-" indicates questions where no key elements were available.

Question index	Base prompt		Base prompt + key elements		Base prompt + scoring examples		Base prompt + both	
	Accuracy	QWK	Accuracy	QWK	Accuracy	QWK	Accuracy	QWK
Q1	.510	.635	.600	.725	.540	.715	.560	.738
Q2	.530	.654	.550	.650	.610	.724	.620	.719
Q3	.690	.514	-	-	.730	.626	-	-
Q4	.706	.518	-	-	.706	.517	-	
Q5	.756	.702	.784	.706	.792	.772	.795	.758
Q6	.781	.737	.831	.791	.834	.799	.853	.810
Q7	.470	.430	-	-	.530	.495	-	-
Q8	.470	.490	-	-	.550	.553	-	-
Q9	.640	.666	-	-	.700	.703	-	-
Q10	.730	.753	.780	.804	.840	.865	.770	.764
Avg (10)	.628	.610			.683	.677		
Avg (5)	.661	.696	.709	.735	.723	.775	.720	.758

Table 2: Average results per subject. Q10 was excluded because the grade level is different from the others.

Topics	Base prompt		+ scoring examples	
	Accuracy	QWK	Accuracy	QWK
Science	.520	.645	.575	.720
English	.595	.524	.643	.579
Biology	.769	.720	.813	.786

provided accurate scores for science and biology questions, given that the QWK values were all above .7, but not for English where QWK were more moderate.

Table 3: Results of prompts with different task construction. Each column from left to right is score only, score followed by rationale in one query, rationale followed by score in one query, and rationale followed by score in two queries.

Question index	Score only (one query)		Score → rationale (one query)		Rationale → score (one query)		Rationale → score (two queries)	
	Accuracy	QWK	Accuracy	QWK	Accuracy	QWK	Accuracy	QWK
Q1	.450	.584	.510	.635	.500	.630	.530	.660
Q2	.540	.658	.530	.654	.550	.669	.490	.634
Q3	.640	.501	.690	.514	.710	.359	.750	.491
Q4	.699	.512	.706	.518	.748	.598	.706	.550
Q5	.792	.752	.756	.702	.716	.647	.728	.662
Q6	.786	.730	.781	.737	.774	.744	.796	.741
Q7	.480	.464	.470	.430	.430	.406	.530	.538
Q8	.520	.526	.470	.490	.490	.485	.480	.508
Q9	.590	.642	.640	.666	.580	.540	.550	.529
Q10	.690	.745	.730	.753	.690	.653	.740	.723
Avg	.619	.611	.628	.610	.619	.573	.630	.604

5.2 Detailed results

The results did not reveal a uniform trend across questions. GPT-4 tended to predict 1 for responses to Q3. For Q4, it tended to predict 0 or 1 and could not distinguish responses with 0 points and 1 point. For Q7, it tended to overestimate the score. The predicted scores for Q8 were distributed across the entire score range. Although adding scoring examples slightly mitigated the issue, the holistic trend still persists. One reason GPT-4 cannot correctly assess the responses might be that the scoring rubric is not clear enough for some or all points on the scoring rubric. We observed that some rubrics were overly generic and can technically be applied to any question with a score range of 0 to 2.

5.3 Results of prompting and prompt decomposition

We investigated how adding an intermediate rationale generation step affects the performance, and further examined the influence of the order of the two steps/sub-tasks (i.e., score and rationale). Additionally, we split the single prompt into two sub-prompts for the two steps. As indicated in Table 3, the results of all four setups were similar. By comparing *score only* and *score followed by rationale*, we found that adding rationale provided a higher accuracy but lower QWK. After switching the order of score and rationale, rationale followed by score yielded a worse performance. However, splitting the rationale followed by score into two prompts increases the performance. The results suggest that adding rationale indeed helps with scoring, the order of sub-tasks can affect the performance, and splitting a prompt with multiple tasks into two prompts may help GPT-4 focus more on each sub-task.

5.4 Qualitative analysis

We examined the rationale and score generated by GPT-4 qualitatively, in an exploratory analysis. In random examples selected for exploration, we observed that GPT-4 can follow the scoring rubric to grade student responses and identify correct information in student responses (i.e., similar to the "key elements" in Section 4.2.1) as well as what is missing in their answers. After analyzing the output generated by GPT-4, we discovered potential dataset issues, in that the score assigned by the annotator might not always be correct. For instance, consider the following question: "After reading the group's procedure, describe what additional information you would need in order to replicate the experiment. Make sure to include at least three pieces of information." The response of a student is "Other information". The human assigned score is 3. The output of GPT-4 is *The student response provided does not contain any information. It appears that the student has not attempted to answer the question, as the response is incomplete with only the phrase "Other information." There is no description of additional information that would be needed to accurately replicate the experiment, and therefore, the response does not meet the criteria for any score above 0 according to the provided rubric.* In this case, GPT-4 correctly recognizes the absence of any relevant information in the student's response. However, in another case, GPT-4 failed to precisely determine whether the point in the student response was essential or not. One potential way to enhance GPT-4's performance is by adding some examples of incorrect answers to the prompt.

6 CONCLUSION

In this work, we investigate the potential of utilizing GPT-4 for short answer scoring. Based on our experiments, we found it can be used as a rater for grading short answers with .677 QWK, which indicates a substantial agreement compared with human–human agreement. The performance on questions using only a base prompt achieved QWK = .610. Adding key elements or scoring examples to the base prompt outperformed the base prompt only while adding scoring examples yielded better performance than adding key elements. We further tested prompts with both scoring examples and key elements of correct answers, yet this failed to exceed the performance of prompts with only scoring examples. Additionally, we examined whether the order of rationale and score generation impacts performance, as might be expected from prior research in other domains [29]. However, depending on the metrics used to evaluate, the best prompt differed; predicting rationale first and then score in separate queries is the best in terms of accuracy, while predicting score only is the best in terms of QWK. Thus, on average, the best way to perform short answer scoring appears to be employing a prompt with scoring examples, predicting score first before rationale (or no rationale at all if not needed) in one query setup, or rationale first in two queries.

The current study has limitations in that the ten questions covered only biology, science, and English subjects, mostly at the tenth-grade level. We also found that the scoring rubric might affect the performance. Furthermore, the utility of the rationale provided by GPT-4 remains uncertain. In the future, we plan to explore these possible directions to enhance the automatic short-answer scoring, performance as well as determine in which way the provided rationale might be most helpful for students and instructors.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report.

[2] Enrique Alfonseca and Diana Pérez. 2004. Automatic assessment of open ended questions with a bleu-inspired algorithm and shallow nlp. In *Advances in Natural Language Processing: 4th International Conference*. Springer, Alicante, Spain, 25–35.

[3] Lyle F Bachman, Nathan Carr, Greg Kamei, Mikyung Kim, Michael J Pan, Chris Salvador, and Yasuyo Sawaki. 2002. A reliable approach to automatic assessment of short answer free responses. In *The 17th International Conference on Computational Linguistics: Project Notes*. Association for Computational Linguistics, aipei, Taiwan, 1–4.

[4] Barbara, Ben Hamner, Jaison Morgan, lynnvandev, and Mark Shermis. 2012. The Hewlett Foundation: Short Answer Scoring. https://kaggle.com/competitions/asap-sas

[5] Isaac I Bejar, David M Williamson, and Robert J Mislevy. 2006. *Human scoring*. Lawrence Erlbaum, Mahwah, NJ. 49–81 pages.

[6] Randy Elliot Bennett. 2006. Moving the field forward: Some thoughts on validity and automated scoring. *Automated scoring of complex tasks in computer-based testing* (2006), 403–412.

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

[8] Jill Burstein, Susanne Wolff, and Chi Lu. 1999. *Using lexical semantic techniques to classify free-responses*. Vol. 10. Springer, Dordrecht, DE, 227–244.

[9] Philip G Butcher and Sally E Jordan. 2010. A comparison of human and computer marking of short free-text student responses. *Computers & Education* 55, 2 (2010), 489–499.

[10] Li-Hsin Chang and Filip Ginter. 2024. Automatic Short Answer Grading for Finnish with ChatGPT. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. AAAI Press, Palo Alto, CA, 23173–23181.

[11] Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 1835–1845.

[12] Laurie Cutrone, Maiga Chang, et al. 2011. Auto-assessor: computerized assessment system for marking student's short-answers automatically. In *2011 IEEE International Conference on Technology for Education*. IEEE, 81–88.

[13] Larry Davis and Spiros Papageorgiou. 2021. Complementary strengths? Evaluation of a hybrid human-machine scoring approach for a test of oral academic English. *Assessment in Education: Principles, Policy & Practice* 28, 4 (2021), 437–455.

[14] Christian Gütl. 2007. e-Examiner: towards a fully-automatic knowledge assessment tool applicable in adaptive e-learning systems. In *Proceedings of the Second International Conference on Interactive Mobile and Computer Aided Learning*. Citeseer, Amman, Jordan, 1–10.

[15] Wen-Juan Hou and Jia-Hao Tsao. 2011. AUTOMATIC ASSESSMENT OF STUDENTS'FREE-TEXT ANSWERS WITH DIFFERENT LEVELS. *International Journal on Artificial Intelligence Tools* 20, 02 (2011), 327–347.

[16] Sally Jordan and Tom Mitchell. 2009. e-Assessment for learning? The potential of short-answer free-text questions with tailored feedback. *British Journal of Educational Technology* 40, 2 (2009), 371–385.

[17] Chinmay E Kulkarni, Richard Socher, Michael S Bernstein, and Scott R Klemmer. 2014. Scaling short-answer grading by combining peer assessment with algorithmic scoring. In *Proceedings of the first ACM conference on Learning@ scale conference*. Association for Computing Machinery, New York, NY, USA, 99–108.

[18] Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. 2022. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems* 64, 12 (2022), 3197–3234.

[19] Guangming Ling, Pamela Mollaun, and Xiaoming Xi. 2014. A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing* 31, 4 (2014), 479–499.

[20] Nitin Madnani, Jill Burstein, John Sabatini, and Tenaha O'Reilly. 2013. Automated Scoring of Summary-Writing Tasks Designed to Measure Reading Comprehension. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Atlanta, Georgia, 163–168.

[21] Tom Mitchell, Terry Russell, Peter Broomhead, and Nicola Aldridge. 2002. Towards robust computerised marking of free-text responses. In *Proceedings of the 6th CAA Conference*. Loughborough University, Loughborough.

[22] Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics* 2, 2 (2023), 100050.

[23] Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada, 394–403.

[24] Diana Pérez, Enrique Alfonseca, Pilar Rodríguez, Alfio Gliozzo, Carlo Strapparava, and Bernardo Magnini. 2005. About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment. *Revista signos* 38, 59 (2005), 325–343.

[25] Donald E Powers, Jill C Burstein, Martin S Chodorow, Mary E Fowles, and Karen Kukich. 2002. Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research* 26, 4 (2002), 407–425.

[26] Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chungmin Lee. 2017. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th workshop on innovative use of NLP for building educational applications*. Association for Computational Linguistics, Copenhagen, Denmark, 159–168.

[27] Johannes Schneider, Bernd Schenk, Christina Niklaus, and Michaelis Vlachos. 2023. Towards LLM-based autograding for short textual answers. *arXiv preprint arXiv:2309.11508* (2023).

[28] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 2609–2634.

[29] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[30] Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada, 576–584.

[31] Lishan Zhang, Yuwei Huang, Xi Yang, Shengquan Yu, and Fuzhen Zhuang. 2022. An automatic short-answer grading model for semi-open-ended questions. *Interactive learning environments* 30, 1 (2022), 177–190.

[32] Mo Zhang. 2013. Contrasting automated and human scoring of essays. *R & D Connections* 21, 2 (2013), 1–11.