Discover Artificial Intelligence

Brief Communication

Performance of the pre-trained large language model GPT-4 on automated short answer grading

Gerd Kortemeyer[1]

Received: 9 December 2023 / Accepted: 25 June 2024
Published online: 05 July 2024

© The Author(s) 2024 OPEN

Abstract
Automated Short Answer Grading (ASAG) has been an active area of machine-learning research for over a decade. It promises to let educators grade and give feedback on free-form responses in large-enrollment courses in spite of limited availability of human graders. Over the years, carefully trained models have achieved increasingly higher levels of performance. More recently, pre-trained Large Language Models (LLMs) emerged as a commodity, and an intriguing question is how a general-purpose tool without additional training compares to specialized models. We studied the performance of GPT-4 on the standard benchmark 2-way and 3-way datasets SciEntsBank and Beetle, where in addition to the standard task of grading the alignment of the student answer with a reference answer, we also investigated withholding the reference answer. We found that overall, the performance of the pre-trained general-purpose GPT-4 LLM is comparable to hand-engineered models, but worse than pre-trained LLMs that had specialized training.

Keywords Automated short answer grading · Large language model · SciEntsBank · Beetle · GPT

1 Introduction

Providing meaningful feedback to learners is one of the most important tasks of instructors [1], yet it can also become one of the most work-intensive or even tedious tasks. Particularly for large-enrollment courses, lack of grading personnel can limit this feedback to automatically gradable closed-answer formats such as multiple-choice or numerical inputs. This limitation might be overcome by using Artificial Intelligence (AI) solutions [2]; it is therefore not surprising that when it comes to the use of AI in higher education, assessment and evaluation are the most prominent topics [3], and acceptance of this technology for education is increasing based on its perceived usefulness [4]. In particular, studies on Automated Short Answer Grading (ASAG) [5, 6] are highly relevant for educators to extend the limits of what can be assessed at large scales.

It is impossible to do justice to the spectrum of sophisticated ASAG methods in this short study; Burrows, Gurevych, and Stein provide an excellent overview up to 2015 [5]; Haller, Aldea, Seifert, and Strisciuglio look at later developments up to 2022 [6]. The latter survey notes a particular shift in recent years as models are moving from hand-engineered features to representation-learning approaches, which draw their initial training data from large text corpora [6] ("pre-trained"). However, most models used for ASAG still have in common that they are explicitly trained or fine-tuned for particular grading tasks, and datasets used in competitions such as SemEval [7] thus include training and testing items. By contrast, recently publicly released Large Language Models (LLMs) such as GPT-4 [8] and Bard [9] have not only been pre-trained from large text corpora, but subsequently extensively fine-tuned following general instead of task-specific

✉ Gerd Kortemeyer, kgerd@ethz.ch | [1]Rectorate and AI Center, ETH Zurich, 8092 Zurich, Switzerland.

strategies. Their users are neither expected nor actually able to further train or fine-tune the model, and an intriguing question is how these out-of-the-box general-purpose tools perform compared to specially trained or fine-tuned models.

In this study, GPT-4 is prompted to grade the items from two standard datasets, SciEntsBank and Beetle [7], which allows comparison of precision, recall, and F1-score (or weighted F1-score in case of 3-way items) to legacy and state-of-the-art ASAG models. SciEntsBank covers general science questions for 3rd to 6th grade, while Beetle covers questions and student answers from a tutorial system for basic electricity and electronics.

The standard judgment method is to compare the student answer to a reference answer, but in addition, it was also investigated if GPT-4 can adequately grade the student answers based on the question alone. For the latter task, the model would need to draw on its own pre-training from its text corpus to independently judge the correctness of the student answer.

In summary, this study

- compares the performance of an out-of-the-box Large Language Model in ASAG of standardized problem sets to that of hand-engineered AI systems and previous-generation Large Language Models that have been specifically trained or fine-tuned for the task, and it does so
- with and without providing the reference answer to the out-of-the-box model, in the latter case completely relying on the knowledge of the pre-trained model.

The motivation of this study is not to prove that general-purpose, out-of-the box models perform better than state-of-the-art specialized systems (and as it will turn out, they do not), but to

- provide a benchmark data point of how these systems compare to previous and current research systems, and
- explore the possibility of deploying these systems as low-threshold tools for instructors to provide meaningful ASAG for learners.

2 Related work

The idea of using machines for grading was arguably first introduced by Sidney Pressey in the late 1920 s with the "Automatic Teacher" [10], which was able to pose automatically graded multiple-choice questions. In the 1960th, computers took the place of such mechanical devices, connecting to mainframes using Teletypes [11], which allowed for numerical answers; these integer answers were checked for equality to the programmed correct solution. With the advent of the web also came more sophisticated systems that could grade ranking or mix-and-match problems, as well as numerical solutions with tolerances and physical units, algebraic answers based on symbolic mathematical equivalence, and string responses based on exact matching or regular expressions [12, 13]. The latter responses are usually limited to individual words or short phrases in the context of cloze ("fill-in-the-blank") questions [14, 15]. What is common to these approaches is that they are based on deterministic algorithms; answers that are basically closed-ended, so there is a limited set of correct answers or hard criteria to judge a responses as correct or incorrect [16]. The system then grades with perfect accuracy, making it fit for use in high-stakes summative assessment; the assessment of the desired learning outcomes might still not be reliable in terms of psychometrics [17–19], but the responses themselves will be judged reliably according to the provided criteria.

At the opposite end of the spectrum are completely open-ended questions, which allow for free expression of ideas, concepts, and solutions approaches without necessarily imposing a particular structure [16]. Typical responses are essays or extended solution derivations. Experimentally, these are increasingly graded by artificial intelligence [20, 21], however, here the accuracy is much lower, resulting in limited reliability and thus limited trust [22, 23]. Current results indicate that artificial intelligence is capable of providing meaningful formative feedback to completely open-ended assessment responses, but might not be ready yet for high-stakes summative assessment [20].

ASAG offers a middle road between closed-ended and completely open-ended responses, which is why they are sometimes referred to as "semi-open-ended" questions [24]. These questions generally expect one to three free-form sentences that directly answer a specific question; as opposed to cluze, scoring systems need to recognize paraphrasing and equivalent meaning [25] (which generally requires machine learning approaches), but as opposed to essays or mathematical derivations, they do not need to deal with a wide-open answer space and the complications of following multi-step arguments or judging matters of style. As machine-learning based systems, to achieve high accuracy, ASAG engines generally need a phase of subject-matter specific training or fine-tuning before being used to judge student work [26]; for transformer or transformer-ensemble approaches, a phase of generating various reference answers improves their performance [27]. Here, we explore the performance of a single general-purpose engine being used out-of-the-box without additional preparation, and we do so with and without providing a reference answer.

3 Methodology

The SciEntsBank and Beetle datasets [7] were downloaded from kaggle [28]. They included both training and test data. The training data were discarded, while the test data included the 504 items and 14, 186 student answers and their reference grading that were used for this study. As no training took place, the distinction between unseen answers (UA), unseen questions (UQ), and unseen domains (UD) that the dataset provided was dropped for this study, since all items were "unseen."

Each item in the datasets contains a question, a reference answer, and student answers including their reference grade. The items came in two versions:

- a 2-way version, where each student answer is either *correct* if it is complete and correct paraphrase of the reference answer or *incorrect* otherwise, and
- a 3-way version, where an additional judgment of *contradictory* replaces some of the *incorrect* labels if the student answer explicitly contradicts the reference answer.

Using a Python script, the XML-coded items were translated into prompts for the GPT-4 API, see Fig. 1 for an overview of the whole process and Fig. 2 for an example. Each item was graded with and without providing a reference answer. The definitions of the judgment criteria for grading were taken from SemEval-2013 [7].

The API was provided by Azure AI Services [29]. For each item, the role and prompt (Fig. 2) were sent via a Python script to be processed by GPT-4 as shown in Fig. 3. Also shown in Fig. 3 is an example of the output, which the Python script wrote to disk as a CSV-file for further analysis.

For 6 of the 504 items, errors occurred during evaluation, which led to 58 of the 28, 372 student statements receiving no or invalid grades (that is, missing or invalid entries in the output table shown in Fig. 3). The invalid grades were *unclear*, *creative*, *epoch*, *accurate*, and *correc* [sic]. These missing or invalid student grades were counted as neither positives nor negatives.

Subsequently, the precision, recall, and (weighted) F1-score were calculated:

Precision Out of all the *correct* grades given by a model, how many were actually correct?
Recall (or Sensitivity) Out of all the actual correct student answers, how many were graded as *correct*?

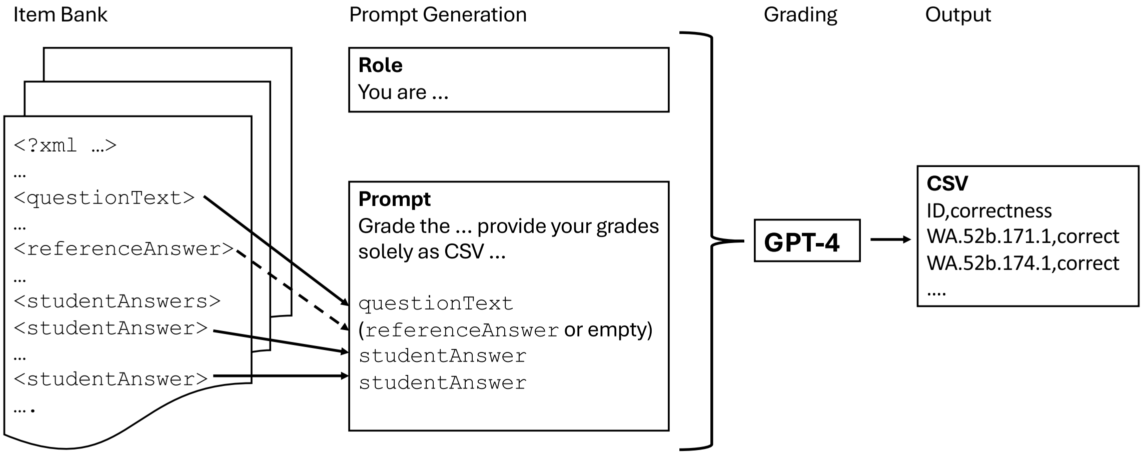
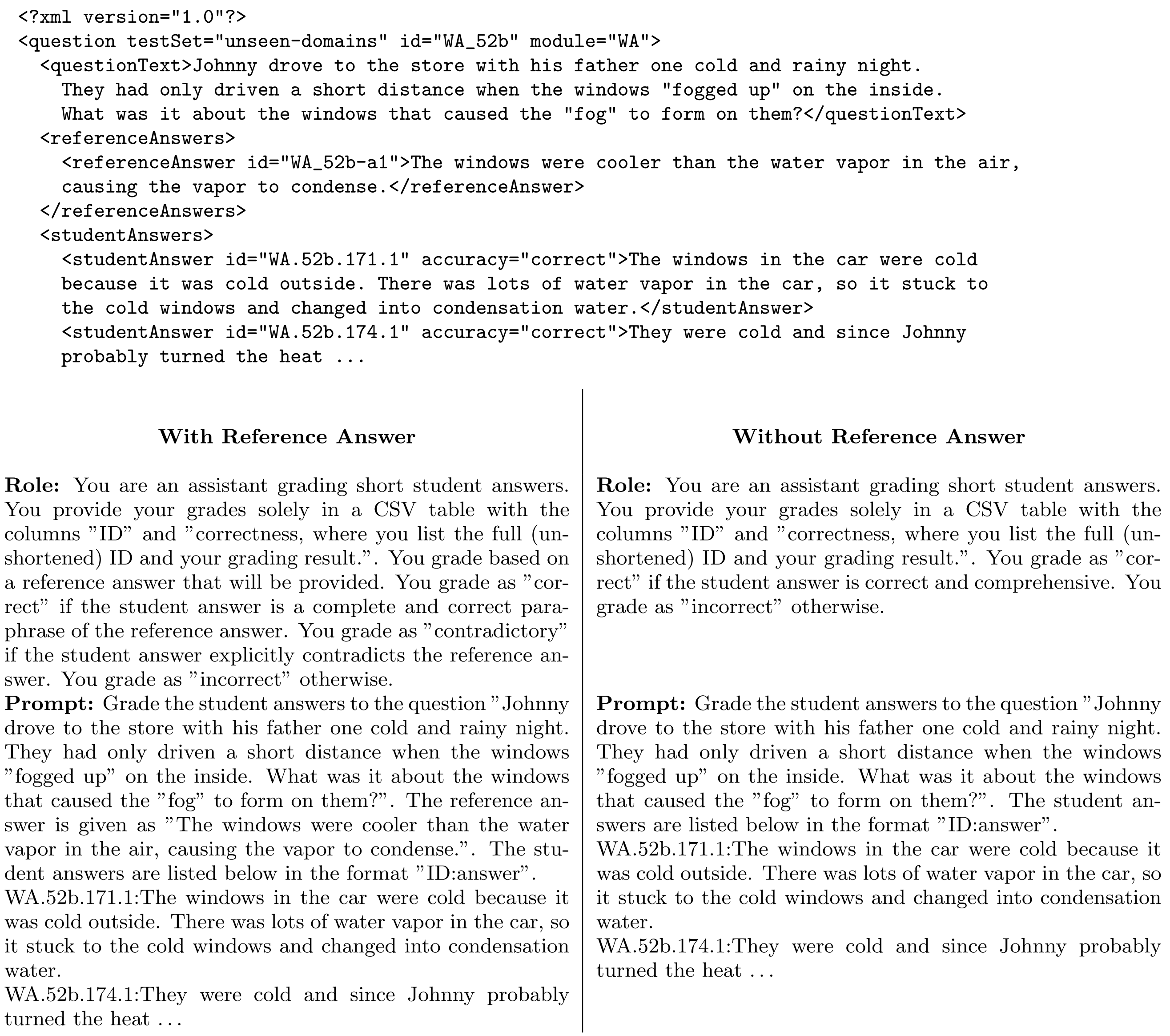


Fig. 1 Overview of the grading process



```
<?xml version="1.0"?>
<question testSet="unseen-domains" id="WA_52b" module="WA">
  <questionText>Johnny drove to the store with his father one cold and rainy night.
    They had only driven a short distance when the windows "fogged up" on the inside.
    What was it about the windows that caused the "fog" to form on them?</questionText>
  <referenceAnswers>
    <referenceAnswer id="WA_52b-a1">The windows were cooler than the water vapor in the air,
    causing the vapor to condense.</referenceAnswer>
  </referenceAnswers>
  <studentAnswers>
    <studentAnswer id="WA.52b.171.1" accuracy="correct">The windows in the car were cold
    because it was cold outside. There was lots of water vapor in the car, so it stuck to
    the cold windows and changed into condensation water.</studentAnswer>
    <studentAnswer id="WA.52b.174.1" accuracy="correct">They were cold and since Johnny
    probably turned the heat ...
```

With Reference Answer

Role: You are an assistant grading short student answers. You provide your grades solely in a CSV table with the columns ”ID” and ”correctness, where you list the full (unshortened) ID and your grading result.”. You grade based on a reference answer that will be provided. You grade as ”correct” if the student answer is a complete and correct paraphrase of the reference answer. You grade as ”contradictory” if the student answer explicitly contradicts the reference answer. You grade as ”incorrect” otherwise.

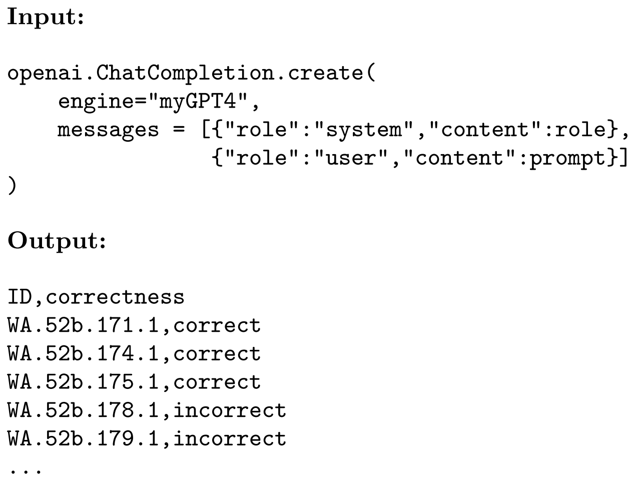
Prompt: Grade the student answers to the question ”Johnny drove to the store with his father one cold and rainy night. They had only driven a short distance when the windows ”fogged up” on the inside. What was it about the windows that caused the ”fog” to form on them?”. The reference answer is given as ”The windows were cooler than the water vapor in the air, causing the vapor to condense.”. The student answers are listed below in the format ”ID:answer”.
WA.52b.171.1:The windows in the car were cold because it was cold outside. There was lots of water vapor in the car, so it stuck to the cold windows and changed into condensation water.
WA.52b.174.1:They were cold and since Johnny probably turned the heat . . .

Without Reference Answer

Role: You are an assistant grading short student answers. You provide your grades solely in a CSV table with the columns ”ID” and ”correctness, where you list the full (unshortened) ID and your grading result.”. You grade as ”correct” if the student answer is correct and comprehensive. You grade as ”incorrect” otherwise.

Prompt: Grade the student answers to the question ”Johnny drove to the store with his father one cold and rainy night. They had only driven a short distance when the windows ”fogged up” on the inside. What was it about the windows that caused the ”fog” to form on them?”. The student answers are listed below in the format ”ID:answer”.
WA.52b.171.1:The windows in the car were cold because it was cold outside. There was lots of water vapor in the car, so it stuck to the cold windows and changed into condensation water.
WA.52b.174.1:They were cold and since Johnny probably turned the heat . . .

Fig. 2 Original XML-code of a 3-way item and the generated prompts for its evaluation with and without providing a reference answer



Input:

```
openai.ChatCompletion.create(
    engine="myGPT4",
    messages = [{"role":"system","content":role},
                {"role":"user","content":prompt}]
)
```

Output:

```
ID,correctness
WA.52b.171.1,correct
WA.52b.174.1,correct
WA.52b.175.1,correct
WA.52b.178.1,incorrect
WA.52b.179.1,incorrect
...
```

Fig. 3 Format of sending the role and prompt in Fig. 2 to be processed by GPT-4 (“myGPT4” is the name of the deployment of the model used), as well as example output from GPT-4

F1-score Harmonic mean of precision and recall; a way to balance the trade-off between precision and recall.

In the 3-way scenario, the above characteristics are correspondingly calculated for the classes *contradictory* and *incorrect*, and a weighted average is calculated for these class F1-scores to form the weighted F1-score (w-F1).

4 Results

4.1 Precision, recall, and F1-scores

Table 1 shows the precision, recall, and F1-scores for SciEntsBank and Beetle for the 2-way and 3-way items, as well as for the scenario where the reference answer was withheld. For the 3-way scenario, the individual-class results and the weighted F1-score (w-F1) are provided.

Looking at the precision and recall, an outlier is the recall on *contradictory* in the 3-way Beetle dataset: a large number of student answers that were labelled as *contradictory* were not recognized as such, but simply as *incorrect* (as evidenced by the low precision on *incorrect*).

GPT-4 generally performs better on SciEntsBank than on Beetle. For SciEntsBank, the model showed its highest performance on the 2-way task (F1=0.744), followed closely by the no-reference scenario (F1=0.731), with the 3-way scenario in last place (w-F1=0.729). For Beetle, the no-reference scenario showed the highest performance (F1=0.651), followed by the 2-way (F1=0.611) and 3-way (w-F1=0.516) scenarios. In other words, for Beetle, providing a reference answer lowered its performance on correctly judging the student answers.

4.2 Comparison to specialized ASAG models

Table 2 shows a comparison of specifically trained models versus the out-of-the-box GPT-4 model. At the time of the SemEval-2013 competition [7], had GPT-4 been around, it would have won the competition for 3-way SciEntsBank, and it would have outperformed all but one competing models in the unseen questions (UQ) category. In these specifically trained models, performance strongly depends on what was "seen" and what was "unseen."

Newer systems perform better, in particular those of the BERT [30] LLM family. These models are pre-trained and then specifically trained for SciEntsBank and Beetle using for example PyTorch [31]. Unfortunately, for the highly successful roberta-large model [32], the performance was not separately reported for the different 'unseen" categories, and no 3-way grading was performed.

Overall, the performance of the pre-trained general-purpose GPT-4 LLM is comparable to hand-engineered models, but worse than pre-trained LLMs that had specialized training.

Table 1 Results for precision, recall, and F1-scores for SciEntsBank and Beetle in the 2-way, 3-way, and no-reference-answer scenarios

	2-way			3-way										No reference answer		
				correct			*contradictory*			*incorrect*						
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	w-F1	Prec.	Rec.	*F1*
SciEnts-Bank	0.788	0.705	0.744	0.717	0.825	0.767	0.696	0.581	0.633	0.754	0.679	0.715	0.729	0.697	0.768	0.731
Beetle	0.657	0.572	0.611	0.635	0.672	0.653	0.680	0.199	0.308	0.426	0.672	0.522	0.516	0.581	0.739	0.651

For the 3-way scenario, the individual-class results and the weighted F1-score (w-F1; also referred to as micro-averaged F1-score) are provided

Table 2 Comparison of (weighted) F1-scores for different ASAG systems and GPT-4

		SciEntsBank							Beetle				
		2-way			3-way			No Ref	2-way		3-way		No Ref
Model	Year	UA	UQ	UD	UA	UQ	UD		UA	UQ	UA	UQ	
CoMeT [7]	2013	0.77	0.58	0.67	0.71	0.52	0.55		0.83	0.70	0.73	0.49	
SoftCardinality [7]	2013	0.72	0.74	0.71	0.65	0.63	0.62		0.77	0.64	0.62	0.45	
Sultan et al. [33, 34]	2016	0.69	0.70	0.71	0.57	0.62	0.60						
Saha et al. [34]	2018	0.79	0.70	0.72	0.71	0.64	0.61						
GCN-DA [35]	2020			0.73			063						
SFRN+ [36]	2021	0.78	0.64	0.67	0.65[a]	0.49[a]	0.47[a]		0.89	0.70	0.67[a]	0.55[a]	
BERT [37]	2022				0.73	0.60	0.62				0.71	0.57	
roberta-large [32]	2021	0.81[b]							0.91[b]				
GPT-4	2023	0.74			0.73			0.73	0.61		0.52		0.65

[a] not stated if macro-averaged F1 or weighted (micro-averaged) F1 was reported

[b] the model was specifically trained, but no separate information on UA, UQ, and UD was provided

5 Limitations

Since GPT is a probabilistic model, running it again, possibly at a different temperature, is likely going to yield different results. However, due to the already large amount of computing required for one run, and in light of the high statistics gained from over 500 items, only one run was considered here. Also, different prompts from the ones shown in Fig. 2 may result in higher or lower performance.

OpenAI, the company behind GPT, does not release information about what constituted the text corpus used for training. Though unlikely, since the datasets are only available as ZIP-files and in XML-format, there is still a possibility that SciEntsBank and Beetle had been used for training. When asked about SciEntsBank, though, the model stated that it was not familiar with a dataset or source by that name; GPT-4 performed better on SciEntsBank than on Beetle, for which it stated that it is a known dataset in the domain of natural language processing and educational research. The model, however, demonstrated ignorance when asked about any specific details regarding Johnny, his father, and the windows in the scenario quoted in Fig. 2, making it unlikely that it had seen the text before.

6 Discussion

The last five years saw the strong emergence of Deep-Learning-based models for ASAG. These models generally exhibit higher performance than hand-engineered models, but still strongly depend on training, which may be pre-training or task-specific. LLMs usually come pre-trained, but the extend of that pre-training greatly varies: while details on GPT-4's text corpus are proprietary, it can be assumed that it was trained and fine-tuned with orders of magnitude more data than for example BERT [30]. However, as this study shows, the difference in pre-training can be more than made up by the BERT-family's openness to additional task-specific training by the user.

At least for the grade-school educational content covered by the datasets in this study, GPT-4 performs ASAG at a performance level comparable to hand-engineered systems from five years ago. It does so even without the need for providing reference answers. There are strong indications that this ability would extend to higher education, for example university-level physics content [38], and that automated grading of open-ended assessment content is possible beyond short answers [20]. In addition, a general-purpose LLM can give more tailored feedback than simple *correct/incorrect* judgments, which has high potential for learning from short answer grading [39].

The low-overhead nature of this out-of-the-box approach opens up the possibility of integrating ASAG into commodity learning management systems (LMSs) as a component of their quizzing engines, particularly for formative assessment. Instructors could write these assessment items with little to no technical support and immediately deploy them. A problem with general-purpose tools like GPT-4 [8] and Bard [9] is that they are running in the cloud. When it comes to grade-relevant student data, the question of data security and privacy cannot be ignored, which may limit the applicability of this approach to ASAG unless additional contractual agreements are in place. An alternative for a model that might also not need additional training, but which could be locally installed, is Llama 2 [40], However, preliminary studies by the author indicate that Llama 2 does generally not perform as well as GPT-4.

7 Conclusion

The performance of the general-purpose Large Language Model GPT-4 on Automated Short Answer Grading does not reach that of specifically trained Deep-Learning models, but it is comparable to that of earlier hand-engineered ASAG models. A clear advantage of GPT-4 is that it does not need to be specifically trained for the task and can be used out-of-the-box, which has the potential to turn it into a commodity for educators as part of learning management systems. In addition to not needing additional training, GPT-4 can also perform ASAG without the need for providing reference answers, at least at the grade-school level covered by the datasets used in this study and likely at the introductory higher-education level.

Acknowledgements The author would like to thank Julia Chatain for her help in connecting to the GPT API.

Author contributions The sole author of this study carried out the work himself.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich.

Data availability The benchmark datasets SciEntsBank and Beetle [7] are available from kaggle [28]. Code and calculated data are made available as supplemental material alongside this paper from https://www.polybox.ethz.ch/index.php/s/mByv0od7uscm3VV (the file readme.txt in the downloadable package explains the code and data files).

Declarations

Competing interests There are no conflicting or 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

1. Bransford JD, Brown AL, Cocking RR, et al. How people learn. Washington, DC: National academy press; 2000.
2. Seo K, Tang J, Roll I, Fels S, Yoon D. The impact of artificial intelligence on learner-instructor interaction in online learning. Int J Educ Technol Higher Educ. 2021;18(1):1–23.
3. Crompton H, Burke D. Artificial intelligence in higher education: the state of the field. Int J Educ Technol Higher Educ. 2023;20(1):1–22.
4. Zhang C, Schießl J, Plößl L, Hofmann F, Gläser-Zikuda M. Acceptance of artificial intelligence among pre-service teachers: a multigroup analysis. Int J Educ Technol Higher Educ. 2023;20(1):49.
5. Burrows S, Gurevych I, Stein B. The eras and trends of automatic short answer grading. Int J Artif Intell Educ. 2015;25:60–117.
6. Haller S, Aldea A, Seifert C, Strisciuglio N. Survey on automated short answer grading with deep learning: from word embeddings to transformers. arXiv preprint arXiv:2204.03503, 2022.
7. Dzikovska MO, Nielsen R, Brew C, Leacock C, Giampiccolo D, Bentivogli L, Clark P, Dagan I, Dang HT. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 263–274, 2013.
8. OpenAI. GPT-4. https://openai.com/gpt-4.
9. Google. Bard. https://bard.google.com/.
10. Petrina S. Sidney pressey and the automation of education, 1924–1934. Technol Cult. 2004;45(2):305–30.
11. Suppes P, Jerman M, Groen G. Arithmetic drills and review on a computer-based teletype. Arith Teach. 1966;13(4):303–9.
12. Sangwin CJ. Assessing elementary algebra with stack. Int J Math Educ Sci Technol. 2007;38(8):987–1002.
13. Kortemeyer G, Kashy E, Benenson W, Bauer W. Experiences using the open-source learning content management and assessment system lon-capa in introductory physics courses. Am J Phys. 2008;76(4):438–44.
14. Jonz J. Another turn in the conversation: what does cloze measure? Tesol Quarterly. 1990;24(1):61–83.

15. Chapelle CA, Abraham RG. Cloze method: what difference does it make. Lang Testing. 1990;7(2):121–46.
16. R Pate. Open versus closed questions: what constitutes a good question. Educational research and innovations, pages 29–39, 2012.
17. Lord FM, Novick MR. Statistical theories of mental test scores. Information Age Publishing, 2008.
18. James Dean Brown. My twenty-five years of cloze testing research: so what. Int J Lang Stud. 2013;7(1):1–32.
19. Kortemeyer G. Extending item response theory to online homework. Phys Rev Special Topics-Phys Educ Res. 2014;10(1): 010118.
20. Kortemeyer G. Toward ai grading of student problem solutions in introductory physics: a feasibility study. Phys Rev Phys Educ Res. 2023;19(2): 020163.
21. Jamil F, Hameed IA. Toward intelligent open-ended questions evaluation based on predictive optimization. Expert Syst Appl. 2023;231: 120640.
22. Jackson Stephen, Panteli Niki. Trust or mistrust in algorithmic grading? an embedded agency perspective. Int J Inf Manag. 2023;69: 102555.
23. Conijn R, Kahr P, Snijders CC. The effects of explanations in automated essay scoring systems on student trust and motivation. J Learn Anal. 2023;10(1):37–53.
24. Zhang Lishan, Huang Yuwei, Yang Xi, Shengquan Yu, Zhuang Fuzhen. An automatic short-answer grading model for semi-open-ended questions. Int Learn Environ. 2022;30(1):177–90.
25. Leacock Claudia, Chodorow Martin. C-rater: automated scoring of short-answer questions. Comput Hum. 2003;37:389–405.
26. Ahmed A, Joorabchi A, Hayes MJ. On deep learning approaches to automated assessment: strategies for short answer grading. CSEDU (2), pages 85–94, 2022.
27. Akila Devi TR, Javubar Sathick K, Abdul Azeez Khan A, Arun Raj L. Novel framework for improving the correctness of reference answers to enhance results of asag systems. SN Computer Science, 2023; 4(4): 415.
28. Kerneler, Kaggle: semeval 2013 2 and 3 way. https://www.kaggle.com/datasets/smiles28/semeval-2013-2-and-3-way.
29. Microsoft. Azure ai services. https://azure.microsoft.com/en-us/products/ai-services.
30. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
31. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 2019; 32.
32. Andrew Poulton and Sebas Eliens. Explaining transformer-based models for automatic short answer grading. In Proceedings of the 5th International Conference on Digital Technology in Education, pages 110–116, 2021.
33. Sultan MA, Salazar C, Sumner T. Fast and easy short answer grading with high accuracy. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1070–1075, 2016.
34. Saha S, Dhamecha TI, Marvaniya S, Sindhgatta R, Sengupta B. Sentence level or token level features for automatic short answer grading?: Use both. In Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part I 19, pages 503–517. Springer, 2018.
35. Tan Hongye, Wang Chong, Qinglong Duan YuLu, Zhang Hu, Li Ru. Automatic short answer grading by encoding student responses via a graph convolutional network. Int Learn Environ. 2023;31(3):1636–50.
36. Li Z, Tomar Y, Passonneau RJ. A semantic feature-wise transformation relation network for automatic short answer grading. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6030–6040, 2021.
37. Filighera A, Tschesche J, Steuer T, Tregel T, Wernet L. Towards generating counterfactual examples as automatic short answer feedback. In International Conference on Artificial Intelligence in Education, pages 206–217. Springer, 2022.
38. Kortemeyer Gerd. Could an artificial-intelligence agent pass an introductory physics course? Phys Rev Phys Educ Res. 2023;19(1): 010132.
39. Jordan Sally, Mitchell Tom. e-assessment for learning? the potential of short-answer free-text questions with tailored feedback. Br J Educ Technol. 2009;40(2):371–85.
40. Meta. Llama 2. https://ai.meta.com/llama/.

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