Automatic Short Answer Grading for Finnish with ChatGPT

Li-Hsin Chang, Filip Ginter

University of Turku
{lhchan, figint}@utu.fi

Abstract

Automatic short answer grading (ASAG) seeks to mitigate the burden on teachers by leveraging computational methods to evaluate student-constructed text responses. Large language models (LLMs) have recently gained prominence across diverse applications, with educational contexts being no exception. The sudden rise of ChatGPT has raised expectations that LLMs can handle numerous tasks, including ASAG. This paper aims to shed some light on this expectation by evaluating two LLM-based chatbots, namely ChatGPT built on GPT-3.5 and GPT-4, on scoring short-question answers under zero-shot and one-shot settings. Our data consists of 2000 student answers in Finnish from ten undergraduate courses. Multiple perspectives are taken into account during this assessment, encompassing those of grading system developers, teachers, and students. On our dataset, GPT-4 achieves a good QWK score (0.6+) in 44% of one-shot settings, clearly outperforming GPT-3.5 at 21%. We observe a negative association between student answer length and model performance, as well as a correlation between a smaller standard deviation among a set of predictions and lower performance. We conclude that while GPT-4 exhibits signs of being a capable grader, additional research is essential before considering its deployment as a reliable autograder.

Introduction

Short answer question is a common form of constructed-response questions, which can test for deeper knowledge understanding (Hancock 1994; Kuechler and Simkin 2010). While widely used in educational settings, its drawbacks include being time-consuming and requiring prerequisite knowledge to evaluate. The task of automatic short answer grading (ASAG) seeks to use computational methods to reduce human efforts in the evaluation of short answers. There has been much research effort dedicated to it (Burrows, Gurevych, and Stein 2015; Bonthu, Sree, and Prasad 2021; Putnikovic and Jovanovic 2023).

Pre-trained language models have brought revolutionary progress in natural language processing in recent years. As the size of the pre-trained models gets larger (e.g. above tens of billions of parameters), they gain the ability to perform certain tasks where the performance of smaller models cannot exceed random baselines (Wei et al. 2022a). These capabilities are called emerging abilities (Wei et al. 2022a) and models of such sizes large language models (LLM) (Zhao et al. 2023). LLMs had been steadily gaining popularity until the release of ChatGPT [1], whose conversational ability impresses the world and led to a surge of research work on LLMs (Zhao et al. 2023). In terms of its application in the educational domain, there has been various efforts to investigate how LLMs can be used and how they may impact the education system. These include the use of LLMs in intelligent tutoring systems (Cao 2023), educational chatbots (Dan et al. 2023), teacher-response generation in educational dialogues (Tack et al. 2023), automated evaluation of student text (Hackl et al. 2023), among others. Nonetheless, various challenges, including low technological readiness, concerns about replicability and transparency, and inadequate privacy and beneficence considerations, currently impede the widespread integration of LLMs in educational settings (Yan et al. 2023). For the automated evaluation of student text, there have been more research on the use of LLMs in automated essay scoring, such as discourse coherence prediction (Naismith, Mulcaire, and Burstein 2023) and second language learner essay evaluation (Yancey et al. 2023; Mizumoto and Eguchi 2023). The research on LLMs advancing ASAG is in its early stages; Yoon (2023) uses one-shot prompting for LLM to identify justification keys, or key phrases in student-written short answers and compare the extracted key phrases with those of reference answers. They find that the GPT-3.5 model has an accuracy of 90.3% in extracting these justification keys. The identified justification keys are compared with a reference scoring rubric, which is utilized to compute the final score. Hackl et al. (2023) evaluate the consistency of GPT-4 in text-rating in terms of style and content, and find high self-consistency (interclass correlation coefficient 0.94-0.99) in GPT-4. Matelsky et al. (2023) build a tool to automatically give feedback to answers to open-ended questions. Concurrent to our research, Schneider et al. (2023) compare GPT-3.5 and humans in assessing bachelor-level German and master-level English short answers, reporting observed issues with the grading of the LLM.

This paper explores the suitability of LLMs for ASAG,

[1] https://openai.com/blog/chatgpt/

testing ChatGPT based on GPT-3.5 and GPT-4 on 2K Finnish short-question answers from ten bachelor-level courses. The experimental design seeks to elucidate whether direct use of ChatGPT in summative assessments by educators is feasible. We find that while one-shot GPT-4 achieves QWK scores considered good on 44 of the 100 tested questions, further research is required before deploying an LLM-based short-answer grader.

Related Work

The definition of short answers vary across studies (Burrows, Gurevych, and Stein 2015; Haller et al. 2022). Typical criteria for short answers are their being under a certain length and evaluation being mainly focused on the semantic instead of the syntax of the answer. ASAG research, spanning three decades (Burrows, Gurevych, and Stein 2015), remains aligned with current NLP developments, as many ASAG systems now employ deep learning techniques (Bonthu, Sree, and Prasad 2021; Haller et al. 2022). Additionally, recent advancements include generalizable ASAG, wherein models are trained to generalize to target domains that do not overlap with their training domain (Zeng et al. 2023). Aside from systems primarily centered on score prediction, other research directions encompass using machines as a second grader (Kulkarni et al. 2014), resorting to human evaluators for more challenging questions (Li et al. 2023), exploring adversarial attacks on grading systems (Filighera, Steuer, and Rensing 2020; Filighera et al. 2022), and generating explainable predictions (Tornqvist et al. 2023), among other areas of investigation.

Data

We use "prompt" to refer to the input to LLM-based chatbots, "question" to refer to the exam task description initiating student responses, and "answer" to refer to the student responses.

We possess a database dump of 24K student exam answers from diverse disciplines at University of Turku. For each answer, we have information about the exam question, the course, the maximum possible score and the evaluated score. No information about the grading criteria is available. To assess LLM performance in content-based scoring, we focus on courses with short-answer questions to narrow down the scope of this study, as longer content-based answers caused the models to mostly exclusively predict high grades in initial experiments. Our selection process aims for ten courses, each with adequate questions and answers. We prioritize questions with a greater number of answers, assuming that a set of answers with a balanced grade distribution is more likely to be selected from a larger answer pool. When multiple questions meet this criterion, the number of possible scores and the average answer length of these questions are taken into account during selection to observe the impact of these factors on model performance.

The final selected answers are written in Finnish by native speakers taking bachelor-level courses. The answers are typically a few sentences to a paragraph long. A summary of ten selected courses from seven disciplines are shown in Table 1. Ten questions are chosen for each course. In total, 377 student answers are selected to be the grading examples shown to the model, and 2,000 student answers are selected as the test data. Out of this 2,000 answers, 200 has a binary word-based grading scale ('pass' or 'failed'), while the rest has varying numerical grading scales.

Experimental Settings

We aim to evaluate if direct use of OpenAI's LLM-based chat models for autograding is feasible for educators, and if so, which types of questions and what disciplines may benefit most. Due to the nature of our data, our experiments align with summative assessments, prioritizing grade predictions over feedback for students. Prompt design is limited by the lack of scoring criteria, and we operate in a single-prompt, single-response setting, assuming no series of interactions with the models.

Initial Explorative Experiments

The initial experiments were carried out on the user interface of OpenAI using the GPT-3.5 due to cost considerations and the then lack of GPT-4 access. We tested several types of questions with abundant numbers of answers in our data to identify ideal candidates for further experiments. These include questions from advanced writing courses asking for correction of language flaws and citations, content-focused short answer questions, and questions asking for longer content-focused essays. We found that, under our setting, questions initiating content-focused short answers is the most feasible type of question for automatic scoring by LLM-based chatbots.

We next did prompt engineering on a small subset of data from various disciplines. We aimed for easily understandable prompts for both human and machine, as well as efficient token usage without compromising performance. We optimized on the following aspects: (1) the presentation of the questions, short answers, and possibly evaluated scores to the model, (2) the language(s) of the prompt, and (3) the number of examples and short answers for evaluation in a prompt.

For data point presentation, we tried presenting the questions, answers, and scores as jsonl, in natural language with or without detailed descriptions about the setup. We tried minor changes in the format such as use of delimiters and punctuation. We also tested varying the presentation of the evaluated scores, such as numeric scores or ordinal grading scale (e.g. fail, satisfactory, good, excellent). Overall, no format consistently outperformed others across the questions. For further experiments, we use natural language descriptions as it most resemble a conversation. We also use the original grading scales of the individual questions, which varies by course and question.

For the language of the prompt, although both models are mainly English models, they have good command of many other languages, including Finnish. Their tokenization is, however, optimized for English. We tried two settings: the code-switching setting, with English instructions and Finnish data, and the monolingual setting, where everything

Course	Discipline	Answer length (in character)	Grading scale	Number of definitions
C1	Biology	∼150 / ∼600	(0,1,1) / (0,5,1)	10
C2	Software Engineering	∼250 / ∼550	(0,1,0.5) / (0,3,0.5)	5
C3	Business & Administration	∼250 / ∼600	(0,1,1)	6
C4	Economics	∼300	(0,2,0.5)	10
C5	Educational Sciences	∼400	(0,1,0.25)	10
C6	Medicine	∼400	pass, fail	0
C7	Medicine	∼50 / ∼400	various	0
C8	Psychology	∼275	(0,2,1)	10
C9	Psychology	∼375	(0,2,1)	0
C10	Psychology	∼300	(0,2,1)	0

Table 1: Summary of the selected courses. For the grading scale, the numbers in the parenthesis are in the format of (lowest score, highest score, interval). "Number of definitions" refers to how many of the selected questions out of ten questions ask for term definitions.

Jäljittelet opettajan arviointia. Arvostelet tenttejä ja kysymys, johon on vastattava, on seuraava: """<tenttikysymys>""" Mahdolliset arvosanat ovat: hylätty, hyväksytty *Esimerkkiarvosanat ovat muodossa "arvosana: opiskelijan kirjoittama vastaus" alla:* *hylätty: <esimerkki hylätystä vastauksesta>* *hyväksytty: <esimerkki hyväksytystä vastauksesta>* Arvioitava opiskelijan vastaus: <arvioitava vastaus> Anna vastauksesi muodossa: """ Arvosana: <arvosana> """	You are simulating the teacher's assessment. You are grading exams, and the question that needs to be answered is as follows: """<exam question>""" Possible grades are: failed, passed *Example grades are given in the format "grade: student's written answer" below:* *failed: <example of a failed answer>* *passed: <example of a passed answer>* Student's answer to be evaluated: <answer to be evaluated> Provide your response in the following format: """ Grade: <grade> """

Figure 1: The prompt template and its English translation. The angle brackets and the text within are replaced by the respective items of the actual data point. The italicized text are only included in the one-shot settings.

was Finnish. We found that code-switching only marginally damaged model performance. We choose the monolingual setting for further experiments to emphasize a language-specific context and facilitate clearer understanding for the model.

The model performance suffers a visible toll when more than one answer to be evaluated is included in a prompt. The model is shown one example per possible score in the one-shot settings. Performance under these settings are on average better than that under zero-shot settings, and one-shot setting performance is affected by the examples shown. For further experiments, we test both one-shot and zero-shot settings to investigate the effect examples have, and only include one answer for evaluation in each prompt.

Main Experiments

We choose 10 courses, each with 10 questions, and evaluate 20 answers per question. These 2000 examples are evaluated on chatbots based on GPT-3.5 and GPT-4 under zero-shot and one-shot settings. The design of the final prompt template takes into account the results from the initial experiments. The template and its English translation is shown in Figure 1. The models are called through the official OpenAI API. The chat model versions used are `gpt-3.5-turbo-0301` and `gpt-4-0613` respectively. The temperature of the models are set to zero for deterministic output, so that the predictions adhere better to the given template and that unwanted noises are otherwise reduced.

The sampling of answers for each question is stratified by grade. For each question, an instance is randomly selected for every possible grade. To avoid leaking gold standards to the model, the example answers are sampled once and used for all 20 prompts. They are excluded from the answers to be graded.

Baselines Given the absence of comparable baselines in the in-context learning setting, we calculate the majority

baselines for all metrics and settings by taking the most frequently occurring grade in the gold standards for each question. Training traditional machine-learning models is impractical due to the limited number of answers per question, sometimes with only a couple of answers beyond the 20 test answers and the few shown examples in the experiments.

Metrics

Our aim is to make the results as understandable to end users as possible, and we acknowledge that the conventional QWK metric can be abstract for end users. We therefore include two additional self-developed metrics based on accuracy and relative answer merits.

Quadratic-Weighted Kappa Quadratic-Weighted Kappa (QWK) is a statistical measure used to assess inter-rater agreement for categorical ratings. It is a standard ASAG metric and takes into account both the actual agreement between raters and the potential for chance agreement. Due to space limits, we refer readers to Section 4 of Bonthu, Sree, and Prasad (2021) for the definition of QWK. QWK values range between -1 and 1, where 0 means random agreement. Thus, the QWK of a majority baseline is 0. Expected QWK value rises as the number of categories increases (Brenner and Kliebsch 1996). If the grading scale includes decimal numbers, the score ranges are converted into categorical ratings. That is, all of the scores are multiplied by the smaller integer factor which converts all the scores to integers. Scaling scores uniformly does not affect the QWK value. For binary grading scale, pass is considered 1 and failed 0.

Tolerance-Adjusted Accuracy While QWK calculates the inter-rater agreement, for end users it can be more interpretable and useful to know the percentage of answers accurately scored. In addition, sometimes predictions within a certain threshold from the gold standards are acceptable. We calculate the percentage of accurately scored answers within a set tolerance, referred to as Tolerance-Adjusted Accuracy (TAA) thereafter. The tolerance is not meaningful for binary grading scale, so it is always set to 0.

For a set S of n answers $\{s_1, s_2, ..., s_n\}$, let their gold standard scores be $G = \{g_1, g_2, ..., g_n\}$ and their predicted scores be $P = \{p_1, p_2, ..., p_n\}$. The tolerance τ is determined by the user unless the grading scale is binary, in which case $\tau = 0$. The correctness of a prediction is determined by

$$C_i = \begin{cases} 1 & \text{if } |g_i - p_i| \leq \tau \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

TAA is then computed by

$$TAA = \frac{\sum_{i=1}^{n} C_i}{n} \times 100\%, \tag{2}$$

TAA with zero tolerance is equivalent to standard accuracy. An issue with TAA is that the gain in TAA as tolerance increases is unequal for grading scales of distinct granularity. TAA of questions with fewer number of possible scores increases more drastically as tolerance is loosened, but the increase in TAA does not reflect the improvement of the automatic scoring. Rather, it is a result of the low granularity of the scoring scale, and accepting such TAA in actual uses reflects the given level of score differentiation being sufficient for purpose. The tolerance, in turn, represents the allowable margin of error in scoring considered acceptable.

Relative Merit Consensus Relative Merit Consensus (RMC) recognizes that evaluators might hold divergent score margins. However, a consensus often emerges when assessing the relative merit of two answers. This metric considers student contentment, assuming students would dispute a lower-quality response receiving a higher score than theirs or scoring on par with theirs. RMC is calculated as the percentage of pairs of assigned scores that students would find satisfactory. Parity is assumed for equal-score answers. This perspective empowers teachers to establish score cutoffs based on observed quality of answers.

For a set A of n answers $\{a_1, a_2, ..., a_n\}$ to be evaluated, let its gold standard scores S_G be $\{s_{g1}, s_{g2}, ..., s_{gn}\}$ and predicted scores S_P be $\{s_{p1}, s_{p2}, ..., s_{pn}\}$. For every possible pair of answers (a_i, a_j) in A where $i \neq j$, it is possible that $s_{gi} = s_{gj}$ or $s_{pi} = s_{pj}$, but there must exist at least two distinct values within both sets S_G and S_P, i.e. neither S_G nor S_P is uniformly scored. The correctness of the pair of predicted scores (s_{pi}, s_{pj}) is defined by the binary function:

$$S(s_i, s_j) \begin{cases} 1 & \text{if } s_{gi} > s_{gj} \text{ and } s_{pi} > s_{pj} \\ 1 & \text{if } s_{gi} < s_{gj} \text{ and } s_{pi} < s_{pj} \\ 1 & \text{if } s_{gi} = s_{gj} \text{ and } s_{pi} = s_{pj} \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

RMC is defined as the fraction of correctly scored pairs out of all possible pairs:

$$RMC = \frac{\sum_{n}^{i=1} \sum_{n}^{j=i+1} S(s_i, s_j)}{\frac{n(n-1)}{2}}. \tag{4}$$

While students arguably should not contest a lower-quality answer scoring on par with theirs, because the difference in quality may not be big enough to receive distinct teacher-assigned scores, we calculate RMC by considering that students exclusively accept preserved relative scores. This approach is driven by the fact that the majority of questions in the dataset offer a very limited range of possible scores, which often results in RMC approaching 1 when a less elaborated answer scoring in parity is accepted. Similar to TAA, the RMC value is notably influenced by the number of potential scores an answer can attain.

Results

Instruction Compliance

The models are instructed to follow a specified output format (Figure 1), but their compliance to this format varies under different settings. Table 2 shows the number of outputs that cannot be parsed successfully without further processing. Despite this format issue, all of the outputs contain predicted grades that are within the respective sets of possible grades. GPT-3.5 occasionally adds explanations, especially to answers with the binary word-based grading scale. According to the ground truth, 59 out of 200 answers were assigned the grade 'failed' by teachers. In the zero-shot setting,

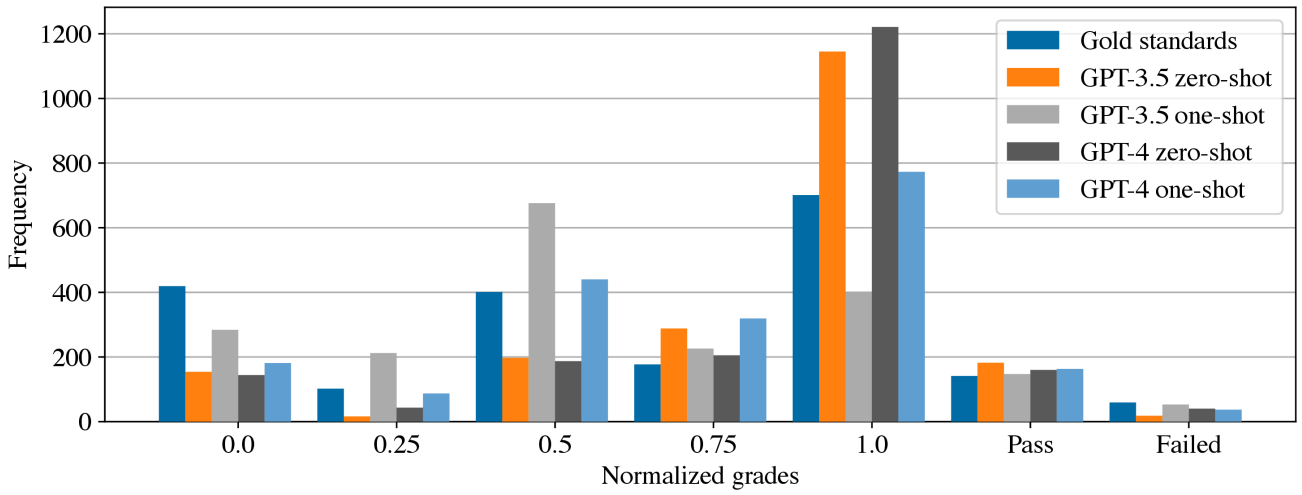

Figure 2: Distribution of normalized grades in gold standards and experimental conditions. The grades have been normalized based on the maximum achievable grade for each question. To enhance readability and account for infrequent occurrences, these grades have been rounded to the nearest 0.25.

Setting	**Incorrect output formats (no. of cases)**	**Total**
GPT-3.5 zero-shot	Has explanations (16)	16
GPT-3.5 one-shot	Missing 'Arvosana: ' (2) Has explanations (17)	19
GPT-4 zero-shot	Includes the answer (14) Includes the triple quotes (645)	659
GPT-4 one-shot	None	0

Table 2: Number of LLM outputs that do not conform to the given output format under each setting. There are 2000 outputs for each setting.

GPT-3.5 only predicted 18 as 'failed', 11 of which have explanations added to the predictions. In the one-shot setting, 53 answers were assigned 'failed', and seven have explanations. Many of the answers receiving explanations are assigned low scores by the models. GPT-4 wraps the answers within triple quotation marks, akin to the format shown in Figure 1, in 32% of its outputs in the zero-shot setting. In the one-shot setting, this confusion does not seem to occur, perhaps due to the use of delimiters in the examples.

Quantitative Performance

The models tend to assign higher scores than human evaluators. The distributions of normalized grades, or the evaluated scores divided by the highest possible scores, in gold standards and across the experimental conditions are depicted in Figure 2. Across all scenarios, the models are more lenient than human evaluators for questions with binary word-based grading scales. This pattern also holds for questions featuring numerical grading scales, except for the GPT-3.5 one-shot setting. Both models are stricter in the one-shot settings compared with the zero-shot setting, but still assign fewer failing grades to answers than human graders.

Unsurprisingly, one-shot GPT-4 attains the best performance. We refer to the curve that depicts the percentage of questions with equal or better performance on a given metric as the survival curve of the metric. Across all metrics (Figures 3-6), zero-shot GPT-3.5 performs worse than zero-shot GPT-4, which nearly matches one-shot GPT-3.5 in performance but is clearly outperformed by one-shot GPT-4. In a recent survey on ASAG systems (Haller et al. 2022), QWK scores were predominantly reported in the range of 0.6 to 0.8 across various datasets. In a related task, automated essay scoring, reported QWK scores ranged between 0.68 and 0.83, with a few outliers at 0.53 and 0.90 (Ramesh and Sanampudi 2022). On our dataset containing 100 questions, GPT-3.5 achieves QWK values of at least 0.6 in both zero-shot and one-shot settings for 8 and 21 questions, respectively. In the case of GPT-4, these numbers rise to 19 and 44, with 28 of these questions achieving QWK values of at least 0.7, and 13 reaching at least 0.8. No patterns emerge with regard to the discipline or type of the questions.

With a one-shot GPT-4 model, TAA with zero tolerance yields suboptimal results on our dataset, achieving an accuracy of at least 0.8 only 17% of the time. Nevertheless, the model performs well when assessing questions with a binary grading scale. Out of 25 such questions in the dataset, the model consistently achieves an accuracy of at least 0.55 for all questions and at least 0.75 for 21 questions. In Figure 5, the TAA survival curve with a tolerance setting of 1 on their own scale is presented. This curve indicates that, assuming similar data distribution to our dataset, educators providing one example per possible grade to GPT-4 can anticipate that, more than 60% of the answers will receive scores within a one-point difference of their actual score 95% of the time.

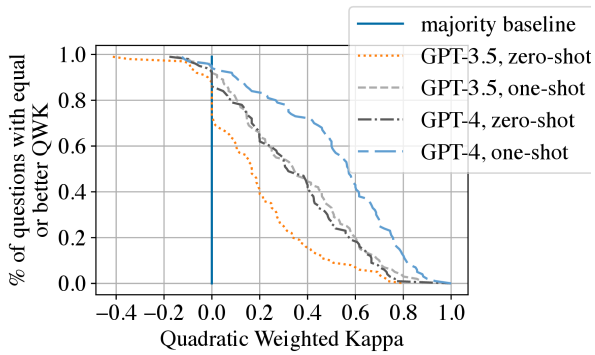

Figure 3: Quadratic Weighted Kappa across the settings: Percentage of courses meeting or exceeding threshold values.

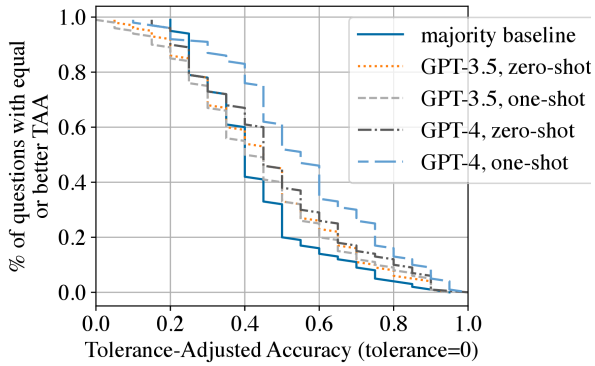

Figure 4: Tolerance-Adjusted Accuracy with tolerance=0 across the settings: Percentage of courses meeting or exceeding threshold values.

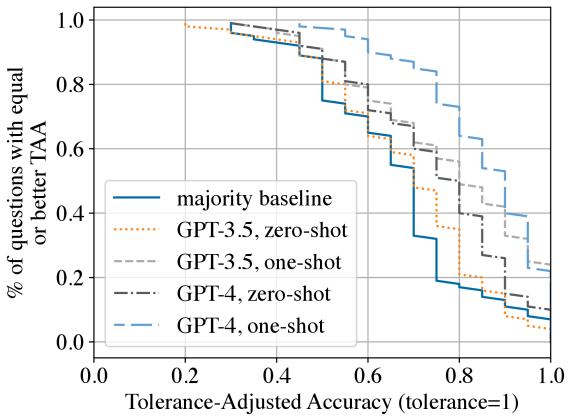

Figure 5: Tolerance-Adjusted Accuracy with tolerance=1 across the settings: Percentage of courses meeting or exceeding threshold values.

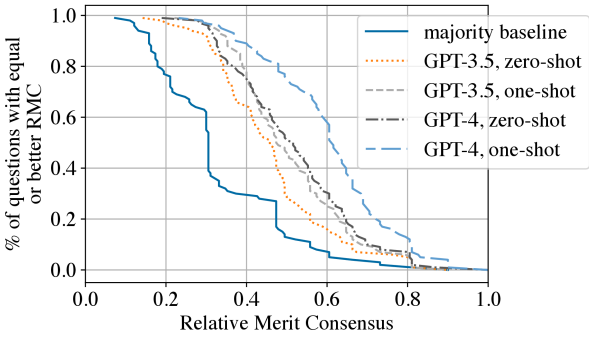

Figure 6: Relative Merit Consensus across the settings: Percentage of courses meeting or exceeding threshold values.

For those seeking a minimum accuracy threshold of 0.8, this level is attained 74% of the time, while a question-wise accuracy of at least 0.9 is realized 54% of the time.

The RMC survival curve is shown in Figure 6. The drastic decrease for the majority baseline occurring at approximately 0.3 and 0.45 can be primarily attributed to questions with three and two possible scores, of which there are 35 and 25 in our dataset, respectively. Assuming a threshold of 0.6 for acceptable correct relative scoring (meaning that at least 60% of the time two randomly selected answers will have a satisfactory relative scoring for students), one-shot GPT-4 meets this criterion in 59 out of 100 questions. When the threshold is increased to 0.8, 13 questions still meet the criteria, and 5 questions achieve an RMC of at least 0.9.

Top-Performing and Challenging Questions

Zero-shot GPT-3.5 tends to assign one single high score to all the answers when the answer lengths are too long. In the QWK survival curve (Figure 3), zero-shot GPT-3.5 takes a sharp decline at QWK=0. This is due to the model assigning a single score for 14 questions, with an additional question receiving predictions that are statistically akin to random values. This remaining question has an average answer length of 271 characters. The questions receiving uniform score predictions have an average answer length of 425 characters compared with the dataset average of 336. With the sole exception of one question with answers receiving a score of 1.5 out of 2, GPT-3.5 assigns full scores to all the answers for the remaining 13 questions. Other settings exhibit fewer questions with QWK=0; zero-shot GPT-4 has seven, one-shot GPT-3.5 has three, and one-shot GPT-4 has two such questions.

Figure 7 illustrates a weak, inverse correlation between model performance in terms of QWK and the length of answers. Table 3 shows the five top-performing and most challenging questions across all conditions and their average answer lengths. While this limited sample of five questions per category may not fully illustrate the trend, shorter-answer questions generally receive more accurate scores from models, irrespective of whether they request definitions. The top 10 scoring questions average 307-character answers, com-

	Top-performing questions
320	Ramsey rule
500	Non-covalent interactions
291	Proximal development zone
283	What are mirror neurons and to what cognitive phenomena are they linked?
579	What do the terms 'cohesion and 'coupling' mean? How are these concepts utilized in achieving software modularity (i.e., software partitioning)?
	Most challenging questions
281	How and when did Johan Haartman influence the history of medicine in Finland?
190	Hierarchical organization
527	In the planning phase, compromises are often necessary. Mention two objectives and provide an example of a compromise between them.
491	The development of antibiotics in the first half of the 20th century.
274	The relative stability of personality traits.

Table 3: The top-performing and most challenging questions for the models and their average answer length in character. All questions are ranked by QWK and selected based on their average ranking across all conditions.

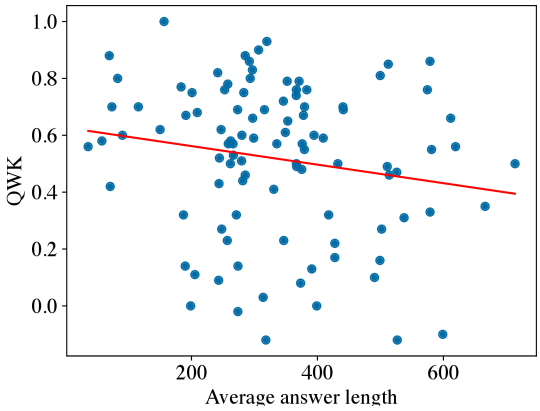

Figure 7: QWK values obtained by one-shot GPT-4 and the average length of answers in characters.

pared to 352-character answers for the bottom 10. From Table 3, it can also be observed that the questions exhibiting superior performance are characterized by a narrower scope, while the challenging ones encompass a broader range. This observation prompts further investigation to discern whether the broader-ranging questions elicit longer answers or if the content of the responses is inherently more challenging to grade.

Error Analysis

When examining questions with extreme QWK values, a notable trend emerges: the standard deviation of predicted scores tends to be larger for questions with higher QWK values. One-shot GPT-4 yields a Pearson correlation coefficient of 0.39 between the QWK values and the standard deviation of predicted scores, and this relationship is statistically significant with a p-value of 6.60e-05.

The few number of incorrect predictions with explanations offer an angle for error analysis. The following answer to the question "Who was Ambroise Paré and when did he have an impact?" is graded failed by zero-shot GPT-3.5:

> Ambroise Paré had an impact in the 17th century and he noticed that rose oil and turpentine worked better for gunshot wounds than pouring boiling oil into them. Boiling oil was the established method for treating gunshot wounds at that time and had a high mortality rate.

The reason for this grade is, as explained in the output, that Ambroise Paré had an impact in the 16th century, not the 17th century. In reality, Ambroise Paré lived from 1510 to 1590 and is considered the Father of Modern Surgery. Interestingly, the quoted answer is graded as a pass in other settings, as well as by the human evaluator. This could be due to the examiner putting focus on Ambroise Paré's achievements, along with the recognition that individuals can impact their field posthumously. This is also reflected in the human-assigned grades, as exemplified by the 'pass' example provided in the one-shot setting, which answered the 18th century.

Discussions

The overall results suggest that, even the best performing one-shot GPT-4 cannot be directly deployed for ASAG. Despite this, these LLM models are worth further exploration. This result aligns with concurrent research (Schneider et al. 2023). Regarding student satisfaction, research has indicated that, given an autograder of approximately 90% accuracy, students tend to overestimate the likelihood of an autograder marking a correct answer as wrong (Hsu et al. 2021). This overestimation is associated with student dissatisfaction and perceptions of unfairness. Hsu et al. (2021) offer several explanations for this overestimation: (1) algorithm aversion (Dietvorst, Simmons, and Massey 2015), where individuals tend to distrust algorithms after seeing them err, even when they surpass human performance, (2) students' perceptions being influenced by complaints from their peers, and (3) some students not being able to distinguish between true and false negatives. With the best-performing setting in this experiment, only 10 questions achieve a TAA with zero tolerance value of at least 0.9. This result falls short of making the models immediately usable. This aligns with the results analyzed using RMC, where only five questions achieve an RMC of at least 0.9. This means that in most of the questions, more than 10% of two randomly selected answers will have a relative ranking that is unsatisfactory to students.

This exploratory study does have certain limitations. The three metrics used are sensitive to the number of possible scores, and the questions have varying grading scales. There are no results available from a few-shot baseline using smaller language models like BERT for comparison. The ab-

sence of grading criteria and reference answer in the dataset constrains the content of the prompt. Due to the data collection method, it is possible that answers to the same question were graded by different evaluators. The extraction of data from its original educational context also poses challenges in assessing whether the model's accuracy truly reflects the accuracy of measuring underlying learning outcomes. The experiments are not conducted on an established English dataset. The dataset being non-English may have a negative impact on the reported results, as ChatGPT has been shown to perform less optimally on non-English data (Lai et al. 2023). On our dataset, one token from a dedicated tokenizer for Finnish[2] corresponds to around 2.11 tokens when processed by `cl100k_base`, the tokenizer of GPT-3.5 and GPT-4.

Numerous research directions await exploration in LLM-based ASAG. In terms of prompt engineering, aside from refining wording, the effects of increasing shots and integrating grading criteria merit investigation. Considering chain-of-thought prompting (Wei et al. 2022b) to elicit explanations before grade prediction, in contrast to the approach adopted in this study, also offers potential advancement in this task. The predicted grades have various applications. The model can complement human grading by acting as a second grader (Kulkarni et al. 2014). Based on the observation that the standard deviation of predictions for difficult-to-auto-grade questions tends to be low, human intervention can be considered when it is detected among a set of answers. Apart from direct grade prediction, other angles of using LLM include ranking or comparison of answers, as well as keyword extraction (Yoon 2023). Concerning the implementation of such systems, it is conceivable that certain students might try to exploit the system if they are aware of the presence of an automatic grading agent. Consequently, preventing cheating becomes a crucial area of research.

Conclusions

This study examined the feasibility of directly using LLM-based chatbots for the assessment of short answers. While immediate deployment presents certain challenges, the performance exhibited by one-shot GPT-4 justifies a more in-depth exploration across multiple dimensions. These avenues encompass investigating the impact of employing additional shots, enhancing question clarity, and providing the model with comprehensive information, including grading criteria and reference answers. Lastly, investing more computational resources to include explanation generation alongside grading and exploring alternative scoring methods are all examples of eligible avenues for future research in this field.

Acknowledgements

This research was supported by the Academy of Finland and the University of Turku Graduate School UTUGS. Computational resources were provided by *CSC — the Finnish IT Center for Science*. We warmly thank Totti Tuhkanen and Kaapo Seppälä for data collection and administrative support. Besides its reported use in the experiments, ChatGPT was used to proofread and improve the English wording of this paper.

[2]FinBERT tokenizer was used for calculation https://huggingface.co/TurkuNLP/bert-base-finnish-cased-v1

References

Bonthu, S.; Sree, S. R.; and Prasad, M. H. M. K. 2021. Automated Short Answer Grading Using Deep Learning: A Survey. In Holzinger, A.; Kieseberg, P.; Tjoa, A.; and Weippl, E., eds., *Machine Learning and Knowledge Extraction (CD-MAKE 2021)*, volume 12844 of *Lecture Notes in Computer Science*, 61–78.

Brenner, H.; and Kliebsch, U. 1996. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, 7(2).

Burrows, S.; Gurevych, I.; and Stein, B. 2015. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, 25(1): 60–117.

Cao, C. 2023. Leveraging Large Language Model and Story-Based Gamification in Intelligent Tutoring System to Scaffold Introductory Programming Courses: A Design-Based Research Study. arXiv:2302.12834.

Dan, Y.; Lei, Z.; Gu, Y.; Li, Y.; Yin, J.; Lin, J.; Ye, L.; Tie, Z.; Zhou, Y.; Wang, Y.; Zhou, A.; Zhou, Z.; Chen, Q.; Zhou, J.; He, L.; and Qiu, X. 2023. EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education. arXiv:2308.02773.

Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1): 114.

Filighera, A.; Ochs, S.; Steuer, T.; and Tregel, T. 2022. Cheating Automatic Short Answer Grading: On the Adversarial Usage of Adjectives and Adverbs. arXiv:2201.08318.

Filighera, A.; Steuer, T.; and Rensing, C. 2020. Fooling It - Student Attacks on Automatic Short Answer Grading. In *European Conference on Technology Enhanced Learning*.

Hackl, V.; Müller, A. E.; Granitzer, M.; and Sailer, M. 2023. Is GPT-4 a reliable rater? Evaluating Consistency in GPT-4 Text Ratings. arXiv:2308.02575.

Haller, S.; Aldea, A.; Seifert, C.; and Strisciuglio, N. 2022. Survey on Automated Short Answer Grading with Deep Learning: from Word Embeddings to Transformers. arXiv:2204.03503.

Hancock, G. R. 1994. Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of experimental education*, 62(2): 143–157.

Hsu, S.; Li, T. W.; Zhang, Z.; Fowler, M.; Zilles, C.; and Karahalios, K. 2021. Attitudes surrounding an imperfect AI autograder. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–15.

Kuechler, W. L.; and Simkin, M. G. 2010. Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test.

Decision Sciences Journal of Innovative Education, 8(1): 55–73.

Kulkarni, C. E.; Socher, R.; Bernstein, M. S.; and Klemmer, S. R. 2014. Scaling short-answer grading by combining peer assessment with algorithmic scoring. In *Proceedings of the first ACM conference on Learning@ scale conference*, 99–108.

Lai, V. D.; Ngo, N. T.; Veyseh, A. P. B.; Man, H.; Dernoncourt, F.; Bui, T.; and Nguyen, T. H. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. arXiv:2304.05613.

Li, Z.; Zhang, C.; Jin, Y.; Cang, X.; Puntambekar, S.; and Passonneau, R. J. 2023. Learning When to Defer to Humans for Short Answer Grading. In *International Conference on Artificial Intelligence in Education*, 414–425. Springer.

Matelsky, J. K.; Parodi, F.; Liu, T.; Lange, R. D.; and Kording, K. P. 2023. A large language model-assisted education tool to provide feedback on open-ended responses. arXiv:2308.02439.

Mizumoto, A.; and Eguchi, M. 2023. Exploring the Potential of Using an AI Language Model for Automated Essay Scoring. *SSRN Electronic Journal*.

Naismith, B.; Mulcaire, P.; and Burstein, J. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Workshop on Innovative Use of NLP for Building Educational Applications*.

Putnikovic, M.; and Jovanovic, J. 2023. Embeddings for Automatic Short Answer Grading: A Scoping Review. *IEEE Transactions on Learning Technologies*.

Ramesh, D.; and Sanampudi, S. K. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3): 2495–2527.

Schneider, J.; Schenk, B.; Niklaus, C.; and Vlachos, M. 2023. Towards LLM-based Autograding for Short Textual Answers. arXiv:2309.11508.

Tack, A.; Kochmar, E.; Yuan, Z.; Bibauw, S.; and Piech, C. 2023. The BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 785–795. Toronto, Canada: Association for Computational Linguistics.

Tornqvist, M.; Mahamud, M.; Guzman, E. M.; and Farazouli, A. 2023. ExASAG: Explainable Framework for Automatic Short Answer Grading. In *Workshop on Innovative Use of NLP for Building Educational Applications*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022a. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Yan, L.; Sha, L.; Zhao, L.; Li, Y.; Martinez-Maldonado, R.; Chen, G.; Li, X.; Jin, Y.; and Gašević, D. 2023. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*.

Yancey, K.; LaFlair, G. T.; Verardi, A.; and Burstein, J. 2023. Rating Short L2 Essays on the CEFR Scale with GPT-4. In *Workshop on Innovative Use of NLP for Building Educational Applications*.

Yoon, S.-Y. 2023. Short Answer Grading Using One-shot Prompting and Text Similarity Scoring Model. arXiv:2305.18638.

Zeng, Z.; Li, L.; Guan, Q.; Gašević, D.; and Chen, G. 2023. Generalizable Automatic Short Answer Scoring via Prototypical Neural Network. In *International Conference on Artificial Intelligence in Education*, 438–449. Springer.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.-Y.; and Wen, J.-R. 2023. A Survey of Large Language Models. arXiv:2303.18223.