Quality-Conditioned Agreement in Automated Short Answer Scoring: Mid-Range Degradation and the Impact of Task-Specific Adaptation

Abigail Victoria Gurin Schleifer[1] **Moriah Ariely**[1] **Beata Beigman Klebanov**[2]
Asaf Salman[1] **Giora Alexandron**[1]
[1] Weizmann Institute of Science, Rehovot, Israel
[2] ETS, Princeton, USA
{abigail.gurin-schleifer,moriah.ariely,giora.alexandron, asaf.salman}
@weizmann.ac.il
bbeigmanklebanov@ets.org

Abstract

Automated short answer scoring (ASAS) is shifting from discriminative, fine-tuned models to large language models (LLMs) used in few-shot settings. This paradigm leverages LLMs' broad world knowledge and ease of deployment, but limited task-specific data may reduce alignment on complex scoring tasks. In particular, its impact on scoring partially correct responses that require nuanced interpretation remains underexplored. We investigate the relationship between the degree of task-specific adaptation of different models and quality-conditioned scoring agreement. We compare three LLMs (GPT-5.2, GPT-4o, Claude Opus 4.5) in few-shot mode, a fine-tuned BERT-based encoder, and a human expert on two open-ended biology items, using several hundred student responses and ground truth scores provided by a biology education expert. The results show that human–human agreement is highest and stable across the full quality spectrum. All AI models perform well on fully correct and fully incorrect responses, but exhibit substantial degradation on mid-range responses. This mid-range degradation is conditioned on task-specific adaptation: It is most severe in few-shot LLMs with few examples and decreases as task-specific data increases, with fine-tuned encoder models performing best. This mid-range degradation may lead to inequitable evaluation of responses produced by students with developing understanding. Our findings highlight the importance of quality-conditioned fairness, with particular attention to mid-range responses.

1 Introduction

Automated short answer scoring (ASAS) of open-ended responses is a central application of natural language processing (NLP) in education (Bonthu et al., 2021; Haller et al., 2022). ASAS approaches can be classified into reference-based – scoring based on similarity to reference graded responses – and example-based, where the model learns to map student responses to rubric scores using task-specific labeled data. The fundamental trade-off is that reference-based approaches require only a few scored examples but their accuracy is typically very limited (Bexte et al., 2023), while example-based ones can reliably learn to mimic expert grading but require ample training data (Gurin Schleifer et al., 2023) and are still prone to gaming (Ding et al., 2020).

Early example-based systems relied on handcrafted linguistic features and machine learning models to approximate human scoring (Haller et al., 2022). More recent work has leveraged neural machine learning that can capture deeper semantic relationships in student responses. Such neural-based automated assessment systems primarily relied on discriminative machine learning implemented on top of encoders that transferred raw responses into vectorized embeddings (Condor, 2020; Li et al., 2021; Gurin Schleifer et al., 2023).

The emergence of generative LLMs introduced a new paradigm for ASAS. Decoder-based language models can evaluate student responses based on their extensive pretraining and general world knowledge, relying on instructions to guide in-context interpretation (e.g., according to given rubrics), rather than on adapting their parameters using task-specific datasets (Lin et al., 2023).

However, because generative models are not explicitly trained on the application of a specific assessment framework to student responses, their scoring may be less aligned with expert grading, particularly in assessment contexts that have domain-specific standards and require subtle interpretation (Wu et al., 2025; Yacobson et al., 2025). Few-shot prompting, which is a prompting strategy that provides a few labeled examples, was developed to help the model align its interpretation with that of experts (Lin et al., 2023). This raises the question of the amount of task-specific data that

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 432–442

various types of architectures require to align with expert grading.

The alignment of automated scoring systems with human grading is typically evaluated using metrics like κ or Pearson's correlation between model predictions and human raters (Beigman Klebanov and Madnani, 2022; Williamson et al., 2012). That said, it is also recognized that overall scoring accuracy may not be sufficient for validating an automated scoring system, as its performance can vary systematically across student subpopulations, raising fairness concerns (Williamson et al., 2012). While the psychometric theory discusses sub-populations to be examined quite generally, e.g., as "identifiable and relevant" subgroups (Xi, 2010), much of the literature on fairness in educational AI tends to focus on demographics-based sub-groups (Loukina et al., 2019; Madaio et al., 2022) and on high-stakes, large-scale assessment (Johnson and McCaffrey, 2023). Discussing fairness in the formative classroom assessment specifically, Camilli (2006) noted that the use of the assessment is generally "to 'locate' the presenting proficiencies of students in order to guide instruction along the contours of their strengths and weaknesses", and help instructors "see if there is anything that, if not addressed, sets certain students up to fail or otherwise miss significant opportunities to learn" (p. 248). To help provide the right learning opportunities for learners at different current states of knowledge, these states of knowledge need to be detected accurately and equitably, as otherwise students whose knowledge state tends to be mis-recognized would be in danger of being provided inappropriate learning support.

In the recent literature on automated analysis of student responses, there are indications that states of partial knowledge may be harder to detect automatically than completely correct knowledge states (Gurin Schleifer et al., 2025; Kortemeyer, 2023) or than both completely correct and completely incorrect ones (Grévisse, 2024). In such cases, students demonstrating partial mastery of a concept (mid-range responses) may receive less reliable evaluation than those whose responses are clearly correct or clearly incorrect. We refer as "quality-conditioned fairness" to the requirement that ASAS system performs equally well across the entire range of response quality.

Due to the nuanced interpretation required for mid-range responses, we expect that task-specific labeled data will be particularly important for AI grading systems when grading such answers, and that its quantity will be associated with the systems' accuracy on these responses.

However, little is known about agreement patterns across the response quality spectrum, and how this relates to model architectures – encoder-based discriminative models and decoder-based generative ones – and across different levels of task-specific adaptation (few-shot with varying levels of 'few', and fine-tuning). Thus, our research is guided by the following research question:

RQ: *What is the association between response quality and inter-rater agreement for human-human vs human-AI pairings for different levels of task-specific adaptation of the AI models? (few-shot with different numbers of examples vs fine-tuned)?*

The data used for this study included responses from about 800 high-school students to two open questions in biology requiring students to provide scientific explanations in the domain of cellular respiration. The responses were graded according to an analytic grading rubric by two biology education experts. To answer the RQs, we compared three generative decoder-based models in a variety of few-shot modes, a discriminative fine-tuned encoder model and a human grader in terms of their agreement with the expert human grader overall and across different levels of response quality.

Previewing the results, we show that disagreements between human and AI raters concentrate among mid-quality responses, while agreement remains high at the extremes of the quality spectrum. They also demonstrate a consistent pattern of **mid-range degradation of AI scoring accuracy**, meaning that AI graders, but not the second human expert grader, tended to be less aligned with the expert on the mid-range responses, and that this tendency was partially mitigated as the AI system received more task-specific data.

These findings highlight the importance of task-specific adaptation for accurately evaluating partially correct responses and suggest that evaluation frameworks for AI-assisted assessment should explicitly consider quality-conditioned, and specifically mid-range, assessment fairness.

2 Related Work

Fine-tuning the parameters of the model on task-specific data is a common technique for adapting encoder-based pre-trained language models such

as BERT (Devlin et al., 2019) to the given task. The more recent models, while still utilizing the Transformer architecture, use orders of magnitude more parameters (hence their designation as *large* language models, or LLMs), and demonstrate in-context learning, namely, the ability to perform well on tasks that the model has not been trained for with only a small number of examples (few-shot). While larger models generally demonstrate stronger in-context learning, empirical studies suggest that parameters other than size, such as the design of the prompt, the model architecture, the diversity of the training data, and the selection of the few-shot examples can impact the model's in-context learning ability (Berti et al., 2025).

A number of studies compared the performance of fined-tuned encoders and few-shot LLMs on ASAS. Chamieh et al. (2024) evaluated ASAS performance on datasets from various domains, including mathematics, science, and English language arts. They found that fine-tuned models generally performed better, especially for tasks that require more complicated reasoning or domain-specific knowledge. Kortemeyer (2024) compared the performance of GPT-4, BERT, and RoBERTa-large (Liu et al., 2019) on ASAS datasets in varied science domains. They found that the fine-tuned BERT and RoBERTa-large outperform the general-purpose GPT-4 LLM. Henkel et al. (2024) found that few-shot models outperformed zero-shot ones in the context of ASAS, suggesting that more task-specific information may help improve scoring performance. Ferreira Mello et al. (2025) studied ASAS in English in the context of undergraduate computer science and in Brazilian Portuguese in 8th grade biology and found that fine-tuned BERT models outperformed few-shot GPT-4o on both datasets. Across these studies, the evaluation of performance was done using overall summary measures, such as RMSE, F1, Pearson's correlation, or kappa-family metrics. In particular, it remains unclear *what kinds of responses* were easier or harder for the different models to score.

There are indications in the literature on automated scoring of constructed responses that point towards a specific weakness of LLM-based scoring, namely, the tendency of these models to over-score low quality responses. Chang and Ginter (2024) examined the quality of GPT-4 scoring, in zero-shot and one-shot setting, of short answers in a variety of undergraduate courses in Finnish. They found that the models were too lenient, both in binary pass/fail scoring and in scoring on a scale: Too few responses were marked as fail, and too few responses were assigned to the bottom half of the scale. Kortemeyer (2023) reported similar results for scoring short answers in introductory physics – GPT-4 was a more lenient rater. The author noted that agreement between human and AI scores was stronger at the high end of the response quality spectrum. In the context of medical undergraduate courses, Grévisse (2024) observed that different LLMs were more or less severe than human raters on data from different courses. They further observed that there was stronger agreement between the LLM and human ratings for the completely incorrect and completely correct answers than for partially correct answers and noted GPT-4's especially high precision on detecting the fully correct answers. These findings suggest that automated scoring should be examined from the point of view of response quality, as there is evidence that the scoring accuracy may be uneven across the score distribution. However, it is not clear whether few-shot LLM scoring is more or less susceptible to uneven performance than the more extensive task adaptation through fine-tuning. It is also not clear whether any unevenness in AI scoring may have to do with inherent difficulty of discriminating between scores at certain levels, namely, whether human scoring exhibits a similar pattern. Our contribution is a systematic examination of multiple scoring mechanisms – human, fine-tuned pre-trained language models, and few-shot LLMs – in terms of the relationship between scoring accuracy and response quality, in the context of ASAS.

3 Methodology

3.1 Assessment Items and Scoring Rubric

We analyzed student responses to two open-ended items in biology. The items deal with cellular respiration, and specifically, the effect of smoking and anemia on physical exercise. The items are conceptually similar and differ in context and surface features, and students' explanations are expected to have the same structure, and use similar argumentative flow. They are similar to questions that frequently appear on the biology national matriculation exam. They were administered in Hebrew and their translation is presented below.

Smoking (Item 1): Cigarette smoke contains harmful substances, including CO, which binds haemoglobin more strongly than oxygen. Explain

how high CO levels impair exercise ability.

Anemia (Item 2): A man with low red blood cell levels (anaemia) reports weakness and difficulty exercising. Explain how reduced red blood cells make exercise difficult.

Due to their conceptual similarity, the items share the same grading rubric. It is an analytic grading rubric that decomposes a student's scientific explanation into a collection of 10 binary categories, each representing an essential property that the student response should include. The items and the rubric were developed by the 2nd author of this paper. The full description of the grading rubric can be found in Ariely et al. (2025).

3.1.1 Data Collection, Scoring, and Partitioning

Student responses were collected in two cycles conducted in two different academic years, with a two-year gap between the cycles. In the first cycle, student responses to an instrument that contained both items were collected from 669 students in grades 10-12 attending 25 high schools across the country. For each category in the rubric, a student's response was graded as '1' if that category was addressed in the text, and '0' otherwise. Approximately 5% of the 669 responses were annotated jointly by two human raters – Rater1 (the second author) and Rater2. Then, an additional 20% of the 669 responses were annotated independently by both raters establishing inter-rater agreement on $n = 130$ responses (κ ranged between 0.89 and 0.98 across categories). Last, Rater1 coded the remaining responses; see Ariely et al. (2023) for more details on the annotation process. The data was used as **training data** for the models, namely, for fine-tuning the BERT-based models and for picking the few-shot examples for the LLM-based grading.

In the second cycle, responses to the same instrument were collected from 152 students (in total, 304 responses) of similar demographics, and scored by Rater1 (the 2nd author of this paper). We refer to this scoring as the ground truth labels. These responses were used as **test data** for all the experiments described below. A second human rating was obtained for the test items as part of the current study of human vs AI raters and will be described in more detail in section 3.3.2.

3.2 Evaluation Metric for Scoring Differences

The scoring of each response is represented as a 10-dimensional binary vector, where each dimension corresponds to a category, with '1' indicating a correct response on that category. As noted above, we referred to the scoring of Rater1 as the ground truth labels. The overall score of a response is thus defined as the number of categories (i.e., bins) that Rater1 marked as correct, which ranges from 0 to 10, meaning that higher scores indicate higher response quality.

3.2.1 Manhattan Distance for Scoring Differences

For a student response s, we denote the gold labels of s by: $l_s^G = (l_1^G, ..., l_{10}^G)$, and the labels assigned by Rater2 (Human or AI-based) – by: $l_s^{R_2} = (l_1^{R_2}, ..., l_{10}^{R_2})$. l_s^G and $l_s^{R_2}$ are binary 10-dimensional vectors. We measure the absolute difference between the two scorings by the Manhattan-distance (also called the L_1-distance) (Chen and Ng, 2004) between l_s^G and $l_s^{R_2}$:

$$\left\| l^G - l^{R_2} \right\|_1 = \left| l_1^G - l_1^{R_2} \right| + ... + \left| l_{10}^G - l_{10}^{R_2} \right|, \quad (1)$$

which equals the number of categories scored differently by Rater2 compared to the gold labels.

3.3 Experimental Set-Up

3.3.1 The Raters and the Scoring Task

The experimental set-up included comparing the alignment to the ground truth labels of the following scoring mechanisms: (i) a second human expert, (ii) a fine-tuned BERT-based model, and (iii) three LLMs – GPT-5.2 (Singh et al., 2025), GPT-4o (OpenAI, 2023), and Claude-opus-4.5(Anthropic, 2025). At the time of writing, GPT-5.2 and Claude-opus-4.5 are the frontier models of OpenAI and Anthropic, while GPT-4o represents OpenAI's previous generation model, which is the current workhorse for many LLM tasks due to its good combination of capabilities, speed, and cost.

The task of the graders was to score each response of the test set (152 responses for each item) according to each of the categories of the rubric, producing a 10-dimensional binary vector per response. For the LLMs we took the majority result: Each LLM was invoked three times for every category on every response; the result that appeared most of the time is considered the final LLM score.

We then computed the Manhattan distance, per response and rater, between the Rater2's scoring

vector and the ground truth vector, as described in Section 3.2.1. Below, we provide more details about the scoring procedure.

3.3.2 Scoring Procedures by Scoring Mechanism

Below we describe how the scoring of the test set was implemented for each scoring mechanism.

1. **Human**: The test set was labeled according to the rubric by a second biology education expert who scored part of the 1st cycle dataset (that was used as training data in this study) a few years ago, after a short recap with Rater1 on the rubric and its implementation.
2. **Fine-tuned BERT-based Classifier**: The BERT-based model DictaBERT (Shmidman et al., 2023) was fine-tuned on the training data (n = 669). A separate binary classifier was trained for each item and category, meaning that overall, 20 classifiers were fitted. More details can be found in Ariely et al. (2025).
3. **Few-shot LLMs**: The generative models – GPT-4o, GPT-5.2 , and Claude Opus 4.5 – were prompted with assessment instructions. Each category in the rubric had its own prompt for each item, overall $2 \times 10 = 20$ prompts. The prompts followed a straightforward few-shot ASAS strategy and their general structure was: role definition, task definition, per-category scoring instructions, and few-shot examples, split equally between positive and negative examples. An example of a prompt for one of the items and categories is provided in Appendix A. We acknowledge that there are various prompting strategies for ASAS. Our approach was to rely on established prompting strategy developed and tested in previous ASAS research. Thus, we adopted the prompts of Ariely et al. (2025), which were initially developed for a GPT-4o grader. We note that Ariely et al. (2025) tried several combinations of few-shot examples, specifically $2/4/6/8/10$-shot (with an equal split between positive and negative examples).
   Determining the Number of Few-Shot Examples for the Frontier Models. Prior to experimenting with the costly models (GPT-5.2 and Claude Opus 4.5), we analyzed the results of running the $2/4/6/8/10$-shot prompts with GPT-4o. The results showed that 10-shot prompting achieved better results, both in terms of overall (average) agreement and in terms of mid-range fairness, as can be seen in Figure 1. Thus, for the costly models, we ran only the 10-shot version. For the few-shot examples, we randomly chose five positive and five negative examples from the training set. The total cost of running the ten 10-shot prompts on the test set for each item for each of the two LLMs (40 prompts in total) was around $1,000.

4 Results

Table 1 shows the average L1 distance between the gold labels and the scoring of the second human rater (H2) and the AI models, per score level (number of correct categories – #CC – in the gold labels), for each of the two items. Figure 1 shows the same information visually, including standard deviations shown in the relevant background color.

Let us first consider the overall performance of the different models on the two items. On average and across score levels, H2 had the highest agreement with the expert, followed by the fine-tuned BERT-based models, followed by the LLMs in few-shot mode. Among the GPT models, we did not observe an advantage for GPT5.2 over the prior generation GPT4o model with the same number of few-shot examples (10). For GPT4o – the only model that was tested with several numbers of few-shot examples – providing 6-10 examples resulted in approximately similar performance, while providing only four examples resulted in performance degradation on Item 2, and providing only two examples resulted in a substantial degradation in the middle range of the scores – difference of 3 categories or more, on average, for scores with 4-6 correct categories out of 10 – for both items.

Analyzing the results by score level (number of correct categories out of 10), we observe excellent AI-model performance on the extremes – best and worst quality responses – regardless of model-type and amount of task-specific data. In contrast, there is a degradation for partially correct responses, especially for the few-shot models.

While on average fine-tuned models performed better than few-shot ones, the GPT models in all few-shot scenarios – including those with only two examples – performed very well on fully or almost fully correct answers (score 9-10), showing average distance of less than one category from the gold score. On the fully correct responses, the GPT models outperformed the fine-tuned models on both items. The exceptionally strong performance of few-shot models on the best responses

aligns with some similar reports in the literature (e.g., Grévisse (2024); Kortemeyer (2023)).

Next, we observe that all GPT models perform well on the lowest end of the scale – score levels 0 and 1. With the exception of two-shot GPT4o on Item 2, in all other cases the average distance from the gold label is less than one category. Although fine-tuned models do better, with average distance of at most 0.5 category from the gold label for score 0 and 1 across the two items, few-shot models are also fairly accurate.

Table 1 shows that all few-shot models suffer a substantial degradation in performance in the middle range of the scores. For all six GPT models, for each of the four score levels 4-7, for both items, all but one of the $6 \times 4 \times 2 = 48$ results show average distance of more than two categories from the gold label. Fine-tuned models also perform worse in this middle range than in the extremes, but the degradation is not as severe – the maximum average distances from the gold scores is 1.5 categories. In contrast, the human rater H2 does not show a clear pattern of degradation in agreement with the gold labels in the middle of the score range.

Our final observation is that of the erratic behavior of the Claude model. While the overall and the score-level-dependent patterns of errors are highly consistent across the two items for H2, fine-tuned, and GPT family models, Claude's performance oscillates dramatically – it is the best few-shot model for Item 1 but is by far the worst model for Item 2.

5 Discussion & Conclusions

Quality-conditioned agreement. The findings show that agreement is strongly conditioned by response quality and that this relationship differs considerably between human-human and human-AI pairings. Human raters exhibit the highest agreement, which is also stable across the response quality range and shows no mid-range degradation, suggesting that human experts are relatively consistent in interpreting partially correct responses. In contrast, the AI models showed a U-shaped pattern:[1] Agreement is highest at the extremes (fully correct or incorrect responses) and substantially lower in the middle range, indicating that model–human agreement is not uniform across the scoring range. This mid-range degradation pattern complements a recent ASAS study showing that criterion-level agreement declines as cognitive complexity increases (Emirtekin and Özarslan, 2026), and a similar trend found in project-based learning (Usher and Faraon, 2025).

Task Specific Adaptation. The results also showed that task-specific adaptation emerged as a key factor in improving agreement. The fine-tuned model, which had access to the highest amount of task specific data, outperformed scoring based on LLMs with few-shot examples across both items, most notably on mid-range responses. In line with this, the initial experiment on the number of few-shot examples on GPT-4o demonstrated a similar pattern, with the U-shape disagreement curve having a larger 'belly' for lower numbers of few-shot examples (see Figure 1). While the advantage of fine-tuned models on LLMs in few-shot mode on ASAS tasks was shown previously on the aggregated level (Kortemeyer, 2024; Henkel et al., 2024; Ferreira Mello et al., 2025), the present work suggests that this advantage may be mostly concentrated in the mid-range examples. In fact, for the fully correct/incorrect responses, LLMs performed well with very little task-specific data.

Possible explanations for the observed pattern. In trying to explain these results, it may be that extreme responses (fully correct or incorrect) provide clearer signals that are easier to classify, while mid-quality responses require nuanced judgment and interpretation of rubric criteria. Experienced human experts may be better at applying such nuanced educational judgment (Yacobson et al., 2025), and encoder models can learn to distinguish between them using labeled data that changes their internal representation in a way that can separate between them. LLMs, when provided with a rubric and few-shot examples, utilize in-context pattern matching to identify semantic and structural alignments between the criteria, the examples, and the student's input. Our empirical results that show that LLMs are less good than fine-tuned encoders on ASAS tasks are in line with previous findings (e.g., Ferreira Mello et al. (2025)).

Implication for automated scoring in practice. The great promise of few-shot automated scoring is fast deployment of the automated scoring model, thus saving instructors valuable time through automated evaluation of student responses without having to wait until sufficient human-scored data is available for fine-tuning. Our results suggest that such scoring may be inequitable, with substantially inaccurate scores concentrating on the partially in-

[1] Specifically, the graphs in Figure 1 show an inverted U-shape since they show discrepancy rather than agreement.

#CC	H2	FT	Claude	GPT 5.2	GPT 4o10	GPT 4o8	GPT 4o6	GPT 4o4	GPT 4o2	Median
Item 1										
0	0	0.09	0.80	1.10	0.60	0.70	0.80	0.90	0.80	0.80
1	0.12	0.25	0.40	0.90	0.40	0.60	0.50	0.70	0.90	0.50
2	0.08	0.23	0.64	1.32	0.64	0.64	0.86	1.05	1.95	0.64
3	0.67	1.33	1.27	2.45	2.09	1.55	1.91	2.18	3.18	1.91
4	0.45	1.27	1.33	2.25	2.17	2.83	2.67	2.83	3.17	2.25
5	0.12	1.12	1.30	2.70	2.4	3.00	2.60	2.60	3.50	2.60
6	0.43	0.91	1.61	2.96	2.43	2.91	2.64	2.32	3.00	2.43
7	0.12	1.44	1.22	2.44	2.33	2.56	2.22	1.83	2.61	2.22
8	0.12	0.56	0.88	1.38	1.38	1.62	1.44	1.25	1.25	1.25
9	0	1.17	0.50	0.67	0.83	0.83	0.83	0.67	0.33	0.67
10	0.09	0.27	0.09	0.09	0.18	0.18	0.18	0.18	0.09	0.18
Avg.	0.20	0.79	0.91	1.66	1.40	1.58	1.51	1.50	1.89	
Stdev.	0.42	0.86	0.82	0.98	0.79	0.91	0.96	0.95	0.93	
Item 2										
0	0	0.17	5.57	0.71	0.29	0.57	0.43	0.86	0.86	0.57
1	0.38	0.50	4.90	0.90	0.70	0.60	0.60	0.90	1.50	0.70
2	0.50	0.75	5.14	0.57	0.57	1.00	0.71	1.29	2.29	0.75
3	0.45	0.64	4.38	1.43	0.71	0.93	1.21	1.57	2.43	1.21
4	0.14	0.86	4.25	2.50	2.62	2.62	2.38	3.00	3.00	2.62
5	0.33	1.50	4.00	2.33	2.22	2.44	2.33	2.33	3.33	2.33
6	0.50	1.29	4.06	2.61	2.78	2.83	3.00	3.28	3.28	2.83
7	0.11	1.39	3.50	2.59	2.55	2.64	2.68	2.82	2.86	2.64
8	0.42	1.23	2.70	1.32	1.82	1.86	1.89	1.75	1.79	1.79
9	0.25	1.12	1.76	0.71	0.88	0.76	0.94	0.82	0.82	0.82
10	0.09	1.18	2.18	0	0	0	0	0	0	0
Avg.	0.29	0.88	3.86	1.42	1.38	1.48	1.47	1.69	2.01	
Stdev.	0.49	0.85	1.99	0.90	0.94	0.86	0.90	0.97	1.09	

Table 1: Average L1 distance per number of correct categories (#CC), based on the gold labels. H2 is the second human rater. Top panel: Item 1; bottom panel: Item 2. Light gray cells show L1 distance ≥ 1; darker gray shows L1 distance ≥ 2; the darkest gray shows L1 distance ≥ 3.

correct responses. However, the high agreement on fully correct and fully incorrect responses, even for LLM models with very few examples, suggests that automated scoring as a first pass can identify these responses and handle them without human intervention, while responses predicted to be in the middle range would be routed to the human rater for scoring. Gradually, as more human-scored data is accumulated and fine-tuning becomes feasible, the scoring can be automated more fully.

Implications for fairness. Bias is a well-recognized hazard in automated scoring of constructed responses (Madnani et al., 2017; Loukina et al., 2019; Gorgun and Yildirim-Erbasli, 2026). Much of the existing literature has focused on demographic-group–based biases, where students from different groups (e.g., by socioeconomic status (Chinta et al., 2024; Baker and Hawn, 2022) gender (Madnani et al., 2017), or linguistic background (Loukina et al., 2019)) receive systematically different scores despite comparable skill levels. However, bias can also arise at the individual level, where similarly skilled students are treated differently due to consistent features of their responses – for example, students who write in a more concise style may be systematically scored differently (Madnani et al., 2017; Beigman Klebanov and Madnani, 2022). In addition, statistical biases may emerge from properties of the data, for instance, if examples that include certain properties, such as specific misconceptions, are not represented enough in the data for the models to learn

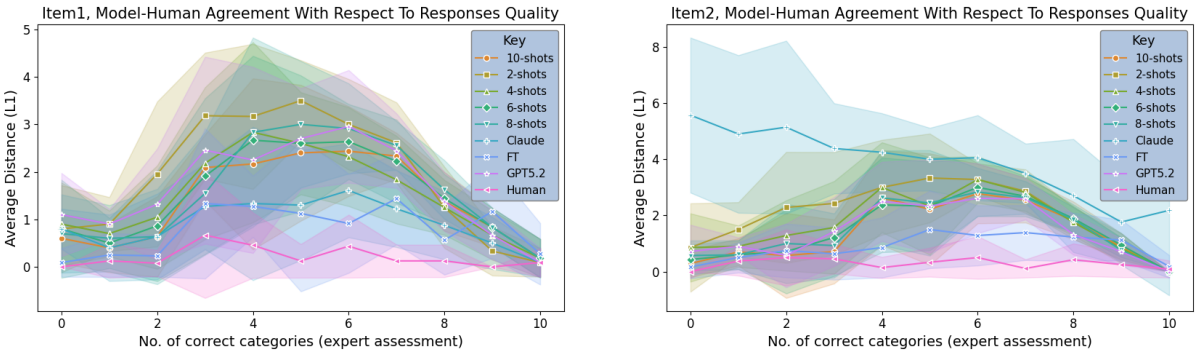

Figure 1: Human-Model Agreement With Respect To Student's Response Quality For All The Models

their linguistic signature.

Statistical bias can also take the form of a representation bias. For example, Gurin Schleifer et al. (2025) identified what they termed an 'Anna Karenina principle', showing that as responses become more incorrect, their representations in the embedding space become more spread, making it harder to automatically cluster together responses of the same error type. This creates a differential measurement error that disproportionately affects lower-quality responses, and may constitute a fairness problem if such clusters are used to design feedback targeted at each knowledge group, as proposed by Ariely et al. (2024). In this case, low-quality responses may receive unsuitable feedback.

Our results contribute to this line of research by identifying a form of measurement bias that appears in ASAS. We observe a U-shaped pattern in which scoring errors concentrate in mid-range responses. These responses are typically produced by students whose knowledge is still developing. Such students may benefit the most from formative feedback, as they are often better able to use feedback to refine their understanding than lower-performing students (Roll et al., 2014; Shute, 2008). Additionally, because mid-range responses contain both correct and incorrect elements, scoring errors – especially misjudging correct components – may be particularly confusing. As a result, reduced accuracy in this range may hinder learning at a critical stage, making this a fairness concern.

We define quality-conditioned fairness as the requirement that automated scoring systems maintain consistent accuracy across the full spectrum of response quality. As we showed, this form of fairness is closely tied to the availability of task-specific data, with greater adaptation reducing mid-range errors. By identifying and characterizing this pattern, we turn it from an "unknown bias" into a "known bias," which is a necessary step toward improving fairness in educational AI systems in general (Baker and Hawn, 2022).

Future work. To better understand the phenomenon of mid-range degradation, it would be useful to extend this research to additional instruments and domains, and to examine the direction of misalignment, as prior work suggests that LLMs may tend to over-score incorrect responses. To improve mid-range alignment, it would be valuable to investigate whether including more mid-quality examples in few-shot prompts can reduce the observed degradation in scoring. A more human-in-the-loop approach worth investigating is whether task-specific prompts can achieve higher alignment, and how to train domain experts or teachers to develop and validate them within their local context.

Contribution. This work identifies mid-range scoring degradation in ASAS and shows that it is strongly governed by the degree of task-specific adaptation. A key takeaway is that LLM-based ASAS systems must be evaluated for quality-conditioned agreement, with particular attention to mid-range responses.

Limitations

The prompting approach used for the LLM-based scoring mechanisms followed a common few-shot strategy drawn from prior work. However, these prompts were originally designed and validated for GPT-4o. It is possible that alternative prompting strategies, or prompts tailored to each specific model, could yield different performance patterns. Additionally, the assessment used in this study consisted of two items in biology. Although these items were selected to represent common question types found in standard assessments on the topic,

they still constitute a limited sample of possible tasks. Thus, the generalizability of the findings to other items and domains remains to be established.

This research does not employ standard evaluation metrics such as Quadratic Weighted Kappa (QWK) or Pearson's correlation, which may hinder direct comparison with existing ASAS benchmarks. Incorporating standard quality-conditioned agreement metrics should be considered in future work.

Ethics statement

We acknowledge that this work complies with the ACL Code of Ethics. The research and its data collection procedures were approved by the Institutional Review Board and the Ministry of Education.

Acknowledgments

The authors thank Cipy Hofman for her contribution to the research. This work was supported by the Knell Family Institute for Artificial Intelligence, Israel.

References

Anthropic. 2025. Claude opus 4.5 system card. Technical report, Anthropic. System card.

Moriah Ariely, Tanya Nazaretsky, and Giora Alexandron. 2023. Machine learning and Hebrew NLP for automated assessment of open-ended questions in biology. *International Journal of Artificial Intelligence in Education*, 33(1):1–34.

Moriah Ariely, Tanya Nazaretsky, and Giora Alexandron. 2024. Causal-mechanical explanations in biology: Applying automated assessment for personalized learning in the science classroom. *Journal of Research in Science Teaching*, 61(8):1858–1889.

Moriah Ariely, Asaf Salman, Anat Yarden, and Giora Alexandron. 2025. Reflective prompt engineering: a new strategy for automated short answer scoring in biology. *International Journal of Science Education*, pages 1–23.

Ryan Baker and Aaron Hawn. 2022. Algorithmic bias in education. *International journal of artificial intelligence in education*, 32(4):1052–1092.

Beata Beigman Klebanov and Nitin Madnani. 2022. *Automated essay scoring*. Springer Nature.

Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci. 2025. Emergent abilities in large language models: A survey. *arXiv preprint arXiv:2503.05788*.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2023. Similarity-based content scoring-a more classroom-suitable alternative to instance-based scoring? In *Findings of the association for computational linguistics: Acl 2023*, pages 1892–1903.

Sridevi Bonthu, S Rama Sree, and MHM Krishna Prasad. 2021. Automated short answer grading using deep learning: A survey. In *International cross-domain conference for machine learning and knowledge extraction*, pages 61–78. Springer.

Gregory Camilli. 2006. Test fairness. *Educational measurement*, 4:221–256.

Imran Chamieh, Torsten Zesch, and Klaus Giebermann. 2024. LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 309–315, Mexico City, Mexico. Association for Computational Linguistics.

Li-Hsin Chang and Filip Ginter. 2024. Automatic short answer grading for finnish with chatgpt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23173–23181.

Lei Chen and Raymond Ng. 2004. On the marriage of lp-norms and edit distance. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 792–803.

Sribala Vidyadhari Chinta, Zichong Wang, Zhipeng Yin, Nhat Hoang, Matthew Gonzalez, T Le Quy, and Wenbin Zhang. 2024. Fairaied: Navigating fairness, bias, and ethics in educational AI applications. *arXiv preprint arXiv:2407.18745*.

Aubrey Condor. 2020. Exploring automatic short answer grading as a tool to assist in human rating. In *International Conference on Artificial Intelligence in Education*, pages 74–79. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuning Ding, Brian Riordan, Andrea Horbach, Aoife Cahill, and Torsten Zesch. 2020. Don't take "nswvtnvakgxpm" for an answer –the surprising vulnerability of automatic content scoring systems to adversarial input. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 882–892, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Emrah Emirtekin and Yasin Özarslan. 2026. Automatic short-answer grading in sustainability education: AI–human agreement. *Journal of Computer Assisted Learning*, 42(1):e70160.

Rafael Ferreira Mello, Cleon Pereira Junior, Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Newarney Costa, Geber Ramalho, and Dragan Gasevic. 2025. Automatic short answer grading in the LLM era: Does GPT-4 with prompt engineering beat traditional models? In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, page 93–103. Association for Computing Machinery.

Guher Gorgun and Seyma N. Yildirim-Erbasli. 2026. Algorithmic bias in BERT for response accuracy prediction: A case study for investigating population validity. *Journal of Educational Measurement*, 63(1):e12420.

Christian Grévisse. 2024. LLM-based automatic short answer grading in undergraduate medical education. *BMC Medical Education*, 24(1):1060.

Abigail Gurin Schleifer, Beata Beigman Klebanov, and Giora Alexandron. 2025. Uncovering measurement biases in LLM embedding spaces: The anna karenina principle and its implications for automated feedback. *International Journal of Artificial Intelligence in Education*, 35(5):2821–2855.

Abigail Gurin Schleifer, Beata Beigman Klebanov, Moriah Ariely, and Giora Alexandron. 2023. Transformer-based hebrew nlp models for short answer scoring in biology. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 550–555.

Stefan Haller, Adina Aldea, Christin Seifert, and Nicola Strisciuglio. 2022. Survey on automated short answer grading with deep learning: from word embeddings to transformers. *arXiv preprint arXiv:2204.03503*.

Owen Henkel, Libby Hills, Adam Boxer, Bill Roberts, and Zach Levonian. 2024. Can large language models make the grade? an empirical study evaluating LLMs ability to mark short answer questions in k-12 education. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, page 300–304, New York, NY, USA. Association for Computing Machinery.

Matthew S Johnson and Daniel F McCaffrey. 2023. Evaluating fairness of automated scoring in educational measurement. In *Advancing natural language processing in educational assessment*, pages 142–164. Routledge.

Gerd Kortemeyer. 2023. Toward AI grading of student problem solutions in introductory physics: A feasibility study. *Phys. Rev. Phys. Educ. Res.*, 19:020163.

Gerd Kortemeyer. 2024. Performance of the pre-trained large language model gpt-4 on automated short answer grading. *Discover Artificial Intelligence*, 4(1):47.

Zhaohui Li, Yajur Tomar, and Rebecca J Passonneau. 2021. 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.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. *arXiv preprint arXiv:2312.01552*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Anastassia Loukina, Nitin Madnani, and Klaus Zechner. 2019. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 1–10.

Michael Madaio, Su Lin Blodgett, Elijah Mayfield, and Ezekiel Dixon-Román. 2022. Beyond "fairness": Structural (in) justice lenses on AI for education. In *The ethics of artificial intelligence in education*, pages 203–239. Routledge.

Nitin Madnani, Anastassia Loukina, Alina von Davier, Jill Burstein, and Aoife Cahill. 2017. Building better open-source tools to support fairness in automated scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 41–52, Valencia, Spain. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ido Roll, Ryan SJ d Baker, Vincent Aleven, and Kenneth R Koedinger. 2014. On the benefits of seeking (and avoiding) help in online problem-solving environments. *Journal of the Learning Sciences*, 23(4):537–560.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. Dictabert: A state-of-the-art bert suite for modern hebrew. *arXiv preprint arXiv:2308.16687*.

Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research*, 78(1):153–189.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. OpenAI gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Maya Usher and Montathar Faraon. 2025. Who grades best? comparing chatgpt, peer, and instructor evaluations across varying levels of student project quality. *Assessment & Evaluation in Higher Education*, pages 1–20.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

X. Wu, P. P. Saraf, G. Lee, and 1 others. 2025. Unveiling scoring processes: Dissecting the differences between LLMs and human graders in automatic scoring. *Technology, Knowledge and Learning*.

Xiaoming Xi. 2010. How do we go about investigating test fairness? *Language testing*, 27(2):147–170.

E. Yacobson, S. Rapp, R. Blonder, and G. Alexandron. 2025. Human experts vs. LLMs: Who is better at explaining student clustering? In *Proceedings of the 2nd Human-Centric eXplainable AI in Education (HEXED) Workshop at EDM 2025*.

A Prompt Example

The general prompt structure is taken from Ariely et al. (2025). The scoring instructions are tuned per category. Each category-specific prompt is then followed by randomly chosen few-shot positive and negative examples.

1. **Role definition:**
   'You are an expert in education and assessment in biology.'

2. **Task definition and general instructions:**
   'You are required to assign a score to category in a student's response titled: "Changes in the rate or amount of energy Production." Note that although the overall context of the response is important, you should focus solely on the students' reference to changes in the rate or amount of energy or ATP production. Anything else mentioned around this category may be significant for the response, but irrelevant for evaluating this specific category.'

3. **Instructions for scoring:**
   'If the category is present in the response, score 1. If not, score 0.'