GUL'S NOTEBOOK · AI RESEARCH · OPEN ACCESS

Grading written answers with AI: what worked, what didn't

Muhammad Gulzaib
https://orcid.org/0009-0000-8709-6989
SynergyFlow Labs · Islamabad, Pakistan

Published	2026-07-12
URL	https://gulzaib.vercel.app/blog/grading-written-answers-with-ai-what-worked-what-didnt
Licence	Creative Commons Attribution 4.0 International (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/

ABSTRACT

AI can propose marks for written answers against a clear rubric, but published results are mixed and agreement on one dataset does not establish reliability in another classroom. In one undergraduate course, GPT-4o exactly matched human scores on 55% of 258 quiz cases despite a high overall correlation. A cautious design keeps the teacher in control, checks transcription separately and measures errors before use. [1]

Keywords: llm, evaluation, edtech, research

This article was prepared in November 2026. The displayed 12 July 2026 date was requested for the post; it is not its original publication date.

The title asks what worked and what did not. The evidence here comes from published research, not an unpublished ZeroExam AI experiment. The rubric, answers and grading output below are illustrative; no model was run on them and no local teacher-agreement result is claimed.

What the studies actually found

Byun, Rajwal and Choi evaluated GPT-4o on short-answer quizzes and project reports in one undergraduate computational linguistics course. Across 258 quiz cases, exact agreement with the human-assigned score was 55%; the model scored lower in 38.8% and higher in 6.2%. The paper reports an overall correlation of 0.98. Its two teaching assistants divided the answers rather than independently marking the same responses, so it could not report a human-to-human agreement baseline. The findings are preliminary and course-specific. [1]

That contrast matters. A model can broadly follow the order of stronger and weaker answers while still assigning an individual student a different mark. A high correlation is not a guarantee that the mark on any particular paper is right.

Chamieh and colleagues tested zero-shot and few-shot approaches across three short-answer datasets and reported weak results, especially on tasks requiring complex reasoning or domain knowledge. A 2026 review of 42 empirical studies found varied results and highlighted the influence of rubrics and prompts; it did not support replacing teachers with generative AI. These results do not establish how a particular model will perform on Pakistani board exams or on Urdu answers. [2] [3]

Start with a marking scheme, not a model answer alone

Suppose a teacher asks: "Explain two ways plant roots help a plant." The following is a fictional four-mark rubric for a practice question, not an official board marking scheme.

Criterion	Marks	Credit rule
Absorption	0–2	1 for absorbing water; 1 for absorbing mineral nutrients
Anchorage	0–2	1 for holding the plant in the soil; 1 for explaining that this keeps it stable

For the illustrative response "Roots take in water and minerals and hold the plant in the soil", the proposed rubric gives **3 out of 4**: two marks for absorption and one for anchorage. The last mark requires an explicit stability explanation under this particular rubric. A teacher could choose a different rule, but it must be decided and applied consistently.

A response such as "Roots drink from the soil and keep the plant from falling over" expresses the water and stability ideas in different words. The grader should assess the meaning rather than demand the wording of an exemplar. It should not award the mineral mark if that idea is absent. For an unclear or contradictory answer, send the response to a teacher rather than guessing what the student intended.

Before testing the model, have teachers agree on partial credit, alternative correct phrasing and how to treat irrelevant or contradictory claims. The goal is not to remove every judgement call. It is to make the judgement being tested visible.

Ask for criterion-level evidence

A proposed grading request should contain the question, the teacher-approved rubric and one student answer. Ask for a score and a short explanation for each criterion, including the relevant words from the answer. Treat the student's text as data, not instructions to change the marking scheme. A prompt boundary is useful but cannot guarantee resistance to manipulation.

The following is an **illustrative target output**, written by hand for the first sample answer. It is not a result produced by a model:

```
{
  "criteria": [
    {
      "id": "absorption",
      "score": 2,
      "max": 2,
      "evidence": "take in water and minerals"
    },
    {
      "id": "anchorage",
      "score": 1,
      "max": 2,
      "evidence": "hold the plant in the soil"
    }
  ],
  "total": 3,
  "review_required": true
}
```

The application should validate the response before displaying it: expected criterion IDs appear exactly once, each score is an allowed value within its criterion's range, evidence is actually present in the submitted text, and the total equals the sum. If parsing or validation fails, do not assign a mark; put the answer in the review queue. Even a valid response can contain a wrong interpretation, so validation does not replace teacher review.

A model's explanation is a starting point for checking its judgement, not proof that its mark is correct. A teacher needs to see the original answer, the rubric and the proposed scores together, and be able to change both the mark and feedback.

Separate handwriting from grading

For scanned scripts, optical character recognition (OCR) converts the image into text before a text-only model can grade it. A missing word, mathematical symbol or negation can change the answer. If the transcription is wrong, a correct rubric cannot repair the lost evidence.

Show the teacher the original image beside the transcription. Allow corrections and regrading, preserve the version that was scored, and route missing or unreadable sections to review. If a system cannot establish a trustworthy transcription, it should abstain rather than silently award zero.

Test handwriting and languages separately from typed English. Urdu, English, mixed-language answers, diagrams and equations may need different handling. Do not assume a vendor's performance on one format transfers to another, and do not assume every subject can be scored from text alone.

Compare against teachers before classroom use

A practical first evaluation can be small, but it needs a clean design. The following is a **proposed protocol**, not a completed experiment:

1. Pick one subject and a teacher-approved question with a fixed rubric. Collect an appropriately authorised sample with names removed, including correct, partial and incorrect answers.
2. Have two qualified teachers mark the same answers independently, without seeing the model's suggestions. Record both marks and disagreements before discussing them.
3. Use a separate development set to refine the rubric and prompt. Keep the evaluation answers and their scores out of prompt examples and model tuning; where possible, separate students across the two sets.
4. Freeze the model version, prompt, rubric and transcription procedure. Grade the held-out answers independently, then compare each proposal with an adjudicated teacher reference.
5. Record mistakes by criterion and review the answer, transcript and rubric. Test again after any model, prompt or rubric change.

Human markers can disagree too. Report their agreement before adjudication and keep the adjudicated decisions rather than treating the first teacher's score as unquestionable truth. A tiny or unrepresentative sample can expose obvious failures, but it cannot establish reliability across schools.

Measure the errors that affect students

Measure	What it reveals
Exact agreement	Percentage of answers with the same total mark as the reference
Mean absolute error	Average size of the mark difference, without letting positive and negative errors cancel
Direction and distribution	Whether marks are generally higher or lower, and how often large errors occur
Human-to-human agreement	How consistently the teachers apply the rubric before adjudication
Borderline and subgroup errors	Whether errors differ near a pass boundary or by question, language or input format

For a hypothetical four-mark question, suppose five adjudicated teacher scores are 4, 3, 2, 1 and 0, while the model proposes 4, 2, 2, 2 and 0. Exact agreement is 3 out of 5, or 60%. The absolute differences sum to 2 marks, so mean absolute error is $2 \div 5 = 0.4$ marks. These

numbers illustrate the calculation only; five examples are far too few to support a deployment decision.

Averages can hide harm near a pass boundary. Inspect how often a suggestion would change a pass to a fail or vice versa, and calculate errors by question. Examine language and input-format groups where sample sizes and privacy safeguards permit; report group sizes and uncertainty rather than treating a handful of papers as conclusive. Decide acceptable error limits with teachers before looking at the evaluation results.

Where I would draw the line

For an early pilot, use AI proposals to support **formative practice**, with a teacher checking every answer and controlling what feedback reaches the student. Measure the total review time as well as accuracy; a plausible explanation that requires repeated checking may not save work. Keep an audit trail of the rubric, model version, proposed mark, corrections and final decision.

For high-stakes exams, a grade should not be automatically finalised or released by this workflow. Provide a way to challenge and correct a mark. Minimise personal data, confirm institutional approval and provider retention terms before sharing any work, and limit access to scripts and logs. UNESCO's guidance calls for privacy protection and human-centred validation of generative AI in education. [4]

A grading system is useful when teachers can inspect and correct it, and when its errors are measured on the answers it will actually see. Until that evidence exists, a confident-looking score is only a proposal.

*[Figure: image]***Grounded exam generation with RAG: retrieval and evaluation**: How ZeroExam AI keeps generated exam questions inside the syllabus: hybrid retrieval, rank fusion, grounding checks and an evaluation plan you can run. *[Figure: image]***Simple wins: designing for the least technical user in the room**: People pick software they enjoy over the longest feature list. A practical guide to designing SaaS for users who would rather not use software at all.

> **Key takeaways**
>
> - Start with a teacher-approved rubric and explicit rules for partial credit.
> - Validate structured output, but treat explanations and scores as fallible proposals.
> - Check OCR against the original script and route uncertain cases to a teacher.
> - Evaluate on held-out answers using human agreement, exact agreement and error size.
> - Keep teachers responsible for consequential grades, corrections and appeals.

Questions readers ask

Can AI grade subjective exam answers accurately?

Published results vary by question, rubric, model and student population. A model can propose criterion-level marks, but its agreement with teachers and its errors must be tested locally. A teacher should review and own consequential grades.

How do you compare AI grades with a teacher's marks?

Use a held-out set of answers scored independently by teachers, then report exact agreement, mean absolute error and the direction and size of disagreements. Compare with human-to-human agreement and examine errors by question and relevant language or transcription groups.

Does good correlation mean AI grades are correct?

No. Correlation can be high even when individual scores differ or the model consistently marks too high or too low. Inspect exact agreement, absolute error and borderline cases as well.

References

1. Byun, G., Rajwal, S. and Choi, J. D. LLM-as-a-Grader: Practical Insights from Large Language Models for Short-Answer and Report Evaluation. Information, 2026. https://doi.org/10.3390/info17050505
2. Chamieh, I., Zesch, T. and Giebermann, K. LLMs in Short Answer Scoring: Limitations and Promise of Zero-Shot and Few-Shot Approaches. Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, 2024.
3. Jukiewicz, M. and Wyrwa, M. Can ChatGPT Replace the Teacher in Assessment? A Review of Research on the Use of Large Language Models in Grading and Providing Feedback. Applied Sciences, 2026. https://doi.org/10.3390/app16020680
4. Miao, F. and Holmes, W. Guidance for generative AI in education and research. UNESCO, 2023. https://doi.org/10.54675/EWZM9535

CITE AS

Gulzaib, M. (2026, July 12). Grading written answers with AI: what worked, what didn't. Gul's Notebook. https://gulzaib.vercel.app/blog/grading-written-answers-with-ai-what-worked-what-didnt

You may share and adapt this text for any purpose, including commercially, as long as you credit the author and link to the original. Licence: https://creativecommons.org/licenses/by/4.0/