REMARKBENCH: BENCHMARKING AGENTIC GRADING OF RAW EXAMINATION DOCUMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous grading agents that assess student work with minimal human intervention can reduce teachers' workload and provide timely, informative feedback at scale. However, existing benchmarks mostly focus on pre-segmented student answers and check whether scores are correct, not whether each score rests on the student's answer and the rubric, or whether its grounds stay the same when the answer is regraded. To address these gaps, we introduce `ReMarkBench`, a benchmark for evaluating end-to-end agentic grading from raw examination documents: complete question papers, grading rubrics, and handwritten answer scripts from a public university-entrance examination. It covers 3,570 grading units in six grading tasks across Mathematics, Physics, Chemistry, Biology, and English. For every grading unit, the grading agent must produce a score and a supporting justification that cites the student's actual answer and explains the score under the rubric. We therefore assess three *complementary dimensions*: *score accuracy*, *justification quality*, and *cross-run consistency*. Across five frontier grading agents, each a model–harness configuration reading the raw documents directly, even the best reaches a joint pass rate of only 41.0%, averaged equally across the six tasks over three independent runs, and many correctly scored records still fail justification quality. Substantial gaps remain even for an agent grading from human-verified structured inputs instead of the raw documents, so the difficulty goes beyond reading the documents. `ReMarkBench` provides a testbed for diagnosing these gaps and advancing accurate, interpretable, and reproducible automated grading.

1 INTRODUCTION

Recent large language models (LLMs) can score student answers in close agreement with human graders (Yancey et al., 2023; Henkel et al., 2024; Chen & Wan, 2025), making autonomous grading agents a promising way to reduce teachers' workload and provide timely, informative feedback (Hattie & Timperley, 2007) at scale. In real examinations, however, grading starts from raw examination documents and involves *more than* a single scoring decision. The score must be **accurate**, faithful to what the student answers and to the grading rubric. Its justification must make the score **interpretable** by pointing to the specific student answer and the corresponding rubric. The grading must also be **reproducible**: regrading the answer must yield the same score with the same justification. Grading agents should be evaluated on all *three essential requirements*: accurate, interpretable, reproducible.

However, previous grading benchmarks *only cover part of the essential requirements*. For example, they evaluate single answers already segmented by question and paired with their rubrics (Khrulev, 2025; Zhao et al., 2026), leaving out the step of recovering each answer from the raw documents and aligning it with its question and rubric. Some benchmarks evaluate rationales by similarity to expert references (Filighera et al., 2022; Zhao et al., 2026), which measures agreement with an expert but does not directly check whether a score rests on the student answer and the rubric. Yu et al. (2026) instead have human reviewers check AI feedback against the rubric on university calculus quizzes. These benchmarks also leave cross-run consistency largely unchecked; even where the same answer is regraded, only scores are compared (Seuti & Ray Choudhury, 2026). A grading agent may therefore award a correct score for the wrong reason or give an answer the same score on different grounds across runs (Figure 1), and neither failure is systematically measured.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

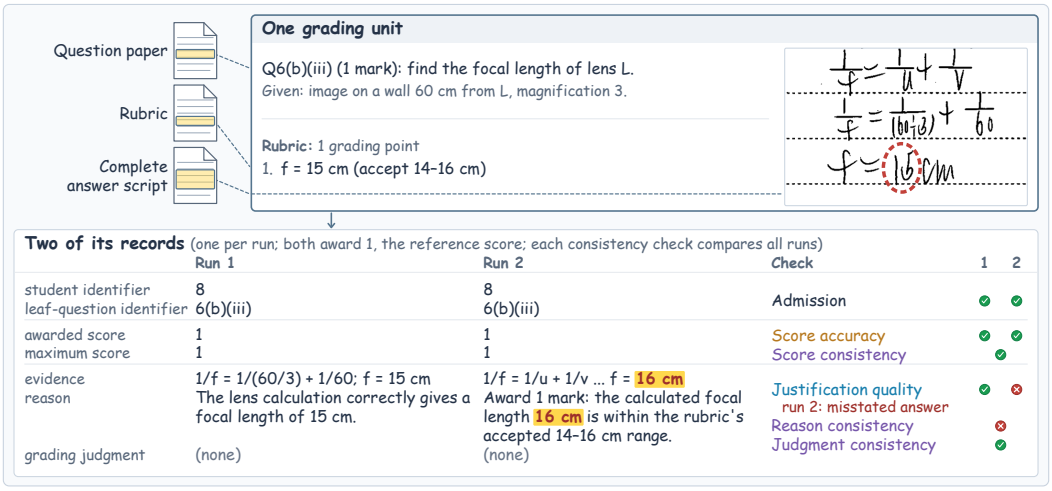


Figure 1: A grading unit and two of its records (question paraphrased; answer cropped). Run 2 misreads the student's 15 cm as 16 cm: its score is still correct, but it fails *justification quality*, and the unit fails *reason consistency*.

To address these gaps, we introduce ReMarkBench, a benchmark for evaluating end-to-end agentic grading from raw examination documents (Figure 2). An agent needs to grade directly from the original documents instead of relying on a hand-engineered multi-stage pipeline. In the five subject tasks, a **grading unit** is one student's answer to one *leaf question* (i.e., a question without a sub-question). In English Writing, it is one assessment dimension (Content, Language, or Organisation) of one composition. For each grading unit, the grading agent submits a grading record with a score and a supporting justification. To meet each of the essential requirements, we define and measure the following three key dimensions. **Score accuracy** compares **awarded scores** (i.e., scores awarded by the agent) with **reference scores** (i.e., scores given by human experts). **Justification quality** checks whether the justification states what the student actually answers, cites the evidence, and derives the score from the rubric point by point. **Cross-run consistency** checks whether each grading unit receives the same score on the same grounds across three independent runs. For the five subject tasks, the **joint pass rate** is the proportion of grading units that satisfy all three dimensions. English Writing uses a script-level definition, detailed in Section 4.

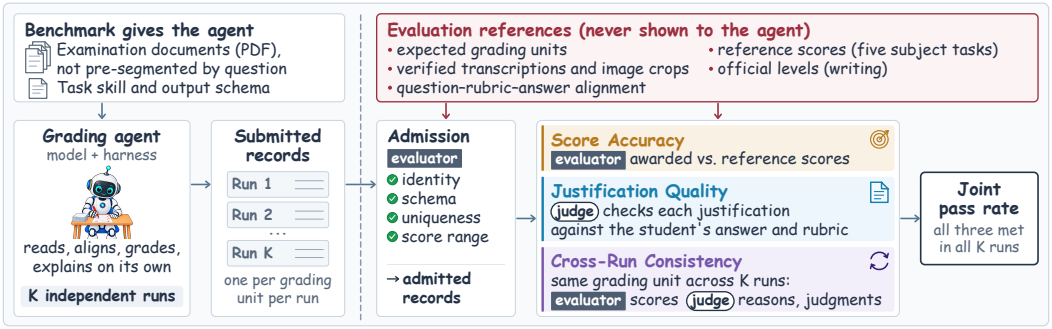


Figure 2: Overview of ReMarkBench. The dashed line separates the agent side (left) from the evaluation side (right). *evaluator*: programmatic rules; *judge*: LLM.

We built ReMarkBench[1] using real question papers, grading rubrics, and real students' handwritten answer scripts sourced from a public university-entrance examination. The benchmark contains 3,570 grading units across six subject tasks: Mathematics, Physics, Chemistry, Biology, English Reading, and English Writing. We compare agent-graded scores with reference scores for each leaf question in

[1] https://anonymous.4open.science/r/ReMarkBench

Table 1: Among the compared studies, only `ReMarkBench` automatically checks reasons against the answer and rubric and compares reasons across runs.

Study	Handwritten answers	Multiple subjects	Not pre-segmented	Answer fidelity	Reasons checked against	Cross-run scores	Cross-run reasons
Benchmarks							
CHECK-MAT (2025)	✓	×	×	Case	Case	NR	NR
EduMARS (2026)	✓	✓	×	✓	Expert ref.	NR	NR
EDU-CIRCUIT-HW (2026)	✓	×	×	✓	NR	NR	NR
GCSE (2026)	✓	✓	×	NR	NR	NR	NR
S-GRADES (2026)	×	✓	×	NR	NR	Per answer	NR
Grading pipelines							
GradeAgentOps (2026)	×	×	×	✓	NR	NR	NR
Perš et al. (2026)	✓	×	✓	NR	NR	Aggregate	NR
`ReMarkBench`	✓	✓	✓	✓	Answer + rubric	Per unit	✓*

✓: included or reported; ×: not included; NR: not reported; Case: qualitative cases only. *Answer fidelity*: student content restated in outputs is checked against the answer. *Cross-run*: across independent runs of the whole workflow under fixed settings. *Five subject tasks; writing compares scores only.

the subject tasks. Justification quality and cross-run consistency rely on LLM-as-a-judge evaluation, whose labels pass the alternative annotator test against three human experts (Section 5.1).

Evaluations with five frontier model–harness configurations show that joint pass rates, averaged equally across the six tasks, range from 23.7% to 41.0% over three runs (Table 3), indicating that current grading agents are far from ready to grade real examinations without human review. Score accuracy alone overestimates agent performance: many records with correct scores still fail justification quality (20.6%–37.8% of correctly scored records in the five subject tasks; Table 26), mostly by misreading the student's answer, citing evidence that the reason never uses, or arguing too weakly from the evidence to the score. Grading is hardest on multi-step questions and where the rubric leaves the score to the grader's own judgment; there, agents differ in whether they write that judgment down, and a written judgment often changes across runs. These failures go beyond reading the documents: substantial gaps remain even when the agent grades from human-verified structured inputs (Section 5.2).

2 Related Work

Table 1 compares `ReMarkBench` with recent grading benchmarks and pipelines on input coverage, output checks, and cross-run comparison. Grading benchmarks have extended to real-world handwritten answers (Khrulev, 2025; Sun et al., 2026; Fox et al., 2026) and to multiple subjects (Zhao et al., 2026; Fox et al., 2026; Seuti & Ray Choudhury, 2026), but most studies still start from answers already segmented by question. Among the compared studies, only Perš et al. (2026) start from scanned examination documents, and they evaluate their own pipeline and judge it by its scores; `ReMarkBench` instead evaluates any grading agent on one common set of such documents, and checks the justifications as well as the scores. For justifications, EduMARS compares rationales with expert references by embedding similarity and, in error analysis, checks their facts against the student's steps (Zhao et al., 2026), and GradeAgentOps matches cited evidence against the answer lexically (Anghel et al., 2026); neither checks, without a reference reason, whether a reason supports the score under the rubric. Justification quality checks this support within each record (Section 4), not whether the record reflects the agent's internal reasoning. Across runs, S-GRADES (Seuti & Ray Choudhury, 2026) and Perš et al. (2026) compare only scores or aggregate metrics, not the reasons behind them. Justification quality and cross-run consistency both rely on an LLM judge, which we validate against human experts (Section 5.1). Further related work is discussed in Appendix I.

3 REMARKBENCH: TASK AND DATA

3.1 TASK DEFINITION

For each task, the grading agent receives the original PDFs of the question paper, rubric, and anonymized answer scripts, together with a minimal instruction, a task skill that describes the grading workflow and required records, and an output schema with an example record. The agent needs to establish the question structure from the documents, align each leaf question with its rubric entry and each student's answer, and score each answer under the rubric, delivering one grading record per grading unit (Figure 2). Within a grading unit, the rubric usually awards marks through several grading points, each awarding or withholding marks on its own.

We evaluate the grading agent using these submitted records. Each record must identify the student and grading unit, report the awarded and maximum scores, and provide a justification comprising three fields: evidence, reason, and grading judgment. The evidence must identify the student's actual wording, mathematical expressions, steps, or diagrams supporting the grading decision. The reason must explain how this evidence supports the awarded score under the rubric. If the agent considers the rubric insufficient and adopts a supplement, interpretation, or departure that affects the score, it must state this explicitly in the grading judgment. Figure 1 shows two such records for the same grading unit from separate runs.

3.2 CHALLENGES OF THE GRADING TASK

Grading real-world student examination papers naturally raises three key challenges.

Reading raw documents. A question may contain several levels of nested parts whose hierarchy is difficult to discern from the original layout, a student's answer to one question may span several pages, and the content ranges over handwriting, mathematical notation, tables, and drawings. Misreading or misalignment can invalidate the justification even when the awarded score is correct.

Connecting answers to grading points. Student answers rarely follow the rubric point by point: steps may be reordered, merged, skipped, or worded differently from the reference answer. An answer that is wrong or incomplete must be graded as written, not as the student probably meant it. To justify a score, the agent must decide which parts of the answer bear on each grading point, including the points the answer fails to meet, and state point by point why marks are awarded or withheld. The more reasoning steps a question requires, the more such connections the agent must make, and the more places the justification can go wrong.

Resolving what the rubric leaves open. Rubrics may be ambiguous or incomplete, and often do not specify how to credit equivalent expressions, partially correct answers, and other borderline answers. The agent must then make its own grading judgment, declare it, and make the same judgment each time it grades the same answer; otherwise, repeated grading can differ in score or in justification.

3.3 BENCHMARK DATA

The benchmark uses data sourced from a public university-entrance examination, including question papers, official grading rubrics, and anonymized answer scripts of real candidates classified into five **performance levels** (two answer scripts per level are released by the examination authority for all subjects except English Reading and Writing which have four scripts per level). The benchmark thus spans the full range of student performance levels, and each script is long, containing many grading units (Table 2). We prepare human-verified transcriptions of student answers for justification quality checks. Except English Writing, experienced markers of different subject tasks assigned **reference scores** using the official grading rubric and checked one another's annotations. These scores are used only for evaluation.

4 EVALUATION METHODOLOGY

As we introduced in Section 1 and illustrated in Figure 2, we evaluate whether grading is accurate, interpretable, and reproducible through score accuracy, justification quality, and cross-run consistency. These dimensions check whether scores match the references, whether justifications support the

Table 2: The six grading tasks. Percentages are over grading units with non-empty answers. Full composition is given in Appendix B.1.

	Math.	Phys.	Chem.	Biol.	Eng. Reading	Eng. Writing
Answer scripts	10	10	10	10	20	20
Pages per answer script	24	19	19	19–20	8	4–10
Leaf questions (per script)	42	48	53	40	121 (80–82)	5 (2)
Maximum question depth	3	3	4	3	2	1
Grading units (non-empty)	420 (337)	480 (452)	530 (494)	400 (369)	1,620 (1,565)	120
Answers containing images	0.0%	20.6%	22.7%	5.1%	0.0%	0.0%
Answers earning partial credit	13.1%	12.6%	14.6%	25.7%	0.8%	—

scores, and whether grading decisions remain consistent across independent runs. The checks corresponding to each grading challenge are listed in Appendix H.

Joint pass rate. The joint pass rate combines the three dimensions. We run each grading agent K times under fixed settings (the same model, harness, inputs, and instructions). A record is admitted only if it is well formed and matches one expected grading unit, and units without an admitted record count as undelivered (Appendix A.4). Except English Writing, a grading unit passes jointly when it is delivered, awarded the reference score, passes justification quality in every run, and also passes cross-run consistency. Undelivered units count as failures. The joint pass rate is the proportion of units that meet these conditions.

English Writing lacks unit-level reference scores and is counted per script: a script's joint score is the justification quality pass rate of all its records over the K runs if its performance level (mapped from its awarded scores) is correct in every run, and zero otherwise. Thus, its joint score requires consistency of the performance level, rather than equality of the individual dimension scores (Appendix B.2). For the main six-task summary, each metric is computed within each task, excluding units with empty answers, and the six task-level results are averaged with equal weight.

Score accuracy. Score accuracy compares the score of each admitted record with the corresponding reference score and reports the exact-match rate over admitted records. English Writing only has official performance levels for the whole scripts; thus, each script's scores in a run are combined into a script score, converted to a level, and compared with the official level (Appendix B.2).

Justification quality. A score is interpretable only if its record shows how the score was awarded. Justification quality checks this for every record. Unlike a score, a justification is not compared with a reference, since many different justifications can support the same score. Instead, we treat each justification as an argument for its score. The argument rests on two kinds of premises: the evidence which states what the student wrote, and the grading basis which includes the rubric and any grading judgment. From these, the reason explains how the score is awarded point by point.

Subject to the English Reading selection-question exemptions in Appendix D, a record fails justification quality if its argument fails in any of the following three categories (corresponding to the three challenges in Section 3.2). (1) **Material reading** concerns the evidence premise: in a misstated answer, the evidence, reason, or grading judgment misstates or misquotes what the student wrote, as Run 2 in Figure 1 does. (2) **Evidence extraction** concerns whether the evidence and the reason correspond: the reason may leave extracted evidence unused, rely on part of the answer that the evidence does not contain (missing evidence), or cite evidence that bears on no grading point (irrelevant evidence). (3) **Score explanation** concerns the step from the premises to the score: the reason may leave the score for a grading point without a basis in the rubric (ungrounded score), pass over part of the answer (ignored answer part), apply a rule found in neither the rubric nor the grading judgment (unsupported rule), contradict itself or the score (contradiction), or argue too weakly from the evidence to the score (weak argument). We name a failure by category and subcategory, such as Evidence extraction – unused evidence. Whether a record writes a grading judgment, and whether that judgment departs from the rubric, is reported but not counted as a failure. These checks use LLM-as-a-judge evaluation: each check is evaluated separately using only its designated fields and materials, never the reference score. Records may be batched, with instructions to judge each independently (Appendix D). Section 5.1 validates it against human experts.

Cross-run consistency. For each grading unit, we compare its records across runs (see Figure 1 and Appendix E). Only delivered runs are compared, and a unit delivered in fewer than two runs is left out. A unit passes score consistency if the scores are equal, reason consistency if the reasons make the same claims about the answer and the rubric, and judgment consistency if the grading judgments reach the same decisions on the same grounds or no run writes one. A unit passes cross-run consistency only if it passes all three consistency checks. Differences in wording and length are ignored, but any change in what is claimed counts, even when the claims of different runs do not contradict one another. For English Writing, only score consistency is compared per assessment dimension.

5 Benchmark Instantiation and Empirical Validation

This section runs the grading tasks defined in Section 3 with five grading agents and applies the evaluation method proposed in Section 4 to examine that method empirically.

5.1 Experimental Setup

We select five frontier models that read images natively, as grading from the original PDFs requires: the closed-source GPT-5.6 Sol, Claude Opus 5, and Grok 4.6, and the open-weight MiniMax M3 and Qwen3.8-Max. Each model runs in its developer's own agent harness, except MiniMax M3: GPT-5.6 Sol with Codex, Claude Opus 5 and MiniMax M3 with Claude Code, Grok 4.6 with Grok Build, and Qwen3.8-Max with Qoder. All run at the default reasoning effort, and each task is run three times. All five grading agents are evaluated on the original PDFs (the *PDF route*). We denote configurations by model/route, for example GPT/PDF. The main summary gives the six task-level results equal weight (Section 4). Model and harness details are given in Appendix G.7.

Judge validation. The LLM judge of Section 4 is GPT-5.6 Sol run at low reasoning effort. Three experts blindly relabeled 115 records for justification quality, 100 reason comparisons, and 60 judgment comparisons from GPT/PDF, where their majority labels serve as references (Appendix F). Across the seven labels reported there, the judge passes the alternative annotator test (Calderon et al., 2025) at $\varepsilon = 0.2$, but held-out agreement is lower for Evidence extraction and overall justification quality. Design-weighted rates differ from the experts' by at most 2.5 points, except for Evidence extraction (-6.4 points; 95% interval $[-16.8, 2.5]$). Judge replacement largely preserves the configuration ranking, with no clear evidence of self-preference in these checks. Human validation covers only GPT/PDF. All 15 expert-reviewed English Writing records fail justification quality, leaving recognition of passing English Writing records unvalidated.

5.2 Overall Results

Grading agents score accurately but are still far from being reliable graders. On the equally weighted six-task average in Table 3, Qwen ties for the highest average score accuracy with Claude at 88.0%, yet its average joint pass rate is only 41.0%. GPT's average score accuracy is 83.0%, but its average justification quality pass rate is only 50.2%. MiniMax has the lowest joint pass rate at 23.7%. Even Grok, which leads in average cross-run consistency at 72.8%, reaches only 34.9% on the joint pass rate. Together, these results show that accurate scoring alone does not ensure justifications that pass review or consistent decisions across runs.

Task difficulty varies with answer form and rubric structure. English Writing and Biology rank near the bottom on almost every metric for all five agents. Their free-text answers and largely descriptive rubrics leave more scoring decisions to the grader. Mathematics has a lower average joint pass rate than Physics and Chemistry (33.7% versus 40.0% and 39.7%). Nearly all its grading units are calculations, and more than a quarter are multi-step calculations. The complex reasoning required by mathematics tasks might be the reason why the agent performs worse. Chemistry has one of the highest average joint pass rates among the five subjects (39.7%). Its rubrics specify what earns each score for most grading units, leaving fewer scoring decisions to the grader. Section 5.3 examines the reasons for these task differences in greater depth.

Justification quality failures show different patterns across material reading, evidence extraction, and score explanation. Figure 3 shows the most frequent problems.

Table 3: Main results of the five grading agents on the six grading tasks (PDF route, $K = 3$, %). Bold marks the best grading agent on each task within a metric.

	Score accuracy ↑					*Justification quality pass rate* ↑				
Task	GPT	Claude	MiniMax	Grok	Qwen	GPT	Claude	MiniMax	Grok	Qwen
Math.	89.8 ±1.5	87.9 ±0.5	85.1 ±1.0	90.4 ±0.3	**91.0** ±0.7	65.1 ±1.2	65.8 ±1.9	40.4 ±9.8	62.8 ±3.7	**71.8** ±2.6
Phys.	87.7 ±0.7	87.0 ±0.5	82.6 ±0.8	87.4 ±0.7	**89.2** ±1.0	66.0 ±1.9	74.5 ±0.7	54.3 ±6.4	68.9 ±1.0	**79.3** ±0.7
Chem.	91.6 ±0.2	92.4 ±1.1	86.4 ±0.9	91.2 ±0.4	**93.5** ±1.0	63.1 ±1.5	75.2 ±1.4	59.4 ±2.7	71.9 ±1.5	**77.5** ±0.8
Biol.	76.1 ±1.2	**86.7** ±1.2	76.1 ±2.6	84.7 ±1.8	83.8 ±0.6	48.4 ±2.2	58.4 ±3.0	43.3 ±2.0	60.5 ±1.1	**66.4** ±2.0
Eng. Reading	96.2 ±0.9	**97.4** ±0.2	94.2 ±1.1	96.0 ±0.1	95.3 ±0.5	56.3 ±2.3	67.4 ±6.0	81.3 ±7.2	77.2 ±5.4	**87.6** ±3.1
Eng. Writing	56.7 ±10.4	**76.7** ±7.6	55.0 ±5.0	73.3 ±2.9	75.0 ±5.0	2.2 ±1.0	0.6 ±0.5	0.3 ±0.5	0.0 ±0.0	**2.5** ±0.8
Average	83.0 ±1.9	**88.0** ±1.4	79.9 ±0.9	87.2 ±0.5	**88.0** ±0.7	50.2 ±0.8	57.0 ±1.1	46.5 ±1.5	56.9 ±0.9	**64.2** ±0.9

	Cross-run consistency pass rate ↑					*Joint pass rate* ↑				
Task	GPT	Claude	MiniMax	Grok	Qwen	GPT	Claude	MiniMax	Grok	Qwen
Math.	**78.0**	73.9	63.2	74.8	71.2	39.8	33.2	11.9	39.2	**44.5**
Phys.	72.1	67.5	60.4	68.1	**77.2**	35.6	43.8	21.9	42.7	**56.0**
Chem.	68.2	**72.9**	62.3	71.9	69.8	30.8	**48.0**	28.9	44.3	46.6
Biol.	49.1	65.3	48.8	**68.3**	57.7	23.8	29.3	20.6	32.2	**33.1**
Eng. Reading	83.7	70.4	75.3	**85.3**	82.3	37.0	32.8	58.6	50.7	**64.3**
Eng. Writing	43.3	61.7	15.0	**68.3**	57.5	0.8	0.0	0.0	0.0	**1.7**
Average	65.7	68.6	54.2	**72.8**	69.3	28.0	31.2	23.7	34.9	**41.0**

GPT, Claude, MiniMax, Grok, and Qwen denote GPT-5.6 Sol, Claude Opus 5, MiniMax M3, Grok 4.6, and Qwen3.8-Max, respectively. Values after ± are sample standard deviations across the three runs. Metric definitions, including the per-script scoring of English Writing, are given in Section 4. Student-level bootstrap 95% intervals of the six-task average joint pass rate are within ±2.5 points; all pairwise differences are significant (Appendix G.2).

Material reading failures misstate what the student wrote. MiniMax does so in 18.2% of records, compared with 8.9%–9.9% for the other agents. This reflects a weakness in visual reading: misreading handwritten text or equations makes the grading unfaithful to the student's actual answer.

Agents do not consistently pinpoint the answer content that supports their scores. Across the five agents, unused evidence appears in 14.2% of records on average, compared with 5.0% for missing evidence. This imbalance suggests a tendency to extract more of the student's answer than the agent actually uses to explain the awarded scores.

Score explanation failures leave individual scores insufficiently justified. Agents may fail to ground a score in the rubric or give only a weak argument for it. Each problem affects about 10% of GPT and MiniMax records, but at or below 4% of Grok and Qwen records. GPT and MiniMax in particular still struggle to explain clearly why each score is awarded.

	GPT-5.6 Sol	Claude Opus 5	MiniMax M3	Grok 4.6	Qwen3.8-Max	Mean
Evidence extraction – unused evidence	16.9	13.2	17.1	17.2	6.7	14.2
Material reading – misstated answer	9.9	9.3	18.2	9.8	8.9	11.2
Score explanation – weak argument	10.4	6.4	9.8	3.3	3.9	6.8
Score explanation – ungrounded score	10.2	5.6	9.7	3.3	1.4	6.0
Evidence extraction – missing evidence	9.3	1.7	7.4	3.2	3.6	5.0
Score explanation – unsupported rule	2.0	1.9	6.1	3.3	2.4	3.1

Figure 3: The six most frequent problems in single-run records (%). Problems can co-occur, so rates are not additive. Model rates average the five subjects equally. English Writing failures are broken down by assessment dimension in Figure 8 and discussed in Appendix G.4. The full ranking is given in Table 28.

Do the agent's own transcription and human-verified structured inputs improve justification quality and cross-run consistency? To examine the role of document reading, we run GPT on two diagnostic input routes (Figure 4). On the *OCR route*, the agent itself first transcribes the examination documents from page images into structured text and then grades from its own transcription. It consults the original pages only where text cannot capture the content. On the *JSON route*, the agent grades from human-verified structured inputs.

Across the six equally weighted tasks, GPT's joint pass rate rises from 28.0% on PDF to 29.4% on OCR and 39.4% on JSON. Moving from PDF to OCR nearly halves Score explanation failures, from 24.9% to 13.6%, even though the agent's own transcription remains imperfect. JSON reduces them further to 8.1%. Evidence extraction failures fall by about five points at each step, while score mismatch changes only from 8.4% to 7.5% and 6.6%. Score consistency failures barely change on OCR (8.6% versus 8.7% on PDF) but halve to 4.3% on JSON, partly through repeated errors: among units mis-scored in at least one run, the share with the same wrong score in all three runs rises from 28.4% on OCR to 50.2% on JSON. For the visual-content subset defined in the appendix, score mismatch stays near 15% on all three routes (Table 29).

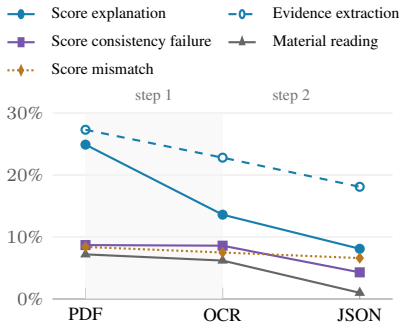


Figure 4: Failure rates of GPT on the three input routes (the same 3,197 grading units). Step 1 grades from the agent's own transcription instead of page images; step 2, from a human-verified one.

5.3 Additional Analysis

More reasoning steps and greater scoring discretion are associated with lower joint pass rates and more written grading judgments. Figure 5(a) shows question types by answer form and reasoning length (Appendix C.1). At the stated-standard level in panel (b), the joint pass rate falls from 41.5% on short calculations to 16.1% on multi-step calculations, and from 28.7% on short answers to 15.0% on multi-step explanations.

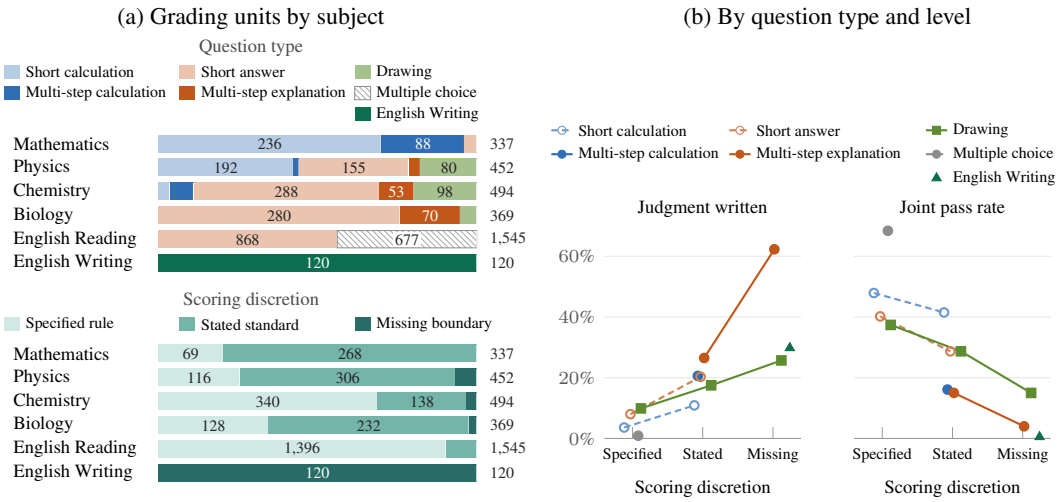


Figure 5: Question type and scoring discretion. (a) Each subject's grading units by question type (upper group) and scoring discretion (lower group). (b) Share of records that write a grading judgment, and joint pass rate, by question type and scoring discretion (five subjects pooled); English Writing's joint pass rate is per script.

Figure 5(a) also groups grading units by the scoring discretion their rubrics leave. In panel (b), the joint pass rate for short answers falls from 40.2% under a specified rule to 28.7% under a stated standard; for drawing questions, it falls from 37.4% under a specified rule to 15.0% with a missing boundary.

These two dimensions account for how grading agents perform on Biology. Biology has the largest shares of multi-step explanations (19%) and short answers (76%), and two-thirds of the grading units in Biology fall under stated standard or missing boundary, so short answers in Biology lag behind those in Physics and Chemistry, and grading agents write a grading judgment for a larger share of the grading units in Biology than in any other subject (31.1%; see the appendix). English Reading is the exception: on short answers it still lags behind Physics and Chemistry because its justification quality is weaker (see the appendix).

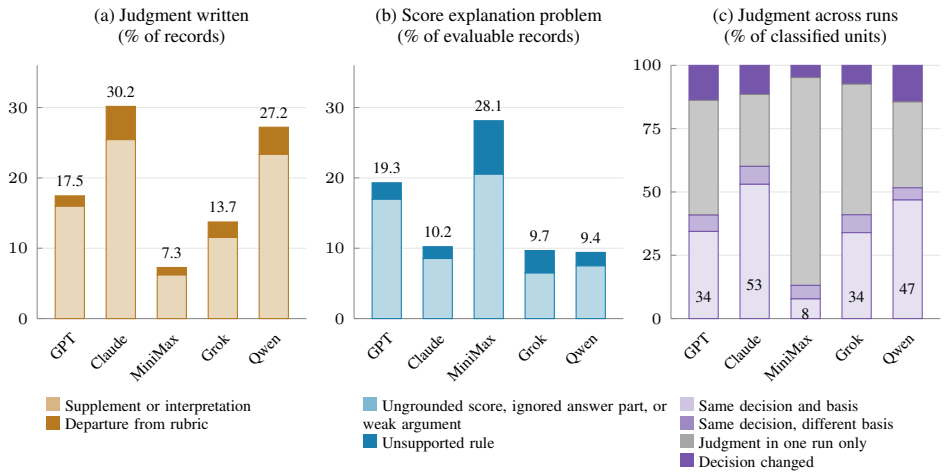


Figure 6: Judgment writing, Score explanation problems, and cross-run judgment outcomes on 1,148 grading units with stated standard or missing boundary (English Writing excluded). In (b), overlapping problems are assigned to unsupported rule. Panel (c) conditions on units with at least one written judgment and a reviewed cross-run class; a judgment need not appear in every run. One MiniMax, one Grok, and five Qwen units lacking a class are excluded from (c).

On grading units whose rubrics leave scoring discretion, agents differ in how they justify scores; their written grading judgments also vary across runs and sometimes depart from the rubric. Figure 6 compares the agents on the 1,148 grading units classified as stated standard or missing boundary, where applying the rubric can require judging whether a differently worded or partly complete answer earns credit. Claude and Qwen most often make a grading judgment explicit (30.2% and 27.2% of records), and both have relatively few Score explanation failures, suggesting that they more often explain how their interpretation supports the score. GPT writes fewer judgments than Claude and Qwen and tends to justify scores by referring to broadly worded rubric criteria without explaining how the student's answer meets them, leaving the reasoning behind the score unclear.

In Figure 6(b), 17.0% of GPT's records fall in the light Score explanation segment: an ungrounded score, an ignored answer part, or a weak argument, excluding records assigned to the overlapping unsupported-rule segment. Grok appears the most conservative about accepting alternative wording as equivalent to the rubric's answer, showing greater reluctance to award credit on that basis. The corresponding exclusive segment is 6.5% for Grok, below GPT's. MiniMax shows the weakest pattern: it often fails to state a needed grading judgment or presents its own interpretation in the reason as though it came from the rubric. It writes the fewest judgments (7.3% of records), and both Score explanation rates in Figure 6(b) are highest for it. Although Claude and Qwen are the most consistent in giving grading judgments across runs, they maintain the same decision on the same basis in merely 53.0% and 46.8% of the cases where they provide a judgment, respectively (Figure 6c).

6 Conclusion

We introduce `ReMarkBench` to evaluate end-to-end agentic grading from raw examination documents along three dimensions: score accuracy, justification quality, and cross-run consistency. Across five grading agents, correctly scored records often fail justification quality, and grading units scored correctly in all three runs can still fail cross-run consistency. The benchmark distinguishes justification quality failures in material reading, evidence extraction, and score explanation, with different patterns across agents. Where the rubric leaves scoring discretion, agents differ in whether they record a grading judgment; recorded judgments can change across runs or depart from the rubric. Together, these findings show how `ReMarkBench` exposes grading failures that score accuracy alone cannot capture. We discuss the limitations of this study in Appendix J.

AI USE STATEMENT

We used AI assistants for literature search, feedback on study design and experiments, code development, result analysis, language editing, and manuscript revision. The authors checked sources, code, results, and text, and take responsibility for the final work. The grading agents and judge used in the experiments are described in the methodology.

ETHICS STATEMENT

The answer scripts in `ReMarkBench` were written by real candidates; we use only the anonymized scripts published by the examination authority and collected no other data from students. The experienced markers who produced the reference scores and the three experts who relabeled records for the judge validation were paid for their work. Errors by grading agents used in practice would affect students' results and the feedback they receive, and our results indicate that current grading agents are far from ready to grade real examinations without human review: even the best configuration achieves a joint pass rate of only 41.0%, averaged equally across the six tasks. A high score on `ReMarkBench` would not by itself justify deploying a grading agent either: the benchmark covers one examination and few candidates, and it does not measure whether grading errors differ across groups of students.

REPRODUCIBILITY STATEMENT

We will provide a complete experimental package for reproducing the experiments and results reported in this paper. The package will include the required data, code, experiment configurations, and instructions for running the experiments and evaluation.

REFERENCES

Catalin Anghel, Andreea Alexandra Anghel, Marian Viorel Craciun, Adina Cocu, Diana-Elena Vulpe, Constantin Adrian Andrei, Calina Maier, Cristian Scheau, Serban Dragosloveanu, and Romica Cergan. GradeAgentOps: A verification-first framework for evidence-anchored LLM exam grading. *AI*, 7(6):198, 2026. doi: 10.3390/ai7060198. URL https://doi.org/10.3390/ai7060198.

Nitay Calderon, Roi Reichart, and Rotem Dror. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16051–16081. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.782/.

Zhongzhou Chen and Tong Wan. Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy. *Physical Review Physics Education Research*, 21(1):010126, 2025. doi: 10.1103/PhysRevPhysEducRes.21.010126. URL https://doi.org/10.1103/PhysRevPhysEducRes.21.010126.

Myroslava O. Dzikovska, Rodney D. Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 263–274. Association for Computational Linguistics, 2013. URL https://aclanthology.org/S13-2045/.

Anna Filighera, Siddharth Singh Parihar, Sebastian Ochs, Tim Steuer, and Tobias Meuser. Your answer is incorrect... Would you like to know why? Introducing a bilingual short answer feedback dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8577–8591. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.587. URL https://aclanthology.org/2022.acl-long.587/.

Malachy Fox, Kavi Samra, and Paul Jung. LLM performance on a real, double-marked GCSE benchmark. *arXiv preprint arXiv:2606.24973*, 2026. URL https://arxiv.org/abs/2606.24973v1. Version 1.

Sebastian Gombert, Zhifan Sun, Fabian Zehner, Jannik Lossjew, Tobias Wyrwich, Berrit K. Czinczel, David Bednorz, Sascha Bernholt, Knut Neumann, Ute Harms, Aiso Heinze, and Hendrik Drachsler. Report on the BEA 2026 shared task on rubric-based short answer scoring for German. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pp. 1179–1192. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.bea-1.85. URL https://aclanthology.org/2026.bea-1.85/.

John Hattie and Helen Timperley. The power of feedback. *Review of Educational Research*, 77 (1):81–112, 2007. doi: 10.3102/003465430298487. URL https://doi.org/10.3102/003465430298487.

Owen Henkel, Libby Hills, Adam Boxer, Bill Roberts, and Zach Levonian. 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)*, pp. 300–304. Association for Computing Machinery, 2024. doi: 10.1145/3657604.3664693. URL https://doi.org/10.1145/3657604.3664693.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4198–4205. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Ruslan Khrulev. CHECK-MAT: Probing the mathematical reasoning and rubric-alignment of vision-language models on handwritten solutions. In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025)*, pp. 77–94. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.mathnlp-main.6. URL https://aclanthology.org/2025.mathnlp-main.6/.

Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. No free labels: Limitations of LLM-as-a-judge without human grounding. *arXiv preprint arXiv:2503.05061*, 2025.

Peichao Lai, Kexuan Zhang, Yi Lin, Linyihan Zhang, Feiyang Ye, Jinhao Yan, Yanwei Xu, Conghui He, Wentao Zhang, Yilei Wang, and Bin Cui. SAS-Bench: A fine-grained benchmark for evaluating short answer scoring with large language models. *Neural Networks*, 203: 109199, 2026. doi: 10.1016/j.neunet.2026.109199. URL https://doi.org/10.1016/j.neunet.2026.109199.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. doi: 10.48550/arXiv.2307.13702. URL https://arxiv.org/abs/2307.13702v1. Version 1.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

Michelle Meadows and Lucy Billington. A review of the literature on marking reliability. Report, AQA, Manchester, May 2005. URL https://filestore.aqa.org.uk/content/research/CERP_RP_MM_01052005.pdf. Report commissioned by the National Assessment Agency.

Gérôme Meyer, Philip Breuer, and Jonathan Fürst. ASAG2024: A combined benchmark for short answer grading. In *Proceedings of the 2024 ACM Virtual Global Computing Education Conference (SIGCSE Virtual), Volume 2*, pp. 322–323. Association for Computing Machinery, 2024. doi: 10.1145/3649409.3691083. URL https://doi.org/10.1145/3649409.3691083.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 752–762. Association for Computational Linguistics, 2011. URL https://aclanthology.org/P11-1076/.

Oikantik Nath, Hanani Bathina, Mohammed Safi Ur Rahman Khan, and Mitesh M. Khapra. Can vision-language models evaluate handwritten math? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14784–14814. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.720. URL https://aclanthology.org/2025.acl-long.720/.

Ofqual. Quality of marking: Description of the marking process used in external exams in general qualifications. Report Ofqual/14/5368, Office of Qualifications and Examinations Regulation, Coventry, February 2014.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Praveen Pathak, Siddharth Tiwary, Charudatt Kadolkar, Vijay Singh, David Rakestraw, Shirish Pathare, and Anwesh Mazumdar. Large scale AI grading of handwritten physics assessments: Score agreement and olympiad team selection outcomes. *arXiv preprint arXiv:2608.20521*, 2026. URL https://arxiv.org/abs/2608.20521.

Janez Perš, Jon Muhovič, Andrej Košir, and Boštjan Murovec. Grading handwritten engineering exams with multimodal large language models. In *Proceedings of the 29th Computer Vision Winter Workshop*, pp. 61–70, 2026. URL https://cmp.felk.cvut.cz/cvww2026/assets/pdfs/CVWW2026-45-final.pdf.

Tasfia Seuti and Sagnik Ray Choudhury. S-GRADES – studying generalization of student response assessments in diverse evaluative settings. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 4684–4701. ELRA Language Resources Association, 2026. doi: 10.63317/2hybhgodz2en. URL https://aclanthology.org/2026.lrec-1.368/.

Weiyu Sun, Liangliang Chen, Yongnuo Cai, Huiru Xie, Yi Zeng, and Ying Zhang. EDU-CIRCUIT-HW: Evaluating multimodal large language models on real-world university-level STEM student handwritten solutions. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15281–15314. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.751. URL https://aclanthology.org/2026.findings-acl.751/.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 74952–74965, 2023. doi: 10.52202/075280-3275. URL https://doi.org/10.52202/075280-3275.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 9440–9450, 2024.

Kevin P. Yancey, Geoffrey T. LaFlair, Anthony R. Verardi, and Jill Burstein. Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pp. 576–584. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.bea-1.49. URL https://aclanthology.org/2023.bea-1.49/.

Zhiqi Yu, Xingping Liu, Haobin Mao, Mingshuo Liu, Long Chen, Jack Xin, and Yifeng Yu. Evaluating AI grading on real-world handwritten college mathematics: A large-scale study toward a benchmark. *arXiv preprint arXiv:2603.00895*, 2026. URL https://arxiv.org/abs/2603.00895v1. Version 1; preprint version of the ICML 2026 paper.

Xuan Zhao, Jiashun Chen, Wanting Xu, Huiyuan Yan, Chaowei Fang, and Xing Wei. EduMARS: Can vision-language models grade like teachers? Benchmarking multimodal, rubric-based assessment on Chinese K-12 answers. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9561–9583. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.466. URL https://aclanthology.org/2026.findings-acl.466/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, 2023.

Appendix guide. Appendix A formalizes the run-level grading task, task manifest, and submission-record specification. Appendices B and C document benchmark composition, canonical alignment, and task-characteristic feature axes. Appendices D and E specify the justification-quality codebook and cross-run consistency comparison rules. Appendix F reports expert validation of the judge. Appendix G provides supplementary experimental results, including question-type analyses and run robustness. Appendix H maps grading challenges to evaluator checks. Appendix I provides extended related work, and Appendix J discusses limitations.

A Formal Grading Task and Record Contract

A.1 Run-Level Task, Agent, and Submission

A grading task is defined at the level of a complete run. Its source materials form the PDF document package

$$\mathcal{D} = \big(P^Q, P^R, \{P_i^A\}_{i\in\mathcal{I}}\big),$$

where P^Q is the question paper, P^R is the rubric, and P_i^A is student i's complete answer script, each as a PDF. For input route λ, let $\mathbf{X}^{(\lambda)}$ denote the complete package actually supplied for one run. On the PDF and OCR routes it consists of the source PDFs; on the JSON route it consists of human-verified structured JSON derived from them (Section 5.2).

Let $\mathcal{Q}_{\text{leaf}}$ be the set of independently marked leaf questions and let $\mathcal{U} \subseteq \mathcal{I} \times \mathcal{Q}_{\text{leaf}}$ be the finite set of expected grading units. The required output is one run-level grading submission containing one uniquely addressable record for every $u = (i, q) \in \mathcal{U}$. The benchmark interface therefore takes whole documents as input and returns records addressed by grading unit. The notation follows the five subject tasks; in the English Writing task, a grading unit is one dimension of one part-level composition (Appendix B.2).

A *grading agent* is the complete mark-generating construction that performs this task: a model running in a harness, which may split the work across subagents, process documents, use the permitted tools, review its own output, and retry within the run. Let C denote a configuration and r one independent run of that unchanged configuration. The raw grading submission returned in run r is

$$\mathbf{Y}_C^{(r)} \sim G_C(\mathbf{X}^{(\lambda_C)}).$$

Thus $\mathbf{X}^{(\lambda_C)}$ and $\mathbf{Y}_C^{(r)}$ are the complete run-level input and output at the benchmark interface. Internal document processing or delegation to per-student subagents remains inside G_C and does not change this boundary.

Before grading begins, the evaluation protocol fixes a task manifest that enumerates $\mathcal{U}$, records each canonical maximum M_u, and provides the authoritative identity-to-material join used after submission. This evaluator-side object is distinct from $\mathbf{X}^{(\lambda_C)}$ and is not inferred from the records the agent happens to return. A raw submission conforms to the task contract only if it contains, for every expected unit, one seven-field record that satisfies the output contract. When a submission is incomplete or malformed, evaluation remains indexed by the unchanged $\mathcal{U}$: post-run Admission produces either one canonical record or one typed placeholder at each expected identity. A placeholder is an evaluator-side representation of nondelivery, not a record returned by the agent.

Table 4 summarizes the objects that are fixed or varied in the evaluation.

Table 4: Formal objects defining a configuration, a run, and a grading unit.

object	definition	evaluation status
C	Model, harness, and reasoning-effort setting; input route; skill, grading instructions, record contract, and the materials mounted in the workspace; orchestration, subagent roles, tools, and any validation, review, or retry rule that can change a mark.	Fixed across runs; identifies the evaluated agent.
r	One independent run of an unchanged configuration C.	Varies across runs; identifies one raw grading submission.
$\mathcal{D}$	The source question paper, rubric, and complete answer scripts as PDFs.	Common source object from which the tested input routes are constructed.
$\mathbf{X}^{(\lambda_C)}$	The complete package actually supplied for one run under configuration C's input route.	Consumed at the run level; its representation varies with the tested route.
$\mathcal{U}$	The finite set of expected grading-unit identities enumerated by the task manifest.	Fixed before outputs are inspected; determines membership and denominators.
Task manifest	The pre-run specification that enumerates $\mathcal{U}$, records M_u, and links each identity to its canonical aligned materials.	Evaluator-side join authority; neither reconstructed from a submission nor supplied to the agent as a separate input.
$\mathbf{Y}_C^{(r)}$	The raw run-level grading submission returned by configuration C in run r.	Checked after the agent boundary; may conform to or violate the output contract.
g_u or writing level	The frozen reference score (five subject tasks) or official level (English Writing task) used for score accuracy.	Evaluator-only and shared across configurations and runs.

The agent boundary ends when the raw grading submission is frozen. Deterministic formatting or collation belongs to G_C only if it can alter or regenerate a mark. Post-run Admission, semantic evaluation, comparison with the reference, and metric aggregation belong to the benchmark evaluator. This boundary allows a route comparison to change where extraction and alignment occur while holding the model, the grading task, and the downstream measurements fixed.

A.2 Source Package, Expected Units, and Canonical Alignment

The source package and task manifest have different organizations and roles. $\mathcal{D}$ is organized as complete examination documents. The task manifest is organized by the identities at which output and evaluation occur. It is fixed before a grading submission is read, so expected membership and denominators do not depend on which records an agent returns. The manifest contains neither an agent configuration nor generated grading output.

For each expected unit $u = (i, q) \in \mathcal{U}$, the evaluator maintains the canonical alignment in Table 5. These fields define the identity, score scale, and educational materials shared across routes. They do not define the physical serialization of $\mathbf{X}^{(\lambda)}$ or a separate invocation of the agent for each unit.

Table 5: Canonical alignment fields associated with one expected grading unit.

alignment field	notation	role
student_id	i	Identifies the student whose answer is to be graded.
question_id	q	Identifies the leaf question to which an independent mark is assigned.
question	Q_u	The source leaf-question content and any question-side figures.
max_score	M_u	The canonical maximum score for the leaf question.
rubric	R_u	The rubric entries that apply to the leaf question.
reference_material	T_u	Any reference answer or other grading material the agent may see; the field may be empty.
student_answer	A_u	The student's answer, including any linked image crops.

Each tested route serializes the source documents differently while preserving the same underlying educational task. The JSON route exposes the alignment of questions, rubric entries, and answers directly. On the PDF and OCR routes, the agent must identify i and q, locate the associated question, rubric entries, and answer, and recognize the maximum score from the PDFs. That recovery is part of G_C. Regardless of route, the generated record must declare the exact canonical identity $u = (i, q)$.

Only material serialized in $\mathbf{X}^{(\lambda_C)}$ is supplied to the agent. In particular, the task manifest is not supplied as a separate structured index on the PDF or OCR routes. For the five subject tasks, the reference score g_u is held in a separate evaluator-only table keyed by the same exact identity. The English Writing task instead holds one official level per complete answer script, which is compared with the combined marks of the script's two part-level compositions (Appendix B.2). Either reference is joined only after the grading submission has been frozen. Reference material T_u that the agent may see belongs to the task materials; the student-specific reference does not.

A.3 Grading-Submission Contract

The output contract requires the raw grading submission to contain one record for every expected manifest entry, using the seven fields in Table 6. These records are generated outputs, not entries in the pre-run manifest.

Formally, the submitted unit record is

$$Y^{(r)}_{u,C} = (i,\ q,\ s^{(r)}_{u,C},\ m^{(r)}_{u,C},\ e^{(r)}_{u,C},\ h^{(r)}_{u,C},\ j^{(r)}_{u,C}).$$

The three explanatory fields have different roles. $e^{(r)}_{u,C}$ identifies the key score-bearing content used in the decision and must be a faithful excerpt or meaning-preserving description of A_u. $h^{(r)}_{u,C}$ explains how that evidence satisfies, fails, or partially satisfies the applicable grading points and how this supports $s^{(r)}_{u,C}$. The conditional field $j^{(r)}_{u,C}$ records a case-specific, score-affecting supplement, interpretation, or departure introduced by the agent when its grading basis is not simply the rubric as written. Ordinary rubric application remains in the reason field; a grading judgment does not replace evidence or explanation.

Table 6: Required fields of a generated grading record.

output field	contract
student_id	Must equal the student identifier i in an expected identity $u = (i, q)$.
question_id	Must equal the leaf-question identifier q. Together, the two identifiers must match exactly one expected grading unit.
score	The awarded numeric mark $s^{(r)}_{u,C}$, bounded by $0 \le s^{(r)}_{u,C} \le M_u$.
max_score	Must agree with the manifest's canonical M_u.
evidence	A faithful excerpt or meaning-preserving description of the key score-bearing student work used in the decision.
reason	A non-empty explanation connecting the observed work, applicable grading basis, and awarded score.
grading_judgment	A string recording a score-affecting supplement, interpretation, or departure introduced by the agent when applicable; otherwise the empty string.

The two identifier fields bind a record to its grading unit. The evaluator joins a record to the manifest only by the exact identity $u = (i, q)$, never by approximate content matching or by the agent's output order. After this join, all semantic evaluation refers to the manifest-bound Q_u, R_u, T_u, and A_u, while score accuracy separately joins the evaluator-only reference score.

The visible audit trace is therefore

$$A_u \longrightarrow e_{u,C}^{(r)} \xleftrightarrow{h_{u,C}^{(r)}} (R_u, T_u, j_{u,C}^{(r)}) \longrightarrow s_{u,C}^{(r)}.$$

This trace is a property of the submitted unit record. It does not claim that generated prose reveals the model's private computational process. It asks whether an external reviewer can inspect the basis that the agent itself offers for the mark.

These seven fields define conformity to the output contract. The narrow normalization rules below can make a nonconforming raw record safe to evaluate, but they do not make the original submission contract-conforming.

A.4 Evaluator-Side Admission of Nonconforming Submissions

A.4.1 Admission Checks and Permitted Normalization

A raw submission may violate the output contract. The evaluator therefore decides Admission independently for each expected identity. A candidate is admitted only when the checks in Table 7 leave exactly one safe record bound to that identity. The canonical admitted representation contains exactly the seven output fields in Table 6; container structure, extraneous fields, and source order do not become grading evidence.

Table 7: Deterministic Admission checks and permitted normalization.

check	deterministic condition	permitted handling	failure disposition
Record schema	The candidate resolves to a record object containing student_id, question_id, score, evidence, and reason.	A bare list of records may be wrapped, and fields outside the contract are discarded. A missing grading_judgment is treated as a repairable contract violation and becomes the empty string; max_score is handled below.	A non-object candidate or one missing any non-defaultable field is rejected.
Identity type	student_id is an integer of at least 1, with no conversion from other types, and question_id is a string.	If the pair is not an expected identity, surface-only canonicalization first trims whitespace and matching outer quotation marks, and then ignores asterisks in the leaf-question identifier; each step is accepted only when the result names exactly one expected identity of that student.	An invalid or still-unknown identity is dropped and cannot create a new expected unit.
Exact membership	The identity pair equals one and only one $u = (i, q) \in \mathcal{U}$.	No approximate content match or generated row order is used.	A record outside $\mathcal{U}$ is dropped; the expected universe is unchanged.
Uniqueness	Exactly one safe candidate claims the expected identity.	None.	If two or more candidates claim the identity, all are rejected; the evaluator never selects a preferred score or explanation.
Field types	score is finite and numeric; evidence and reason are strings; grading_judgment is a string when supplied; and reason is not the empty string.	The empty string is allowed for evidence and grading judgment; max_score is handled under the canonical maximum.	A type violation or empty reason rejects the record.
Score bounds	The awarded score satisfies $0 \le s_{u,C}^{(r)} \le M_u$.	None; the awarded score is never rounded, clipped, or recomputed.	An out-of-range or non-finite awarded score rejects the record.
Canonical maximum	The admitted record's maximum equals authoritative M_u.	A missing, non-numeric, or mismatched submitted maximum is treated as a repairable contract violation and replaced by M_u, provided that the awarded score itself is valid.	The awarded score is never changed; an invalid score follows the score-bounds rule.
Authoritative joins	The identity uniquely binds to manifest-side (Q_u, R_u, T_u, A_u), including any permitted visual answer material.	Joined educational material and authoritative order are restored from the manifest.	If the expected unit lacks a safe authoritative join, no candidate grading content is admitted for that unit.

Table 7: Deterministic Admission checks and permitted normalization (continued).

check	deterministic condition	permitted handling	failure disposition
Coverage and order	Every expected identity occurs once in manifest order after admission.	Safe records are reordered without altering identity or generated content.	An expected identity with no admitted record receives one typed missing/invalid placeholder.

In the English Writing task, one submitted record covers one composition and holds its three dimension records. Admission applies the same identity, membership, and uniqueness checks, with stricter content rules: each dimension mark is an integer from 0 to 7 with a stated maximum of 7, the composition score equals the sum of the three marks with a stated maximum of 21, evidence is a list of non-empty strings, and a grading judgment, when present, is non-empty. A wrong maximum is not repaired, and if any dimension breaks these rules, the whole composition record is rejected, so its three grading units are undelivered.

These operations are structural. Defaulting a missing grading_judgment is a deterministic normalization of the conditional field and supplies no judgment content. Replacing max_score treats the maximum as manifest-owned metadata. It never repairs the awarded score: if misrecognizing the maximum changed the mark, that change remains in $s_{u,C}^{(r)}$ and is reflected in score accuracy; a score outside the canonical bounds is rejected. More generally, Admission may canonicalize a declared identifier, restore manifest-owned values, or remove material outside the output contract. It may not infer an unknown question, choose among duplicates, invent evidence, reason, or judgment text, infer a score from explanatory prose, or change an awarded score. A retry capable of changing a mark belongs to the grading configuration before submission, not to post-run Admission.

A.4.2 Admission States and Denominator Consequences

Table 8 records the distinction between rejection of a candidate record and retention of its expected identity. Retaining expected identities prevents malformed or missing output from improving the joint pass rate by shrinking its denominator. Score accuracy and justification quality are conditional on Admission; cross-run consistency is conditional on comparability.

Table 8: Admission state and downstream denominator consequences.

state	unit-level representation	measurement consequence
Admitted	One canonical seven-field record bound to the expected identity.	Enters score accuracy and justification quality within their stated scopes; counts as delivered for cross-run consistency and the joint pass rate.
Missing or rejected record	One typed placeholder carrying identity and no generated grading content.	Retained in the expected units, but excluded from score accuracy and justification quality; counts as a failure in the joint pass rate. Cross-run consistency compares only the delivered records of the same grading unit: a unit with at least two delivered runs enters the cross-run consistency denominator, and a unit with fewer is not comparable and is not counted as unstable (Appendix E).
Unknown submitted record	No expected-unit representation; the record is dropped.	Cannot enlarge $\mathcal{U}$. If an expected identity consequently has no safe record, that expected identity receives the placeholder above.
Conflicting duplicates	No duplicate is selected; the expected identity receives a placeholder.	Same downstream treatment as a missing or rejected record.

A.4.3 Score-Comparison Reporting

Every comparable admitted record in run r is joined to the same frozen evaluator-only g_u; neither another run nor model consensus serves as its reference. A missing or rejected record carries no candidate score and therefore does not enter the exact-match rate. Its effect remains visible in the

joint pass rate, whose denominator in the five subject tasks is every expected grading unit with a non-empty answer, whether or not a record was delivered for it.

A.4.4 Schematic Binding Example

For a manifest entry $u = (17, \texttt{q3b})$ with $M_u = 4$, a conforming record declares the same two identifiers, an awarded score in $[0, 4]$, a maximum of 4, and the three explanatory fields. If the raw record omits the grading-judgment field or misstates only the maximum, Admission supplies the canonical value without changing the score; the raw record remains nonconforming. This example is schematic and contains no evaluator-only reference score.

B Benchmark Composition and Alignment

B.1 Composition of the Six Tasks

The benchmark is built from real-world examination documents (question papers, rubrics, and anonymized student answer scripts) from one examination system, covering five subject tasks and one English Writing task. Each grading unit aligns a leaf question, its maximum score, the applicable rubric entries, any reference material the agent may see, the student's answer, and an evaluator-only reference under one stable identity. Table 9 gives the composition of all six tasks. Unless stated otherwise, the results for the five subject tasks in Section 5 use the 3,217 grading units with non-empty answers.

Not every leaf question applies to every script. In English Reading, every candidate answers Part A and one of two alternative Part B sections, so the 121 leaf questions give 80–82 grading units per script; 10 of the 20 scripts take each Part B section. In the English Writing task, every candidate writes the Part A composition and one of four Part B compositions, each chosen on five scripts. In the other four tasks, every leaf question defines an expected grading unit for every script. One of the 121 English Reading leaf questions was deleted from grading, and the rubric awards its mark to every candidate; its 20 grading units are counted here but are not annotated with the feature axes of Appendix C. Questions are numbered hierarchically, for example 14(b)(i); a leaf question at depth three sits two levels below its top-level question.

In the English Writing column, a leaf question is one composition task; each composition is marked out of 21 (three dimensions of 0–7) and forms three grading units.

B.2 English Writing Task

The English Writing task contains 20 complete answer scripts. Each script contains one Part A and one Part B composition, giving 40 compositions. Each composition is judged as a whole on three dimensions, Content, Language, and Organisation, so the task has 120 grading units. Each dimension is marked on a single-marker scale of 0–7, so a composition receives 0–21 marks; the 42-mark maximum printed on the source paper reflects double marking and is not used. The official reference is one level from 1 to 5 for each complete answer script, with four scripts per level; the examination publishes no reference score for individual grading units and no mark boundaries for the levels. The compositions and grading units are therefore counted at the part level, whereas the official levels belong to the 20 complete answer scripts. The task keeps the record fields and justification quality checks of the five subject tasks, but its score accuracy and joint pass rate are defined per script, and its cross-run consistency per grading unit.

For script i in run r, let $t_{i,\mathrm{A}}^{(r)}$ and $t_{i,\mathrm{B}}^{(r)}$ be the Part A and Part B totals, each the sum of three dimension marks. The evaluator combines them as $z_i^{(r)} = 0.4\, t_{i,\mathrm{A}}^{(r)} + 0.6\, t_{i,\mathrm{B}}^{(r)}$, following the 40%/60% weighting prescribed by the examination, and converts $z_i^{(r)}$ to a level by a least-squares linear fit of the official level on the combined score, rounded to the nearest level. Each configuration has one conversion, shared by its runs, so a script's level changes across runs only when its combined score does. The script being converted is left out of the fit that converts it, so its own official level never enters its conversion. On the converted level:

Table 9: Composition of the six grading tasks. Percentages in the question rows are over the leaf questions of a task; the other percentages are over its grading units with non-empty answers, except the share row and the E1 row. E1 and E2 are the recognition-demand axes of Appendix C; E1 is an OCR-based proxy defined only for answers without images, and its row is over those answers. English Writing has no unit-level reference scores; its reference is one official level (1–5) per script, four scripts per level. —: not applicable.

	Mathematics	Physics	Chemistry	Biology	English Reading	English Writing
Documents						
Answer scripts	10	10	10	10	20	20
Pages: question paper	17	20	19	20	21	12
Pages: rubric	12	9	11	9	5	2
Pages per answer script	24	19	19	19–20	8	4–10
Questions						
Leaf questions	42	48	53	40	121	5
Leaf questions per script	42	48	53	40	80–82	2
Maximum depth	3	3	4	3	2	1
Leaf questions at depth ≥ 3	28.6%	58.3%	52.8%	32.5%	0.0%	0.0%
Maximum score per leaf question	1–5	1–4	1–6	1–11	1–2	21
Total marks per script	105	84	84	84	84	42
Grading units						
Grading units	420	480	530	400	1,620	120
Share of the five subject tasks	12.2%	13.9%	15.4%	11.6%	47.0%	—
With non-empty answers	337	452	494	369	1,565	120
Answers containing images	0.0%	20.6%	22.7%	5.1%	0.0%	0.0%
Grading reads figures or tables (E2)	8.9%	20.4%	26.1%	18.7%	0.6%	—
OCR loses $>20\%$ of characters (E1 $= 2$)	9.2%	2.8%	3.9%	27.1%	1.9%	—
Reference scores (non-empty answers)						
Full marks	69.7%	56.9%	52.0%	47.2%	60.1%	—
Partial marks	13.1%	12.6%	14.6%	25.7%	0.8%	—
Zero	17.2%	30.5%	33.4%	27.1%	39.1%	—

- the exact-match rate compares the converted level with the official level, over scripts and runs;
- cross-run consistency is the share of the 120 grading units that receive the same mark in every run; only score consistency is compared;
- justification quality is checked per record as in the five subject tasks, each script contributing six dimension records per run;
- a script's joint score is the justification quality pass rate of all its dimension records over the K runs when its converted level is correct and identical in every run, and zero otherwise; the joint pass rate averages this score over the 20 scripts.

In Table 3, the writing row enters the average over the six tasks.

As supplementary measures of rank agreement, rank-constrained quadratic weighted kappa (QWK) and Somers' D compare the ordering of the combined scores with the official levels. Both depend only on this ordering, so they are invariant to strictly increasing transformations of the combined score. Each is computed per run and averaged over the runs.

Rank-constrained QWK uses the known frequency of four scripts per official level in this sample. The 20 scripts are ordered by $z_i^{(r)}$ from lowest to highest and assigned predicted levels 1 through 5 in successive groups of four, matching the official level frequencies. When a group of tied scripts crosses a level boundary, the squared level error is averaged analytically over all equally likely orderings within the group. Under this sample's uniform five-level frequencies, the metric simplifies to

$$\kappa = 1 - \frac{1}{80} \sum_{i=1}^{20} \mathbb{E}[(\hat{\ell}_i - \ell_i)^2],$$

where ℓ_i is the official level, $\hat{\ell}_i$ is the level assigned by the ranking rule, and the expectation is only over orderings within groups of tied scripts.

B.3 Canonical Alignment Across Input Routes

In the five subject tasks, canonical identities preserve the student and leaf-question coordinates through extraction, grading, Admission, and evaluation. The same grading unit may reach the agent through more than one input route. The routes need not present identical strings, because how the materials are read is itself what the routes vary, but each must denote the same question, student's answer, rubric, and score scale, and each must map back to the same expected grading units. The PDF route is specified in Section 5.1; the OCR and JSON diagnostic routes are described in Section 5.2.

The evaluator's canonical alignment supplies the key for the post-run join. It records the educational materials represented across routes but remains separate from the student-specific reference. Outputs from every route are joined by the exact identity $u = (i, q)$, never by approximate text matching or by the order in which the agent writes its records. Writing outputs keep their script, part, and dimension coordinates, and the evaluator combines the part totals of each script as above before comparing them with the script-level ordinal reference.

C Task Characteristics and Feature Axes

Section 5.3 compares grading units by question type and scoring discretion. These are independent question-level annotations, not transformations of the ten axes below. This appendix defines them and then documents a separate set of ten supplementary feature axes. Among them, E1 and E2 support the composition and visual-content analyses in Appendices B and G. The annotations are assigned before examining the grading agents' outputs.

C.1 Scope and Value Rules

Question type and scoring discretion. In the five subject tasks, both annotations are assigned to each leaf question and shared by its grading units. They cover 304 questions. One English Reading question whose rubric awards the mark to every candidate is excluded from the analyses, leaving 303 questions and 3,197 grading units. English Writing is shown separately from the six question types in Figure 5(a).

Question type first follows the form in which credit is earned: calculation, written answer, drawing, or selection from given options. Calculations and written answers are then split by the number of successive reasoning steps or links. At most two give *short calculation* or *short answer*; three or more give *multi-step calculation* or *multi-step explanation*. A step must build on the preceding one: parallel points, rubric entries, marks, and answer length are not counted as steps. Operations listed only in chronological order do not add steps; neither does a result from an earlier question or the final choice of an answer. Drawing questions form their own type. *Multiple choice* covers selection from given options, including True/False/Not Given; word filling, filling a text-only table, and labeling a diagram remain short answers because accepted wording or placement must still be assessed. Mixed questions follow the form that earns most of the marks, with drawing taking precedence.

Scoring discretion is assigned by checking every scoring point in the complete rubric and taking the highest level that could affect the mark on a plausible answer; it does not depend on the observed student answer or an agent's score. Under *specified rule*, the rubric states the accepted content, marks, and any needed point allocation. Under *stated standard*, it gives the scoring target and allocation, but the grader must decide whether a substantively different wording, solution path, or partial step meets that target. Under *missing boundary*, the rubric leaves a consequential point split or qualitative mark boundary unresolved, so the grader must supply a scoring rule. A list of accepted synonyms, a method mark, or general permission to use professional judgment alone does not raise the level; "or equivalent" raises it only when equivalence requires substantive judgment. A two-mark criterion that clearly states two independent requirements specifies one mark per requirement. The 120 English Writing grading units share band descriptors rather than point-specific award rules and are assigned missing boundary. This question-level annotation is distinct from the older J1 answer-form axis below.

For each question, the annotator examined the question and complete rubric. The question-type annotation records the reasoning-step count or classification basis; the scoring-discretion annotation records the relevant scoring point, rubric excerpt, and reason for its level. Programmatic checks verify question coverage and agreement between recorded step counts and question-type labels, and verify the scoring-discretion question IDs. Each missing-boundary label names the unresolved boundary. Three Physics graph or diagram questions and one English Reading question were additionally checked against the original rubric PDF. Neither annotation has an independent second human annotation, so we report no inter-annotator agreement.

Supplementary feature axes. The ten axes below use the same 303-question, 3,197-unit five-subject scope. English Writing is not annotated on these axes. They separately describe recognition demand, reasoning demand, and judgment latitude; J1 answer form is not a proxy for the scoring-discretion levels used in Figure 5.

Four axes are question-level (E3, A1, A2, J1): each takes one value per leaf question, shared by all answers to that question. The other six are unit-level (E1, E2, B1, B2, B3, J2) and take one value per grading unit. Every value is determined by the question paper, the rubric, and the student's answer, except that E1 uses an independent OCR system as a proxy; no value depends on the output of an evaluated agent. For the reasoning axes, the criterion is subject logic: the rubric shows what a correct solution contains, but its number of entries or marks is not counted as steps. Wherever possible, the annotator records a countable quantity and the level follows from it by a fixed cut-off (Table 10); only J1 and J2 are rated directly. E1, E3, and part of J2 are assigned mechanically, without human judgment. An axis that does not apply is recorded as not applicable rather than at its lowest level, because the lowest level asserts something that would be false there (for example, that a chain of reasoning skips no step when the answer contains no chain). Each axis is therefore analyzed only over the grading units to which it applies.

Table 10: Feature axes, their annotation level, and their value rules.

axis	level	quantity and levels	not applicable when
Recognition demand			
E1 Text and symbols	unit	Share of the characters in the human-verified transcription that an independent OCR system fails to recover, after notation is normalized; the maximum over four to five OCR runs. 0: none lost; 1: at most 20%; 2: more than 20%. Answers in closed notation (option letters, True/False/Not Given) are set to 0, and answers transcribed as illegible to 2.	The answer is kept only as an image (read under E2).
E2 Non-text material	unit	Number of elements (objects, curves, labels, connections, axis assignments, data rows) that grading must read from figures or data tables, on the question and answer sides together, and combine. 0: at most 3; 1: 4–5; 2: more than 5. Answer tables that only provide blanks do not count.	Grading needs no such element.
E3 Rubric structure	question	Number of structural features in the rubric: several entries; separate method and accuracy marks; alternative solutions; starred marks. 0: none; 1: one; 2: two or more. Accepted ranges and "any N of" rules are not counted; they belong to judgment latitude.	—
Reasoning demand			
A1 Dependency depth	question	Number of successive reasoning steps the question requires; a step is an independently checkable assertion or transformation that builds on the previous one. Parallel points add no step, and a method mark with its accuracy mark is one step. 0: one step; 1: two; 2: three or more.	English Reading.
A2 Joint conditions	question	Largest number of specific quantities or constraints that a single step must use together; general formulas, theorems, and rules are not counted. 0: at most one; 1: two; 2: three or more.	English Reading.
B1 Answer chain length	unit	Number of successive links the student writes out, counted along the student's own solution path. 0: at most one; 1: two; 2: three or more.	Single-step question; answer without two or more mutually supporting assertions; English Reading.
B2 Skipped steps	unit	Number of places where the answer states an assertion without writing the step that supports it. 0: none; 1: at least one.	As for B1.
B3 Error carry-forward	unit	Whether later steps build on the first error, where an error is judged by subject knowledge regardless of the marks the rubric awards. 0: no later step builds on it; 1: later steps build on it.	The answer contains no error; single-step question; English Reading.
Judgment latitude			
J1 Answer form	question	Rated directly. 0 closed form: a value, unit, formula, equation, or designated term, decided by exact match; 1 equivalence list: the rubric lists equivalent expressions or allows "or equivalent" or "any one of"; 2 open wording: an explanation, reason, or mechanism for which the rubric gives a single reference wording.	—

axis	level	quantity and levels	not applicable when
J2 Deviation from the rubric	unit	Rated directly; exact matches and wrong closed options are assigned mechanically. 0 matching form: the answer corresponds to an identifiable rubric entry in the same form; 1: it corresponds to an identifiable entry in a different form, recorded as a wording, path, or completion deviation; 2 content deviation: it corresponds to no entry.	—

C.2 Distribution over Subjects

Table 11 gives the distribution of each axis within each subject. Because English Reading contributes about half of the unit-level denominators, subjects should be compared through these within-subject distributions rather than through pooled level counts.

Table 11: Within-subject distribution of the feature axes (% at levels 0/1/2; 0/1 for the two-level axes B2 and B3). Parentheses give the number of applicable grading units for unit-level axes and of leaf questions for question-level axes. —: not applicable.

axis	Mathematics	Physics	Chemistry	Biology	English Reading	total
E1	46.0/44.8/9.2 (337)	69.1/28.1/2.8 (359)	74.9/21.2/3.9 (382)	25.4/47.4/27.1 (350)	93.7/4.3/1.9 (1,545)	2,973
E2	0.0/100.0/0.0 (30)	45.7/27.2/27.2 (92)	42.6/41.1/16.3 (129)	49.3/47.8/2.9 (69)	0.0/100.0/0.0 (10)	330
E3	0.0/16.7/83.3 (42)	0.0/41.7/58.3 (48)	62.3/28.3/9.4 (53)	52.5/47.5/0.0 (40)	97.5/2.5/0.0 (120)	303
A1	26.2/31.0/42.9 (42)	75.0/20.8/4.2 (48)	66.0/22.6/11.3 (53)	47.5/32.5/20.0 (40)	—	183
A2	14.3/35.7/50.0 (42)	29.2/45.8/25.0 (48)	54.7/28.3/17.0 (53)	87.5/7.5/5.0 (40)	—	183
B1	5.1/19.1/75.7 (235)	7.3/44.5/48.2 (110)	8.5/45.1/46.3 (164)	7.5/33.9/58.6 (174)	—	683
B2	96.6/3.4 (235)	90.9/9.1 (110)	90.2/9.8 (164)	81.6/18.4 (174)	—	683
B3	37.1/62.9 (70)	33.3/66.7 (39)	68.3/31.7 (82)	53.6/46.4 (84)	—	275
J1	81.0/0.0/19.0 (42)	62.5/10.4/27.1 (48)	50.9/17.0/32.1 (53)	12.5/12.5/75.0 (40)	68.3/23.3/8.3 (120)	303
J2	35.6/49.0/15.4 (337)	38.9/36.9/24.1 (452)	28.7/46.6/24.7 (494)	11.9/63.7/24.4 (369)	46.2/15.8/38.0 (1,545)	3,197

C.3 Limitations

No second annotator has independently re-annotated the human-rated quantities, so we report no inter-annotator agreement; the mechanically assigned values are reproducible from the stored materials. E1 is a behavioral proxy: it measures what an independent OCR system fails to recover, not what a human reader finds hard, and taking the maximum over OCR runs counts the run-to-run variation of that system as difficulty. E3 is strongly associated with subject (Cramér's $V = 0.665$), because each subject writes its rubrics in its own format, so it supports only within-subject comparisons; it is nearly constant in mathematics and English Reading. E2 applies to only 330 grading units, with mathematics and English Reading entirely at level 1, and B2 places fewer than one tenth of its units at level 1. A1 remains closely associated with the maximum score of a question, which should be stated whenever results are stratified by A1.

J2 is not ordered by latitude. At level 0 the reference score is almost always full (98.7% of units), and at level 2 it is almost always zero (93.8%), so both ends leave little discretion, and the latitude the axis is meant to capture lies mainly at level 1. Level 1 also combines answers that are correct in a different wording or by another path with answers that are only partly complete. J2 is therefore retained as a nominal grouping by deviation type rather than an ordered scoring-discretion axis. Path deviation occurs only in mathematics and physics, whose rubrics list alternative solutions.

C.4 Material and Grading Characteristics of the Six Tasks

Table 12: Material and grading characteristics of the six tasks.

Task	Material characteristics	Grading characteristics
Mathematics	The materials contain extensive mathematical notation. The rubric uses solution–marks–remarks columns, with labels for method and accuracy marks, boxes, and remarks specifying how marks are awarded for different solutions and steps.	Mainly calculations; grading requires checking each step and the reasoning that links the steps.
Physics	The materials include experimental apparatus diagrams, circuit diagrams, and graphs. The rubric uses solution–marks–remarks columns, with some mark labels embedded directly in diagrams.	Answers include calculations, short answers, experimental explanations, and drawings; grading requires checking calculation steps, explanations, and diagrams in their physical context.
Chemistry	Chemical equations, electron diagrams, structural formulae, and calculations appear together. The rubric uses symbols, underlining, and notes to specify acceptable answers and conditions for awarding marks.	Answers include short answers, calculations, drawings, and extended explanations; grading requires checking chemical representations and calculations and assessing whether reasoning based on structures, reaction conditions, and experimental observations is valid.
Biology	Some scans combine two pages side by side. The materials include handwriting, micrographs, tables, and graphs. The rubric uses columns for grading concepts, examples, and marks.	Mainly short answers, alongside drawings and extended explanations; grading draws on biological mechanisms and experimental data to assess whether different wordings cover the grading points and whether explanations and inferences are valid.
English Reading	Reading passages are printed in booklets separate from the questions and answer spaces. Part A is compulsory, with a choice between Parts B1 and B2. Answer pages mix selection marks, table completion, true/false/not-given matrices, image options, and dialogue boxes.	Mainly structured and short answers; grading requires checking information retrieval, referents, and inferences in the context of the passage, distinguishing words copied from the text from acceptable paraphrases, and checking requirements such as word count and grammar.
English Writing	Part A provides visual and textual prompts and specified subheadings; Part B offers a choice of four tasks in different genres. Handwritten compositions span multiple pages and include crossings-out, interlinear insertions, and marginal additions. The rubric consists of band descriptors.	Each composition is assessed as a whole in light of its purpose and genre, and receives separate marks for Content, Language, and Organisation, each on a single-marker 0–7 scale.

D Justification Quality: Formal Model and Check Codebook

This appendix formalizes what a grading record must show for its score to be checked and then operationalizes the thirteen checks of justification quality. Table 13 maps the checks to the failures of Section 4. For each check, the appendix specifies the information available to the judge, the information deliberately withheld, the allowed result, and the boundary with adjacent checks. The checks inspect the submitted record; they neither reconstruct hidden reasoning nor independently re-grade the student's answer.

Table 13: Justification quality checks and the failures of Section 4. A record passes justification quality when none of these failures occurs; the two judgment checks only describe grading judgments.

failure	checks whose failure counts toward it
Material reading – misstated answer	*Answer misstated in the reason*: the reason or judgment misstates the student's answer; *evidence misquotes the answer*: the evidence misquotes it.
Evidence extraction – unused evidence	*Extracted evidence unused*: extracted evidence is left unused by the reason.
Evidence extraction – missing evidence	*Answer used by the reason not extracted*: answer content used by the reason is not extracted.
Evidence extraction – irrelevant evidence	*Evidence unrelated to any grading point*: extracted evidence bears on no grading point.
Score explanation – ungrounded score	*Grading point not specifically explained*: a grading point is not specifically explained by the reason.
Score explanation – ignored answer part	*Grading point without evidence or explanation*: an attempted grading point has neither evidence nor explanation.
Score explanation – unsupported rule	*Criterion outside the rubric and judgment*: the reason uses a grading basis found in neither the rubric nor the grading judgment.
Score explanation – contradiction	*Self-contradictory reason*: the reason contradicts itself; *reason contradicts the score*: the reason contradicts the awarded score.
Score explanation – weak argument	*Grading argument rating*: the grading argument is rated 0 or 1, or is not assessable.
Descriptive only	*Judgment–rubric relation*: how the grading judgment relates to the rubric; *judgment applied in the reason*: whether the reason applies the handling declared in the judgment.

D.1 Formal Evaluation Object and Atomic Audit Trace

For a grading unit u with an admitted record, let A_u denote the student's answer, R_u the rubric, J_u the submitted grading judgment, and s_u the awarded score. The effective grading basis is

$$\mathcal{B}_{J,u} = \mathrm{EffectiveBasis}(R_u, J_u),$$

with $\mathcal{B}_{J,u} = R_u$ when J_u is empty. Among the answer paths permitted by this basis, let

$$\Pi_u^\star(A_u) \coloneqq \mathrm{ClosestApplicablePath}(A_u, \mathcal{B}_{J,u})$$

denote the path most closely aligned with the student's actual attempt. This notation denotes a rubric-constrained alignment for one grading unit; it assumes neither a numerical distance nor a unique surface representation of a path. It determines which grading points require explanation but does not authorize the judge to invent a path, select a preferred answer, or recompute a preferred score. Let

$$\mathcal{R}_u^\star(A_u) = \mathrm{AtomicPoints}(\Pi_u^\star(A_u)) = \{R_{u1}, \ldots, R_{un}\}$$

be the atomic grading points on that path that the current grading decision must address. Further, let

$$A_{ui} = \mathrm{WorkFor}(A_u, R_{ui})$$

denote the answer component directed at R_{ui}. Any directed attempt makes $A_{ui} \neq \varnothing$, even when the attempt is incorrect or incomplete; $A_{ui} = \varnothing$ only when no directed attempt is present.

The record supplies declared evidence units $\widehat{\mathcal{E}}_u$ and reason units $\widehat{\mathcal{P}}_u$. Evidence may represent correct, incorrect, or incomplete work. For each attempted R_{ui}, an ideal audit trace contains a faithful evidence unit e_{ui} and a reason unit r_{ui} that explains why the observed work satisfies, fails, or partially satisfies the point and how the decision contributes to s_u:

$$\mathcal{C}_{ui} \coloneqq (R_{ui}, r_{ui}, e_{ui}), \qquad R_{ui} \longleftrightarrow r_{ui} \longleftrightarrow e_{ui},$$

where

$$\mathrm{Faithful}(e_{ui}, A_u), \qquad \mathrm{Bridge}(e_{ui}, R_{ui} \mid r_{ui}), \qquad \{r_{ui}\}_{i=1}^{n} \rightsquigarrow s_u.$$

The indices come from the applicable grading points, not from the way the agent organizes its prose. The one-point–one-reason–one-evidence correspondence is an atomic representation used for

evaluation, not a formatting requirement on the raw fields. One sentence may express several reason units and one evidence field may contain several excerpts; evaluation decomposes and aligns them relative to R_{ui}. For an attempted grading point, failure to identify a corresponding r_{ui} or e_{ui} is a structural omission, not an empty chain.

An unattempted component is an exception only to the evidence requirement. An attempted point with $A_{ui} \neq \varnothing$ requires corresponding evidence and reason, whereas $A_{ui} = \varnothing$ requires no synthetic evidence but must be explicitly acknowledged by the reason. Its structural form is

$$\mathrm{TracePresent}(R_{ui}) = \begin{cases} \exists\, e_{ui} \in \widehat{\mathcal{E}}_u, r_{ui} \in \widehat{\mathcal{P}}_u : \\ \mathrm{Corresponds}(e_{ui}, r_{ui}, R_{ui}) & A_{ui} \neq \varnothing, \\ \exists\, r_{ui} \in \widehat{\mathcal{P}}_u :\ \mathrm{AcknowledgesAbsent}(r_{ui}, R_{ui}), & A_{ui} = \varnothing. \end{cases}$$

Here $\mathrm{Corresponds}$ states only that the three atomic units can be assigned to the same grading point; it prejudges neither evidence fidelity nor connection strength. A complete ideal trace must also satisfy the $\mathrm{Faithful}$, Bridge, and reason-to-score conditions above. The content and logic checks below test whether the declared content is trustworthy, and the judgment and explanation-structure checks test whether the required structure exists and forms substantive connections. Their outputs remain independent: no check rewrites the decision of another.

D.2 Judge Inputs and Output Format

For compactness, let Q denote the question, including the stems of its parent questions and any question-side figures; A the student's answer, including any linked image crops; B the rubric entry of the leaf question (R_u above), which also states its reference answer and maximum score; E the declared evidence; P the submitted reason; J the submitted grading judgment; s the awarded score; and M the canonical maximum. Each judge request also carries the student and leaf-question identifiers, which provide no grading information. In the English Writing task, Q is the writing prompt, B is the original band-descriptor rubric with the 0–7 mark scale, and A includes the pages of the student's answer script as images, with an instruction to use only the composition being graded; no reference level is exposed.

Table 14 covers every check. For each row, the inputs not named in the judge-visible column are withheld from that check. The evaluator-only reference score g, the outputs of all other checks, and the results of other configurations or runs are always withheld.

Table 14: Inputs visible to the judge for each justification quality check. M is visible wherever B is, because the rubric entry states the maximum.

check	judge-visible inputs	withheld inputs
Answer misstated in the reason	Q, A (text and image crops), P, J	B, E, s, M
Evidence misquotes the answer	Q, A, E	B, P, J, s, M
Criterion outside the rubric and judgment	Q, B, P, J	A, E, s
Self-contradictory reason	P	Q, A, B, E, J, s, M
Reason contradicts the score	s, M, P	Q, A, B, E, J
Judgment–rubric relation	Q, B, J	A, E, P, s
Judgment applied in the reason	B, E, s, M, P, J	Q, A
Extracted evidence unused	E, P	Q, A, B, J, s, M
Answer used by the reason not extracted	A, E, P	Q, B, J, s, M
Grading point not specifically explained	Q, A, B, P, J	E, s
Grading point without evidence or explanation	Q, A, B, E, P, J	s
Evidence unrelated to any grading point	Q, B, E, J	A, P, s
Grading argument rating	Q, B, E, s, M, P, J	A

The five content and logic checks and the five coverage checks each return a pass or fail result, a subtype, a decisive quote, and a rationale. The check of whether the reason applies the judgment has the same form and may also return `not_applicable`. The judgment–rubric relation check returns one relation label, a decisive quote, and a rationale. The grading argument rating returns an assessability status, an integer grade in $\{0, 1, 2\}$ when assessable, diagnostic findings, and a rationale. Every rationale must be specific to the grading unit. A failure of the two grading-point checks, which can be caused solely by omission, may have an empty decisive quote because no supplied text can faithfully stand for the missing content.

Only the two judgment checks use `not_applicable`. Deterministic assembly applies

$$J = \varnothing \Longrightarrow \text{both judgment checks} = \texttt{not_applicable},$$

without a judge request; a non-empty J requires a substantive decision for both checks. The grading argument rating uses the distinct status `not_assessable` when the record contains no complete evidence–grading-point–score connection to rate: the check remains applicable, its grade is undefined, and the record counts under Score explanation – weak argument. Records whose student's answer is empty, and placeholders for undelivered records, are removed before any check runs, so every check, and every rate in Section 5, covers grading units with non-empty answers.

D.3 Check Definitions

D.3.1 Content and Logic Checks

The content and logic checks test whether the content and claims declared in the record are trustworthy. Their five failure events have the following logical skeleton:

Answer misstated in the reason:

$$\exists x \in \mathrm{StudentWorkClaims}(P, J) : \neg\, \mathrm{SupportedBy}(x, A),$$

Evidence misquotes the answer:

$$\exists e \in E : \neg\, \mathrm{Faithful}(e, A),$$

Criterion outside the rubric and judgment:

$$\mathrm{BasisUsed}(P) \not\subseteq \mathcal{B}_J,$$

Self-contradictory reason:

$$\exists c_p, c_q \in \mathrm{Claims}(P) : c_p \perp c_q,$$

Reason contradicts the score:

$$\exists c \in \mathrm{ScoreConclusions}(P) : c \not\sim s.$$

These expressions identify the formal condition violated by each check. The operational failure boundaries, exclusions, and edge cases used by the judge follow below.

Table 15: Operational boundaries for the content and logic checks.

check	failure boundary	exclusions and edge cases
Answer misstated in the reason	P or J attributes to the student content that the student's answer does not contain, reverses its meaning, or imports from another answer.	Meaning-preserving compression and visibly supported inference pass. A visual placeholder directs the judge to the associated image crop; it is not evidence that the answer was empty. Rubric use, score preference, and writing quality are excluded.
Evidence misquotes the answer	E fabricates, materially rewrites, reverses, or cross-question misattributes student work.	An empty E passes because absence is not fabrication. Partial coverage is tested by the coverage checks, and a faithful excerpt need not reproduce the whole answer.
Criterion outside the rubric and judgment	P treats a criterion, condition, exception, penalty, or rule found in neither B nor J as part of the grading basis.	This check asks only whether the invoked basis exists. The judgment–rubric relation check, not this one, classifies how J relates to B; omissions and weak connections belong to the explanation-structure checks.
Self-contradictory reason	Two score-relevant claims inside P are irreconcilable about the work, criterion satisfaction, or the grading conclusion.	Comparison with the externally awarded score belongs to the next check. The judge does not add unstated rubric rules to manufacture a contradiction.
Reason contradicts the score	The score consequence explicitly stated by P conflicts with s, including credit said to be earned but not awarded or credit awarded despite a stated denial.	The judge checks the record's own reason-to-score account. It does not calculate a preferred score from A or B, and a terse but compatible account may pass.

A failed content or logic check reports its check-specific subtype and the smallest faithful excerpt that exposes the defect; a passing check reports no subtype. All five checks remain separate: for example, invented student work in P can fail the answer-misstated check while the numerical conclusion stated by that same P remains consistent with s, so the reason-contradicts-the-score check passes.

D.3.2 Judgment and Explanation-Structure Checks

Judgment checks. For non-empty J, the judgment–rubric relation check selects exactly one relation in Table 16. The label describes how the judgment relates to the rubric and does not affect whether the

record passes justification quality. The check of whether the reason applies the judgment separately asks whether the declared handling operates in the score account formed by P and J; the reason need not repeat a canonical term, accepted equivalence, exception, or score consequence already stated in J.

Table 16: Relation labels of the judgment–rubric relation check for a non-empty grading judgment.

label	**anchor**
Supplement	B leaves the relevant case uncovered or incomplete and J supplies concrete handling, including resolution of ambiguity in the student's wording, notation, or diagram.
Interpretation	B itself has more than one reasonable reading or contains an explicit allowance whose scope must be applied, such as equivalence, an alternative method, rounding, or follow-through.
Departure	J uses a basis that conflicts with a clear requirement in B. This is a descriptive relation, not an automatic quality failure.

The check of whether the reason applies the judgment passes when the combined $P + J$ account applies the declared case-specific handling, whether explicitly or implicitly. It fails only when the account ignores, bypasses, or contradicts that handling. It does not relabel the judgment, re-grade the answer, or fail merely because information is divided between P and J. Like the relation check, it is descriptive and does not enter the justification quality pass condition.

Coverage checks. Let $\mathrm{Uses}(P, x)$ mean that the reason uses evidence or student work in a locatable, meaning-preserving way; let $\mathrm{Extracted}(A_i, E)$ mean that the answer component is declared or locatable in the declared evidence; $\mathrm{Explained}(P, R_i)$ that the reason explains the treatment of the grading point; $\mathrm{HasEvidence}(E, R_i)$ that some declared evidence addresses the answer component governed by that point; and $\mathrm{Relevant}(e, R_i)$ that the evidence addresses the answer slot governed by that point. The index i ranges over the grading points in $\mathcal{R}^\star(A)$. Because the check for evidence unrelated to any grading point does not see the student's answer, let

$$\mathcal{R}^\star_E = \mathrm{AtomicPoints}\big(\mathrm{ClosestApplicablePath}(E, \mathcal{B}_J)\big)$$

denote the atomic points on the path most applicable to the supplied evidence as a whole. The five structural failure events are then

Extracted evidence unused:

$$\exists e \in E :\ \mathrm{NonEmpty}(e) \wedge \neg\, \mathrm{Uses}(P, e),$$

Answer used by the reason not extracted:

$$\exists i :\ A_i \neq \varnothing \wedge \mathrm{Uses}(P, A_i) \wedge \neg\, \mathrm{Extracted}(A_i, E),$$

Grading point not specifically explained:

$$\exists R_i \in \mathcal{R}^\star(A) :\ \neg\, \mathrm{Explained}(P, R_i),$$

Grading point without evidence or explanation:

$$\exists i :\ A_i \neq \varnothing \wedge \neg\, \mathrm{HasEvidence}(E, R_i) \wedge \neg\, \mathrm{Explained}(P, R_i),$$

Evidence unrelated to any grading point:

$$\exists e \in E :\ \mathrm{NonEmpty}(e) \wedge \forall R_i \in \mathcal{R}^\star_E,\ \neg\, \mathrm{Relevant}(e, R_i).$$

Here $\mathrm{Explained}(P, R_i)$ includes an explicit acknowledgment of an unattempted component. Table 17 operationalizes these logical skeletons as five independent binary questions. They are not mutually exclusive subtypes, and one record may fail more than one. Formally, every failure of the grading-point-without-evidence-or-explanation check is also a failure of the grading-point-not-specifically-explained check; the former is nevertheless judged in a separate request that also sees the evidence, so the two results can differ, and each counts under its own subcategory.

Table 17: Operational boundaries for the coverage checks.

check	failure boundary	adjacent-check exclusion and edge case
Extracted evidence unused	At least one non-empty declared evidence unit in E is not mentioned or used by P in a locatable, meaning-preserving way.	Empty E passes. The check does not inspect A, test evidence fidelity or completeness, map evidence to B, or permit J to replace reason use.
Answer used by the reason not extracted	P uses non-empty student work that is present in A but is neither contained nor located in E.	A statement in P cannot replace E. Evidence equal to `[IMAGE_DATA]` locates the complete image answer but does not establish the fidelity of a claim about its visual content. A reason that addresses only an absent component can pass.
Grading point not specifically explained	P fails to explain the treatment of at least one atomic grading point on the effective path most applicable to A.	A clear acknowledgment covers an absent component; a directed attempt counts as present even if wrong or incomplete. A bare rubric number, score, or generic verdict is not an explanation. J helps define the basis but cannot replace P.
Grading point without evidence or explanation	At least one attempted component has neither corresponding declared evidence nor a reason explanation.	Only attempted, non-empty components can fail; an absent component requires no synthetic evidence and cannot fail this check. The check is narrower than the previous one and may overlap the answer-not-extracted check.
Evidence unrelated to any grading point	At least one non-empty evidence unit is unrelated to every grading point on the path most applicable to E under $\mathcal{B}_J$.	Empty E passes. Incorrect evidence can still be relevant when it attempts the answer slot governed by a grading point. Fidelity belongs to the evidence-misquotes check, use by P to the extracted-evidence-unused check, and full-answer coverage to the other three coverage checks.

Grading argument rating. The grading argument rating rates only a structurally complete atomic chain $\mathcal{C}_i = (R_i, r_i, e_i)$, whose solidity is

$$\mathrm{Solid}(\mathcal{C}_i) = \mathrm{Bridge}(e_i, R_i \mid r_i).$$

The bridge asks whether r_i substantively explains why the work in e_i satisfies, fails, or partially satisfies R_i, and how that treatment affects the score. Operationally, the check first identifies each independent connection that bears on the score and is structurally complete:

$$E \longrightarrow \text{effective-basis grading point} \longrightarrow \text{consequence for } s.$$

Incomplete chains are not silently rated as weak; they are excluded from the strength rating and remain visible through the coverage checks. Each complete connection is *strong* when all three endpoints and their relation are explicit, *moderate* when one limited semantic or terminological gap remains, and *weak* when a key step is missing, relevance is low, or the account substitutes a conclusion for a substantive bridge.

If no complete connection exists, the check returns `not_assessable`, a null grade, no findings, and a rationale. Otherwise it returns `assessable`: grade 2 when every connection is strong; grade 1 when exactly one is moderate and all others are strong; and grade 0 when any connection is weak or at least two are moderate. Findings record every non-strong connection needed to justify the assigned grade. A grade of 0 or 1, or a `not_assessable` result, counts as Score explanation – weak argument. A judgment can supply only the grading-basis endpoint; it cannot supply student work or substitute for E. Incorrect work can nevertheless participate in a strong connection when the record clearly relates the work, the requirement it misses, and the resulting score.

D.4 Assembly and Reporting

Each check is judged separately. A request carries the instructions of one check and the inputs of Table 14 for up to 16 records of one student, or up to four when the records carry images; the judge is instructed to judge every record independently and never to use neighbouring records, and no request sees the definition or the result of another check. A batch whose output is malformed is split into smaller batches, down to single records, and single records are retried. The evaluator then verifies that each record has exactly one result per check with the required identity and form, and it rejects a batch in which at least 80% of the rationales, over five or more records, are identical; the two grading-point checks, whose omission failures can share a rationale, are exempt. If a result is still missing, duplicated, malformed, or mismatched, the evaluation of the whole run is incomplete and cannot enter the reported rates until it is repeated; an incomplete result is never interpreted as a pass or a failure. The `not_applicable` results of the two judgment checks for an empty J are

generated deterministically. Assembly never feeds the decision of one check into another: a content failure remains visible but does not rewrite an explanation-structure result.

A record passes justification quality when none of the failures in Table 13 occurs. The justification quality pass rate is the share of evaluated records that pass, and the rate of each failure is the share of evaluated records in which it occurs; a record can count under several failures at once. In English Reading, failures of the extracted-evidence-unused and grading-point-not-specifically-explained checks are not counted on the 51 selection-type leaf questions, which are answered by choosing among given options; the rates in Section 5 follow this rule. Each check is also reported with its own denominator, together with the decisive quote, findings, or rationale needed to locate the result in a specific record.

D.5 Schematic Boundary Cases

Table 18 contains boundary illustrations only. The rows are not benchmark observations, model outputs, or evidence that the judge is reliable.

Table 18: Schematic boundary cases for justification quality.

schematic condition	**check-specific treatment**
Answer represented by an image crop	The answer-misstated check inspects the crop before deciding whether P or J invents student work. For the answer-not-extracted check, `[IMAGE_DATA]` can locate the answer; this location rule alone says nothing about whether a textual visual claim is faithful.
Empty declared evidence	The evidence-misquotes, extracted-evidence-unused, and unrelated-evidence checks pass because nothing is fabricated, left unused, or irrelevant. The answer-not-extracted and the two grading-point checks can still fail when an attempted component used or required by the account lacks the relevant evidence or explanation.
Unattempted grading component	The grading-point-not-specifically-explained check can pass when P clearly acknowledges the absence, and the grading-point-without-evidence-or-explanation check cannot fail on that component.
Incorrect but slot-directed excerpt	The evidence-misquotes check can pass when the excerpt is faithful, and the unrelated-evidence check can pass because it addresses an applicable point. The grading argument rating may rate a complete explanation of why it is wrong as strong.
One underlying defect crosses boundaries	The checks remain independent. For example, attempted work that the reason uses but the evidence omits fails the answer-not-extracted check, and if its grading point is also left unexplained, the record fails the grading-point-not-specifically-explained check as well; assembly retains both rather than selecting a single diagnosis.

E Cross-Run Consistency: Comparison Rules

This appendix gives the rules behind score consistency, reason consistency, judgment consistency, and their conjunction, cross-run consistency. On the PDF route, each of the five configurations is run $K = 3$ times, as are the GPT configurations on the OCR and JSON contrast routes. Cross-run consistency asks whether the records of the same grading unit agree across the K_u^{eff} runs that actually delivered one; this covers all K runs only when delivery is complete. A grading unit with fewer than two delivered runs is not comparable, and cross-run consistency does not measure agreement with the evaluator-only reference score. Score consistency checks whether the scores are equal; reason consistency whether the reasons make the same claims about the answer and the rubric; and judgment consistency whether the grading judgments reach the same decisions on the same basis, or no run writes one. Cross-run consistency is their conjunction for each grading unit. Reason and judgment consistency apply to the five subject tasks; for the English Writing task, only score consistency is compared, per grading unit (Appendix B.2).

E.1 Expected Grading Units, Alignment, and Delivery

For configuration C and one grading task, the expected set $\mathcal{U}_C$ is fixed before any raw grading submission is inspected. The configuration fixes the evaluated model, harness, input route, instructions, context, orchestration, tools, and any retry rule capable of regenerating a mark. Changing any of these produces a different configuration rather than another run of the same one. Every run has a distinct ordered identifier but is evaluated against the same task manifest.

Alignment uses the exact identity of the expected grading unit, never output position or approximate answer content. Each $u \in \mathcal{U}_C$ is bound to one student's answer and one maximum M_u. Admission yields one admitted record or one typed placeholder for every run and grading unit. The expected set remains $\mathcal{U}_C$, not an observed union, intersection, or delivered-only subset. The cross-run consistency denominator is the comparable subset $\mathcal{U}_C^{\mathrm{cmp}} \subseteq \mathcal{U}_C$ defined below, and both sets are reported with the result.

Let $d_{u,r} = 1$ if run r has an admitted record for u, and $d_{u,r} = 0$ otherwise. Define the number of delivered runs and the comparability indicator

$$K_u^{\mathrm{eff}} = \sum_{r=1}^{K} d_{u,r}, \qquad \eta_u = \mathbf{1}\left[K_u^{\mathrm{eff}} \geq 2\right].$$

Cross-run consistency compares runs rather than scoring a single run, so each grading unit is compared on the runs that actually delivered it. Two delivered runs are the minimum that admits any comparison; a unit below that threshold is not comparable, and no undelivered run is imputed. Table 19 records the downstream treatment.

Table 19: Deterministic alignment and delivery treatment for cross-run consistency.

aligned state	**representation of the grading unit**	**consequence for consistency**
K admitted records	The K raw scores, reasons, and judgment strings remain attached to their ordered run identifiers.	$K_u^{\mathrm{eff}} = K$, $\eta_u = 1$; all three components compare all K runs.
At least one but fewer than $K-1$ missing or rejected records	Each unfilled run contains a typed placeholder, not an imputed score, reason, or judgment; the delivered records keep their ordered run identifiers.	$2 \leq K_u^{\mathrm{eff}} < K$, $\eta_u = 1$; the three components compare the K_u^{eff} delivered runs, and K_u^{eff} is reported with the result.
Fewer than two delivered records	The undelivered runs contain typed placeholders, and the grading unit supports no comparison across runs.	$\eta_u = 0$; c_u^{S}, c_u^{R}, c_u^{J}, and c_u^{all} are undefined rather than zero. The unit leaves the consistency denominators instead of counting as unstable, and no semantic label is fabricated for absent content.
Unknown submitted identity	The record is dropped because it cannot bind to $\mathcal{U}_C$.	It cannot enlarge the denominator; any unfilled expected position receives a placeholder.
Duplicate or invalid identity	No candidate is selected for that grading unit in that run.	The position receives a placeholder and is treated as undelivered.
K genuinely awarded zeros	All are admitted grading decisions rather than placeholders.	They are eligible for $c_u^{\mathrm{S}} = 1$; zero is not evidence of nondelivery.

E.2 Score Consistency: Equality and Range

For $\eta_u = 1$, score consistency requires exact equality of the K_u^{eff} delivered raw scores. The normalized range

$$\Delta_u^{\mathrm{S}} = \max_{r:\, d_{u,r}=1} (s_{u,r}/M_u) - \min_{r:\, d_{u,r}=1} (s_{u,r}/M_u)$$

only describes the size of a disagreement. Neither normalization nor a tolerance can turn unequal raw scores into $c_u^{\mathrm{S}} = 1$.

E.3 Reason Consistency: Semantic Comparison

Reasons are compared in judge requests of their own. A request carries up to eight comparable grading units of one student, and the judge is instructed to judge every unit independently; for each unit it sees the grading-unit identifier, the student's answer, and the K_u^{eff} ordered pairs $\langle$run identifier, reason text$\rangle$. Awarded scores, the maximum score, the evidence and judgment fields, the rubric, the evaluator-only reference score, and the outputs of other checks are withheld, so that agreement or disagreement between scores cannot decide whether the reasons make the same grading claims.

The result contains exactly one label, one claim summary per run in run order, a statement of the decisive common point or difference, and a concise rationale. Summaries may compress the supplied reasons but may not infer a score or an unstated claim.

Table 20: Labels for the semantic comparison of reasons across the K_u^{eff} delivered runs.

label	decision anchor
`same_rationale`	All K_u^{eff} delivered reasons make materially the same claims about what the student did and why the work satisfies, partially satisfies, or fails the relevant requirements. Wording, language, order, formatting, length, and harmless paraphrase do not change the label.
`compatible_shift`	Reasons add, omit, or emphasize different evidence, criteria, or details, but their grading claims can all be true together.
`substantive_change`	At least two reasons make incompatible grading claims, such as satisfied versus not satisfied, valid versus invalid method, or harmless versus score-blocking error.
`unclear`	The supplied reasons are too vague or incomplete to distinguish same, compatible, and incompatible claims without inventing content.

Only `same_rationale` yields $c_u^{\text{R}} = 1$. A compatible shift is still a failure of reason consistency, because a reader would then find two explanations for one score. The results report `compatible_shift` as a compatible basis change and `substantive_change` as conflicting claims. A score phrase inside a reason remains part of the supplied text; the judge does not use it to reconstruct the withheld score.

E.4 Judgment Consistency: Presence and Semantic Comparison

For delivered records, let $p_{u,r} = 1$ exactly when the grading judgment of run r is non-empty after trimming whitespace. Presence is determined before any semantic comparison. If at least two runs write a judgment, the judgments are compared in a separate judge request, batched in the same way as reasons; for each unit, the judge sees the grading-unit identifier, the student's answer, the rubric, and the ordered pairs of run identifier and written judgment only. Reasons, evidence, scores, the evaluator-only reference score, and the outputs of other checks are withheld.

The result contains one label, one decision summary for each written judgment, a statement of the decisive common point or difference, and a concise rationale. The judge never infers a judgment for a run that wrote none.

Table 21: Labels for the semantic comparison of two to K_u^{eff} written judgments.

label	decision anchor
`same_decision_same_basis`	All written judgments make the same accept, reject, allow, or disallow decision and rely on materially the same interpretation, exception, scope, and focal issue.
`same_decision_different_basis`	The final grading treatment is compatible, but the judgments rely on different interpretations, exceptions, scopes, or focal issues.
`decision_changed`	At least two written judgments imply incompatible treatment, such as accepting versus rejecting a method or treating an error as minor versus score-blocking.
`unclear`	The written judgments are too vague, conditional, or incomplete to determine whether decisions and bases match.

Over K_u^{eff} delivered runs there are $2^{K_u^{\text{eff}}}$ binary presence patterns, which reduce to the four classes in Table 22 (three when $K_u^{\text{eff}} = 2$). The table applies only to comparable grading units.

Table 22: Judgment presence, semantic comparison, and c_u^{J}.

runs writing a judgment	semantic comparison	deterministic status	c_u^{J}
0	Not run	`no_judgment`; no delivered run writes a judgment.	1
1	Not run	`single_expressed_not_comparable`; presence is intermittent, and one judgment cannot be compared.	0
2 to $K_u^{\text{eff}} - 1$	Run on the written judgments	The semantic label is kept for diagnosis, but presence remains intermittent regardless of the label.	0
K_u^{eff}	Run on all delivered judgments	Passes only with `same_decision_same_basis`; every other label records a changed decision or basis, or an unclear comparison.	1 only for the passing label

Passing judgment consistency because no run writes a judgment says nothing about the quality of judgments. Judgment consistency is therefore read together with judgment presence: Section 5 reports how often records carry a non-empty grading judgment. At the level of grading units, the share of comparable units in which at least one run writes a judgment,

$$\frac{1}{|\mathcal{U}_C^{\text{cmp}}|} \sum_{u \in \mathcal{U}_C^{\text{cmp}}} \mathbf{1}\left[\sum_{r:\, d_{u,r}=1} p_{u,r} > 0 \right],$$

measures whether a judgment appears at all. Figure 6(b) reports presence patterns on the 1,148-unit discretion subset. Its semantic classes compare only the judgments that were written; unlike strict judgment consistency, the matching class can include units with intermittent judgment presence.

E.5 Assembly and Reporting

Assembly verifies that every expected grading unit is covered exactly once, that run identifiers are unique and ordered, and that each semantic result has the required form. A reason comparison must contain exactly K_u^{eff} claim summaries, and a judgment comparison exactly one summary for each written judgment. Missing, extra, malformed, misordered, or mismatched results make the evaluation incomplete; they are never converted to a passing label or included in a partial aggregate.

For a complete evaluation,

$$c_u^{\mathrm{S}} = \begin{cases} 1, & \eta_u = 1 \text{ and all } s_{u,r} \text{ with } d_{u,r} = 1 \text{ are equal,} \\ 0, & \eta_u = 1 \text{ otherwise,} \\ \text{undefined}, & \eta_u = 0, \end{cases}$$

$$c_u^{\mathrm{R}} = \begin{cases} 1, & \eta_u = 1 \text{ and the label is } \texttt{same_rationale}, \\ 0, & \eta_u = 1 \text{ otherwise,} \\ \text{undefined}, & \eta_u = 0, \end{cases}$$

and

$$c_u^{\mathrm{J}} = \begin{cases} 1, & \eta_u = 1 \text{ and } \sum_{r:\, d_{u,r}=1} p_{u,r} = 0, \\ 1, & \eta_u = 1,\ \sum_{r:\, d_{u,r}=1} p_{u,r} = K_u^{\mathrm{eff}}, \text{ and the label is} \\ & \quad \texttt{same_decision_same_basis}, \\ 0, & \eta_u = 1 \text{ otherwise,} \\ \text{undefined}, & \eta_u = 0. \end{cases}$$

Their conjunction is

$$c_u^{\mathrm{all}} = c_u^{\mathrm{S}} c_u^{\mathrm{R}} c_u^{\mathrm{J}}.$$

All four rates use the same denominator, the comparable subset $\mathcal{U}_C^{\mathrm{cmp}} = \{u \in \mathcal{U}_C : \eta_u = 1\}$:

$$\rho^{\mathrm{X}}(\mathcal{U}_C) = \frac{1}{|\mathcal{U}_C^{\mathrm{cmp}}|} \sum_{u \in \mathcal{U}_C^{\mathrm{cmp}}} c_u^{\mathrm{X}}, \qquad \mathrm{X} \in \{\mathrm{S}, \mathrm{R}, \mathrm{J}, \mathrm{all}\}.$$

The cross-run consistency rate ρ^{all} averages the per-unit conjunction c_u^{all}; it is not the product of the three component rates. The sizes $|\mathcal{U}_C|$ and $|\mathcal{U}_C^{\mathrm{cmp}}|$, the distribution of K_u^{eff}, label distributions, presence patterns, and normalized score ranges are kept as diagnostics. Comparing fewer runs can hide disagreements. Excluding units with fewer than two delivered records also changes the evaluated population, so consistency rates should be interpreted alongside delivery coverage and the distribution of K_u^{eff}.

F Expert Validation of the Judge

This appendix reports how the labels behind justification quality and cross-run consistency were checked against human experts. The judge is GPT-5.6 Sol run at low reasoning effort. Experts relabel the same items the judge labels; they do not re-grade the students' answers.

F.1 Sample and Annotation

All validated items come from the ten GPT/PDF runs, which include the three runs of the main results. The sample is fixed in a frozen sampling list before any expert label is collected.

Justification quality records. Each of the five subject tasks contributes 20 records. Seven form a quota panel that over-draws records for which the judge reported a rare but consequential problem; the other 13 are a simple random sample (anchors) from the records outside the quota panel. The writing task contributes 15 records, five per dimension, taken from 15 different compositions (12 answer scripts) with the same two panels. Each record carries every applicable check of Appendix D, for 1,295 item labels.

Cross-run consistency units. For reason consistency, 20 grading units per subject are drawn at random from the subject's sampling frame (Table 24). For judgment consistency, a grading unit enters the population only if at least two of its delivered runs wrote a non-empty grading judgment, since fewer leave nothing to compare. Units are drawn at random within two strata per subject: *full*, where every delivered run wrote one, and *partial*, where at least two but not all did; 60 units in total. The writing task has no reason or judgment consistency labels to validate (Appendix E).

Annotation. Three experts label every item independently, with the same decision rules as the judge. For a justification quality record they see the question, the rubric, the student's answer with any images, and the record; for a cross-run consistency unit they see the reasons or grading judgments of all runs. The judge's labels are hidden. The majority label of the three experts is the reference. Three items of the grading argument rating received three different labels and are left out of comparisons with the reference.

F.2 Statistics

Agreement in the sample. We report Cohen's κ between experts (mean of the three pairs) and between the judge and the reference. To compare the judge with an expert on equal terms, we hold out one expert at a time, keep the items on which the other two agree, and compare the held-out expert and the judge with that label; the difference is averaged over the three held-out experts.

Alternative annotator test. Following Calderon et al. (2025), for each held-out expert j and item i we score how well the judge and expert j match the remaining two experts, and test whether the judge's advantage probability falls short of the expert's by more than ε (one-sided paired t-test, Benjamini–Yekutieli correction over the three experts). The judge passes when it beats at least half of the experts. We report the smallest ε on a grid of 0.01 at which it passes; the recommended value for expert annotators is $\varepsilon = 0.2$.

Design weighting. Sample agreement over-represents the quota panel. To estimate rates over all GPT/PDF records in the sampling frame, each quota record represents only itself (weight 1), each anchor represents N'/n' records, where N' is the number of records outside the quota panel and n' the number of anchors in its stratum, and each cross-run consistency unit represents N/n units of its stratum. Confidence intervals are 95% percentile intervals from 2,000 bootstrap resamples of records or units, drawn within each stratum and panel; strata sampled in full are held fixed.

F.3 Results

Table 23 gives agreement in the sample. The judge passes the alternative annotator test for every label in the table at the recommended $\varepsilon = 0.2$, and for all but Evidence extraction and justification quality as a whole at $\varepsilon = 0.11$ or below. The weakest single check is the grading argument rating, on which the experts also agree only moderately: on its 0–2 grade, over the 113 items that all four raters found assessable, the quadratically weighted κ is 0.48 between experts and 0.28 between the judge and the reference, and the judge passes the alternative annotator test from $\varepsilon = 0.12$. The full four-way judgment consistency label passes from $\varepsilon = 0.11$.

Table 23: Judge–expert agreement in the validation sample (Cohen's κ). The first four rows group the failure subcategories of Section 4; a record counts under a row if any of them occurs. The first two numeric columns give mean pairwise expert–expert agreement and judge–reference agreement, respectively. "Held-out" columns compare a held-out expert, or the judge, with the two remaining experts on the items where they agree; the gap is judge minus expert with a 95% bootstrap interval, computed before rounding. $\varepsilon_{\min}$ is the smallest margin at which the judge passes the alternative annotator test.

	sample agreement		**held-out comparison**			
label	**expert pairs**	**judge vs. reference**	**expert**	**judge**	**gap**	$\varepsilon_{\min}$
Misstated answer or unsupported rule	0.85	0.88	0.92	0.92	$0.00\ [-0.09, 0.08]$	0.04
Evidence extraction	0.81	0.58	0.89	0.64	$-0.25\ [-0.41, -0.12]$	0.19
Ungrounded score or ignored answer part	0.69	0.67	0.81	0.78	$-0.03\ [-0.18, 0.10]$	0.06
Weak argument or contradiction	0.52	0.52	0.66	0.56	$-0.09\ [-0.32, 0.13]$	0.10
Justification quality fail	0.84	0.72	0.91	0.76	$-0.15\ [-0.27, -0.03]$	0.14
Reason consistency fail	0.77	0.82	0.88	0.88	$0.00\ [-0.10, 0.10]$	0.07
Judgment consistency fail	0.93	0.83	0.97	0.87	$-0.09\ [-0.22, 0.01]$	0.11

Table 24 gives design-weighted rates over the five subject tasks. As in the reported rates, failures of the extracted-evidence-unused and grading-point-not-specifically-explained checks are not counted on the 51 English selection-type questions (Appendix D); both the judge's and the experts' labels follow this rule. For every label except Evidence extraction, the judge's rate is within 2.5 points of the experts'; on Evidence extraction it is 6.4 points lower, but the interval includes zero and also permits substantial under-detection. The judge fails slightly more of these GPT/PDF records than the experts do (45.4% against 44.0%). This comparison does not establish whether the judge favors its own model family relative to other agents' records.

Table 24: Design-weighted rates over the GPT/PDF records of the five subject tasks (sampling frame: the 3,155 grading units with non-empty answers that at least two GPT/PDF runs delivered, excluding three chemistry answers that were recorded as empty when the sample was drawn; judgment consistency: 525 of these units in which at least two runs wrote a grading judgment). Rows as in Table 23. Rates are the share of records or units with the problem, or failing the consistency component. Sensitivity and specificity treat the expert reference as truth. Differences are computed before rounding.

label	judge	experts	difference (pp)	sens.	spec.
Misstated answer or unsupported rule	8.7%	7.8%	+0.9 [0.1, 2.7]	100.0%	99.0%
Evidence extraction	29.1%	35.4%	−6.4 [−16.8, 2.5]	68.6%	92.7%
Ungrounded score or ignored answer part	19.1%	16.6%	+2.5 [−8.4, 13.2]	71.2%	91.3%
Weak argument or contradiction	13.8%	12.1%	+1.7 [−13.7, 16.3]	39.4%	89.7%
Justification quality fail	45.4%	44.0%	+1.4 [−13.1, 16.7]	78.7%	80.7%
Reason consistency fail	22.1%	21.0%	+1.1 [−3.4, 7.1]	88.7%	95.6%
Judgment consistency fail	57.3%	58.7%	−1.4 [−7.5, 4.3]	93.4%	94.1%

Two subsets need separate reading. English Reading holds about half of the weighted records; there the judge's justification quality failure rate matches the experts' (46.1% against 46.2%), although its record-level agreement with the reference is the lowest of the five subjects ($\kappa = 0.38$). On the writing task, the reference finds a justification quality problem in all 15 records and the judge in 12, so these records carry no information about how the judge separates passing from failing records.

Judge replacement. To test whether the results depend on the judge model, we re-judge the run-1 records of six students (277 records per configuration, the same grading units for all five) with Claude Opus 5, keeping the checks, prompts, and pass rule unchanged. The ranking of the five configurations on justification quality agrees with that under GPT-5.6 Sol (Kendall $\tau = 0.8$, against 0.6 between two runs of GPT-5.6 Sol on the same records): Qwen, Grok, and MiniMax keep first, second, and last place, and no pass rate moves by more than 5.9 points. On GPT-generated records, GPT-5.6 Sol's 5.9-point margin over Claude Opus 5 shrinks to 1.3 points when GPT-5.6 Sol is run again. These limited comparisons provide no clear evidence of self-preference, but do not rule it out.

Limitations. The validation covers one configuration, GPT/PDF, and the writing task only for justification quality. Record-level agreement is low for weak argument or contradiction: the design-weighted κ between the judge and the reference is 0.28, and the judge flags 39% of the records the experts flag, although its overall rate is close to theirs (13.8% against 12.1%). Bootstrap intervals are wide for rare checks: the self-contradictory-reason, reason-contradicts-the-score, and evidence-unrelated-to-any-grading-point checks each have at most four failing labels from any rater and are read only through the rows they are grouped into. The descriptive judgment–rubric relation check passes the alternative annotator test from $\varepsilon = 0.08$, but on only 15 items.

G Supplementary Experimental Results

G.1 Joint Pass Rate by Question Type

Figure 7 gives the subject-specific rates underlying the question-type comparison in Section 5.3. The main-text figure shows how the six question types are distributed within each subject.

Subject	short calculation	multi-step calculation	short answer	multi-step explanation	drawing	five types
Mathematics	38.9	15.7	61.5			33.7
Physics	47.7	17.5	42.6	16.5	23.8	40.0
Chemistry	40.0	16.8	45.8	15.1	43.7	39.7
Biology			33.9	9.1	7.4	27.8
English Reading			34.0			34.0
All five subjects	42.7	16.1	37.1	12.3	32.1	35.2

Figure 7: Joint pass rate (%) by subject and question type. The color strips match the question-type bars in Figure 5(a). Multiple choice (English Reading only) is excluded from the "five types" column; English Writing is not shown. One deleted English Reading question (20 grading units) is omitted. Empty cells indicate that the subject has no questions of that type.

G.2 Run Robustness and Sampling Uncertainty

Number of runs. GPT/PDF was run ten times. For each K, we average over all $\binom{10}{K}$ subsets of these runs the share of grading units that are scored correctly, or scored correctly and pass justification quality, in every run of the subset (Table 25; averaged over the five subject tasks). Reason and judgment consistency are labeled once over all ten runs and cannot be recomputed for a subset, so the second row is an upper bound on the joint pass rate, not the rate itself.

Table 25: Share of grading units passing in every one of K runs (GPT/PDF, mean over all subsets of the ten runs, %).

K	1	2	3	5	10
Scored correctly	86.8	81.7	78.2	73.0	63.2
Scored correctly and passes justification quality	54.0	41.2	34.2	26.0	15.7

Sampling uncertainty. We resample students with replacement within each of the six tasks (2,000 replicates), holding the writing level conversion fixed, and recompute each model's joint pass rate averaged over the six tasks, as in Table 3; a pairwise difference uses the same resampled students for both models. The 95% percentile intervals are GPT [25.8, 30.0], Claude [29.3, 33.0], MiniMax [21.3, 25.8], Grok [32.4, 37.1], and Qwen [39.1, 42.9]. Every pairwise difference excludes zero. With 10–20 students per task and one examination paper per subject, these intervals do not reflect variation across papers and are likely too narrow.

G.3 Failures Conditional on Correct Scores

Tables 26 and 27 report failures conditional on correct scores for the five grading agents on the PDF route with $K = 3$. English Writing is excluded because it lacks unit-level reference scores. Each percentage is computed within a subject; the macro average gives the five subjects equal weight. These analyses retain non-empty answers to the deleted English Reading question, including its 20 units in Claude's all-runs-correct denominator. This scope differs from the task-characteristic analyses, which exclude that question.

Table 26 conditions on an admitted record whose awarded score matches the reference, with records from the three runs pooled within each subject. Table 27 conditions on a grading unit delivered and scored correctly in all three runs, with cross-run consistency evaluable. Justification quality failure means that at least one run fails the checks in Section 4; an unassessable grading argument also counts as failure. The adopted exemptions for English Reading selection questions apply (Appendix D). A unit scored correctly in every run necessarily has score consistency, so its cross-run consistency failures arise from reason or judgment consistency. These can overlap. The quality-or-consistency row counts their union at the unit level, rather than adding failure rates.

Table 26: Justification quality failures among correctly scored records (PDF route, $K = 3$). Each cell gives the percentage and failure count / correct-record count.

Subject	GPT	Claude	MiniMax	Grok	Qwen
Mathematics	32.5 (295/908)	32.5 (289/889)	53.8 (463/860)	34.9 (319/914)	26.6 (245/920)
Physics	32.2 (383/1,189)	22.9 (270/1,180)	38.9 (435/1,119)	27.7 (328/1,185)	17.5 (211/1,209)
Chemistry	34.6 (470/1,357)	22.4 (307/1,369)	35.0 (448/1,281)	26.8 (362/1,351)	21.0 (291/1,385)
Biology	41.7 (351/842)	36.9 (354/960)	46.4 (390/840)	35.1 (329/938)	26.8 (249/928)
English Reading	43.4 (1,877/4,328)	32.9 (1,446/4,401)	14.9 (630/4,232)	21.7 (941/4,339)	10.9 (469/4,307)
Macro average	36.9	29.5	37.8	29.2	20.6

GPT, Claude, MiniMax, Grok, and Qwen denote GPT-5.6 Sol, Claude Opus 5, MiniMax M3, Grok 4.6, and Qwen3.8-Max, respectively.

Table 27: Failures among units scored correctly in all three runs with cross-run consistency evaluable. Each cell gives the percentage and failure count / conditional unit count; macro averages give subjects equal weight. The deleted English Reading question is retained.

Subject	GPT	Claude	MiniMax	Grok	Qwen
Justification quality failure in any run					
Mathematics	51.7 (151/292)	52.7 (154/292)	83.8 (212/253)	53.5 (159/297)	43.1 (128/297)
Physics	52.1 (195/374)	37.1 (141/380)	65.3 (215/329)	41.9 (155/370)	29.6 (115/389)
Chemistry	55.7 (238/427)	36.0 (160/444)	57.9 (224/387)	41.0 (174/424)	36.7 (164/447)
Biology	54.7 (122/223)	54.1 (164/303)	63.1 (135/214)	50.0 (141/282)	45.7 (133/291)
English Reading	55.6 (790/1,422)	57.9 (844/1,458)	28.2 (372/1,320)	40.4 (578/1,431)	19.2 (271/1,415)
Macro average	54.0	47.6	59.7	45.4	34.8
Cross-run consistency failure					
Mathematics	13.7 (40/292)	22.6 (66/292)	18.6 (47/253)	18.5 (55/297)	22.6 (67/297)
Physics	18.7 (70/374)	25.8 (98/380)	21.6 (71/329)	20.8 (77/370)	14.1 (55/389)
Chemistry	24.1 (103/427)	22.1 (98/444)	22.7 (88/387)	19.1 (81/424)	25.1 (112/447)
Biology	23.3 (52/223)	27.4 (83/303)	18.7 (40/214)	17.7 (50/282)	30.9 (90/291)
English Reading	13.7 (195/1,422)	28.9 (422/1,458)	15.3 (202/1,320)	12.2 (174/1,431)	15.3 (217/1,415)
Macro average	18.7	25.4	19.4	17.7	21.6
Reason consistency failure					
Mathematics	7.2 (21/292)	8.9 (26/292)	16.2 (41/253)	12.1 (36/297)	14.5 (43/297)
Physics	11.0 (41/374)	10.3 (39/380)	17.6 (58/329)	12.7 (47/370)	3.1 (12/389)
Chemistry	13.3 (57/427)	7.7 (34/444)	18.1 (70/387)	12.3 (52/424)	5.8 (26/447)
Biology	15.2 (34/223)	6.3 (19/303)	15.4 (33/214)	3.9 (11/282)	3.1 (9/291)
English Reading	10.7 (152/1,422)	25.2 (367/1,458)	14.0 (185/1,320)	9.2 (132/1,431)	9.8 (138/1,415)
Macro average	11.5	11.7	16.3	10.0	7.2
Judgment consistency failure					
Mathematics	7.5 (22/292)	17.1 (50/292)	4.3 (11/253)	8.4 (25/297)	11.4 (34/297)
Physics	10.2 (38/374)	18.7 (71/380)	6.1 (20/329)	10.3 (38/370)	11.3 (44/389)
Chemistry	12.6 (54/427)	17.1 (76/444)	9.0 (35/387)	7.1 (30/424)	20.8 (93/447)
Biology	10.3 (23/223)	21.8 (66/303)	4.2 (9/214)	14.5 (41/282)	28.5 (83/291)
English Reading	3.6 (51/1,422)	5.7 (83/1,458)	1.5 (20/1,320)	3.9 (56/1,431)	6.8 (96/1,415)
Macro average	8.8	16.1	5.0	8.8	15.8
Justification quality or cross-run consistency failure					
Mathematics	54.1 (158/292)	61.6 (180/292)	84.2 (213/253)	55.6 (165/297)	49.5 (147/297)
Physics	57.0 (213/374)	47.9 (182/380)	69.9 (230/329)	47.8 (177/370)	35.0 (136/389)
Chemistry	64.4 (275/427)	46.6 (207/444)	63.0 (244/387)	48.3 (205/424)	48.5 (217/447)
Biology	60.5 (135/223)	64.4 (195/303)	64.5 (138/214)	57.8 (163/282)	58.1 (169/291)
English Reading	59.3 (843/1,422)	64.7 (944/1,458)	30.5 (403/1,320)	44.6 (638/1,431)	28.8 (408/1,415)
Macro average	59.1	57.1	62.4	50.8	44.0

GPT, Claude, MiniMax, Grok, and Qwen denote GPT-5.6 Sol, Claude Opus 5, MiniMax M3, Grok 4.6, and Qwen3.8-Max, respectively.

G.4 JUSTIFICATION QUALITY FAILURES IN ENGLISH WRITING

Figure 8 reports the three failure categories separately for Content, Language, and Organisation on English Writing, using the five grading agents on the PDF route with $K = 3$. Each model contributes 120 records per dimension (40 compositions across three runs).

Evidence extraction failures occur in 75.8%–98.3% of records across the model–dimension combinations, while Score explanation failures occur in 52.5%–97.5%. For every model, Material reading failures are more frequent in Language than in Content or Organisation; the Language rates range from 33.3% to 79.2%. These results show that the low justification quality pass rates in English Writing accompany frequent evidence extraction and score explanation failures across all three assessment dimensions.

Category	**Dimension**	GPT	Claude	MiniMax	Grok	Qwen
Material reading	Content	7.5	21.7	21.7	11.7	13.3
	Language	33.3	58.3	79.2	51.7	50.0
	Organisation	2.5	38.3	35.0	15.0	19.2
Evidence extraction	Content	93.3	96.7	97.5	97.5	88.3
	Language	81.7	98.3	89.2	98.3	75.8
	Organisation	93.3	97.5	78.3	91.7	90.8
Score explanation	Content	80.8	66.7	90.8	75.8	70.8
	Language	82.5	52.5	97.5	83.3	66.7
	Organisation	63.3	61.7	95.8	90.8	75.8

Figure 8: Failure categories on English Writing (%, PDF route, $K = 3$). Each rate uses the 120 records of one dimension (40 compositions $\times$ 3 runs) as its denominator; the three categories can overlap and are not additive. Shading runs on a 0–100% scale in a different hue from Figure 3, whose scale is 0–30%. GPT, Claude, MiniMax, Grok, and Qwen denote GPT-5.6 Sol, Claude Opus 5, MiniMax M3, Grok 4.6, and Qwen3.8-Max, respectively.

G.5 FULL RANKING OF RECORD-LEVEL FAILURES

Table 28 gives all nine failure subcategories underlying Figure 3, for the five grading agents on the PDF route with $K = 3$. English Writing is reported separately in Appendix G.4. Non-empty answers to the deleted English Reading question are excluded. For each model, records from the three runs are pooled within each subject, and the five subject rates are averaged equally. The final column averages the five model rates and determines the ranking.

Rates use records evaluable for justification quality, except that Score explanation – weak argument is conditional on an assessable grading argument: ratings of 0 or 1 are counted among records rated 0, 1, or 2. Unassessable arguments are excluded from this ranking entry but count as failures in the overall justification quality pass rate. Material reading – misstated answer and Score explanation – contradiction each count the union of their constituent checks. Failure subcategories can overlap, so their rates must not be added.

G.6 SCORE MISMATCH ON THE VISUAL-CONTENT SUBSET

Table 29 compares GPT-5.6 Sol across the PDF, OCR, and JSON routes with $K = 3$. The analysis uses the 3,197 grading units shared by the three routes across the five subject tasks, after excluding non-empty answers to the deleted English Reading question. The visual-content subset contains units with E2$\geq$ 1 or a drawing question type, using the pre-run annotations attached to the PDF units (Appendix C.1). E2$\geq$ 1 means that grading requires reading and combining at least four elements from figures or data tables. There are 199 such units and 197 drawing units, with an overlap of 102, giving 294 units in the subset. The remaining units are labelled "Other units"; they can still involve non-text material below the E2 threshold.

Rates pool records across the three runs and all five subjects. The numerator counts records whose score differs from the reference score; the denominator counts records with comparable scores. Each route has all 882 expected records for the visual-content subset, and all of these scores are

Table 28: All nine failure subcategories ranked by the five-model mean (%). Each model column is an equal-weight average of the five subject rates. The weak argument row is conditional on an assessable argument, as described above.

Rank	Failure subcategory	GPT	Claude	MiniMax	Grok	Qwen	Mean
1	Evidence extraction – unused evidence	16.87	13.19	17.15	17.15	6.74	14.22
2	Material reading – misstated answer	9.95	9.32	18.19	9.77	8.88	11.22
3	Score explanation – weak argument	10.45	6.36	9.85	3.33	3.85	6.77
4	Score explanation – ungrounded score	10.25	5.63	9.68	3.26	1.42	6.04
5	Evidence extraction – missing evidence	9.29	1.68	7.43	3.16	3.57	5.03
6	Score explanation – unsupported rule	1.97	1.93	6.15	3.25	2.43	3.15
7	Evidence extraction – irrelevant evidence	2.67	3.20	3.35	3.50	2.34	3.01
8	Score explanation – ignored answer part	1.53	0.63	2.22	0.83	0.71	1.18
9	Score explanation – contradiction	0.12	0.32	0.99	0.21	0.26	0.38

GPT, Claude, MiniMax, Grok, and Qwen denote GPT-5.6 Sol, Claude Opus 5, MiniMax M3, Grok 4.6, and Qwen3.8-Max, respectively.

comparable. Across all units, PDF and OCR each have 177 records with non-comparable scores, which are excluded from their denominators; JSON has none. No expected records are missing on any route.

Table 29: Score mismatch by input route and subset for GPT-5.6 Sol ($K = 3$). Cells show percentages with mismatched/comparable record counts in parentheses. Drawing units are included in the visual-content subset.

Subset	Units	PDF	OCR	JSON
All units	3,197	8.4 (790/9,414)	7.5 (704/9,414)	6.6 (630/9,591)
Visual-content subset	294	15.1 (133/882)	15.1 (133/882)	15.4 (136/882)
Drawing units	197	12.4 (73/591)	11.7 (69/591)	13.4 (79/591)
Other units	2,903	7.7 (657/8,532)	6.7 (571/8,532)	5.7 (494/8,709)

G.7 Model and Harness Details

The model identifiers are `gpt-5.6-sol`, `claude-opus-5`, `minimax-m3`, `grok-4.6`, and `Qwen3.8-Max`. GPT uses Codex; Claude and MiniMax use Claude Code 2.1.260; Grok uses Grok Build 1.0.40; and Qwen uses Qoder CN 1.1.59. The result configurations do not record the Codex version. All five models are evaluated on the PDF route. GPT is additionally evaluated on the OCR and JSON routes described in Section 5.2.

G.8 Short-Answer Results by Subject

Table 30: Short-answer results by subject, averaged equally over the five grading agents on the PDF route ($K = 3$, %).

Subject	Units per model	All scores correct	Justification quality passes in all runs	Cross-run consistency	Joint pass rate
Physics	155	86.1	50.8	73.8	42.6
Chemistry	288	89.7	55.5	76.3	45.8
English Reading	868	85.5	41.4	69.7	34.0

Short answers include non-selection closed-format responses such as fill-in and table completion. Non-empty answers to the deleted English Reading question are excluded. All rates use grading units as the denominator. The score and justification quality columns require every run to pass; the latter is not a per-record pass rate. Each model has the same units within a subject, so equal averaging over models is equivalent to pooling 775, 1,440, and 4,340 model–unit pairs for Physics, Chemistry, and English Reading, respectively.

H From Grading Challenges to Checks

Table 31 maps each grading challenge of Section 3.2 to the failures it produces in the records and to the checks of Section 4 that detect them.

Table 31: Where each grading challenge of Section 3.2 shows in the records, and which check detects it. Each challenge corresponds to one justification quality category; admission and cross-run consistency add checks for the first and third challenges.

Failure visible in the records	Detected by
Reading raw documents → Material reading	
A grading unit is missing, duplicated, or assigned an invalid or unexpected identifier	Admission
The answer is misread, or restated as the student probably meant it, and the evidence, reason, or score follows	Material reading – misstated answer; score accuracy
Connecting answers to grading points → Evidence extraction	
The reason leaves extracted evidence unused or relies on answer content that the evidence does not contain, or the evidence bears on no grading point	Evidence extraction – unused, missing, or irrelevant evidence
Resolving what the rubric leaves open → Score explanation	
A score rests on neither the rubric nor a declared grading judgment	Score explanation – ungrounded score or unsupported rule
Where the rubric does not settle the score, the reason argues weakly, passes over part of the answer, or contradicts itself	Score explanation – weak argument, ignored answer part, or contradiction
The same answer is handled differently across runs	Score, reason, and judgment consistency

I Extended Related Work

From text answers to examination documents. Automated grading has moved from short answers and essays submitted as text (Mohler et al., 2011; Dzikovska et al., 2013; Meyer et al., 2024; Yancey et al., 2023) to real-world handwritten answers, as in CHECK-MAT, EDU-CIRCUIT-HW, and the GCSE benchmark (Khrulev, 2025; Sun et al., 2026; Fox et al., 2026), and to multiple subjects (Zhao et al., 2026; Lai et al., 2026; Gombert et al., 2026). Beyond the studies in Table 1, Pathak et al. (2026) grade 520 whole scanned physics submissions; like Perš et al. (2026), they evaluate their own pipeline and judge it by its scores. Reading such documents is itself hard for multimodal models, from general text recognition (Liu et al., 2024) to handwritten mathematics (Nath et al., 2025); `ReMarkBench` makes this reading part of grading rather than a separate recognition task.

Evaluating scores and justifications. For scores, grading studies report QWK, correlation, or error metrics against human scores (Yancey et al., 2023; Seuti & Ray Choudhury, 2026; Fox et al., 2026; Meyer et al., 2024), which cannot show whether a correct score rests on a correct basis. Justification quality checks whether each record supports its score, not whether the record reflects the agent's internal reasoning (Jacovi & Goldberg, 2020; Turpin et al., 2023; Lanham et al., 2023).

Repeated grading and LLM judges. Human marking standards require that the same answer be marked the same way (Meadows & Billington, 2005; Ofqual, 2014); cross-run consistency asks the same of repeated runs (Section 4). Justification quality and cross-run consistency both rely on LLM-as-a-judge evaluation, and such judges are known to favor particular answer positions, longer outputs, and their own generations (Zheng et al., 2023; Wang et al., 2024; Panickssery et al., 2024); we therefore validate the judge against human experts (Section 5.1) instead of assuming that its labels are correct (Krumdick et al., 2025; Calderon et al., 2025).

J Limitations

This study has limitations. It does not include human graders as a performance baseline. It uses one examination and the answer scripts published to illustrate its performance levels. Experts validated the judge only on GPT/PDF runs. The official writing levels are public. Each model was run with one

harness, so model and harness effects cannot be fully separated. Only GPT was run on the diagnostic routes, and the main results use three runs per task.