ExamOCRBench: A Multi-Task Benchmark for Real-World Exam Document Understanding and Automated Grading

Anonymous authors
Paper under double-blind review

Abstract

Existing evaluations of vision-language models cover text recognition, document parsing, and visual question answering, yet end-to-end grading of student-completed examination pages remains underexplored. This setting requires faithful interpretation of handwritten answers and corrections in the context of questions and page layouts to support grading. We introduce ExamOCRBench, built from photographed primary-school examination pages in Chinese, Mathematics, and English, with linked structural, transcription, and human-verified grading annotations. It defines four tasks: exam paper grading, structure parsing, content grounding, and handwritten text recognition. End-to-end grading forms the core evaluation, while the remaining tasks independently diagnose related capabilities. Across five evaluated models, complete-page grading accuracy ranges from 33.22% to only 58.18%, highlighting the difficulty of consistently assessing all student responses on a page. Structure recovery, content grounding, and handwriting recognition also remain challenging. Qualitative analysis also reveals that misreading an incorrect student response as the expected answer can lead to undeserved credit. ExamOCRBench provides a unified framework for evaluating real-world examination grading and identifying related capability gaps.

1 Introduction

A student may overwrite a word, place an answer between two lines, or cross out one option before writing another. On a completed examination page, these marks determine the response to be graded. A model that silently corrects a misspelling can assign credit to an answer the student did not write. A response also cannot be graded reliably when essential question material is missing, such as the media materials required to answer an English listening-comprehension question. These situations make examination-page assessment a useful setting for studying how visual interpretation affects a consequential document-understanding output.

OCRBench (Liu et al., 2024), OCRBench v2 (Fu et al., 2025), OmniDocBench (Ouyang et al., 2025), and CC-OCR V2 (Peng et al., 2026) evaluate text recognition, layout understanding, structured extraction, or visual question answering. Short-answer grading (Mohler et al., 2011) and the SemEval student-response analysis task (Dzikovska et al., 2013) evaluate the correctness of student answers, often starting from text representations. ExamOCRBench brings these perspectives together at the level of a photographed, student-completed page. Its central object is the association between the question, the final visible response, and the grading decision.

We evaluate this endpoint with Exam Paper Grading. Three independently evaluated tasks provide complementary evidence: Exam Paper Structure Parsing recovers the question and response-slot hierarchy; Content Grounding localizes a queried region; and Handwritten Text Recognition transcribes student answers. Their outputs are not supplied to the grading model. Together, these diagnostics reveal capability differences that page-level grading alone would conceal.

The current evaluation shows substantial room for improvement. The highest mean structure-parsing TEDS is 59.58%, and the highest handwritten recognition EM is 33.15%. These observations motivate analysis at the task and content levels. They do not establish that any single diagnostic weakness causes all grading errors.

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.

Our contributions are threefold:

1. An end-to-end multimodal evaluation benchmark centered on grading real, student-completed primary-school examination pages, with linked structure, transcription, and human-verified grading annotations.
2. Four task protocols that connect page-level grading with structure, grounding, and handwriting diagnostics while specifying each task's inputs and scoring units.
3. An evaluation of five multimodal models and an analysis of their grading, parsing, grounding, and handwriting-recognition results.

Figure 1 provides an overview of the benchmark: dataset scale and question-type composition, model performance profiles, and representative examples of the four tasks. The examples connect the visual evidence on a completed examination page to the different outputs evaluated by the benchmark.

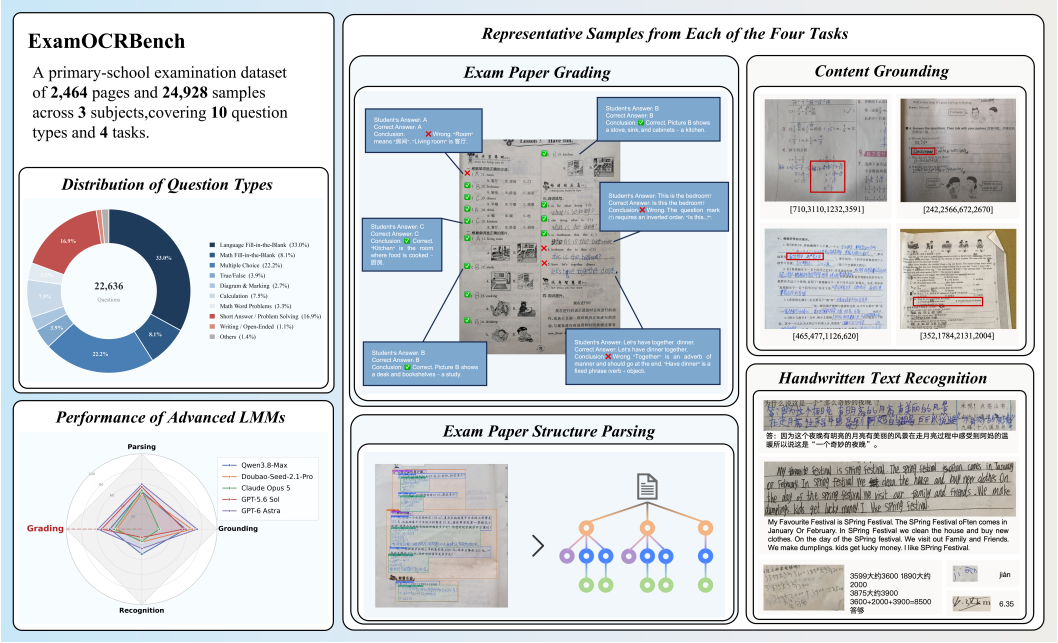


Figure 1: Overview of ExamOCRBench. Left: dataset scale, the distribution of 22,636 grading-task questions across ten refined categories, and a radar visualization of model performance. Right: representative grading, grounding, structure-parsing, and handwriting-recognition examples. The 24,928 task instances comprise pages, queries, and crops; the donut chart counts questions. Grading marks, boxes, and transcriptions illustrate outputs or annotations.

2 Related Work

2.1 OCR and document understanding

OCRBench (Liu et al., 2024) and OCRBench v2 (Fu et al., 2025) organize text-centric multimodal evaluation across recognition, localization, extraction, parsing, and reasoning tasks. CC-OCR (Yang et al., 2024) and CC-OCR V2 (Peng et al., 2026) further emphasize challenging document scenarios and task-specific output formats. OmniDocBench (Ouyang et al., 2025) provides region, content, attribute, and reading-order annotations, and separates output extraction, matching, and metric computation. These works inform our task decomposition and the separation of geometry from content evaluation.

Document models also differ in how they obtain textual evidence. LayoutLMv3 (Huang et al., 2022) learns aligned text, image, and layout representations. Donut (Kim et al., 2022) generates structured document outputs without an external OCR stage, while Nougat (Blecher et al., 2023) converts

academic document images into markup. TrOCR (Li et al., 2021) uses pretrained image and text transformers for recognition, including handwritten text. ExamOCRBench evaluates the resulting capabilities through common task interfaces; these methods are related approaches, not additional baselines evaluated in this release.

2.2 Visual evidence and question answering

DocVQA (Mathew et al., 2021) and InfographicVQA (Mathew et al., 2022) require answers grounded in document content and layout. ST-VQA (Biten et al., 2019) emphasizes reading scene text, whereas ChartQA (Masry et al., 2022) and CharXiv (Wang et al., 2024) examine chart elements and visual reasoning. MMLONGBENCH-DOC (Ma et al., 2024) extends document question answering to long multimodal contexts. AI2D (Kembhavi et al., 2016) and BabyVision (Chen et al., 2026) provide complementary perspectives on diagram understanding and fine visual distinctions.

These benchmarks demonstrate why accurate answers can require spatial and visual evidence beyond a text string. Our grounding task makes the target region an explicit output, distinguishing handwriting, corrections, question stems, and section material. A document-QA task may require layout understanding without directly evaluating a predicted layout or bounding box.

2.3 Student response assessment and benchmark positioning

Automatic short-answer grading (Mohler et al., 2011) has used semantic similarity and dependency-graph alignment between student and instructor answers. The SemEval student-response analysis task (Dzikovska et al., 2013) studies correctness and entailment-related categories in educational answers. Those formulations clarify the role of reference answers, but do not require recovering the student's response from a complete photographed sheet. More recent work evaluates multimodal models on handwritten mathematics exams (Caraeni et al., 2024) and end-to-end engineering-exam grading (Perš et al., 2026). ExamOCRBench focuses on real primary-school examination pages across subjects and combines page-level grading with independent structure, grounding, and recognition diagnostics.

ExamOCRBench focuses on that visual recovery and its relation to grading. The frozen annotations cover several response types, including open responses; the reported grading endpoint is agreement with human-verified binary credit rather than holistic essay-quality assessment. Human verification is applied to model-assisted grading candidates, and the measured endpoint is agreement with those verified labels.

Here, Grading denotes assessment of student responses; Parsing denotes structural parsing; Grounding denotes query-driven region localization; and Recognition denotes handwritten-text recognition.

Table 1 compares the directly evaluated tasks of the reference benchmarks. A check denotes a directly evaluated task. Size preserves each benchmark's sample count.

3 ExamOCRBench

3.1 Benchmark overview and task design

The benchmark centers on grading real, student-completed examination pages. As illustrated in Figure 1, grading combines understanding the question, identifying the student's final response, and assigning credit from the available evidence. Three independently evaluated tasks probe related capabilities: structure parsing recovers how questions and response positions are organized, grounding locates a requested region, and recognition transcribes the student's handwriting. Their outputs are not supplied to the grading model.

Table 2 describes the tasks and their evaluation units. The 2,464 source pages yield 24,928 task instances: 2,464 grading pages, 2,464 parsing pages, 10,000 grounding queries, and 10,000 handwriting crops. Source pages are shared across tasks, and a page can produce multiple queries or crops; the instance total therefore does not count unique documents.

Table 1: Directly evaluated task coverage. Size retains each benchmark's own counting unit; ExamOCRBench sums pages, queries, and crops, not unique pages.

Benchmark	Grading	Recognition	Parsing	Grounding	Size
AI2D (Kembhavi et al., 2016)	—	—	✓	—	15,000
DocVQA (Mathew et al., 2021)	—	—	—	—	50,000
InfographicVQA (Mathew et al., 2022)	—	—	—	—	30,035
CharXiv (Wang et al., 2024)	—	—	—	—	11,615
MMLONGBENCH-DOC (Ma et al., 2024)	—	—	—	—	1,082
OCRBench (Liu et al., 2024)	—	✓	—	—	1,000
OCRBench v2 (Fu et al., 2025)	—	✓	✓	✓	10,000
OmniDocBench (Ouyang et al., 2025)	—	✓	✓	—	981
CC-OCR (Yang et al., 2024)	—	✓	✓	—	7,058
CC-OCR V2 (Peng et al., 2026)	—	✓	✓	✓	7,093
BabyVision (Chen et al., 2026)	—	—	—	—	388
ExamOCRBench	✓	✓	✓	✓	24,928

Table 2: Task statistics and descriptions.

Track	Target	Samples	Description
Grading	Exam paper grading	2,464	2,408 gradable pages; 2,473 safety units requiring abstention
Parsing	Structure parsing	2,464	Parse Page, Section, Instruction, Question, Table, Figure, and ResponseSlot nodes
Recognition	Handwritten text recognition	10,000	Transcribe only the student's final handwritten answer
Grounding	Handwritten answers	5,893	Locate the queried student handwritten-answer region
Grounding	Student corrections	1,396	Locate crossed-out or rewritten answer regions
Grounding	Question stems	1,978	Locate the queried question-stem region
Grounding	Section material	733	Locate section instructions or shared material
Total	—	24,928	—

3.2 Data construction and quality control

Figure 2 summarizes the data construction process. We collect 2,464 photographed primary-school examination pages in Chinese, Mathematics, and English, covering varied capture conditions, handwritten corrections, incomplete responses, and complex layouts. Pages are screened for image and annotation quality, and sensitive regions are masked before evaluation.

Annotations link page regions, text, and grading decisions through a page–section–question–response hierarchy. Transcriptions preserve the student's final visible answer, including misspellings and corrections. GPT-5.5 (OpenAI, 2026a) assists with reference-answer and grading candidates; humans then verify the answer, binary credit, scoring point, and justification.

Initial labeling, quality checks, and expert review resolve ambiguous handwriting, corrections, and disputed scores. Consistent identifiers link questions and responses across tasks. The reviewed annotations yield grading units, structural nodes, grounding queries with target boxes, and handwriting crops; private labels remain separate from model inputs.

3.3 Subject and question-type coverage

Table 3 summarizes the subject distribution of the 2,464 examination pages. Subject proportions describe dataset composition rather than relative difficulty.

The question-type chart in Figure 1 summarizes 22,636 questions associated with the grading task using ten refined categories. Language fill-in-the-blank questions account for 33.0%, multiple choice

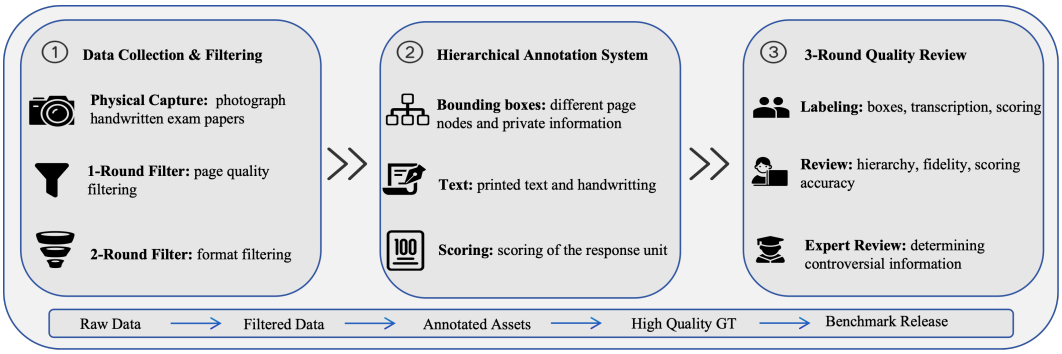


Figure 2: ExamOCRBench data construction pipeline. Data collection and filtering, linked annotation, and multi-stage quality review produce the task-specific evaluation data.

Table 3: Subject distribution of the 2,464 examination pages.

Subject	Pages	Share
Chinese	492	20.0%
Mathematics	694	28.2%
English	1,278	51.9%

for 22.2%, and short-answer or problem-solving questions for 16.9%. The remaining categories cover mathematical fill-in-the-blank, calculation, true/false, diagrams and marking, mathematical word problems, writing or open-ended responses, and other questions. These reporting categories refine the annotation labels using classification rules. Their question-level denominator is distinct from both the response-unit count used in grading and the cross-task instance total.

4 Tasks and Evaluation

The examples on the right of Figure 1 illustrate the four task interfaces. Grading and parsing receive a complete privacy-safe page; grounding receives a privacy-safe page and a natural-language query; recognition receives a single student-answer crop. The grading marks, region boxes, structural tree, and transcriptions visualize outputs or annotations rather than extra information supplied to the models. Grounding examples may be enlarged for display, while evaluation uses the full page. The following subsections define each task and its metrics.

4.1 Exam Paper Grading

Exam Paper Grading assesses responses with human-verified binary credit. From the page, a model identifies the relevant question, interprets the final student answer, derives an answer from the available question evidence, and produces a grading decision. It receives neither a human reference answer nor a gold response box. The evaluator matches the prediction to the frozen grading units and compares the awarded credit with human-verified labels.

Let P contain pages with at least one normal grading unit, and I_p be those units on page p. Let $\hat{s}_{p,i}$ and $s_{p,i}$ be predicted and gold binary credit. PageScoreAccuracy requires agreement on every normal unit:

$$\text{PageScoreAccuracy} = \frac{1}{|P|} \sum_{p \in P} \prod_{i \in I_p} \mathbf{1}[\hat{s}_{p,i} = s_{p,i}]. \tag{1}$$

This measures complete per-unit agreement, rather than agreement of summed page totals. Missing or invalid predictions cannot receive credit for a normal unit.

Score agreement does not necessarily imply that the model has understood the answer correctly. For example, a response may deserve one point, yet the model misreads the student's answer and also derives an incorrect reference answer. If those two incorrect answers coincide, the model can still award one point and match the gold score by accident. PageScoreAccuracy records the matching credit, but does not certify the intermediate transcription or reasoning. This motivates examining answer fidelity and the diagnostic tasks alongside grading accuracy.

Let U be the set of safety units whose answering conditions are insufficient for normal grading. Safety-unit recall measures the proportion of these units that the model correctly identifies as non-evaluable. Under the evaluation protocol, let $r_u = 1$ only when a matched prediction explicitly preserves the non-evaluable status and supplies no substantive reference answer; otherwise, let $r_u = 0$. The metric is:

$$\text{SafetyUnitRecall} = \frac{1}{|U|} \sum_{u \in U} r_u. \tag{2}$$

Low recall indicates that the model fails to recognize many units with insufficient answering conditions. If it nevertheless produces an answer or awards credit, it hallucinates an unsupported grading decision. This metric evaluates the recovery of safety units, rather than all forms of model hallucination.

4.2 Exam Paper Structure Parsing

Exam Paper Structure Parsing predicts an ordered page hierarchy using seven node types: Page, Section, Instruction, Question, Table, Figure, and ResponseSlot. A ResponseSlot is the smallest independent position where a student fills in, chooses, or writes an answer; it is not a bounding box around the ink alone. Correction and privacy regions are not output node types. Because the locations of response regions for matching questions are disputed during human annotation, their ResponseSlot nodes are excluded from evaluation.

We evaluate tree recovery and geometry separately. Following Image-based Table Recognition (Zhong et al., 2020), we adapt TEDS to the exam-page tree. Let T and $\hat{T}$ be the reference and predicted trees after evaluation filtering, including their synthetic root. Their normalized tree-edit similarity is:

$$\text{TEDS}(T, \hat{T}) = \max\left(0, 1 - \frac{d_{\text{tree}}(T, \hat{T})}{\max(|T|, |\hat{T}|, 1)}\right). \tag{3}$$

Here $d_{\text{tree}}(T, \hat{T})$ is the minimum total cost of an edit sequence transforming the reference ordered tree into the predicted ordered tree. Allowed operations are node insertion, deletion, and replacement; child order is preserved. Insertion and deletion each cost 1. Replacing a node with a different type costs 1. For nodes of the same type, replacement costs 0 if the reference node is not text-evaluable; otherwise it uses the edit distance between normalized texts, divided by the reference-text length (with a minimum denominator of 1) and capped at 1. Thus the distance accounts for missing or extra nodes, incorrect types, and eligible text mismatches, rather than simply counting different node labels.

The reported TEDS is averaged over the formal pages. The use of an exam hierarchy and eligibility rules distinguishes this implementation from the HTML-table TEDS formulation (Zhong et al., 2020) and its use in OmniDocBench (Ouyang et al., 2025).

Geometric overlap is evaluated using standard intersection over union (IoU) (Rezatofighi et al., 2019). Layout mAP@0.5 averages the per-class average precision over eligible node categories at an IoU threshold of 0.5, following the object-detection evaluation convention of the PASCAL VOC challenge (Everingham et al., 2010).

Together, TEDS and layout mAP distinguish hierarchy recovery from region localization.

4.3 Content Grounding

Content Grounding asks a model to locate a region described by a natural-language query on the full page. There are four target types: **HandwrittenAnswer (H)** is the student's handwritten answer area; **Question (Q)** is a subquestion's stem; **Instruction (I)** is the instructions or shared material for a question group; and **StudentCorrection (S)** is an area where the student crossed out or revised an answer.

A prediction is counted as a hit when it is valid, reports the target as found, and its box has IoU of at least 0.5 with the reference box. For the query set Q_t of target type t, let a_q indicate this event. We refer to the arithmetic mean of the four per-type accuracies as Macro Accuracy:

$$\text{Macro Accuracy} = \frac{1}{4} \sum_{t \in \{H,Q,I,S\}} \frac{1}{|Q_t|} \sum_{q \in Q_t} a_q. \tag{4}$$

This IoU-based evaluation follows the localization setting of OCRBench v2 (Fu et al., 2025) and CC-OCR V2 (Peng et al., 2026).

4.4 Handwritten Text Recognition

Handwritten Text Recognition transcribes the final effective student handwriting in one answer crop. Printed question text, crossed-out obsolete content, and content outside the target response are excluded. For 2,468 crops containing overlapping response regions, target-preserving masks suppress non-target handwriting; the remaining 7,532 crops retain their original input images. Normalization consistently handles Unicode full-/half-width forms, whitespace, and case; removes formatting-only mathematical wrappers; and canonicalizes comparable mathematical notation and judgment markers. Model outputs and student handwriting undergo the same normalization before comparison.

Let $\hat{y}_j$ and y_j be normalized prediction and reference strings. ExactMatch (EM) is defined as:

$$\text{EM} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}[\hat{y}_j = y_j]. \tag{5}$$

Character Error Rate (CER) aggregates edit distance across the reference characters:

$$\text{CER} = \frac{\sum_{j=1}^{n} d_{\text{edit}}(\hat{y}_j, y_j)}{\sum_{j=1}^{n} |y_j|}. \tag{6}$$

Unlike EM, CER can exceed one when predictions contain many insertions. Its use follows the transcription evaluation setting represented by TrOCR (Li et al., 2021). We do not adopt the answer-containing or semantic matching rules used for some VQA questions in OCRBench (Liu et al., 2024) and OCRBench v2 (Fu et al., 2025), because accepting a plausible paraphrase would obscure the student's actual spelling or symbols.

5 Experiments

5.1 Experimental setup

We compare Qwen3.8-Max (Qwen Team, 2026), Doubao-Seed-2.1-Pro (ByteDance, 2026), GPT-5.6 Sol (OpenAI, 2026a), Claude Opus 5 (Anthropic, 2026), and GPT-6 Astra (OpenAI, 2026b). Grading and parsing use the same 2,464 pages. Recognition uses 10,000 crops with target-preserving masks where required. Grounding uses 10,000 queries.

Task-specific prompts request structured grading, hierarchical regions, bounding boxes, or transcription. Grounding may batch queries from the same page, but each query is evaluated independently. The diagnostic outputs are not supplied to the grading model.

5.2 Overall results

The radar panel in Figure 1 is illustrative; all quantitative comparisons use the metrics in Table 4, which distinguishes primary from secondary measures. GPT-6 Astra achieves the highest PageScoreAccuracy (58.18%) and SafetyUnitRecall (41.04%), mean parsing TEDS (59.58%), layout mAP@0.5 (36.52%), and Macro Accuracy (73.44%). Qwen leads recognition EM (33.15%), while Doubao has the lowest CER (30.44 edits per 100 reference characters). Astra ranks second on both recognition metrics.

Table 4: Results by task and model, shown to two decimal places. PageScoreAccuracy and SafetyUnitRecall are the grading metrics; all metrics are percentages except CER (edits per 100 reference characters). Bold indicates the best result; arrows indicate the preferred direction.

Task	Metric	Qwen 3.8-Max	Doubao 2.1-Pro	GPT-5.6 Sol	Claude Opus 5	GPT-6 Astra
Grading	PageScoreAccuracy ↑	42.32	33.22	53.41	36.75	**58.18**
	SafetyUnitRecall ↑	9.06	30.04	26.77	13.99	**41.04**
Parsing	TEDS ↑	50.06	47.01	56.04	51.59	**59.58**
	mAP@0.5 ↑	30.01	35.15	29.97	11.21	**36.52**
Grounding	Macro Accuracy ↑	58.46	66.70	55.91	23.25	**73.44**
Recognition	EM ↑	**33.15**	17.11	10.88	16.09	25.78
	CER ↓	99.78	**30.44**	36.70	43.71	31.71

The strongest grading run correctly assesses all normal response units on 1,401 of 2,408 eligible pages. This stringent page-level endpoint is not directly comparable to crop recognition or query localization accuracy. Independently evaluated diagnostic tasks describe associated capabilities without establishing causal explanations for grading errors.

5.3 Grading accuracy and SafetyUnit diagnostics

Astra leads both PageScoreAccuracy and SafetyUnitRecall, but other models retain different strengths. GPT-6 Astra reaches 58.18% PageScoreAccuracy and 41.04% SafetyUnitRecall. GPT-5.6 Sol ranks second on PageScoreAccuracy (53.41%), whereas Doubao ranks second on SafetyUnitRecall (30.04%) despite having the lowest PageScoreAccuracy (33.22%). Qwen has higher PageScoreAccuracy than Doubao (42.32% versus 33.22%) but the lowest SafetyUnitRecall (9.06%). Astra's joint lead shows that these capabilities can improve together; the remaining rank differences still motivate evaluating them separately.

Unsupported grading remains a major failure on safety units. These include listening questions with missing media materials and questions missing essential context. Even Astra reaches only 41.04% SafetyUnitRecall. Across the five models, recorded unsupported-grading rates on safety units range from 58.96% to 88.96%. Thus, the strongest model still frequently grades despite insufficient evidence. Missing outputs produce additional recall failures for some models but are not counted as hallucinated grading.

5.4 Structure and grounding diagnostics

Astra leads aggregate structure metrics, while fine-grained localization remains difficult. Astra attains the highest TEDS (59.58%) and layout mAP@0.5 (36.52%). GPT-5.6 Sol ranks second on TEDS (56.04%), while Doubao ranks second on layout mAP (35.15%). Claude ranks third on TEDS (51.59%) but last on mAP (11.21%); its response-slot AP is only 0.4%. Even Astra's response-slot AP is just 19.6%. High relative performance on tree recovery therefore does not remove the difficulty of locating the precise answer positions needed for grading.

Astra improves correction grounding, but corrections remain its weakest target type. Its Macro Accuracy is 73.44%, compared with Doubao's 66.70%. The hit rate for student corrections reaches 30.8%; this exceeds Qwen's 20.1%, but remains low.

Astra's joint lead in grading, parsing, and grounding does not make the diagnostic tasks interchangeable. Other models change rank across metrics, and the most difficult region types remain weak even for the leader. The results support separately measuring hierarchical organization, target identity, and box extent; they do not identify which localization error caused an individual grading failure.

5.5 Handwriting recognition diagnostics

Astra balances whole-answer recovery and character-level fidelity without leading either metric. Qwen has the highest EM (33.15%) but the highest CER (99.78 edits per 100 reference characters). Doubao has the lowest CER (30.44), with an EM of 17.11%. Astra ranks second on both, reaching 25.78% EM and 31.71 CER. GPT-5.6 Sol and Claude also reverse their relative order between the two metrics: Claude has higher EM, whereas GPT-5.6 Sol has lower CER. The number of completely recovered answers and the total character-edit cost capture different properties of a recognition system.

Qwen's edit errors are concentrated in particular handwriting groups. Corrected and cross-line responses together account for 17.8% of the recognition crops but 65.9% of Qwen's total edit distance. This concentration helps explain how the highest EM can coexist with the worst CER: exact recovery on many crops can be offset by high edit costs on a subset of difficult responses. The groups also differ in text length and content, so the imbalance cannot be attributed to handwriting style alone.

Grading success does not imply the best transcription fidelity. Astra leads page grading but ranks second on recognition EM. GPT-5.6 Sol ranks second on grading but last on EM, while Qwen leads EM despite lower grading accuracy. These differences motivate distinguishing reasoning about answer correctness from preserving what a student actually wrote. Because grading and recognition use different inputs and evaluation units, rank reversals are not evidence that weaker recognition improves grading; they support evaluating these capabilities separately.

6 Conclusion

ExamOCRBench evaluates student-completed examination pages through end-to-end grading and independent structure, grounding, and handwriting diagnostics. The benchmark comprises 24,928 task instances derived from 2,464 pages with linked annotations. Among the five evaluated models, the highest grading accuracy is 58.18%, indicating that multimodal models still have substantial limitations in end-to-end grading of real, handwritten primary-school examination pages.

Reliable assessment requires preserving what students actually wrote, associating each response with its question, and recognizing when essential evidence is missing. Future work can extend the data to additional subjects and grade levels, and investigate alternative task decompositions that better characterize the capabilities required for end-to-end exam grading.

AI use statement

In this work, generative AI was used primarily to assist with manuscript polishing and data pre-annotation. During manuscript preparation, AI was used to refine word choice and consolidate sections. For data pre-annotation, AI was used to generate preliminary reference answers for the grading task.

References

Anthropic. Models overview. Official model documentation, 2026. URL `https://platform.claude.com/docs/en/about-claude/models/overview`. Accessed 2026-09-24.

A. F. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusinol, E. Valveny, C. V. Jawahar, and D. Karatzas. Scene Text Visual Question Answering. In *ICCV*, 2019. URL `https://arxiv.org/abs/1905.13648`.

L. Blecher, G. Cucurull, T. Scialom, and R. Stojnic. Nougat: Neural Optical Understanding for Academic Documents. *arXiv:2308.13418*, 2023. URL `https://arxiv.org/abs/2308.13418`.

ByteDance. Doubao model service. Official model documentation, 2026. URL https://www.volcengine.com/product/doubao. Accessed 2026-09-24.

A. Caraeni, A. Scarlatos, and A. Lan. Evaluating GPT-4 at Grading Handwritten Solutions in Math Exams. *arXiv:2411.05231*, 2024. URL https://arxiv.org/abs/2411.05231.

L. Chen, W. Xie, Y. Liang, H. He, H. Zhao, Z. Yang, Z. Huang, H. Wu, H. Lu, Y. charles, Y. Bao, Y. Fan, G. Li, H. Shen, X. Chen, W. Xu, S. Si, Z. Cai, W. Chai, Z. Huang, F. Liu, T. Liu, B. Chang, M. Wu, X. Hu, K. Chen, Y. Ren, Y. Liu, Y. Gong, and K. Li. BabyVision: Visual Reasoning Beyond Language. *arXiv:2601.06521, version 2*, 2026. URL https://arxiv.org/abs/2601.06521.

M. O. Dzikovska, R. D. Nielsen, C. Brew, C. Leacock, D. Giampiccolo, L. Bentivogli, P. Clark, I. Dagan, and H. T. Dang. SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In *SemEval*, 2013. URL https://aclanthology.org/S13-2045/.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. URL https://www.robots.ox.ac.uk/~vgg/projects/pascal/VOC/pubs/everingham10.html.

L. Fu, Z. Kuang, J. Song, M. Huang, B. Yang, Y. Li, L. Zhu, Q. Luo, X. Wang, H. Lu, Z. Li, G. Tang, B. Shan, C. Lin, Q. Liu, B. Wu, H. Feng, H. Liu, C. Huang, J. Tang, W. Chen, L. Jin, Y. Liu, and X. Bai. OCRBench v2: An Improved Benchmark for Evaluating Large Multimodal Models on Visual Text Localization and Reasoning. *arXiv:2501.00321*, 2025. URL https://arxiv.org/abs/2501.00321.

Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. In *ACM Multimedia*, 2022. URL https://arxiv.org/abs/2204.08387.

A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi. A Diagram Is Worth A Dozen Images. In *ECCV*, 2016. URL https://arxiv.org/abs/1603.07396.

G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park. OCR-free Document Understanding Transformer. In *ECCV*, 2022. URL https://arxiv.org/abs/2111.15664.

M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei. TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. *arXiv:2109.10282*, 2021. URL https://arxiv.org/abs/2109.10282.

Y. Liu, Z. Li, M. Huang, B. Yang, W. Yu, C. Li, X.-C. Yin, C.-L. Liu, L. Jin, and X. Bai. OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models. *arXiv:2305.07895*, 2024. URL https://arxiv.org/abs/2305.07895.

Y. Ma, Y. Zang, L. Chen, M. Chen, Y. Jiao, X. Li, X. Lu, Z. Liu, Y. Ma, X. Dong, P. Zhang, L. Pan, Y.-G. Jiang, J. Wang, Y. Cao, and A. Sun. MMLONGBENCH-DOC: Benchmarking Long-context Document Understanding with Visualizations. In *NeurIPS Datasets and Benchmarks*, 2024. URL https://arxiv.org/abs/2407.01523.

A. Masry, X. L. Do, J. Q. Tan, S. Joty, and E. Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of ACL*, 2022. URL https://aclanthology.org/2022.findings-acl.177/.

M. Mathew, D. Karatzas, and C. V. Jawahar. DocVQA: A Dataset for VQA on Document Images. In *WACV*, 2021. URL https://arxiv.org/abs/2007.00398.

M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, and C. V. Jawahar. InfographicVQA. In *WACV*, 2022. URL https://arxiv.org/abs/2104.12756.

M. Mohler, R. Bunescu, and R. Mihalcea. Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In *ACL*, 2011. URL https://aclanthology.org/P11-1076/.

OpenAI. Models: OpenAI API documentation. Official model documentation, 2026a. URL https://developers.openai.com/api/docs/models. Accessed 2026-09-24.

OpenAI. GPT-6 Astra. Official model documentation, 2026b. URL https://developers.openai.com/api/docs/models/gpt-6-astra. Accessed 2026-09-24.

L. Ouyang, Y. Qu, H. Zhou, J. Zhu, R. Zhang, Q. Lin, B. Wang, Z. Zhao, M. Jiang, X. Zhao, J. Shi, F. Wu, P. Chu, M. Liu, Z. Li, C. Xu, B. Zhang, B. Shi, Z. Tu, and C. He. OmniDocBench: Benchmarking Diverse PDF Document Parsing with Comprehensive Annotations. In *CVPR*, 2025. URL https://arxiv.org/abs/2412.07626.

C. Peng, Z. Xu, Y. Xiong, Z. Chen, J. Ji, Q. Liu, Z. Liu, Z. Qin, B. Zhao, J. Wan, J. Tang, Z. Yang, S. Bai, D. Liu, and M. Sun. CC-OCR v2: Fine-Grained Attribution of LMM Failures in Real-World Visual Document Understanding. *arXiv:2605.03903*, 2026. URL https://arxiv.org/abs/2605.03903.

J. Perš, J. Muhovič, A. Košir, and B. Murovec. Grading Handwritten Engineering Exams with Multimodal Large Language Models. *arXiv:2601.00730*, 2026. URL https://arxiv.org/abs/2601.00730.

Qwen Team. Qwen: Official model documentation and releases. Official model documentation, 2026. URL https://qwenlm.github.io/. Accessed 2026-09-24.

H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 658–666, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Rezatofighi_Generalized_Intersection_Over_Union_A_Metric_and_a_Loss_for_CVPR_2019_paper.html.

Z. Wang, M. Xia, L. He, H. Chen, Y. Liu, R. Zhu, K. Liang, X. Wu, H. Liu, S. Malladi, A. Chevalier, S. Arora, and D. Chen. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. In *NeurIPS*, 2024. URL https://arxiv.org/abs/2406.18521.

Z. Yang, J. Tang, Z. Li, P. Wang, J. Wan, H. Zhong, X. Liu, M. Yang, P. Wang, S. Bai, L. Jin, and J. Lin. CC-OCR: A Comprehensive and Challenging OCR Benchmark for Evaluating Large Multimodal Models in Literacy. *arXiv:2412.02210*, 2024. URL https://arxiv.org/abs/2412.02210.

X. Zhong, E. ShafieiBavani, and A. Jimeno Yepes. Image-based Table Recognition: Data, Model, and Evaluation. In *ECCV*, 2020. URL https://arxiv.org/abs/1911.10683.