LiveK12Bench: Have Large Multimodal Models Truly Conquered High School-level Examinations?

Xiaohan Wang[1*†], Mingze Yin[1,2*], Yilin Zhao[1], Gang Liu[1], and Dian Li[1†]

[1] Tencent PCG, Beijing, China
[2] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[3] {shawnbywang,goodli}@tencent.com

Abstract. Advanced Large Multimodal Models (LMMs) have demonstrated impressive performance in K-12 reasoning tasks, exhibiting great promise as intelligent tutors. Realizing this potential requires models to navigate real-world examinations effectively, yet most existing benchmarks fail to capture the complexity of authentic testing environments. Specifically, most datasets are static, prone to data contamination, and are often confined to restricted modalities, disciplines, and evaluation criteria. To address these issues, we introduce **LiveK12Bench**, a dynamic, holistic, multi-disciplinary benchmark designed to evaluate the reasoning abilities of LMMs in realistic examination scenarios. LiveK12Bench comprises 2K+ verified questions spanning Mathematics, Physics, Chemistry, and Biology, sourced from the latest real-world exam papers and designed to grow over time. Our framework features several core innovations: 1) featuring an automated pipeline that continuously ingests and parses latest examination papers to mitigate data leakage; and 2) proposing a novel 'Mock Exam' evaluation scheme, which assesses the ability to complete end-to-end exams autonomously with accurate and efficient reasoning paths. Extensive experiments on 12 LMMs reveal that advanced models suffer substantial performance degradation under exam-realistic constraints: GPT-5's score drops from 79 to 53 (out of 100) when process rigor and efficiency are jointly evaluated. Our findings expose critical vulnerabilities, such as sensitivity to complex visual layouts, highlighting the gap between idealized reasoning capabilities and true educational readiness. Both code and dataset are publicly available.

Keywords: LLM Evaluation · Multimodal Reasoning · Disciplinary Problem Solving · AI for Education

1 Introduction

Generative AI is rapidly transforming the educational landscape. As large language models continue to push the boundaries of their reasoning capabilities, they have achieved near-perfect performance on high school-level mathematical

∗ Equal contribution † Corresponding authors

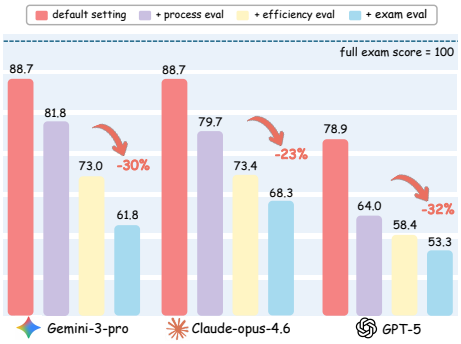


Fig. 1: Performance degradation of cutting-edge LMMs in authentic exam scenarios.

benchmarks such as MATH [12] and AIME [30, 35]. However, to truly serve as effective and reliable tutors for human students, AI must first demonstrate the ability to successfully navigate authentic human examinations. While recent news frequently highlight that advanced LMMs can achieve impressive scores on college entrance examinations, a critical question remains: *Have large multimodal models truly conquered high school-level examinations?*

To drive the evolution of AI reasoning, mainstream research has predominantly focused on highly logic-dependent disciplines such as mathematics and programming. To isolate reasoning capabilities from other confounding factors, traditional benchmarks typically provide meticulously parsed question information and are confined to evaluating easily verifiable final answers. Recently, some benchmarks have shifted their focus toward evaluating generative AI in educational contexts, introducing K-12 multi-disciplinary assessments and photo-based problem-solving evaluations [6, 29, 33]. However, these existing benchmarks struggle to comprehensively answer the aforementioned question, as they fundamentally fail to bridge three core gaps between AI evaluation and authentic human testing: 1) **Data Leakage**: Most datasets are static. Once published, they are inevitably ingested into the training corpora of next-generation LLMs, rendering subsequent evaluations unreliable and losing their reference value [23]. 2) **Insufficient Evaluation**: Human examinations assess students under strict time and environmental constraints, comprehensively grading both the final answer and the step-by-step reasoning process across questions of varying importance. In contrast, AI evaluation criteria remain largely one-dimensional. 3) **Human Intervention**: Testing models on K-12 exam papers often involves manual question extraction, image cropping, or providing textual descriptions for visual elements. Consequently, the task input for AI is not equivalent to that of a human student, precluding a genuine end-to-end examination. These gaps make it exceedingly difficult to accurately estimate the practical usability and potential value of mainstream LMMs as intelligent tutors or educational assistants.

To address these limitations, we propose LiveK12Bench, a dynamic, comprehensive, and multi-disciplinary AI examination benchmark designed to systematically investigate the capabilities and limitations of mainstream LMMs in real-world K-12 scenarios. Specifically, to eradicate the issue of test data leakage at its source (and to avoid the pitfalls of AI-generated synthetic questions), LiveK12Bench introduces a highly efficient, automated examination paper parsing pipeline based on structural document extraction and LLM parsing. This pipeline enables the periodic ingestion of fresh questions newly authored by frontline educators, continuously expanding the dataset scale. Concurrently, we propose a

"Mock Exam" evaluation scheme that simulates the multi-dimensional assessment of human exams, evaluating mainstream models across four dimensions: answer accuracy, process correctness, reasoning efficiency, and a weighted comprehensive exam score. Building upon standard text-only and text-image multimodal settings, we introduce an "Image-Only" full-page modality. This setting aligns with an end-to-end testing scenario, significantly reducing manual assistance and intervention.

By evaluating mainstream multimodal reasoning models (illustrated in Figure 1), the performance of leading LMMs degrades across three progressively challenging scenarios: from standard settings to exam scoring incorporating process and efficiency evaluation, and end-to-end "Image-Only" exam modality. These analytical insights provide crucial implications for the future development of generative AI in educational applications.

Our primary contributions are summarized as follows:

- We propose the first comprehensive, multi-disciplinary benchmark that holistically simulates authentic human K-12 examinations.
- We design an automated exam paper ingestion pipeline that facilitates the efficient and dynamic iteration of the dataset, effectively mitigating data contamination.
- We introduce a comprehensive "Mock Exam" evaluation protocol assessing reasoning process and problem-solving efficiency, and incorporating an test paper "Image-Only" input modality to impose real-world layout noise.

2 LiveK12Bench

LiveK12Bench systematically resolves the aforementioned challenges through three core innovations: a holistic dataset encompassing multi-disciplinary and multi-modal scenarios, a dynamic data construction pipeline that continually ingests fresh examination papers, and a novel "Mock-Exam" evaluation protocol that demands end-to-end problem solving with process, efficiency and modality constraints. The overall architecture and workflow of our framework are illustrated in Figure 2. The following subsections detail the design and implementation of each component.

2.1 Dataset Composition and Comparison

The LiveK12Bench dataset currently comprises 2,114 high-quality, manually verified questions covering four core K-12 disciplines that heavily rely on reasoning capabilities: Mathematics, Physics, Chemistry, and Biology. The dataset encompasses diverse question formats, including Multiple-Choice Questions (MCQs), Fill-in-the-Blank (FIB) questions, and Q&A questions. The dataset consists of two timestamp splits of 26-03 and 26-05 (indicating the release time of questions), both with Chinese and English translation versions.

To comprehensively evaluate the robustness of Large Language Models (LLMs) and Large Multimodal Models (LMMs), we curate tasks across three distinct

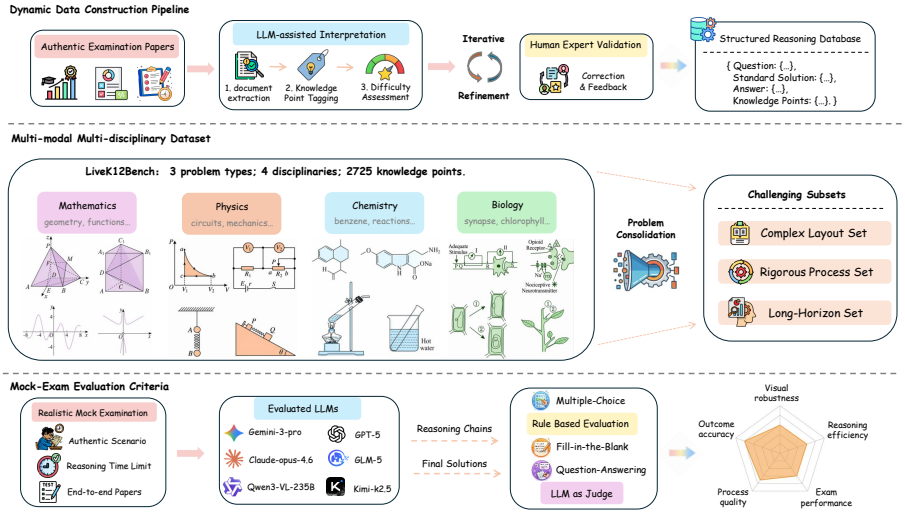


Fig. 2: Overall framework of LiveK12Bench. The framework consists of three interconnected modules that conduct multi-dimensional evaluation starting from raw exam papers: a dynamic data construction pipeline, a comprehensive dataset, and a Mock-Exam evaluation protocol.

modalities, corresponding to three realistic evaluation scenarios. Formally, let f_θ denote the evaluated model, $\mathcal{T}$ denote textual input, and $\mathcal{V}$ denote visual input. The three task modalities are defined as follows:

- **Text-Only (TO):** Both the input and expected output are purely textual, evaluating the foundational linguistic and symbolic reasoning capabilities of LLMs. The task is formulated as $A = f_\theta(\mathcal{T}_q)$, where $\mathcal{T}_q$ represents the textual question stem and options, and A is the textual predicted answer.
- **Text-Image (TI):** The input consists of an interleaved mixture of text and images, assessing the LMM's ability to ground textual concepts in visual representations (*e.g.*, geometric diagrams, circuit schematics, or biological structures). The task is defined as $A = f_\theta(\mathcal{T}_q, \mathcal{V}_q)$, where $\mathcal{V}_q$ denotes the cropped image(s) essential for solving the problem.
- **Image-Only (IO, Exam mode):** The input is an uncropped snapshot of a full examination page alongside a target question index, simulating the authentic task environment of a human student. It intentionally removes human-assisted intermediate steps, such as manual OCR and image cropping. The model must autonomously locate, extract, and interpret the relevant problem information from the page layout. The task is formalized as:

$$A = f_\theta(\mathcal{V}_{pages}, idx) \tag{1}$$

where $\mathcal{V}_{page}$ is the raw exam page images, and idx is the specified question number to be solved.

Table 1: Key statistics of LiveK12Bench.

Category	Overall	Mathematics	Physics	Chemistry	Biology
Task Modality					
- Text-Only (TO)	1,096	617 (56.3%)	65 (5.9%)	240 (21.9%)	174 (15.9%)
- Text-Image (TI)	1,018	155 (15.2%)	331 (32.5%)	292 (28.7%)	240 (23.6%)
- Image-Only (IO, Exam-mode)	2114	772 (36.5%)	220 (10.4%)	532 (25.2%)	414 (19.6%)
Question Type					
- Multiple-Choice (MCQ)	1,473	444 (30.1%)	274 (18.6%)	419 (28.4%)	336 (22.8%)
- Fill-in-the-Blank (FIB)	164	119 (72.6%)	26 (15.9%)	18 (11.0%)	1 (0.6%)
- Question-Answering (Q&A)	477	209 (43.8%)	96 (20.1%)	95 (19.9%)	77 (16.1%)
Total Number	**2,114**	**772 (36.5%)**	**396 (18.7%)**	**532 (25.2%)**	**414 (19.6%)**

Beyond the core inputs, the dataset provides rich metadata for each question, including the ground-truth final answer annotated by professional educators, step-by-step solution paths, question types, score value, subject categories, and fine-grained knowledge point tags.

To probe the unique vulnerabilities of models, prior benchmarks [17, 22] have constructed multiple subsets assessing different aspects of visual reasoning, such as measurement and puzzle tests. In line with this approach, aiming to capture the unique challenges of real-world K-12 examinations, we deliberately establish three subsets (50 questions per subject per subset, 600 in total) from our benchmark as follows:

1. **Complex Layout Set:** This subset specifically targets the visual challenge of real-world test papers, featuring question layouts with highly complex visual formatting. It includes cases where problems span across multiple pages, question stems are spatially detached from their corresponding figures, or images are tightly embedded within text blocks. This set challenges the LMM's capacity to accurately extract reasoning contexts from noisy visual margins and complex layouts (*e.g.*, interpreting function curves and data tables). End-to-end proficiency on this set is a prerequisite for deploying AI educators in real-world, unconstrained input environments.
2. **Rigorous Process Set:** Mainstream benchmarks predominantly evaluate final answer accuracy. However, the design of MCQs in human exams often allows students to guess the correct option through elimination or surface-level heuristics without rigorously deducing the underlying concepts. We specifically compile questions that are assigned with multiple knowledge points and are susceptible to such "lucky guesses" with the feature of excessive premises. This set is designed to evaluate the model's ability to arrive at the correct answer through a logically sound and solid reasoning process (process evaluation methodology is detailed in Section 2.3).
3. **Long-Horizon Reasoning Set:** This subset comprises problems that frequently trap models in excessively long or circular reasoning chains, where the questions typically pose complex objectives and are assigned with higher max points in the papers. The difficulty may stem from intrinsic mathematical complexity, intentionally confounding conditions, or deceptive visual information. It aims to specifically evaluate the *reasoning efficiency* of LMMs. Intuitively, an AI model that requires fewer generated tokens to correctly

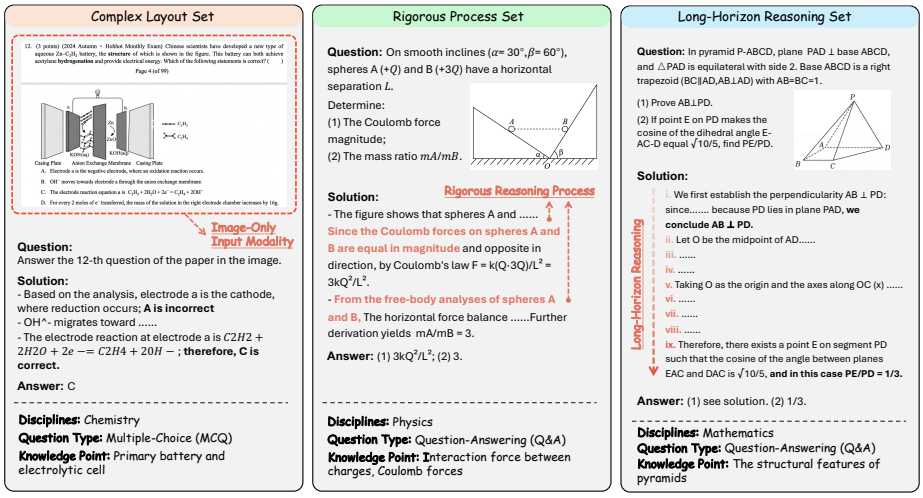


Fig. 3: Examples from the Three Challenging Subsets in LiveK12Bench. The figure illustrates typical inputs and annotated metadata for the Complex Layout Set, Rigorous Process Set, and Long-Horizon Reasoning Set.

solve a complex problem offers superior computational efficiency and user experience.

According to the criteria and features described above, we incorporate advanced LLMs as pre-annotators to excavate typical questions for these subsets from the whole dataset. Then human experts are prompted to verify these annotations and determine the final composition of 3 subsets. Figure 3 presents illustrative examples from these three challenging subsets, highlighting their distinct inputs and corresponding annotations. Figure 4 and Table 1 detail the statistical distribution of the dataset across subjects, modalities, and question types.

2.2 Dynamic Data Construction Pipeline

To address the pervasive issue of data contamination and ensure the continuous relevance of our evaluation, we propose a highly automated data construction pipeline based on structured Optical Character Recognition (OCR) and Large Language Model (LLM) parsing. This pipeline systematically processes raw examination PDFs, categorizes and extracts textual and visual elements, and leverages an LLM to decompose the content into structured fields (*e.g.*, question stems, options, standard answers, and reasoning paths) for archiving and subsequent evaluation. Specifically, the dataset construction consists of the following four stages:

Examination Paper Collection. We collected 200 of the latest (published in 2026) authentic Chinese high school examination papers across four disciplines: Mathematics, Physics, Chemistry, and Biology. These freshly curated papers exhibit an extremely low probability of prior data leakage. To establish rigorous ground truths, professional educators meticulously annotated the standard

answers and step-by-step reasoning processes. This sourcing strategy ensures fairness across all evaluated models. We specifically selected Chinese examination papers because they are constructed upon a highly scientific and systematic knowledge taxonomy, rigorously tested by tens of millions of students, and widely recognized for their strong representativeness of K-12 educational standards.

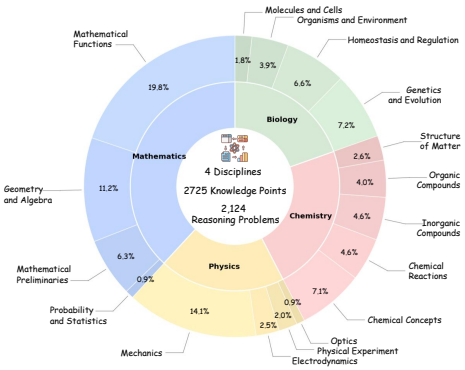


Fig. 4: Dataset Distribution. Proportion of questions across disciplines and knowledge points.

Structural Document Extraction. For PDF and scanned image inputs, we utilize a structural extraction workflow to convert the raw exam papers into Markdown text and cropped images. This workflow encompasses preprocessing, text box and embedded figure detection, multimodal OCR (for text, mathematical formulas, and tables), and image cropping. We implement this using the MinerU framework [26]. To better adapt the framework to Chinese examination papers, we fine-tuned the detection and OCR modules on an in-house exam dataset, significantly improving the recognition accuracy of complex mathematical and chemical formulas.

LLM-based Parsing with Variable Templates. Once the raw Markdown text is extracted, it is necessary to identify and structure the individual questions, their corresponding answers, and associated figures. This structured archiving is essential for constructing the context for the evaluated models and for comparing their outputs against the reference answers. Compared to rigid rule-based parsing, utilizing an LLM offers greater flexibility in comprehending diverse paper formats across different subjects and can semantically correct minor OCR recognition errors.

To handle the structural variance and distinct typographical features of exams from different sources, we introduce a variable template-based parsing method. We inject structured descriptions into the LLM's context, including the examination question type, target output fields (format and characteristics), and specific layout features. Before parsing, the template parameters can be adjusted to help the LLM precisely locate and extract the corresponding content. Formally, the context construction and parsing process can be expressed as:

$$C = \mathcal{T}_{type} \oplus \mathcal{F}_{target} \oplus \mathcal{L}_{layout} \oplus \mathcal{D}_{md} \tag{2}$$

$$S_{db} = \text{LLM}_{\text{parser}}(C) \tag{3}$$

where C represents the constructed prompt context, $\mathcal{T}_{type}$ denotes the question type definition, $\mathcal{F}_{target}$ specifies the target JSON schema and field constraints, $\mathcal{L}_{layout}$ describes the source document's layout characteristics, and $\mathcal{D}_{md}$ is the

raw extracted Markdown text. The operator $\oplus$ denotes string concatenation, and S_{db} is the final structured JSON database generated by the parsing agent.

Verification and De-duplication. To ensure parsing correctness, the structured text is re-rendered into HTML for manual verification against the original exam papers. This human-in-the-loop review focuses strictly on verifying crucial formulas and problem-solving clues to guarantee that the questions are solvable. Subsequently, we compute text similarity across the problem stems to remove duplicate questions. We extract a hierarchical knowledge tree from the official teaching syllabus and train a knowledge classification model based on Qwen3-VL-4B [4] to assign knowledge tags to questions without manual annotations.

Ultimately, this extensively automated data ingestion pipeline empowers LiveK12Bench to routinely refresh its test sets and publish refreshed leaderboards, effectively mitigating evaluation biases caused by data contamination. Furthermore, it allows external users to seamlessly upload their own PDF exams to build customized datasets and conduct comprehensive evaluations.

2.3 Mock-Exam Evaluation Criteria

Standard reasoning benchmarks typically employ simple accuracy metrics (*e.g.*, Pass@1) to evaluate models. However, this simplified approach fails to authentically reflect an AI's comprehensive performance in human-level examinations. To bridge this gap, we propose a multi-dimensional "Mock Exam" evaluation scheme that requires models to complete a full set of test papers under rigorous constraints. Overall, our evaluation criteria encompass four primary dimensions:

Outcome Dimension. For each problem, the evaluated model is prompted to explicitly output its step-by-step reasoning process, followed by the final answer enclosed within a `\boxed{}` command. Formally, the generated response $\hat{Y}$ is structured as $\hat{Y} = \mathcal{R} \oplus \backslash\text{boxed}\{\hat{A}\}$, where $\mathcal{R}$ denotes the reasoning path, $\hat{A}$ is the extracted final answer. In this dimension, we evaluate the standard Accuracy (Pass@1) metric as $Acc = \sum S_i$, where $S_i = \mathbb{I}(\hat{A}_i \equiv A_i^*)$ for the i-th question using an automated evaluator, A_i^* represents the ground-truth answer and $\mathbb{I}(\cdot)$ is the indicator function. Here, the correctness of intermediate steps is ignored. For proof-based questions, the entire generated proof is extracted and compared holistically against the reference.

Reasoning Process Dimension. Rather than merely inspecting the correctness of the final solutions, we rigorously evaluate the reasoning process to identify potential root-cause errors. Outcomes divorced from their underlying reasoning process have limited evidentiary value. Concretely, the model needs to accurately parse the problem statement, ground its analysis in sound logical assumptions, and conduct progressive deductive reasoning to attain precise solutions. Drawing on comprehensive analyses of enormous real-world cases, we distill three categories of critical reasoning process errors: (1)*Condition Interpretation Error*: A systematic misreading of the given data, relevant theorems, and specified constraints, resulting in a distorted grasp of problem statement facts. The error arises when interpreted conditions deviate from the true ones, $\hat{I}(x) \neq I(x)$, typically due to omission, misrepresentation, or the inclusion of spurious conditions. (2)*Logical*

Assumption Error: The insertion and layering of assumptions unsubstantiated by the problem statements. This error occurs whenever aberrant assumptions exceed the bounds of conditional statements warranted by accepted knowledge and valid inference $\Gamma \nsubseteq \mathrm{Cl}(I(x))$. (3)*Deductive Reasoning Error*: Reasoning steps and computational procedures fail to justify the claimed inference, yielding an internally invalid deductive chain that conflicts with established laws, theories, or axioms. This occurs when there exists a transition $s_t \Rightarrow s_{t+1}$ which is not entailed by the preceding steps under sound inference rules $\{s_1, \ldots, s_t\} \not\models s_{t+1}$. We execute a systematic audit of the reasoning process score P_i to detect defined errors, thus proposing a more rigorous standard for evaluation criteria:

$$P_i = V_i - \tau \cdot \sum_{k=1}^{3} x_{i,k} \tag{4}$$

where V_i represents the overall point value for the i-th question and τ denotes the hyperparameter governing the penalty term in the reasoning process evaluation. $x_{i,k} \in \{0, 1\}$ indicates the presence of k-th reasoning process error.

Reasoning Efficiency Dimension. To quantify the problem-solving efficiency of Large Multimodal Models (LMMs), we introduce two novel metrics:
ARL (Accuracy weighted by Response Length): Inspired by recent studies on efficient reasoning, ARL measures whether a model can achieve high accuracy with a more concise generation. It is formulated as:

$$\mathrm{ARL} = \frac{1}{N} \sum_{i=1}^{N} S_i \cdot (1 + \lambda \ln \frac{\bar{L}}{l_{i,j}}) \tag{5}$$

where N is the total number of questions, λ denotes the weight factor of reasoning efficiency,$l_{i,j}$ is the generated token length of model j on question i, and $\bar{L}$ is an empirical constants of average response length. If a model's generation length aligns with the average level, its ARL equals its Accuracy (S_i). An ARL higher than the base Accuracy indicates above-average reasoning efficiency, whereas a lower ARL suggests excessive verbosity.
$Acc_{\leq r}$: This metric evaluates the model's accuracy when its maximum generation length is strictly restricted to a ratio r of the full context window. By counting the total completion tokens (including intermediate "thinking" tokens) as a quantifiable proxy for time, we directly simulate the human capability of finishing an exam within a specific time limit. This approach effectively isolates the impact of varying hardware throughput and model sizes.

Exam Performance Dimension. In this dimension, we meticulously simulate human grading mechanisms to evaluate the authentic test-taking performance of LMMs. The performance is measured by an Overall Exam Score (OES), aggregated from the Exam Score (ES) of individual questions. The ES is composed of a process component (ES_i^P) and an outcome component (ES_i^O):

$$\mathrm{ES}_i = \underbrace{w_p \cdot P_i \cdot (O_i / V_i)}_{\mathrm{ES}_i^P} + \underbrace{(1 - w_p) \cdot O_i}_{\mathrm{ES}_i^O} \tag{6}$$

Table 2: Comparison of LiveK12Bench with existing benchmarks.

Benchmark	Subject(s)	Modality	Kn. Points	Solution	Dynamic	Level	Evaluation
MathVista [17]	Math	TI	✓	✓	✗	K-12,College	A
SciBench [28]	Math,Phy,Chem	TO, TI	✓	✓	✗	College	A
M3Exam [33]	Multi-subj.	TO, TI	✓	✗	✗	K-12	A
GAOKAO-MM [37]	Multi-subj.	TO, TI	✗	✓	✗	High School	A
MMSciBench [29]	Math, Phy	TO, TI	✓	✓	✗	High School	A
K12Vista [16]	Math,Phy,Chem,Bio	TI	✓	✓	✗	K-12	A,P
MDK12-Bench [36]	Multi-subj.	TO, TI	✓	✓	*Synthetic*	K-12	A
LiveK12Bench	Math,Phy,Chem,Bio	TO,TI,**IO**	✓	✓	✓	High School	**A,P,E,Exam**

- TO: Text-Only, TI: Text-Image, IO: Image-Only; A: Accuracy, P: Process eval, E: Efficiency eval; Exam: Exam score eval. *Synthetic* denotes achieving dynamic evaluation through synthesizing new questions, in contrast to ours that ingests authentic questions.

where P_i, O_i are scores for reasoning and outcome, respectively. w_p balances the weight between process and outcome. Specifically, O_i assigns points equally across correctly answered sub-components; hence, O_i/V_i denotes the proportion of correct sub-questions. We apply this ratio to the score calculation to emphasize the outcome correctness as a prerequisite. Ultimately, the OES is normalized to a standard 100-point scale. We similarly derive the Process Exam Score (PES) and Outcome Exam Score (OCS) by aggregating their respective components:

$$\text{OES} = 100 \times \frac{\sum \text{ES}_i}{\sum V_i}, \quad \{\text{PES}, \text{OCS}\} = 100 \times \frac{\sum \{\text{ES}_i^P, \text{ES}_i^O\}}{\sum V_i} \tag{7}$$

By incorporating human-educator-assigned weights and evaluating fine-grained sub-problem correctness, this weighted scoring system distinguishes problem importance and provides a significantly more comprehensive assessment than uniform accuracy.

Finally, to guarantee robust and stable evaluation across all dimensions, we adopt a multi-model arbitration scheme. For all LLM-based evaluators, we aggregate and average the judgments from a panel of multiple advanced models, which substantially mitigates the risk of subjective evaluation bias or comprehension failures from any single judge. Table 2 provides a comprehensive comparison between LiveK12Bench and existing related benchmarks, underscoring our unique contributions in evaluation dimensions and end-to-end simulation.

3 Experiments

In this section, we evaluate a diverse set of Large Multimodal Models (LMMs) on LiveK12Bench. Our experiments are systematically designed to answer four key questions: (i) How do state-of-the-art models perform across different K-12 disciplines? (ii) What is the impact of real-world "snapshot" noise on reasoning capabilities? (iii) To what extent are AI models relying on lucky guesses during exams? (iv) Do LMMs tend to overthink when solving test problems?

3.1 Experimental Setup

We test the performance of current mainstream LMMs on the proposed dataset. Our evaluation covers 12 models of varying parameter sizes, categorized into two groups:

Table 3: Main results on disciplinary performance (26-03 split).

Models	Mathematics				Physics				Chemistry				Biology			
	Acc	ARL	PES	OES	Acc	ARL	PES	OES	Acc	ARL	PES	OES	Acc	ARL	PES	OES
Claude-opus-4.6$_{26\text{-}02}$	87.2	94.6	87.6	90.0	79.1	84.2	80.8	83.7	77.0	81.4	62.3	71.1	88.0	91.9	63.6	74.1
Gemini-3-pro$_{25\text{-}11}$	88.3	82.9	88.3	90.3	86.2	82.5	85.4	87.9	79.4	77.2	71.2	76.7	90.7	89.5	71.3	78.6
GPT-5$_{25\text{-}08}$	82.7	83.1	83.1	85.5	70.9	70.7	68.8	72.7	59.3	57.6	37.9	44.2	71.6	70.3	44.1	53.7
Claude-sonnet-4.6$_{26\text{-}02}$	83.9	89.3	84.5	87.4	77.6	80.4	78.5	78.5	73.8	75.6	58.9	66.4	84.9	84.8	56.1	65.1
Gemini-3-flash$_{25\text{-}11}$	87.2	84.8	88.2	89.8	84.4	82.6	85.3	87.0	80.7	79.7	68.6	74.4	88.0	86.5	68.2	75.9
GPT-5-mini$_{25\text{-}08}$	79.0	85.2	75.0	79.8	60.1	63.9	54.9	60.1	45.2	47.1	24.1	29.6	55.9	58.5	31.7	39.0
GLM-5$_{26\text{-}02}$	79.8	76.9	76.5	79.5	67.5	65.3	56.9	61.9	65.1	63.1	43.3	49.7	77.2	75.4	41.1	52.4
Kimi-k2.5$_{26\text{-}01}$	85.0	86.0	86.2	87.8	81.6	82.9	83.3	85.4	77.2	80.1	64.0	71.8	86.7	89.7	64.3	73.2
Qwen3-VL-235B$_{25\text{-}09}$	82.5	82.5	75.8	81.3	77.3	82.9	67.0	73.8	74.3	79.1	51.6	62.8	85.5	91.1	55.0	65.2
Qwen3-VL-32B$_{25\text{-}09}$	82.7	84.3	76.6	81.1	75.8	79.2	69.2	73.5	73.8	77.2	44.8	76.2	86.7	90.1	45.9	55.3
Qwen3-VL-8B$_{25\text{-}09}$	79.6	78.8	70.9	77.2	66.6	65.3	49.2	56.8	62.2	62.0	30.6	39.3	75.3	74.2	36.1	45.5
GPT-4o$_{24\text{-}11}$	35.0	-	19.2	24.2	29.1	-	17.0	21.0	29.4	-	8.0	11.4	44.1	-	16.6	24.9

- Results of models (grouped into proprietary, open-source, and non-thinking models with release date) in 4 subjects using Accuracy (Acc), Efficiency-weighted Accuracy (ARL), Process Exam Score (PES), and Overall Exam Score (OES). The best and second-best performances are highlighted red and blue.

- **Proprietaries:** GPT-5 [20], GPT-5-mini [19], Gemini-3-pro [10], Gemini-3-flash [9], Claude-opus-4.6 [2], and Claude-sonnet-4.6 [3], GPT-4o [1] (no thinking ability).
- **Open-Source Models:** GLM-5 [31], Kimi-k2.5 [25], Qwen3-VL-235B-A22B [4], Qwen3-VL-32B [4], and Qwen3-VL-8B [4].

For all problem requests, we employ a unified prompt template, strictly instructing the models to output their final answers in a standardized format. For the answer verification process, we utilize a multi-model arbitration panel drawn from four advanced evaluators: GPT-4o, Gemini-3-flash, DeepSeek-V3 [7], and Qwen3-30B. These LLMs are selected as evaluators due to their high consistency with human adjustment (over 90% on average). Specifically, three distinct models are selected for each judgment to prevent any model from evaluating its own generated answers, thereby minimizing self-evaluation bias. Throughout our experiments, the process penalty factor τ is set as 3, the efficiency weight λ and process score weight ω_p are set as 0.15 and 0.5, respectively. The average-level response length constant $\bar{L}$ is set as 4096 posteriorly according to the statistics of models' responses. In this section, both scores of the reasoning process (PES) and outcome (OCS) are normalized to a 100-point scale same as OES for better illustration.

3.2 Results and Analysis

Main results: Disciplinary Performance

Table 3 presents the performance of the evaluated models across different disciplines on the full dataset, assessing LMMs' capabilities in realistic exam distributions. Gemini-3-pro achieves the highest Accuracy (Acc) and Overall Exam Score (OES) across most subjects. Its notable advantage is particularly evident in the Process Exam Score (PES) for Chemistry and Biology, outperforming the second-best model by 2.6 and 3.1 points, respectively. Notably, the smaller Gemini-3-flash also attains highly competitive scores. When considering

Table 4: Performance on challenging subsets.

(a) Complex Layout Set

Models	Acc		OCS	
	Standard	Exam	Standard	Exam
Claude-opus-4.6	87.2	55.0 -32.2	92.8	50.6 -42.2
Gemini-3-pro	88.3	60.0 -28.3	92.6	53.7 -38.9
GPT-5	82.7	50.4 -32.3	88.1	44.1 -44.0
Claude-sonnet-4.6	83.9	45.5 -38.4	90.8	45.8 -45.0
Gemini-3-flash	87.2	55.0 -32.2	91.8	50.0 -41.8
GPT-5-mini	79.0	42.9 -36.1	85.7	39.8 -45.9
GLM-5	79.8	-	85.4	-
Kimi-k2.5	85.0	59.0 -26.0	90.2	51.8 -38.4
Qwen3-VL-235B	82.5	56.0 -26.5	88.4	50.6 -37.8
Qwen3-VL-32B	82.7	55.5 -27.2	85.6	48.2 -37.4
Qwen3-VL-8B	79.6	53.6 -26.0	83.5	46.8 -36.7
GPT-4o	35.0	5.5 -29.5	40.8	6.6 -34.2

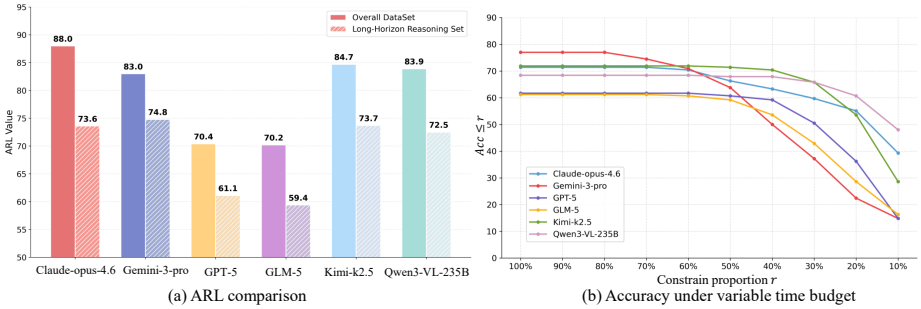
(b) Rigorous Process Set

Models	Exam Score ↑			Process Error ↓		
	OCS	PES	OES	CIE	LAE	DRE
Claude-opus-4.6	89.0	68.7	78.0	36	7	11
Gemini-3-pro	90.9	76.9	81.7	17	7	10
GPT-5	83.0	56.0	61.0	61	13	30
Claude-sonnet-4.6	87.9	63.9	70.8	40	6	20
Gemini-3-flash	91.6	75.4	80.3	17	8	18
GPT-5-mini	75.8	42.7	48.7	78	29	46
GLM-5	77.1	54.1	60.4	46	31	23
Kimi-k2.5	87.5	71.9	77.3	25	7	16
Qwen3-VL-235B	87.0	63.6	70.3	36	12	25
Qwen3-VL-32B	87.6	52.4	59.7	71	18	59
Qwen3-VL-8B	79.7	44.2	50.9	47	11	36
GPT-4o	50.0	13.5	19.8	129	34	111

- (a) OCS: Outcome exam score, Standard modality: text-only or text-image input, Exam modality: Image-only input. The score differences between two modalities are noted in red. (b) PES: Process Exam Score, OES: Overall Exam Score. We report the number of process errors in 3 types (CIE: condition interpretation error, LAE: logical assumption error, DRE: deductive reasoning error).

reasoning efficiency, Claude-opus-4.6 achieves the highest ARL scores across all disciplines, indicating it attains higher accuracy with fewer reasoning tokens. Conversely, Gemini models yield relatively lower ARL scores (ranking 5th to 7th), reflecting a design trade-off that sacrifices reasoning time for enhanced performance. Furthermore, while open-source models like Kimi-k2.5 and Qwen3 exhibit a slight disadvantage in exam scores compared to leading proprietary models, their reasoning efficiency surpasses that of GPT-5 and Gemini. It is also observable that factors such as smaller parameter sizes, earlier release dates, or a lack of explicit "thinking" capabilities cause models like GPT-4o (which is excluded from ARL evaluation due to the absence of thinking tokens) and Qwen3-VL-8B to struggle significantly at lower performance tiers. Comparing across subjects, all models perform worse in Chemistry and Biology compared to Mathematics and Physics—the traditional focus of mainstream reasoning research. This performance drop stems primarily from broader knowledge taxonomies and more complex Q&A structures involving multiple images and sub-questions.

What is the impact of real-world "snapshot" noise on reasoning capabilities? Table 4 (Left) details the model performance on the Complex Layout Set. We contrast performance under standard input modalities (parsed TO/TI) against the Exam modality (IO) to investigate how LMMs handle the complex visual layouts of raw exam papers. Significantly, unparsed layouts and embedded images cause a drastic decline in both Acc and Outcome Score (OCS) across all models. The GPT series is the most severely impacted, with an average drop of 33.7% in Acc and 44.9 points in OCS. In contrast, Gemini and Kimi-k2.5 demonstrate superior visual robustness. According to the LLM-judge analysis, errors predominantly originate from image dislocation and complex text-image interleaving, which lead models to overlook or misinterpret crucial visual information, thereby derailing subsequent reasoning steps.



(a) ARL comparison (b) Accuracy under variable time budget

Fig. 5: Reseults comparison on Long-horizon Reasoning Set.

To what extent do AIs rely on lucky guesses during exams? Table 4 (Right) evaluates the Rigorous Process Set. Beyond exam scores, we tally the frequencies of the three reasoning process errors defined in Section 2.3. Compared to the general distribution in Table 3, all models exhibit lower PES but maintain relatively high OCS. This divergence quantitatively demonstrates that for these specific problems, models frequently arrive at the correct final answer through flawed reasoning—effectively relying on "lucky guesses." Comparing the three error types, models are more prone to Condition Interpretation Errors (CIE) and Deductive Reasoning Errors (DRE), while Logical Assumption Errors (LAE, or assumption hallucinations) are less frequent. Notably, the Gemini series maintains higher process quality, which fundamentally accounts for its leading overall scores in Table 3.

Do LMMs overthink during exams? Figure 5(a) compares the ARL performance of LMMs on the long-horizon reasoning subset with that on the overall dataset. We observe significant ARL score degradation across all models (ranging from -8 to -15), highlighting the reasoning complexity of this subset constructed via the strategy outlined in Section 3. Notably, Gemini-3-pro exhibits the smallest performance drop (-8.2) and achieves the highest score (74.8) on this subset, despite ranking only fourth on the overall dataset. This suggests that Gemini-3-pro is less prone to overthinking than Claude-Opus-4.6 when facing the peak complexity in reasoning tasks, achieving higher accuracy through a more efficient allocation of its reasoning budget.

Figure 5(b) illustrates the accuracy variation of models on the Long-Horizon Reasoning Set as the time limit (proportional threshold of context) decreases. This aims to assess the usability of LMMs under varying speed requirements. As the generation budget is restricted to 10% of the default length (i.e., 3.2k tokens), most models experience significant timeout failures, causing their accuracy to plummet to approximately half of their unconstrained performance. Qwen3-VL exhibits the smallest accuracy drop, indicating it can maintain robust capabilities even under stringent token budgets, making it highly suitable for scenarios requiring rapid responses. Conversely, Gemini-3-pro experiences the earliest and most precipitous decline in accuracy. This suggests that while it excels when allowed to deliberate extensively, its practical utility diminishes in fast-response applications.

Table 5: Accuracy of LLM judges relative to human-expert process-error annotations. Multi-LLM arbitration approaches the human-human upper bound on every metric while substantially outperforming a single-judge baseline.

Evaluator Setup	**Acc.**	**Prec.**	**Recall**	**F1**	**Cohen's κ**
Single LLM Evaluation (GPT-5)	82.0	81.7	84.1	83.5	0.82
Multi-LLM Arbitration (Ours)	**89.6**	**90.4**	**88.1**	**87.7**	**0.87**
Human-Human (Upper Bound)	*93.2*	*94.1*	*92.5*	*93.2*	*0.92*

Table 6: Stability of process-error evaluation on a 200-sample slice of the Rigorous Process Set. The arbitration panel is consistent across repeated runs, alternative judges, and different evaluation partitions.

Dimension	Metric	CIE	LAE	DRE	Overall
Intra-judge	Agreement rate	92.4%	95.5%	92.8%	**91.7%**
Inter-judge	Agreement rate	81.7%	85.9%	83.9%	**84.3%**
Cross-subset Ranking (Kendall's τ, mean±std)	By subject (4)		**0.901±0.025**		
	By type (3)		**0.939±0.049**		
	By difficulty (3)		**0.932±0.045**		
	Random half		**0.914±0.038**		

Is the LLM-as-judge process evaluation reliable? Since our Process Exam Score (PES) relies on a multi-LLM arbitration panel rather than a single judge, we validate it from two angles: agreement with human experts and stability across runs, judges, and data slices. For the former, two PhD students with strong K-12 STEM backgrounds independently annotate process errors on a stratified sample of the Rigorous Process Set as a human reference. As shown in Table 5, the arbitration panel raises Cohen's κ from 0.82 (single GPT-5 judge) to 0.87, recovering most of the gap to the 0.92 inter-human ceiling. We attribute this to three design choices: *reference-grounded* judgment against expert solutions, so only demonstrable contradictions with established facts are flagged; *root-cause focus* on path-independent error types (CIE / LAE / DRE), which removes bias from differing reasoning styles; and an *absolute penalty* by error count ($P_i = V_i - \tau \sum_k x_{i,k}$) rather than step-level accuracy, which removes the output-length confound. For stability, we probe three dimensions on a 200-sample slice of the Rigorous Process Set: *intra-judge* (one judge, 3 runs at temperature 0.6), *inter-judge* (Claude-Sonnet-4.6, Gemini-3-Flash, GPT-5; pairwise agreement), and *cross-subset ranking* (Kendall's τ of PES rankings across subject, type, difficulty slices, and 100 random half-splits over all 2,124 questions). As reported in Table 6, intra-judge repeatability exceeds 91%, inter-judge consensus exceeds 84%, and all four partitions yield $\tau > 0.9$, confirming that PES induces a stable model ordering robust to the evaluator instance and the evaluation slice.

4 Related Works

4.1 Scientific Reasoning Benchmarks for LMMs

The evaluation of Large Multimodal Models (LMMs) on scientific problems initially expanded from the discipline of mathematics to encompass broader scientific domains, including physics, chemistry, informatics and general science problems [8, 13, 15, 24, 28]. In particular, research focusing on high-school and pre-college levels has progressively evolved into capability assessments tailored for K-12 educational scenarios [11, 34, 37]. By benchmarking models against problems designed for human students, these studies serve as a natural testbed for their potential applications in the educational sector, such as intelligent tutoring systems and AI assistants [14].

While some existing benchmarks, such as K12Vista [16], MDK12Bench [36], MMSciBench [29], have proposed evaluation methodologies for K-12 level problems, they typically evaluate questions in isolation. Other exam-oriented benchmarks like M3Exam [33] and Exams-v [6] focus on multi-lingual evaluation. Human education, however, is heavily anchored in comprehensive examination systems that encapsulate massive educational resources and standardized curricula. Existing works lack a systematic evaluation of model capabilities within these authentic K-12 examination scenarios. In contrast, LiveK12Bench addresses this gap by providing a holistic and systematic examination-based evaluation protocol for LMMs.

The evaluation of reasoning capabilities has consistently been a pivotal driving force behind the rapid development of large reasoning models. From early text-based benchmarks like GSM8K [5] and MATH [12] to recent multimodal benchmarks evaluating visual reasoning such as MathVista, MathVision, etc. [18, 21, 27, 32], a crucial dimension among these is assessing whether a model can successfully extract valid and accurate information from visual inputs to construct the premises for subsequent logical deductions. Driven by real-world practical scenarios, we introduce a novel visual reasoning challenge: parsing problem information directly from the visually noisy layouts of raw examination pages, accurately comprehending image properties, and formulating the correct premises for reasoning.

5 Conclusion

In this paper, we introduce LiveK12Bench, a dynamic, comprehensive, and multi-disciplinary benchmark designed to authentically simulate human K-12 examinations. Powered by a highly efficient and automated data ingestion pipeline, our framework continuously exposes state-of-the-art LMMs to the recent and uncontaminated real-world exam questions. By restricting models to an end-to-end "Image-Only" input modality under time constraints, we systematically evaluate their abilities regarding visual robustness, process rigor, and reasoning efficiency. Utilizing multi-dimensional criteria modeled after professional educators' grading standards, we assign a holistic overall exam score to each evaluated model. Our evaluation reveals that even the most advanced models, including the GPT-5 series, exhibit substantial space for improvement when confronted with complex visual layouts and rigorous reasoning process assessments.

References

1. Achiam, J., Adler, S., Agarwal, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Anthropic: Claude 4.6 Opus (February 2026)
3. Anthropic: Claude 4.6 Sonnet (February 2026)
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-VL Technical Report. arXiv preprint arXiv:2511.21631 (2025)
5. Cobbe, K., Kosaraju, V., Bavarian, M., et al.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021)
6. Das, R., Hristov, S., Li, H., Dimitrov, D., Koychev, I., Nakov, P.: Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7768–7791 (2024)
7. DeepSeek-AI: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
8. Evans, M., Soós, D., Landers, E., Wu, J.: Msvec: A multidomain testing dataset for scientific claim verification. In: Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing. pp. 504–509 (2023)
9. Google DeepMind: Gemini 3 Flash (November 2025)
10. Google DeepMind: Gemini 3 Pro (November 2025)
11. He, C., Luo, R., Bai, Y., et al.: Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. arXiv preprint arXiv:2402.14008 (2024)
12. Hendrycks, D., Burns, C., Kadavath, S., et al.: Measuring mathematical problem solving with the math dataset. In: NeurIPS (2021)
13. Jassim, S., Holubar, M., Richter, A., Wolff, C., Ohmer, X., Bruni, E.: Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. arXiv preprint arXiv:2311.09048 (2023)
14. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al.: Chatgpt for good? on opportunities and challenges of large language models for education. Learning and individual differences **103**, 102274 (2023)
15. Li, C., Ye, X., Chen, S., Wei, W., Tang, X.: Mmscibench: Benchmarking language models on chinese multimodal scientific problems. In: Findings of ACL (2025)
16. Li, C., Zhu, C., Zhang, T., Lin, M., Zhou, Z., Xie, J.: K12vista: Exploring the boundaries of mllms in k-12 education. arXiv preprint arXiv:2506.01676 (2025)
17. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In: The Twelfth International Conference on Learning Representations (2024)
18. Lu, P., Bansal, H., Xia, T., et al.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In: International Conference on Learning Representations (ICLR) (2024)

19. OpenAI: Gpt-5-mini (August 2025)
20. OpenAI: Introducing gpt-5 (August 2025)
21. Qiao, R., Tan, Q., Dong, G., MinhuiWu, M., Sun, C., Song, X., Wang, J., GongQue, Z., Lei, S., Zhang, Y., et al.: We-Math: Does your large multimodal model achieve human-like mathematical reasoning? In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 20023–20070 (2025)
22. Qiao, R., Tan, Q., Yang, P., Wang, Y., Wang, X., Wan, E., Zhou, S., Dong, G., Zeng, Y., Xu, Y., et al.: We-Math 2.0: A versatile mathbook system for incentivizing visual mathematical reasoning. arXiv preprint arXiv:2508.10433 (2025)
23. Sainz, O., Campos, J.A., García-Ferrero, I., et al.: Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. arXiv preprint arXiv:2310.18018 (2023)
24. Tarsi, T., Adel, H., Metzen, J.H., Zhang, D., Finco, M., Friedrich, A.: Sciol and mulms-img: Introducing a large-scale multimodal scientific dataset and models for image-text tasks in the scientific domain. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4560–4571 (2024)
25. Team, K., Bai, T., Bai, Y., Bao, Y., Cai, S., Cao, Y., Charles, Y., Che, H., Chen, C., Chen, G., et al.: Kimi k2. 5: Visual agentic intelligence. arXiv preprint arXiv:2602.02276 (2026)
26. Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F., et al.: Mineru: An open-source solution for precise document content extraction. arXiv preprint arXiv:2409.18839 (2024)
27. Wang, K., Pan, J., Shi, W., Lu, Z., Ren, H., Zhou, A., Zhan, M., Li, H.: Measuring multimodal mathematical reasoning with MATH-Vision dataset. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024)
28. Wang, X., Hu, Z., Lu, P., Zhu, Y., Zhang, J., Subramaniam, S., Loomba, A.R., Zhang, S., Sun, Y., Wang, W.: Scibench: Evaluating college-level scientific problem-solving abilities of large language models. arXiv preprint arXiv:2307.10635 (2023)
29. Ye, X., Li, C., Chen, S., Wei, W., Tang, R.: Mmscibench: Benchmarking language models on chinese multimodal scientific problems. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 14621–14663 (2025)
30. Ye, Y., Xiao, Y., Mi, T., Liu, P.: Aime-preview: A rigorous and immediate evaluation framework for advanced mathematical reasoning. `https://github.com/GAIR-NLP/AIME-Preview` (2025), gitHub repository
31. Zeng, A., Lv, X., Hou, Z., Du, Z., Zheng, Q., Chen, B., Yin, D., Ge, C., Xie, C., Wang, C., et al.: Glm-5: from vibe coding to agentic engineering. arXiv preprint arXiv:2602.15763 (2026)
32. Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.W., Qiao, Y., et al.: MathVerse: Does your multi-modal llm truly see the diagrams in visual math problems? In: European Conference on Computer Vision. pp. 169–186 (2024)
33. Zhang, W., Aljunied, M., Gao, C., et al.: M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In: NeurIPS (2023)
34. Zhang, X., Li, C., Zong, Y., et al.: Evaluating the performance of large language models on gaokao benchmark. arXiv preprint arXiv:2305.12474 (2023)
35. Zhang, Y., Math-AI, T.: American invitational mathematics examination (aime) 2025 (2025)

36. Zhou, P., Zhang, F., Peng, X., Xu, Z., Ai, J., Qiu, Y., Li, C., Li, Z., Li, M., Feng, Y., et al.: Mdk12-bench: a multi-discipline benchmark for evaluating reasoning in multimodal large language models. arXiv preprint arXiv:2504.05782 (2025)
37. Zong, Y., Qiu, X.: Gaokao-mm: A chinese human-level benchmark for multimodal models evaluation. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 8817–8825 (2024)