

# mmJEE-Eval: A Bilingual Multimodal Benchmark for Evaluating Scientific Reasoning in Vision-Language Models

**Arka Mukherjee\***  
Kalinga Institute of  
Industrial Technology (KIIT)  
arka.mukherjee078@gmail.com

**Shreya Ghosh**  
Indian Institute of Technology  
(IIT), Bhubaneswar  
shreya@iitbbs.ac.in

## Abstract

Contemporary vision-language models (VLMs) perform well on existing multimodal reasoning benchmarks (78-85% accuracy on MMMU, MathVista). Yet, these results fail to sufficiently distinguish true scientific reasoning articulation capabilities from pattern-matching. To address this gap, we introduce **mmJEE-Eval**, a multimodal bilingual (English and Hindi) benchmark comprising 1,460 questions from India's JEE Advanced examination (2019-2025) spanning pre-college Physics, Chemistry, and Mathematics domains. Our evaluation of 17 state-of-the-art models reveals that while frontier VLMs (GPT-5, Gemini 2.5 Pro/Flash) achieve 77-84% accuracy on held-out 2025 questions, open-source models plateau at 37-45% despite scaling to 400B parameters, a significant difference not observed on existing benchmarks. While closed frontiers from Google and OpenAI show high problem-solving accuracies (up to 100% pass@3 scores), they fully collapse when the reasoning load is increased meta-cognitively (GPT-5 fixes just 5.2% errors). Systematic ablations show mmJEE-Eval's difficulty stems from complexity and reasoning depth rather than memorization. Effectively, our benchmark segregates superior training and reasoning methodologies where alternatives fail. We publicly release our code and data: <https://mmjee-eval.github.io>

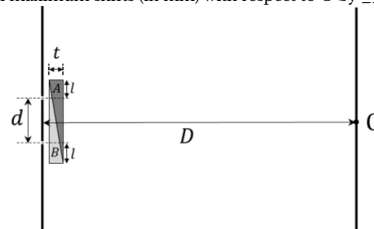
## 1 Introduction

Recent advances in vision-language models (VLMs) have produced systems that perform impressively on broad multimodal leaderboards. To evaluate these systems, numerous benchmarks exist: MMMU (Yue et al., 2024), MathVista (Lu et al., 2024), MMMU-Pro (Yue et al., 2025), and CharXiv (Wang et al., 2024) test multimodal reasoning across diverse domains.

\*Work done while at IIT Bhubaneswar

### JEE Advanced 2025 Physics Problem:

In a Young's double slit experiment, a combination of two glass wedges A and B, having refractive indices 1.7 and 1.5, respectively, are placed in front of the slits, as shown in the figure. The separation between the slits is  $d = 2$  mm and the shortest distance between the slits and the screen is  $D = 2$  m. Thickness of the combination of the wedges is  $t = 12$   $\mu$ m. The value of  $l$  as shown in the figure is 1 mm. Neglect any refraction effect at the slanted interface of the wedges. Due to the combination of the wedges, the central maximum shifts (in mm) with respect to O by \_\_\_\_.



**CORRECT ANSWER: 1.2**



### OpenAI o3:

The additional optical path introduced is only due to the different refractive indices, because each slit's beam traverses the **same physical thickness**  $t$  of glass. *Wrong assumption*

Optical path difference  
 $\Delta = (n_A - n_B) t$   
 $= (1.7 - 1.5) \times 12 \mu\text{m}$   
 $= 0.2 \times 12 \times 10^{-6} \text{ mm}$   
 $= 2.4 \times 10^{-6} \text{ mm} = 2.4 \mu\text{m}$  *Wrong formula usage*

For the central maximum to re-appear, the geometrical path difference at the new point must cancel this:  
 $dy / D = \Delta = y = D \Delta / d$

Given:  $D = 2 \text{ m} = 2000 \text{ mm}$ ;  $d = 2 \text{ mm}$

$y = 2000 \text{ mm} \times 2.4 \mu\text{m} / 2 \text{ mm}$   
 $= 2000 \times 2.4 \times 10^{-6} \text{ mm} / 2$   
 $= 2.4 \text{ mm}$

So the central maximum shifts by  $\boxed{2.4}$  *Wrong answer*

Figure 1: Example problem and response from mmJEE-Eval. Despite mathematical correctness, the model incorrectly assumes uniform thickness ("same physical thickness  $t$  of glass") when the figure clearly shows wedge-shaped glass pieces with varying thickness. This multimodal reasoning failure demonstrates the multiple dimensions our proposed benchmark tests.

These benchmarks show that recent open-source VLMs are generally within single-digit to low-double-digit points of leading closed systems, and occasionally surpass them. On MMMU, Qwen3 VL 235B (78.7%) and Llama 4 Maverick (73.4%) trail Gemini 2.5 Pro (84.0%) and GPT-5 (84.2%) by modest margins. MMMU-Pro shows a slightly wider trend (68.1% vs 78.4%). However, on Math-

Vista, Qwen3 VL 235B achieves 84.9%, edging out Gemini 2.5 Pro’s 84.6% and GPT-5’s 82.7%. Taken together, these results suggest that while divergence slightly widens on multimodal, multi-step compositional reasoning, the gap between open and closed models has narrowed.

Recent efforts like Humanity’s Last Exam (HLE) (Phan et al., 2025) show that pure difficulty widens the difference. While GPT-5 (25.3%), Gemini 2.5 Pro (21.6%), and o3 (20.3%) cluster below 26%, open models score 5-12%. Yet such uniformly low scores provide limited insight into *why* models struggle or which capabilities remain deficient.

To address this, we introduce mmJEE-Eval, a multimodal bilingual benchmark built from seven years (2019-2025) of India’s JEE Advanced examination, a rigorous high school-level entrance test for India’s premier technical institutions, the IITs. JEE Advanced provides several properties valuable for evaluation: exam-style constraints including negative marking for incorrect responses; native bilingual presentation with questions appearing in both English and Hindi (as in the original exam booklets); integrated STEM reasoning across physics, chemistry, and mathematics; and contamination avoidance through annual new questions, provided with our 2025 held-out set (Section 4.5). While individual properties appear in prior work, their combination enables systematic diagnosis of failure modes. Figure 1 illustrates such a multimodal reasoning failure from OpenAI o3, where the model is mathematically correct but misinterprets the visual setup.

To further understand such multimodal reasoning gaps, we answer the following research questions in this study:

**RQ1:** Does mmJEE-Eval’s exam-style multimodal evaluation reveal additional VLM capability gaps compared to standard benchmarks?

**RQ2:** Are VLMs resilient to semantically equivalent input language perturbations?

**RQ3:** Can contemporary VLMs detect and fix errors in their reasoning chains?

**RQ4:** What factors make mmJEE-Eval difficult?

Despite covering pre-college material, mmJEE-Eval reveals dramatic performance gaps absent in prior benchmarks. Through comprehensive evaluation of 17 VLMs (Section 4.1), we show Llama 4 Maverick scores 32.8% while Gemini 2.5 Flash achieves 83.3% on realistic exam marking (a 50.5-point gap), far larger than on MMMU, MMMU-

Pro, and MathVista. This disparity suggests that *problem complexity alone does not limit reasoning capabilities*. Through rigorous experiments and ablations, we attribute this gap to an absence of joint multilingual understanding, complex visual reasoning, and cross-domain concept integration mastery (Section 4.6).

Beyond measuring problem-solving accuracy, we also probe whether models can detect and correct their own reasoning errors (Section 4.2). Here, we find universal metacognitive failure: both open and closed models detect errors in 21-73% of cases but correct only 1.1-5.2% of them, far below the 30-40% typical pass@k gains. This suggests contemporary VLMs lack meta-cognitive capabilities for self-correction, even when they eventually get problems correct, given multiple attempts.

## 2 Related Work

### 2.1 General Reasoning Datasets

Mathematical reasoning benchmarks have been the standard for evaluating LLM capabilities. MATH (Hendrycks et al., 2021) introduced 12,500 challenging competition mathematics problems from AMC, AIME, and Mathematical Olympiads. GSM8K (Cobbe et al., 2021) focused on 8,500 linguistically diverse grade school math word problems. These early datasets are English only, failing to capture the global scale of modern VLMs.

Recent efforts have scaled both in sophistication and LLM trends, such as instruction tuning. MathInstruct (Yue et al., 2023) compiled 260,000 problems from 13 math datasets with intermediate rationales using hybrid Program-of-Thought and Chain-of-Thought approaches. MetaMathQA (Yu et al., 2024) bootstrapped 395,000 mathematical questions using GPT-3.5 Turbo-generated rephrasings to address the performance gap between open-source and closed-source models. However, these datasets primarily focus on computational mathematics rather than the advanced conceptual reasoning required in competitive examinations.

### 2.2 Multilingual Reasoning Datasets

Multilingual mathematical reasoning has emerged as a critical research direction, though most efforts rely on translation-based approaches. Notable datasets include X-CSQA and X-CODAH (Lin et al., 2021), which automatically translated English commonsense reasoning datasets to 16 languages using DeepL Pro. XCOPA (Ponti et al.,

2020) manually translated commonsense reasoning tasks to 11 typologically diverse languages.

More sophisticated approaches include mC-SQA (Sakai et al., 2024), which proposed efficient methods to create multilingual NLU datasets using multilingual generative language models for cross-linguistic commonsense understanding analysis. Further, TyDiQA (Clark et al., 2020) collected question answering data directly in 11 typologically diverse languages without translation, setting the gold standard for multilingual evaluation.

### 3 Data Collection

mmJEE-Eval is collected and verified through a rigorous human annotation process to ensure high-quality data that is sufficiently challenging for state-of-the-art VLMs. Following Arora et al. (2023), we collect JEE Advanced questions from the official archive<sup>1</sup>. After preprocessing, our multimodal dataset comprises 1,460 questions spanning seven editions of the exam (2019–2025).

#### 3.1 Dataset Collection and Annotation Process

Given the inherent complexity of extracting structured information from multimodal PDF documents, our initial attempts using automated PDF parsing tools proved inadequate. We also evaluated automated bounding box detection techniques, including DART (Xin et al., 2024), which failed to handle challenging question types such as matching problems and exhibited poor generalization across years due to formatting variations in the question papers.

To ensure accurate collection of multimodal data from JEE Advanced PDFs, we developed custom manual annotation tools with multiple verification steps for both question and solution collection (detailed in Appendix D).

Our annotation software automatically downloads all question papers and presents them to annotators in a page-by-page interface. The annotation task requires drawing precise bounding boxes around individual questions. We manually processed all 24 question papers, with particular attention to ensuring question images remain identifiable and complete. This process yielded an initial collection of 1,476 question images.

Collecting accurate ground-truth answers for all questions in our dataset is essential for proper eval-

uation. However, we discovered that not all JEE Advanced question papers are accompanied by official answer keys. Given that comprehensive solutions have been archived by third-party educational websites, we implemented a multi-source verification approach by collecting answer keys from five independent nation-wide popular coaching institutes: Aakash<sup>2</sup>, FIITJEE<sup>3</sup>, Allen Kota<sup>4</sup>, Motion Education<sup>5</sup>, and Resonance<sup>6</sup>.

These institutes annually publish detailed solutions for all JEE Advanced papers to assist students in performance assessment, with all analyses carefully archived on their respective websites for current aspirants. We collect six independent answers per question and employ majority voting to determine the ground-truth solution. Solutions receiving greater than 60% consensus are accepted into our dataset. Questions failing to achieve a majority consensus are flagged for manual verification. This discrepancy most commonly occurred in numerical-type questions, where official answer keys accept ranges of values while independent solutions provide singular values or narrower ranges. Only 30 such cases were recorded out of 1,476, for which we deferred to the official answer key when available. Table 1 presents an analysis of confidence levels recorded in the answer collection process.

Consensus Level	Questions	Percentage
100%	1326	90.8
80% – 99.99%	68	4.7
60% – 79.99%	36	2.4
50% – 59.99%	26	1.8
<50%	4	0.3

Table 1: Consensus level statistics for manually collected answers for mmJEE-Eval from six independent sources.

#### 3.2 Data Statistics

Finally, mmJEE-Eval comprises 1,460 questions spanning seven years (2019–2025). As detailed in

<sup>2</sup><https://www.aakash.ac.in/jee-advanced-previous-year-question-papers>

<sup>3</sup>[https://old.fiitjee.com/downloads/jee-advanced-\(earlier-iit-jee\)-solutions](https://old.fiitjee.com/downloads/jee-advanced-(earlier-iit-jee)-solutions)

<sup>4</sup><https://allen.in/jee-advanced/previous-year-papers-with-solutions>

<sup>5</sup><https://motion.ac.in/blog/jee-advanced-previous-year-question-papers>

<sup>6</sup><https://www.resonance.ac.in/answer-key-solutions/JEE-Advanced.aspx>

<sup>1</sup><https://jeeadv.ac.in/archive.html>

Table 2, the dataset maintains a balanced distribution across all three core subjects: Chemistry (492 questions, 33.7%), Mathematics (492 questions, 33.7%), and Physics (476 questions, 32.6%).

Year	Subjects			Question Types				Total
	Chem	Math	Phys	Multi	Single	Num.	Match.	
2019	72	72	72	96	48	72	24	216
2020	72	72	68	72	36	104	0	212
2021	76	76	76	72	48	108	0	228
2022	72	72	66	70	44	96	22	210
2023	68	68	64	24	84	92	24	200
2024	68	68	68	36	72	96	24	204
2025	64	64	62	42	64	84	18	190
<b>Total</b>	<b>492</b>	<b>492</b>	<b>476</b>	<b>412</b>	<b>396</b>	<b>652</b>	<b>112</b>	<b>1460</b>

Table 2: Year-wise distribution of questions by subject and question type. **Num.:** Numerical-type questions, **Match.:** Matching-type questions.

Question type analysis shows that Numerical questions constitute the largest category with 652 instances (44.7%), followed by MCQ-Multiple with 412 questions (28.2%) and MCQ-Single with 396 questions (27.1%). This distribution aligns with the JEE Advanced examination pattern, where numerical answer questions play a prominent role in assessing problem-solving capabilities.

Additionally, 442 questions (30.3%) mandate visual interpretation of mathematical diagrams, chemical structures, or physical illustrations, reflecting the multimodal nature of our task (termed "Image Required"). For the remainder, images are optional, albeit still fed in as screenshots to test the models' OCR capabilities (termed "Image Optional").

## 4 Results

We evaluate 17 state-of-the-art VLMs on mmJEE-Eval: sub-10B models (Aya Vision 8B, Qwen 2.5 VL 7B (Bai et al., 2025), InternVL3 8B (Zhu et al., 2025)), sub-50B models (Kimi VL Thinking 2506 16B, InternVL3.5 14B/30B, Gemma3 27B (Team et al., 2025)), sub-200B models (InternVL3 78B, Llama4 Scout 109B), and large open models (Qwen3 VL 235B, Llama4 Maverick 400B). For closed-source models, we test OpenAI o3<sup>7</sup>, Gemini 2.5 Pro/Flash, GPT-5/GPT-5-mini, and Grok 4 Fast (July-October 2025 API access). All experiments used Nvidia RTX 5080/5090 GPUs. Prompts are detailed in Appendix C.

<sup>7</sup><https://openai.com/index/introducing-o3-and-o4-mini>

### 4.1 Performance analysis of contemporary VLMs

Table 3 presents the Pass@1 accuracy averaged over  $k = 10$  runs (Appendix B justifies this choice) of all evaluated models across subjects and question types (RQ1). Our motivation for selecting Pass@1 is two-fold: (a) OpenAI Simple Evals<sup>8</sup> promotes it, and (b) allows fair comparison with other industry standard vision benchmarks.

Open-source models reach 9-51% on the 2025 held-out set while frontier models achieve 73-80%, a gap of 28-37 percentage points. This frontier advantage holds across all benchmarks tested, but mmJEE-Eval avoids the ceiling effects that compress differences elsewhere: Qwen3 VL scores 85% on MathVista but only 45% on mmJEE-Eval.

**Confidence Thresholding.** Following Arora et al. (2023), we apply confidence-based selective answering to JEE Advanced 2025's 190 MCQs. For single-correct and multiple-correct MCQs, we implement self-consistency (Wang et al., 2022) and generate responses in 10 attempts to measure consistency. Appendix A details ideal threshold measurement. Table 3 shows that weak models benefit dramatically (Qwen 2.5 VL: +117%, InternVL3 8B: +157%), while strong models improve modestly (Gemini 2.5 Pro: +12%, Flash: +0%).

**Human baselines.** Since confidence thresholding to avoid negative marking is part of the human model for JEE Advanced, and only total scores have been released for the exam, we report human baselines in the **Marks w/ CT (%)** column of Table 3. Open models and Grok 4 Fast score below the 50th percentile of human test-takers, with most failing the mmJEE-Eval test (scores below qualifying cutoffs). In contrast, we notice a pattern in closed-frontier models from Google and OpenAI: Gemini Flash/Pro, o3, GPT-5/Mini exceed the 90th percentile. Notably, Gemini 2.5 Flash achieves 300/360 marks (83.3%) despite being a light model, placing 32 points below the human topper (332/360, 92.2%).

### 4.2 Metacognitive Self-Correction

We evaluate whether VLMs can detect and correct their reasoning errors (RQ3) using error presence detection (EP), and error correction (EC) following Li et al. (2024). This accurately captures how humans behave in a realistic exam setting, where students first come up with an initial solution, then

<sup>8</sup><https://github.com/openai/simple-evals>



Model	mmJEE-Eval (Ours)				Other Industry Standard Benchmarks			
	Acc. % (Full)	Acc. % (2025 set)	Marks (%)	Marks w/ CT (%)	MMMU	MMMU Pro	Math Vista	Char Xiv
Random choice	8.3 $\pm$ 0.3	9.3 $\pm$ 1.6	-46 (-12.7%)	-7 (-1.9%)	22.1%	12.6%	—	10.8%
Aya Vision 8B	8.5 $\pm$ 0.6	9.2 $\pm$ 1.0	-45 (-12.5%)	-16 (-4.4%)	39.9%	—	—	—
Kimi VL Thinking 2506 16B	9.8 $\pm$ 0.2	11.2 $\pm$ 2.7	-16 (-4.44%)	-11 (-3.06%)	64.0%	46.3%	68.7%	—
Qwen 2.5 VL 7B	10.9 $\pm$ 0.8	12.4 $\pm$ 1.9	-21 (-5.83%)	5 (1.4%)	58.6%	46.2%	68.2%	42.5%
InternVL3 8B	10.8 $\pm$ 0.5	12.5 $\pm$ 1.3	-22 (-6.1%)	13 (3.6%)	62.7%	—	71.6%	37.6%
InternVL 3.5 14B	15.6 $\pm$ 0.9	16.1 $\pm$ 1.6	18 (5.0%)	26 (7.2%)	73.3%	—	80.5%	—
InternVL 3.5 30B	18.8 $\pm$ 1.2	18.8 $\pm$ 1.1	22 (6.1%)	26 (7.2%)	75.6%	—	80.9%	—
Grok 4 Fast	19.3 $\pm$ 1.1	17.6 $\pm$ 1.5	27 (7.5%)	40 (11.1%)	—	—	—	—
InternVL3 78B	22.0 $\pm$ 0.8	23.6 $\pm$ 3.4	30 (8.33%)	44 (12.22%)	72.2%	—	79.0%	46.0%
Human (qualifying cutoff)	—	—	—	74 (20.6%)	76.2%	73%	—	—
Gemma 3 27B	29.8 $\pm$ 0.3	30.6 $\pm$ 0.7	67 (18.6%)	79 (21.9%)	64.9%	—	63.3%	—
Llama4 Scout 109B	40.4 $\pm$ 0.5	37.1 $\pm$ 2.3	101 (28.1%)	122 (33.9%)	69.4%	—	70.7%	—
Qwen3 VL 235B Instruct	44.6 $\pm$ 0.4	46.2 $\pm$ 4.9	130 (36.1%)	134 (37.2%)	78.7%	68.1%	84.9%	—
Llama4 Maverick 400B	50.9 $\pm$ 0.3	43.8 $\pm$ 2.8	118 (32.8%)	129 (35.8%)	73.4%	—	73.7%	—
OpenAI o3	77.4 $\pm$ 0.7	72.7 $\pm$ 3.7	240 (66.7%)	241 (66.9%)	82.9%	76.4%	81.1%	78.6%
GPT-5-mini ‘Medium’	75.3 $\pm$ 0.7	71.3 $\pm$ 6.7	255 (70.8%)	259 (71.94%)	80.0%	—	—	—
Gemini 2.5 Pro	81.2 $\pm$ 0.7	77.4 $\pm$ 0.8	254 (70.6%)	285 (79.2%)	84.0%	71.2%	84.6%	—
Gemini 2.5 Flash 09-2025	82.6 $\pm$ 0.4	79.8 $\pm$ 1.2	300 (83.3%)	300 (83.3%)	79.7%	—	81.2%	—
GPT-5	83.9 $\pm$ 0.6	79.5 $\pm$ 1.5	291 (80.8%)	295 (81.9%)	84.2%	78.4%	82.7%	81.1%
Human (Top 10 performers)	—	—	—	324 (90.1%)	82.6%	80.8%	—	—
Human (Rank 1, Topper)	—	—	—	332 (92.2%)	88.6%	85.4%	60.3%	80.5%
$\Delta$ over human (Human – Best model)	—	—	—	10.1%	4.4%	7%	-24.3%	-0.6%

Table 3: Comparison of model performance on mmJEE-Eval (ours) and other benchmarks, MMMU, MMMU Pro, Math Vista, and CharXiv. For mmJEE-Eval, **Acc. % (Full)** represents Pass@1 accuracy on the full set of 1,460 questions, **Acc. % (2025 set)** presents Pass@1 accuracy on the held-out 2025 subset, **Marks (%)** represents total score on the two papers of JEE Advanced 2025 following the official marking scheme, **Marks w/ CT (%)** presents confidence thresholded scores. For other benchmarks, we source Pass@1 accuracies from respective leaderboards.

detect errors, and finally fix them. Table 4 compares EP/EC against pass@k sampling. Models, both open and frontier, show 2.2-6.7% improvement in total score, much lower than the 30.4% improvement with Pass@3 on average.

Most strikingly, on average, models detect errors in 52.9% of cases but correct only 8% of those detected. The EP→EC conversion rate averages 3.5% across open models. GPT-5 exhibits an anomaly where it scores only 21.9% in EP (worst result) but 20.8% in EC (best result) simultaneously. However, when chained together, overall correction capabilities degrade (5.2% overall EP→EC accuracy). For some context, the model logged a 14.6% pass@3 improvement. This suggests that while models can eventually get questions correct with sufficient retries, they fail to do so when presented with previous erroneous generations. This is a metacognitive limitation in current-generation models, regardless of size, that mmJEE-Eval uniquely tests compared to standard VLM reasoning benchmarks.

### 4.3 Cross-Lingual Consistency

Table 5 shows model consistency on English-Hindi question pairs from JEE Advanced 2025 by evaluating model responses to identical questions in both Hindi and English, and Figure 2 shows how average performance varies by language (RQ2). We categorize outcomes into four cases: (a) both Hindi and English responses are correct, (b) both are incorrect, (c) only the English response is correct, and (d) only the Hindi response is correct. Frontier models maintain both languages correctly 67-79% of the time, while OSS models achieve only 2-36%.

We also note higher language-specific failure (Eng+/Hin- or Eng-/Hin+) rates on open models, hinting at English-dominant training (Llama4 models show 17-19% Eng+/Hin- but only 9-10% Eng-/Hin+). This asymmetry is more subtle in frontier models (GPT-5: 7.0% vs 3.6%, o3: 9.99% vs 8.37%). Interestingly, Gemini 2.5 Pro scores higher on Eng-/Hin+ (10.64%) than Eng+/Hin- (6.93%), reversing the trend, while Flash maintains the typical English advantage (7.10% vs 5.47%). We

Model	Pass@1	EP(%)	EC(%)	EP→EC (%)		Pass@1 w/ EP/EC	Pass@3	Pass@5
				EP	EC			
<b>Qwen2.5 VL 7B</b>	10.6%	66.10	1.86	73.23	1.06	12.2%	41.2%	52.9%
<b>InternVL3.5 30B</b>	18.8%	79.50	4.96	40.55	4.10	21.9%	44.1%	64.7%
<b>InternVL3 78B</b>	22.2%	30.30	8.51	28.87	1.86	25.8%	44.1%	67.6%
<b>Gemma3 27B</b>	29.8%	47.16	6.45	47.08	5.05	34.2%	64.7%	85.3%
<b>Llama4 Scout</b>	40.4%	72.57	5.13	72.91	3.47	43.5%	79.4%	88.2%
<b>GPT-5</b>	83.9%	21.90	20.77	19.89	5.20	86.3%	100.0%	100.0%

Table 4: Performance comparison of various models across Error Presence (EP) and Error Correction (EC) steps. In EP→EC, we chain the two steps in the same chat instance, triggering EC only if the model successfully detects an error in its previous erroneous instance. We note that performance improvements with EP/EC remain significantly behind those with Pass@k, hinting at severe metacognition gaps in contemporary VLMs.

attribute this Pro-specific gain to a mix of larger model size and Gemini 2.5’s 400-language pretraining, where specific gains have been internally noted in Indic languages by the Gemini team (Comanici et al., 2025).

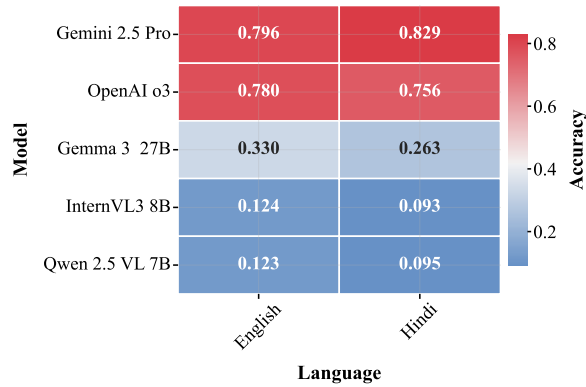


Figure 2: Heatmap comparing accuracy across languages (English, Hindi) for four language models. Gemini 2.5 Pro achieves consistently high performance (>0.80), while smaller models show uniformly low performance (< 0.10). Color scale represents accuracy from 0.1 (light blue) to 0.8+ (dark red).

#### 4.4 Source of error analysis

To understand which factors contribute to mmJEE-Eval’s higher difficulty (RQ4), we review 400 model outputs through a hybrid LLM-as-a-Judge (LaaJ) and human-in-the-loop verification pipeline across five representative models using a Bloom’s taxonomy-inspired<sup>9</sup> framework (Jiang et al., 2024; Arora et al., 2023). LaaJ was implemented as an additional evaluator with Claude 4.5 Sonnet to avoid overlaps with the tested models. It first rates responses on four dimensions (0-10 scale): conceptual understanding, grounding in problem context, computational accuracy, and instruction following, which were designed in accordance with the first

<sup>9</sup><https://teaching.cornell.edu/resource/blooms-taxonomy>

four Bloom’s taxonomy levels (Remember, Understand, Apply, and Analyze). Finally, the authors review the assigned scores and reasoning chains before accepting or rejecting an annotation.

Table 6 shows Llama4 Scout 109B (40.4% accuracy) and GPT-5 (83.9% accuracy) score nearly identically in overall reasoning quality (7.07 vs 7.08). Scout matches or exceeds GPT-5 on conceptual errors (7.43 vs 6.9), grounding (7.71 vs 6.8), and computation (7.29 vs 6.9), but falls substantially behind on instruction following (5.86 vs 7.7). We conclude that accuracy gaps don’t directly correlate with better reasoning quality.

As an illustration, Figure 1 shows a representative error from OpenAI o3, where the model demonstrates correct mathematical reasoning but incorrectly assumes "uniform thickness" for wedge-shaped glass pieces despite the figure showing otherwise. This leads to an incorrect answer despite correct subsequent steps, exemplifying the challenge put forward by mmJEE-Eval (RQ3).

#### 4.5 Dataset contamination tests

Table 7 compares performance on 2019-2024 questions (potentially seen during training) versus the 2025 held-out set (exam conducted May 18, 2025; answers released June 2, 2025). Open models show stable or slightly improved performance on 2025 (+0.2% to +1.9%), likely reflecting sample variance given N=190. Frontier models drop 2-5%: GPT-5 (-5.3%), Gemini 2.5 Pro (-4.3%), OpenAI o3 (-4.7%), keeping training set overlap for 2019-2024 public exam papers modest.

Critically, models released after June 2, 2025 (Qwen3 VL, Grok 4 Fast) show no advantage over earlier models, meaning answer keys did not leak into training. The 28-37 point frontier-open gap persists on 2025 data, further proving that mmJEE-Eval is not a memorization artifact. With every year, fresh JEE Advanced questions can be aug-

Model Name	Both Correct (%)	Both Incorrect (%)	Eng+, Hin- (%)	Eng-, Hin+ (%)
Random Baseline	1.67 $\pm$ 0.36	84.89 $\pm$ 0.89	6.58 $\pm$ 0.53	6.86 $\pm$ 0.66
Qwen 2.5 VL 7B	2.87 $\pm$ 0.51	81.14 $\pm$ 1.34	9.35 $\pm$ 0.69	6.64 $\pm$ 1.03
InternVL 3.5 30B	4.41 $\pm$ 0.79	66.80 $\pm$ 2.22	22.81 $\pm$ 1.84	5.98 $\pm$ 1.36
InternVL 3.5 14B	4.80 $\pm$ 1.01	73.69 $\pm$ 1.55	15.20 $\pm$ 1.71	6.31 $\pm$ 0.53
Grok 4 Fast	5.58 $\pm$ 0.60	67.53 $\pm$ 1.43	21.07 $\pm$ 0.88	5.81 $\pm$ 0.53
InternVL3 78B	7.22 $\pm$ 1.33	62.78 $\pm$ 1.37	21.98 $\pm$ 1.34	8.02 $\pm$ 1.32
Gemma 3 27B	16.42 $\pm$ 0.28	56.67 $\pm$ 0.64	16.85 $\pm$ 0.77	10.07 $\pm$ 0.65
Llama4 Scout 109B	27.31 $\pm$ 0.98	46.48 $\pm$ 0.62	17.03 $\pm$ 1.26	9.18 $\pm$ 0.82
Llama4 Maverick 400B	36.26 $\pm$ 0.24	34.25 $\pm$ 0.46	19.47 $\pm$ 0.96	10.02 $\pm$ 0.83
OpenAI o3	72.24 $\pm$ 0.47	12.22 $\pm$ 0.64	9.14 $\pm$ 0.47	6.39 $\pm$ 0.36
Gemini 2.5 Flash	75.96 $\pm$ 0.91	11.34 $\pm$ 0.45	7.00 $\pm$ 0.89	5.70 $\pm$ 0.63
Gemini 2.5 Pro	72.60 $\pm$ 1.83	9.91 $\pm$ 0.87	6.76 $\pm$ 1.10	10.73 $\pm$ 0.95
GPT-5	79.60 $\pm$ 1.16	9.44 $\pm$ 2.37	7.25 $\pm$ 1.90	3.71 $\pm$ 0.46

Table 5: Model performance comparison across languages in the 2025 edition of JEE Advanced. **Eng+, Hin-**: English correct but Hindi incorrect; **Eng-, Hin+**: English incorrect but Hindi correct.

Model	Conceptual Errors	Grounding Errors	Computation Errors	Instruction Errors	Average
Grok 4 Fast	4.2 $\pm$ 1.92	4.0 $\pm$ 1.73	3.8 $\pm$ 1.79	5.4 $\pm$ 1.52	4.35
InternVL3 78B	5.2 $\pm$ 0.84	4.6 $\pm$ 1.14	4.8 $\pm$ 1.1	7.80 $\pm$ 0.45	5.6
Llama4 Maverick 400B	5.83 $\pm$ 1.33	5.83 $\pm$ 1.83	6.17 $\pm$ 2.14	6.0 $\pm$ 1.55	5.96
Llama4 Scout 109B	7.43 $\pm$ 1.9	7.71 $\pm$ 1.8	7.29 $\pm$ 2.14	5.86 $\pm$ 1.07	7.07
GPT-5	6.9 $\pm$ 2.38	6.8 $\pm$ 2.44	6.9 $\pm$ 2.42	7.7 $\pm$ 1.16	7.08

Table 6: Source of error analysis results of our hybrid LaaJ and human-in-the-loop study on 400 model outputs of a few representative models. All values are reported on a scale of 0-10, with higher being better.

mented to mmJEE-Eval following our methodology (Section 3), limiting contamination concerns.

#### 4.6 Ablations

First, we isolate visual and textual components by testing four conditions: (1) original images with diagrams, (2) EasyOCR text extraction (known to corrupt complex mathematical notation), (3) Gemma3 27B-based OCR (improved symbolic parsing), and (4) cropped images removing all diagrams. Table 8 presents results on the 2025 held-out set.

Frontier models exhibit severe OCR sensitivity: Gemini 2.5 Flash drops 48.5 points with corrupted EasyOCR (79.5% $\rightarrow$ 31.1%), recovering 26.3 points with improved Gemma3 OCR (57.4%), then regaining an additional 8.9 points when diagrams are restored (66.3%). GPT-5 shows similar degradation. This pattern indicates SoTA models don’t simply pattern-match text. Open models exhibit opposite behavior: while EasyOCR corruption drops accuracy, they fully recover with better OCR, with minimal diagram dependence. This suggests open models rely primarily on textual patterns rather than visual-symbolic integration, explaining the ceiling observed in Table 3.

Even with diagrams removed, Flash achieves 66.3% while Maverick reaches 40.5% and Scout 31%. Since these are all MoEs with  $\sim$ 17B active parameters, we hypothesize that training methodol-

ogy supersedes vision capabilities.

#### 4.7 Discussion

Among our results, three insights about contemporary VLM evaluation and capabilities stand out.

First, **benchmark design matters**. In mmJEE-Eval, we observe a near 50-point performance difference between closed OpenAI and Google frontiers (66.7-83.3% total marks) and the next-best OSS VLMs (28.1-36.1% total marks), a separation that does not appear on MathVista or MMMU (RQ1). Through systematic ablations, we attribute this difference to the exam-style evaluation design, multimodal question setup, and bilingual inputs, not just problem complexity (RQ4). Relatedly, OSS models show brittleness to bilingual perturbations, with Llama 4 Maverick is on average 38.84% behind closed VLMs from OpenAI and Google (RQ2). We also find that exam-style scoring reveals calibration behaviors hidden by standard accuracy metrics: weaker models benefit dramatically from selective answering (+117-157%), while strong models show calibration (+0-12%).

Secondly, we find additional proof that **training methodology dominates raw scale**. Gemini Flash’s 30-40% better accuracy than Maverick and Scout (all 17B sparse MoEs) indicates superior proprietary training innovations. Further, pass@k accuracy doesn’t fully capture reasoning quality gaps:

Model	Release Date	Knowledge Cutoff Date	2019-2024 Acc (%)	2025 Acc (%)	$\Delta$
Random baseline	—	—	$8.20 \pm 0.52$	$9.26 \pm 1.56$	+1.06%
Aya Vision 8B	Mar 2025	Late 2024	$8.07 \pm 2.50$	$9.21 \pm 10.03$	+1.14%
Kimi VL A3B Thinking 2506 16B	Jun 2025	Unknown	$9.88 \pm 0.50$	$11.05 \pm 13.37$	+1.17%
Qwen 2.5 VL 7B	Jan 2025	Nov 2024	$10.64 \pm 0.83$	$12.37 \pm 1.96$	+1.73%
InternVL3 8B	Apr 2025	Aug 2024	$10.64 \pm 0.42$	$12.54 \pm 1.28$	+1.91%
InternVL3.5 14B	Aug 2025	Aug 2024	$15.95 \pm 1.11$	$16.11 \pm 1.64$	+0.15%
Grok 4 Fast	Sep 2025	Nov 2024	$19.32 \pm 0.99$	$17.62 \pm 1.46$	-1.70%
InternVL3.5 30B	Aug 2025	Aug 2024	$18.90 \pm 0.90$	$18.84 \pm 1.07$	-0.06%
InternVL3 78B	Apr 2025	Aug 2024	$21.99 \pm 0.61$	$23.60 \pm 3.43$	+1.61%
Gemma 3 27B	Mar 2025	Aug 2024	$29.48 \pm 0.27$	$30.62 \pm 0.72$	+1.14%
Llama4 Scout 109B	Apr 2025	Aug 2024	$40.87 \pm 0.63$	$37.07 \pm 2.28$	-3.81%
Llama4 Maverick 400B	Apr 2025	Aug 2024	$51.97 \pm 0.55$	$43.83 \pm 2.79$	-8.13%
Qwen3 VL 235B A22B Instruct	Sep 2025	Unknown	$44.52 \pm 0.00$	$45.03 \pm 0.00$	+0.50%
OpenAI o3	Apr 2025	Jun 2024	$77.41 \pm 0.66$	$72.70 \pm 3.65$	-4.71%
Gemini 2.5 Pro	Mar 2025	Jan 2025	$81.73 \pm 0.83$	$77.39 \pm 0.79$	-4.33%
GPT-5	Aug 2025	Sep 2024	$84.75 \pm 0.63$	$79.47 \pm 1.49$	-5.27%
Gemini 2.5 Flash 09-2025	Sep 2025	Jan 2025	$82.43 \pm 0.00$	$80.53 \pm 0.00$	-1.90%

Table 7: Data contamination tests on mmJEE-Eval. We evaluate effects with (a) model release dates and (b) performance on older (2019-2024 editions) and held-out (2025) JEE Advanced questions. Notably, the 2025 edition was held on May 12, 2025, with answers released on June 2, 2025 (after the cutoff dates of the tested models). The difference between the two sets averages at 2.79% (range: -8.13%-1.91%), keeping any contamination effects modest. Particularly, we notice large drops on o3, Gemini 2.5 Pro, and GPT-5, meaning closed frontiers suffer more from training data memorization than open alternatives, which report gains on the held-out set.

Model	Images+ Diagrams	EasyOCR (text)	Gemma3 OCR (text)	Diagrams Cropped
Qwen2.5 VL 7B	12.37%	11.58%	11.58%	11.58%
Gemma3 27B	30.62%	14.74%	33.68%	26.32%
InternVL3 78B	23.60%	13.68%	28.42%	21.58%
Llama4 Scout	37.07%	20.53%	31.05%	28.42%
Llama4 Maverick	43.83%	19.47%	40.53%	38.42%
Gemini 2.5 Flash	79.52%	31.05%	57.37%	66.32%
GPT-5	79.47%	27.37%	61.01%	62.63%

Table 8: Ablation study showing model performance on the 2025 subset across different setups and OCR pipelines. Reported numbers are mean accuracies.

Scout outperforms GPT-5 on Bloom’s taxonomy-inspired evaluation despite a 40% pass@1 gap. This suggests problems in pure breadth-style leaderboard benchmarks, motivating our exam style and error analysis-based evaluation.

Finally, addressing RQ3, we find **universal metacognitive limitations in VLMs**. Models can often detect when their reasoning is uncertain (21–73% of cases) but rarely correct themselves when prompted (1.06–5.2% improvements), far below simple sampling gains ( $\approx 30\%$  from pass@3). Even GPT-5 corrects only 5.2% of its errors (vs 14.6% pass@3 gain). This suggests a problematic finding: while VLMs can eventually arrive at correct solutions, they do not reliably recognize when they are wrong.

Together, these findings position mmJEE-Eval as a strong complementary VLM reasoning benchmark besides breadth-first tests like MMMU.

## 5 Conclusion

We introduced mmJEE-Eval, a multimodal and bilingual benchmark built from 7 years of India’s JEE Advanced examination, designed to evaluate exam-style reasoning in VLMs. Through detailed evaluation, we found a near-50% performance gap between the leading closed VLMs and the strongest OSS systems. We further find that open models show limited bilingual consistency, often failing to get the same question correct across English and Hindi. Ultimately, we observe VLMs only correct 1.06-5.05% of their own identified errors. Given typical 30-40% gains with pass@k sampling, this indicates weak meta-cognition even when reasoning capability is present.

In essence, mmJEE-Eval presents the first comprehensive tool for auditing reasoning depth, bilingual robustness, and meta-cognitive behavior within a single framework, revealing gaps that other tests do not fully capture.

## Limitations

Since JEE Advanced is conducted only in English and Hindi, our study is limited to two languages, Hindi and English, in linguistic scope. JEE Advanced comes with additional constraints, such as timing, where all questions must be solved within 3 hours. Our evaluation doesn’t model this. Moreover, human test-takers employ iterative problem-



solving and error correction capabilities, while our implementation focuses on single-pass EP and EC.

Additionally, our evaluation does not capture the cultural and educational context that Indian students receive through years of specialized JEE preparation. Finally, while JEE Advanced represents challenging pre-college STEM reasoning, ceiling effects may limit discrimination among frontier models.

## Acknowledgments

SPARC (Scheme for Promotion of Academic and Research Collaboration) Phase-III (Project ID: 3385) and TIH IIT Tirupati (IITTNiF/TPD/2024-25/P16) partially supported this research work. We also thank Nvidia India for providing Blackwell-based RTX 50-series GPUs to run the experiments.

## References

- Daman Arora, Himanshu Singh, and Mausam. 2023. [Have LLMs advanced enough? a challenging problem solving benchmark for large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, Singapore. Association for Computational Linguistics.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibor Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.
- Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. [Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages](#). *Transactions of the Association for Computational Linguistics*, 8:454–470.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
- Zhuoxuan Jiang, Haoyuan Peng, Shanshan Feng, Fan Li, and Dongsheng Li. 2024. [LLMs can find mathematical reasoning mistakes by pedagogical chain-of-thought](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 3439–3447. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Xiaoyuan Li, Wenjie Wang, Moxin Li, Junrong Guo, Yang Zhang, and Fuli Feng. 2024. [Evaluating mathematical reasoning of large language models: A focus on error identification and correction](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11316–11360, Bangkok, Thailand. Association for Computational Linguistics.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*.
- Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. [Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. [Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts](#). In *The Twelfth International Conference on Learning Representations*.
- Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, and 1090 others. 2025. [Humanity’s last exam](#). *Preprint*, arXiv:2501.14249.
- Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. [XCOPA: A multilingual dataset for causal commonsense reasoning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

- Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.
- Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mcsqa: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. *arXiv preprint arXiv:2406.04215*.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Hao-tian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. [Charxiv: Charting gaps in realistic chart understanding in multimodal llms](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 113569–113697. Curran Associates, Inc.
- Chen Xin, Andreas Hartel, and Enkelejda Kasneci. 2024. [Dart: An automated end-to-end object detection pipeline with data diversification, open-vocabulary bounding box annotation, pseudo-label review, and model training](#). *Expert Systems with Applications*, 258:125124.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. [Metamath: Bootstrap your own mathematical questions for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. [Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi](#). *Preprint*, arXiv:2311.16502.
- Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhao Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhao Chen, and Graham Neubig. 2025. [MMM-pro: A more robust multi-discipline multimodal understanding benchmark](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, Vienna, Austria. Association for Computational Linguistics.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. [Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models](#). *Preprint*, arXiv:2504.10479.

# Appendices

This supplementary material presents additional details on the following aspects:

- **Appendix A:** Confidence Thresholding Tests
- **Appendix B:** Optimal  $k$  Value Selection for Pass@1 Accuracy
- **Appendix C:** Prompts Used
- **Appendix D:** Manual Annotation Software Design
- **Appendix E:** Studying Failure Cases
- **Appendix F:** JEE Advanced Scoring System
- **Appendix G:** Datasheet

## A Confidence Thresholding Tests

Figure 3 illustrates how positive, negative, and total scores vary with confidence thresholds for GPT-5 and InternVL3-78B. For example, in single-correct MCQ questions, GPT-5’s answers are only considered if they appear in at least 35% of the sampled responses. Similarly, for InternVL 3 78B, the threshold is 30%. Answers that fail to form a consensus are skipped to avoid negative marking.

Looking at the characteristics of the plots, we note that the closed-source model (GPT-5) maintains relatively stable performance across experimental settings, while InternVL3-78B exhibits erratic behavior with volatile score fluctuations. This provides further evidence for the inconsistent reasoning capabilities and poor calibration of open-source models.

## B Optimal $k$ Value Selection for Pass@1 Accuracy

To determine the optimal number of inference runs for reliable pass@1 accuracy estimation, we conducted systematic experiments with Gemma3 27B across varying numbers of runs. This analysis is crucial for balancing computational efficiency with statistical precision in our evaluation methodology.

Figure 5 demonstrates the convergence behavior of confidence intervals as the number of runs increases. The 95% confidence interval width decreases rapidly from 14.1% at a single run to 2.97% at 3 runs, eventually stabilizing at 0.56% after 28 runs. The mean accuracy stabilizes around 29.62% after approximately 10 runs, with minimal fluctuation thereafter. This convergence pattern

indicates that statistical precision improves substantially with increased sampling until reaching a plateau.

Our analysis reveals distinct accuracy characteristics between different model capability levels. Gemma3 27B exhibits consistent performance with accuracy values clustering around the mean (29.63%) across 32 runs, showing minimal variance (standard deviation: 0.768%). The maximum deviation spans only 3.35 percentage points (28.36% to 31.71%), indicating stable and reliable measurement characteristics suitable for comparative evaluation.

In contrast, experiments with Granite Vision 3.2 2B (Figure 4) revealed highly erratic accuracy patterns resembling random guessing behavior. Individual trial accuracies fluctuated dramatically between 5.88% and 29.41% across 20 trials, with no discernible convergence pattern. This erratic behavior, characteristic of models with insufficient capability for the task complexity, produces unreliable measurements unsuitable for systematic evaluation. Consequently, we exclude such models from our final analysis to maintain evaluation integrity. Based on this empirical analysis, we selected  $k = 10$  runs as the optimal balance between computational efficiency and statistical reliability. While confidence intervals stabilize at  $k = 7$  (Figure 5), we chose  $k = 10$  to provide additional headroom for optimization and ensure robust estimates across different model capabilities. This configuration yields a 95% confidence interval width of approximately 1.09%, providing sufficient precision for comparative analysis while maintaining practical computational requirements.

## C Prompts Used

We employ different prompting strategies based on the question language and type to ensure optimal model performance across our multilingual dataset. For Hindi questions, we explicitly acknowledge the language barrier while requesting comprehensive analysis, as models often struggle with cross-lingual mathematical reasoning. English questions receive more direct instructions focusing on step-by-step analysis and clear answer presentation.

Our answer extraction methodology uses LaTeX `\boxed{}` formatting for all question types, departing from traditional evaluation scripts. This design choice is motivated by extensive evidence in the literature showing that contemporary lan-

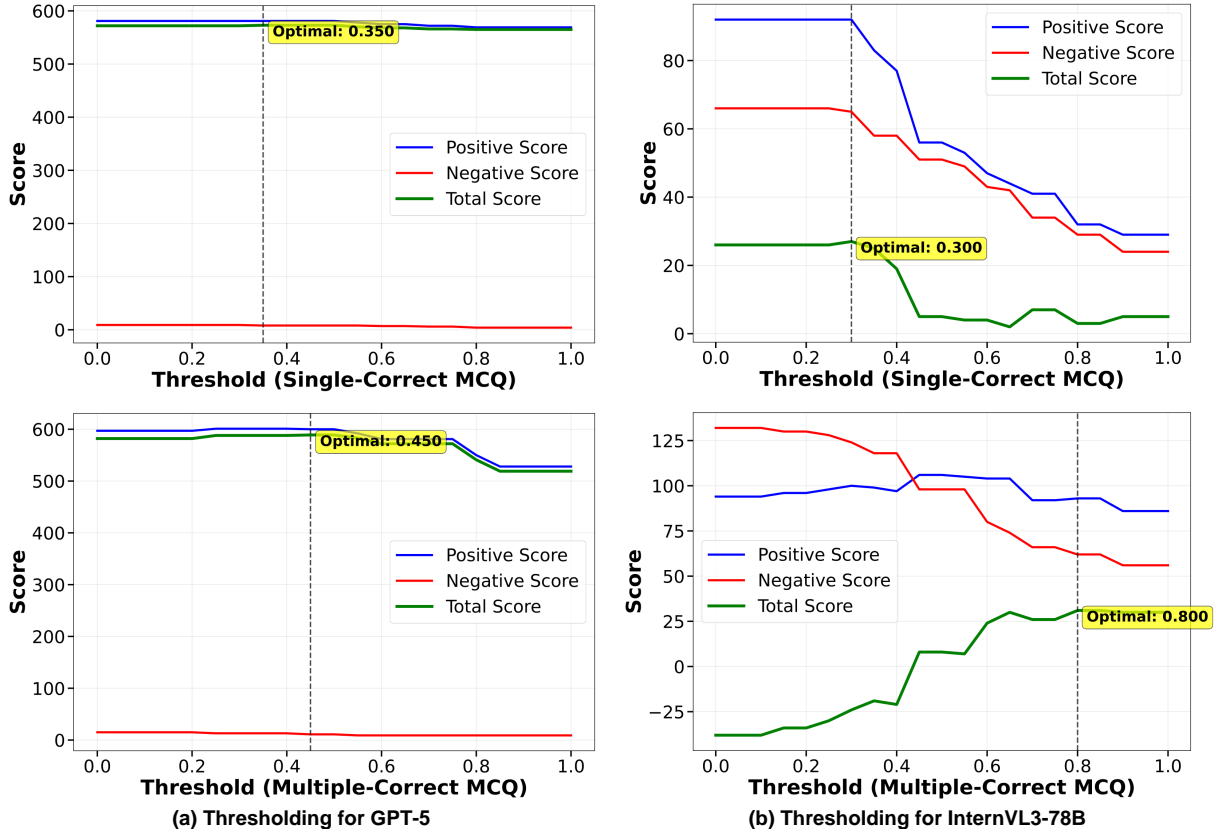


Figure 3: Thresholding analysis comparing GPT-5 and InternVL3-78B performance. (a) shows the thresholding behavior for GPT-5, while (b) demonstrates the corresponding analysis for InternVL3-78B. We note more stable behavior on the OpenAI model than InternVL3, denoting that correct answers were generated more consistently across 10 runs.

guage models are predominantly trained on LaTeX-formatted mathematical content (Lightman et al., 2023; Hendrycks et al., 2021). Additionally, our preliminary experiments using OpenAI’s MATH dataset evaluation script<sup>10</sup> yielded significantly lower success rates due to parsing failures and format inconsistencies, confirming that structured LaTeX formatting provides more reliable answer extraction.

Prompts for MCQ-Single and MCQ-Multiple questions emphasize exact option selection with explicit formatting instructions, while Numerical questions require precise decimal formatting within `\boxed{}` tags. Matching questions, being the most complex, receive detailed instructions about selecting the single best option that correctly maps items between columns.

<sup>10</sup><https://github.com/hendrycks/math/tree/main/modeling>

## C.1 System Prompts by Language

### Hindi System Prompt

This is a question image in Hindi. Please analyze the image carefully, understand the question despite the language barrier, reason step-by-step and provide your answer at the end.

### English System Prompt

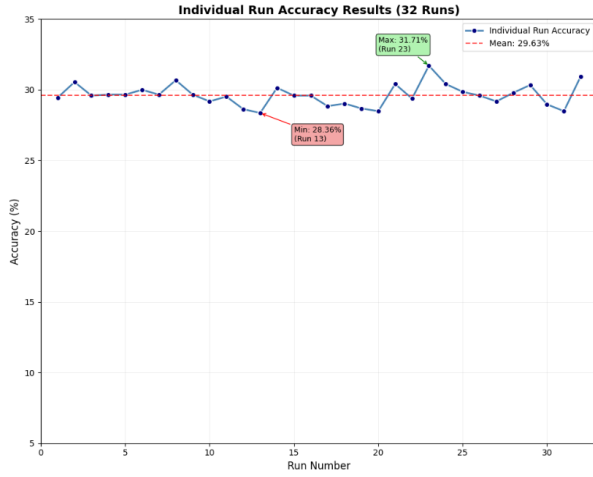
Please analyze the image carefully, reason step-by-step and provide your answer at the end.

## C.2 Question Type-Specific Instructions

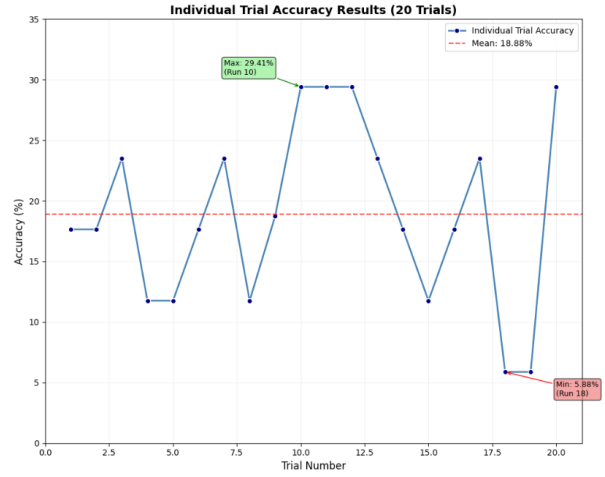
### MCQ-Single Instructions

For this question: - Choose exactly ONE option (A, B, C, or D) - Format your answer in `\boxed{}` as just one letter (e.g., `\boxed{A}`)





(A) Progress for Gemma3 27B



(B) Progress for Granite Vision 3.2 2B

Figure 4: Accuracy characteristics of two open-source models—Gemma3 27B and Granite Vision 3.2 2B—on mmJEE-Eval. Gemma3 27B (left) shows consistent performance with minimal variance around 29.63% mean accuracy. Granite Vision 3.2 2B (right) exhibits erratic behavior resembling random guessing, with accuracy fluctuations between 5.88% and 29.41%, making it unsuitable for reliable evaluation.

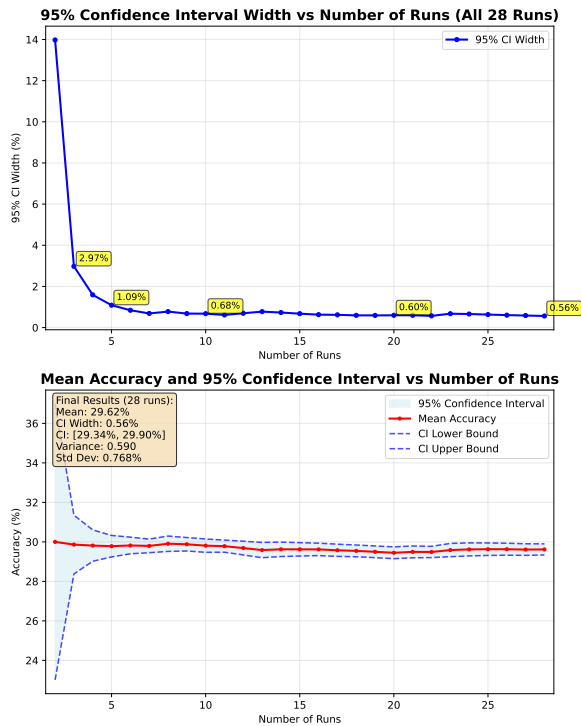


Figure 5: Analysis of confidence interval convergence across multiple experimental runs. The top panel shows how the 95% confidence interval width decreases as the number of runs increases, demonstrating improved statistical precision. The bottom panel displays the mean accuracy with confidence bounds, showing stabilization of the estimate around 29.62% after sufficient runs.

### MCQ-Multiple Instructions

For this question: - Choose ONE OR MORE options (A, B, C, and/or D)  
- Format your answer in `\boxed{}` with letters (e.g., `\boxed{ABC}` or `\boxed{B}`)

### Numerical Instructions

For this question: - Provide a numerical value - Round to appropriate decimal places if needed - Format your answer in `\boxed{}` (e.g., `\boxed{2.5}` or `\boxed{42}`)

### Matching Instructions

For this question: - Choose ONE option that correctly matches items from left column to right column - Format your answer in `\boxed{}` as just one letter (e.g., `\boxed{A}`)

## C.3 Error Presence (EP) Prompts:

For error presence detection, we adapted prompts from (Li et al., 2024) with language-specific modifications as shown in Figure 6.

## C.4 Error Correction (EC) Prompts:

For error correction, we use the prompt in Figure 7.

### English Prompt:

Task Overview: According to the [Question], the attached [Image] and [Solution], please determine if the [Solution] is correct.

Format Requirement: In the [Solution], the model reasons step-by-step, and returns the final answer as `\textbackslash boxed\{answer\}`. If the [Solution] is correct, only output a string of 'yes' as response; while if the [Solution] is incorrect, only output a string of 'no' as response.

Case to be Judged:

[Question] {data['question']}

[Image] Attached

[Solution] {data['solution']}

### Hindi Prompt:

कार्य विवरण: [प्रश्न], संलग्न [चित्र] और [समाधान] के अनुसार, कृपया निर्धारित करें कि [समाधान] सही है या नहीं।

प्रारूप आवश्यकता: [समाधान] में, मॉडल चरण-दर-चरण तर्क करता है, और अंतिम उत्तर `\boxed{{answer}}` के रूप में देता है। यदि [समाधान] सही है, तो केवल 'yes' स्ट्रिंग आउटपुट करें; जबकि यदि [समाधान] गलत है, तो केवल 'no' स्ट्रिंग आउटपुट करें।

मामला जिसका न्याय किया जाना है:

[प्रश्न] {data['question']}

[चित्र] संलग्न

[समाधान] {data['solution']}

Figure 6: Error Presence (EP) prompt

## D Manual Annotation Software Design

To ensure accurate and systematic collection of JEE Advanced questions, we developed three custom annotation utilities that streamline the entire data collection and verification pipeline.

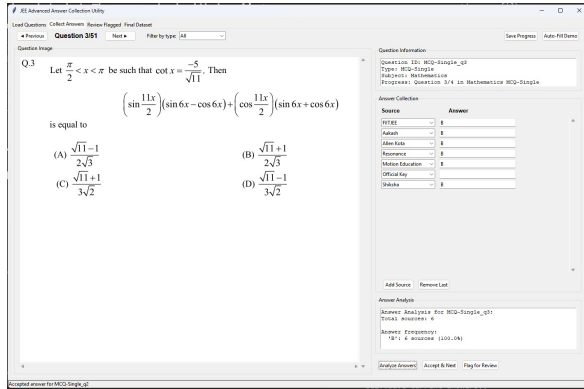


Figure 8: Our custom answer collection utility shows answers collected from five independent coaching institutes (FIITJEE, Aakash, Allen Kota, Resonance, Motion Education), and Official Key. The tool performs automatic majority voting analysis, with answer 'B' received unanimous consensus (100%) from all 6 sources.

**Question Annotation Tool.** Our PDF question annotator (Figure 9) automatically downloads question papers from official sources across all years (2019-2025) and presents them in a page-by-page interface for manual annotation. Annotators can draw precise bounding boxes around individual questions, with support for multi-rectangle questions that span multiple regions or pages. The tool

maintains structured metadata including year, paper number, language, subject, and question type, while automatically generating unique question IDs following our naming convention. Key features include zoom controls for detailed annotation, progress tracking across subjects, and the ability to export annotated questions as high-resolution images.

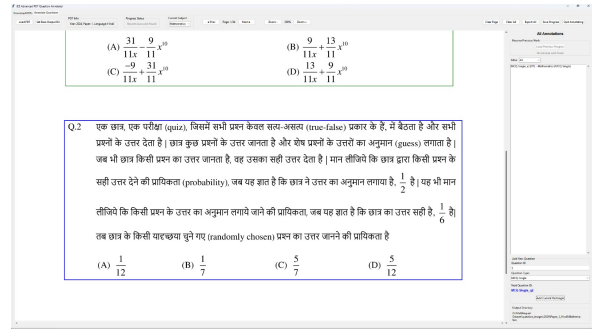


Figure 9: Our custom annotation tool for JEE Advanced question collection automatically downloads question papers and presents them page-by-page to ensure accurate question segmentation and collection.

**Answer Collection Utility.** The answer verification tool (Figure 8) implements our multi-source verification approach by collecting answers from six independent sources: five major coaching institutes (FIITJEE, Aakash, Allen Kota, Resonance, Motion Education) and official answer keys when available. The interface displays each question image alongside a structured form for recording answers from each source. The tool performs au-

### English Prompt:

Task Overview: According to the [Question], its related [Image] and [Solution], if you think the [Solution] is incorrect, fix all errors in the reasoning chain and subsequently derive the correct answer in your [Output].

Format Requirement: Detail your reasoning.

For MCQ-Single questions: Format your answer in `\boxed{}` as just one letter (e.g., `\boxed{A}`)

For MCQ-Multiple questions: Format your answer in `\boxed{}` with letters (e.g., `\boxed{ABC}` or `\boxed{B}`)

For Numerical questions: Format your answer in `\boxed{}` (e.g., `\boxed{2.5}` or `\boxed{42}`)

For Matching questions: Format your answer in `\boxed{}` as just one letter (e.g., `\boxed{A}`)

### Hindi Prompt:

कार्य विवरण: [प्रश्न], उससे संबंधित [चित्र] और [समाधान] के अनुसार, यदि आपको लगता है कि [समाधान] गलत है, तो तर्क श्रृंखला में सभी त्रुटियों को ठीक करें और फिर अपने [आउटपुट] में सही उत्तर प्राप्त करें।

प्रारूप आवश्यकता: अपने तर्क का विवरण दें।

For MCQ-Single: अपना उत्तर `\boxed{}` में केवल एक अक्षर के रूप में प्रारूपित करें (जैसे, `\boxed{A}`)

For MCQ-Multiple questions: अपना उत्तर `\boxed{}` में अक्षरों के साथ प्रारूपित करें (जैसे, `\boxed{ABC}` या `\boxed{B}`)

For Numerical questions: अपना उत्तर `\boxed{}` में प्रारूपित करें (जैसे, `\boxed{2.5}` या `\boxed{42}`)

For Matching questions: अपना उत्तर `\boxed{}` में केवल एक अक्षर के रूप में प्रारूपित करें (जैसे, `\boxed{A}`)

Figure 7: Error Correction (EC) prompt

tomatic majority voting analysis, calculating confidence scores and flagging questions that fail to achieve sufficient consensus ( $>60\%$ ) for manual review.

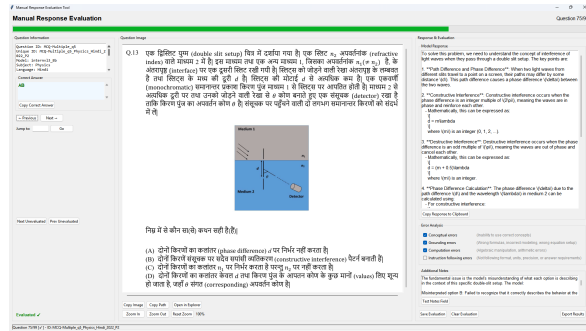


Figure 10: Our manual verification utility loads the question, correct answer, and the model’s response. The human annotator evaluates all presented information and votes on the kind(s) of errors found.

**Manual Evaluation Interface.** Our response evaluation tool (Figure 10) facilitates systematic manual analysis of model responses across four error dimensions: conceptual errors, grounding er-

rors, computation errors, and instruction-following errors. The interface displays question metadata and correct answers, renders question images with zoom and navigation controls, and presents model responses alongside evaluation forms. The tool supports checkpoint-based evaluation sessions, allowing annotators to pause and resume lengthy evaluation processes. Additional features include clipboard functionality for copying images and responses, navigation shortcuts to unevaluated questions, and export capabilities for generating analysis-ready CSV files. This framework enabled our comprehensive error analysis across 400 manually reviewed questions, providing the granular insights presented in Table 6.

## E Studying Failure Cases

Now, we look at a few cherry-picked failure cases to understand where VLMs make mistakes while solving mmJEE-Eval. We present two examples from models with different profiles: the open-source sub-10B InternVL3 10B, and state-of-the-art proprietary Gemini 2.5 Pro. Following is an

analysis of where these models make mistakes to better understand their limitations.

### E.1 Example 1: Conceptual and Grounding Errors

Following is an example of a factual hallucination or mathematical reasoning error. The model incorrectly evaluates the validity of a clearly solvable algebraic equation. Even though the equation is standard and solvable, the model confidently declares it invalid, indicating a lack of basic symbolic reasoning or over-reliance on heuristics rather than step-by-step computation. It reflects failures in math grounding despite surface-level fluency.

#### Numerical\_q2\_Maths\_English\_2022\_P1 + Model Response (InternVL3 8B)

##### Question:

Let  $\alpha$  be a positive real number. Let  $f: \mathbb{R} \rightarrow \mathbb{R}$  and  $g: (\alpha, \infty) \rightarrow \mathbb{R}$  be the functions defined by

$$f(x) = \sin\left(\frac{\pi x}{12}\right)$$

$$\text{and } g(x) = \frac{2 \log_e(\sqrt{x} - \sqrt{\alpha})}{\log_e(e^{\sqrt{x}} - e^{\sqrt{\alpha}})}.$$

Then the value of  $\lim_{x \rightarrow \alpha^+} f(g(x))$  is \_\_\_\_\_.

##### Model Output (InternVL3 8B):

###### Steps 1–2

To solve for  $\lim_{x \rightarrow 0^+} g(f(x))$ , we need to evaluate the composition of functions as  $x$  approaches zero from the positive side.

**Step 1: Evaluate  $f(x)$  as  $x \rightarrow 0^+$**

$$f(x) = \sin\left(\frac{\pi x}{12}\right)$$

$$\Rightarrow f(0) = \sin(0) = 0$$

**Step 2: Evaluate  $g(f(x))$**

$$g(x) = \frac{2 \log_6(\sqrt{x} - \sqrt{\alpha})}{\log_8(e^x - e^{-x})}$$

###### Steps 2–6

$$\Rightarrow \lim_{x \rightarrow 0^+} g(f(x)) = g(0)$$

**Step 3: Simplify  $g(0)$**

$$g(0) = \frac{2 \log_6(0)}{\log_8(0)}$$

**(Wrong formula usage)**

**Step 4: Use Taylor Series Approximations**

$$e^x \approx 1 + x, \quad e^{-x} \approx 1 - x$$

$$\Rightarrow e^x - e^{-x} \approx 2x$$

$$\log_8(e^x - e^{-x}) \approx \log_8(2x) = \frac{\ln(2x)}{\ln(8)}$$

**Step 5: Simplify the numerator**

$$\sqrt{x} - \sqrt[4]{x} = x^{1/2} - x^{1/4}$$

$$= x^{1/4}(x^{1/4} - 1) \approx -x^{1/4}$$

$$\log_6(\sqrt{x} - \sqrt[4]{x}) \approx \log_6(-x^{1/4})$$

**(Wrong assumption)**

**Step 6: Combine numerator and denominator**

$$g(0) \approx \frac{\frac{\ln(x)}{2 \ln(6)}}{\frac{\ln(x)}{3 \ln(2)}} = \frac{3 \ln(2)}{2 \ln(6)} = \boxed{6}$$

**Wrong answer: Interpreted limit at 0 instead of  $\alpha$  and used incorrect domains.**

###### Final Evaluation

**Predicted:**  $\boxed{6}$

**Ground Truth:**  $\boxed{0.5}$

**Marking:** Incorrect due to conceptual and grounding errors

### E.2 Language Control Failure

The model unexpectedly responds in German, despite the prompt being in English and the input image in Hindi. This illustrates a language control failure, where the model drifts from the user's in-



tended language — a critical issue in real-world deployments where understanding cultural cues from the supplied prompt is important before adapting model outputs to fit the user.

### Numerical\_q4\_Physics\_Hindi\_2021\_P2 + Model Response (Gemini 2.5 Pro)

#### Question Image:

प्रश्न 9 और 10 के लिए प्रश्न सत्यम्

एक लोलक (pendulum), द्रव्यमान  $m = 0.1 \text{ kg}$  के एक गोलक (bob) और लम्बाई  $L = 1.0 \text{ m}$  के एक द्रव्यमानरहित (massless) तथा न खींचने वाले (inextensible) धागे से बना है। यह एक घर्षणहीन क्षैतिज फर्श (floor) से  $H = 0.9 \text{ m}$  की ऊँचाई पर एक स्थिर बिंदु से लटका हुआ है। आरंभ में गोलक, फर्श पर निलंबन बिंदु (point of suspension) के ठीक उर्ध्वधर नीचे स्थिरावस्था में है। किसी क्षण, गोलक को  $P = 0.2 \text{ kg}\cdot\text{m/s}$  का एक क्षैतिज आवेग (impulse) प्रदान किया जाता है। इस कारण कुछ दूरी तक फिसलने के बाद, गोलक सतह से ऊपर उठ जाता है और धागा तन जाता है (becomes taut)। गोलक के सतह से उठने के तुरंत पहले, निलंबन बिंदु के सापेक्ष, लोलक का कोणीय संवेग (angular momentum)  $J \text{ kg}\cdot\text{m}^2/\text{s}$  है। सतह से उठने के तुरंत बाद लोलक की गतिज ऊर्जा  $K \text{ Joules}$  है।

Q.10  $K$  का मान \_\_\_\_\_ है।

Correct Answer: 0.16

Predicted Answer: 0.162

Is Correct: No

#### Model's Response (Language: German)

##### Analyse des Problems (Problem Analysis)

Basierend auf dem Hindi-Text lassen sich die folgenden physikalischen Parameter und Bedingungen extrahieren:

- Masse des Pendelkörpers (bob),  $m = 0.1 \text{ kg}$
- Länge des Fadens,  $L = 1.0 \text{ m}$  (masselos, nicht dehnbar)
- Höhe des Aufhängepunktes über dem Boden,  $H = 0.9 \text{ m}$
- Boden: Reibungslose horizontale Fläche
- Anfangszustand: Der Pendelkörper ruht unter dem Aufhängepunkt auf dem Boden ( $L > H$ )
- Ereignis: Ein Impuls  $P = 0.2 \text{ kg}\cdot\text{m/s}$  wirkt horizontal auf den Körper

#### Steps 1-4

- Gesucht: Kinetische Energie  $K$  unmittelbar nachdem der Faden straff wird

#### Schritt 1: Anfangsgeschwindigkeit

$$v_0 = \frac{P}{m} = \frac{0.2}{0.1} = 2.0 \text{ m/s}$$

#### Schritt 2: Geometrie beim Straffen des Fadens

Ein rechtwinkliges Dreieck mit:

Vertikale:  $H = 0.9$

Hypotenuse:  $L = 1.0$

$$\Rightarrow \cos(\theta) = \frac{0.9}{1.0} = 0.9$$

#### Schritt 3: Geschwindigkeit senkrecht zum Faden

Nur der senkrechte Anteil bleibt:

$$v_f = v_0 \cdot \cos(\theta) = 2.0 \cdot 0.9 = 1.8 \text{ m/s}$$

#### Schritt 4: Kinetische Energie

$$K = \frac{1}{2} m v_f^2 = \frac{1}{2} \cdot 0.1 \cdot (1.8)^2 = 0.05 \cdot 3.24 = 0.162 \text{ Joule}$$

0.162

**Note: Although the calculation is correct, it does not match the accepted correct value 0.16. The predicted answer is slightly over the acceptable threshold.**

#### Final Evaluation

Predicted: 0.162

Ground Truth: 0.16

Marking: Incorrect due to precision threshold

Question Type	Response Type	Marks
MCQ (Single Correct)	Correct	+3
	Wrong	-1
	Unanswered	0
MCQ (Multiple Correct)	All correct options chosen	+4
	3 correct options chosen (of 4)	+3
	2 correct options chosen (of 3+)	+2
	1 correct option chosen (of 2+)	+1
	Wrong / Any incorrect option chosen	-2
	Unanswered	0
Numerical Answer Type	Correct	+4
	Wrong / Unanswered	0
Matching Type	Correct	+4
	Wrong	-1
	Unanswered	0

Table 9: JEE Advanced Marking Scheme

## F JEE Advanced Scoring System

Table 9 summarizes the marking scheme for JEE Advanced 2025. The questions involve a mix of positive and negative rewards, which adds to the difficulty of the exam and mmJEE-Eval.

## G Datasheet

This datasheet is based on [Gebru et al. \(2021\)](#). For the code and LaTeX template, check [their GitHub repo](#).

### MOTIVATION

**For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

mmJEE-Eval is constructed for evaluating state-of-the-art vision-language models (VLMs) on STEM reasoning. The dataset was developed to advance evaluation methodologies for vision-language models in multilingual and multimodal scientific reasoning contexts.

**Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

The dataset was created by the authors, Arka Mukherjee and Shreya Ghosh, as part of the NLP research group at IIT Bhubaneswar.

**What support was needed to make this dataset?** (e.g. who funded the creation of the

dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)

Anonymized for peer review

### Any other comments?

N/A

### COMPOSITION

**What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance is a JEE Advanced examination question comprising a question image (containing mathematical notation, diagrams, and text in Hindi or English), metadata (year, paper, subject, question type), and ground truth answers verified through multi-source consensus.

**How many instances are there in total (of each type, if appropriate)?**

The dataset contains 1,460 questions spanning seven years (2019-2025) of JEE Advanced examinations, distributed across Chemistry (492 questions, 33.7%), Mathematics (492 questions, 33.7%), and Physics (476 questions, 32.6%). Question types include Numerical (652 instances, 44.7%), MCQ-Multiple (412 instances, 28.2%),

MCQ-Single (396 instances, 27.1%), and Matching (112 instances).

**Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

This dataset is a subset of all publicly available JEE Advanced questions from 2007-2025. We select editions from 2019-2025 as they specifically incorporate the multilingual aspect (the exam was offered only in English till 2018).

**What data does each instance consist of?** “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance contains: (1) a high-resolution image of the question extracted from official PDF papers, (2) structured metadata including year, paper number, language, subject, question type, and unique identifier, (3) ground truth answers verified through majority voting across six independent sources, and (4) confidence scores for answer verification.

**Is there a label or target associated with each instance?** If so, please provide a description.

Yes, each instance has an associated ground truth answer obtained through multi-source verification. For MCQ questions, labels are option letters (A, B, C, D) or combinations thereof. For numerical questions, labels are precise numerical values.

**Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No essential information is missing from individual instances. All questions include images, metadata, and verified answers. Some instances may lack official answer keys, but these are compensated through our multi-source verification approach

using independent coaching institute solutions.

**Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

Instances are grouped by examination metadata (year, paper, subject) and question type. Questions from the same examination paper may share thematic connections, but each instance is designed to be independently evaluable. No explicit dependency relationships exist between questions.

**Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

No pre-defined training/testing splits are provided, as this is an evaluation-only dataset. Users may create custom splits based on years, subjects, or question types, depending on their evaluation needs. We recommend using temporal splits (e.g., 2019-2023 vs 2024-2025) for contamination-aware evaluation.

**Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

Questions with answer disagreement below 60% consensus were flagged for manual review. Only 30 out of 1,476 initial questions exhibited such disagreements. Such questions were primarily numericals, where official keys accept ranges while coaching institutes provide point values. No significant redundancies exist as each question represents a unique examination item.

**Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained.

**Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.

No, the dataset contains only publicly available examination questions from official JEE Advanced papers.

**Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

No

**Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

No

**Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No

**Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.

No

**Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.

No

**Any other comments?**

N/A

## COLLECTION

**How was the data associated with each instance acquired?** Was the data directly

observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Question images were directly extracted from official JEE Advanced PDF papers using custom annotation tools. Ground truth answers were collected from six independent sources (five coaching institutes plus official keys, detailed in Sec 3.2) and verified through majority voting consensus. All source materials are publicly available through official examination archives.

**Over what timeframe was the data collected?**

Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.

Data collection was conducted between June to July 2025.

**What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** How were these mechanisms or procedures validated?

We developed custom software tools for PDF processing, question extraction, and answer verification (Appendix D).

**What was the resource cost of collecting the data?** (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint.)

Computational costs were minimal, primarily involving PDF processing and image extraction.

**If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

We select questions from the past seven editions of JEE Advanced (2019-2025) as the exam has been offered in Hindi since 2019. The sole criterion for this sample was to ensure an equitable distribution of questions in the two languages.



**Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

Data collection was performed by the authors.

**Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No formal institutional review board process was conducted as the dataset consists entirely of publicly available examination materials with no human subjects involved.

**Does the dataset relate to people?** If not, you may skip the remainder of the questions in this section.

N/A

**Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

All questions were collected from the public JEE Advanced archive. Answers have been sourced from the official archives of the respective coaching institutes.

**Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No, since all data has been publicly released.

**Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

**If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well

as a link or other access point to the mechanism (if appropriate)

N/A

**Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

**Any other comments?**

N/A

### **PREPROCESSING / CLEANING / LABELING**

**Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Questions were extracted from PDFs using custom annotation tools with manual verification for accuracy. Images were processed to ensure high resolution and proper cropping. Answer verification involved majority voting across multiple sources with confidence scoring. Questions failing to achieve >60% consensus were manually reviewed.

**Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the “raw” data.

Yes, original PDF files are preserved and referenced. Our annotation tools maintain links to source materials, enabling reproducibility and verification of extraction accuracy. Moreover, all raw PDF files can be accessed through the official JEE Advanced archive.

**Is the software used to preprocess/clean/label the instances available?** If so, please provide a link or other access point.

We make the [code](#) and [dataset](#) publicly available.

**Any other comments?**

N/A

## USES

**Has the dataset been used for any tasks already?** If so, please provide a description.

This dataset was created for evaluating vision-language models on authentic multimodal scientific reasoning tasks. It has been used to evaluate five state-of-the-art VLMs, including Gemini 2.5 Pro, OpenAI o3, and smaller open-source models.

**Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

A general list of the leaderboard and latest news can be viewed at this [HTTPS URL](#).

**What (other) tasks could the dataset be used for?**

The dataset could be used for multilingual mathematical reasoning research, educational AI development, cross-cultural evaluation of AI systems, multimodal representation learning, and development of specialized educational assessment tools for STEM subjects.

**Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. **Is there anything a future user could do to mitigate these undesirable harms?**

Users should be aware that this dataset reflects the specific cultural and educational context of Indian competitive examinations. Performance gaps between languages or subjects might reflect training data biases rather than inherent model capabilities. Care should be taken when generalizing results to other educational contexts or populations.

**Are there tasks for which the dataset should not be used?** If so, please provide a description.

We advise against including our dataset in the pre-training of VLMs to ensure fair evaluation on future models. This dataset should not be used for commercial assessment tools without proper

licensing, for creating competing examination systems that might disadvantage students, or for any purpose that could harm the integrity of the JEE Advanced examination process.

**Any other comments?**

N/A

## DISTRIBUTION

**Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

We make the [code](#) publicly available. The dataset has been publicly released on [Hugging Face](#).

**How will the dataset be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

Anonymized for peer review

**When will the dataset be distributed?**

The dataset is publicly available from Hugging Face from the date of publication.

**Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is being released under the MIT license permitting academic and research use. Since source materials are public examination papers, no additional licensing restrictions apply beyond standard academic attribution requirements.

**Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No third-party IP restrictions apply, as all source materials are publicly available examination papers.

**Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.  
None

**Any other comments?**  
N/A

## MAINTENANCE

**Who is supporting/hosting/maintaining the dataset?**

The dataset is supported by the SPARC

**How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

The authors can be reached at the following email addresses:

- Arka Mukherjee: [arka.mukherjee078@gmail.com](mailto:arka.mukherjee078@gmail.com)
- Shreya Ghosh: [shreya@iitbbs.ac.in](mailto:shreya@iitbbs.ac.in)

**Is there an erratum?** If so, please provide a link or other access point.  
N/A

**Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Yes, in the June of every year, approximately 190 JEE Advanced questions will be added to the dataset, with leaderboard results publicized on [our website](#).

**If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

N/A

**Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes

**If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

TBD

**Any other comments?**  
N/A