P6Jiggasha: Benchmarking Large Language Models on Bangla Physics Question Answering with Cross-lingual Evaluation

S.M. Shahriar[1], **Md Tahmid Hasan Fuad**[3], **Md Fahim**[2,4,†] **Md. Azad Hossain**[1]

[1]*Chittagong University of Engineering and Technology*
[2]*Center for Computational & Data Sciences, IUB*
[3]*University of Manitoba* [4]*Penta Global Limited*

[†]Team Lead **Correspondence:** `{sayeem26s,fahimcse381}@gmail.com`

Abstract

Understanding scientific concepts in native languages is crucial for educational accessibility and knowledge transfer. In this work, we present a comprehensive evaluation of Large Language Models (LLMs) on Bangla physics questions, introducing `P6Jiggasha`, a novel dataset of 1,500 multiple-choice questions compiled from HSC physics textbooks, supplementary guides, admission preparation books, and past examination papers from various educational boards. We evaluate three state-of-the-art models—GPT-4.1, Gemini-2.5 Pro, and DeepSeek-R1-Distill-Llama-70B—on both native Bangla questions and their English translations. Our results reveal significant performance variations, with GPT-4.1 achieving 86.67% accuracy on Bangla questions in a single inference, while other models show substantial improvement through multiple inference attempts, with Gemini-2.5 Pro reaching 89.52% after four iterations. We introduce a *Cumulative Accuracy@k* metric to evaluate iterative reasoning capabilities and provide comprehensive analysis across six physics topics and six question types. Our error analysis reveals systematic cross-lingual inconsistencies where models produce contradictory answers for identical questions across languages. This study provides valuable insights into the capabilities and limitations of current LLMs for low-resource scientific question answering and establishes benchmarks for future research in Bangla natural language processing.

1 Introduction

The democratization of scientific knowledge through native language understanding represents a fundamental challenge in natural language processing and educational technology. While Large Language Models (LLMs) have demonstrated remarkable capabilities across various domains, their performance on scientific content in low-resource languages remains understudied. This gap is particularly pronounced for languages like Bangla, spoken by over 300 million people worldwide, where scientific education primarily occurs in the native language but computational resources remain limited.

Physics education in Bangladesh follows a structured Higher Secondary Certificate (HSC) curriculum, covering concepts from mechanics to quantum physics. Students typically encounter these concepts first in Bangla before transitioning to English-medium higher education, creating unique challenges for both learners and automated systems. Recent advances in LLMs have shown promising results for English scientific question answering (Hendrycks et al., 2020; Lu et al., 2022), but their applicability to non-English scientific content remains largely unexplored (Ahuja et al., 2023; Lai et al., 2023). While multilingual models like GPT-4 (Achiam et al., 2023) and Gemini (Team et al., 2023) present opportunities to bridge this gap, systematic evaluation on domain-specific content in low-resource languages is lacking.

Our contributions are as follows:

- **Dataset Creation and Curation:** We compile `P6Jiggasha`, the first large-scale Bangla physics MCQ dataset containing 1,500 questions from authentic educational sources with comprehensive English translations and categorizations across six physics domains and six question types.
- **Comprehensive Multi-Model Evaluation:** We systematically evaluate three state-of-the-art LLMs on both Bangla and English versions using the *Cumulative Accuracy@k* metric, analyzing the impact of language modality, question complexity, and iterative inference strategies.
- **Cross-lingual Performance Analysis:** We provide detailed *error analysis* revealing systematic challenges in mathematical reasoning,

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 196–211
December 23, 2025 ©2025 Association for Computational Linguistics

theoretical understanding, and cross-lingual consistency across languages.

Our results reveal nuanced patterns in model performance with significant variations across topics, question types, and languages, providing practical insights for educational technology development and establishing benchmarks for future research in scientific question answering for Bangla and other low-resource languages.

2 Related Work

2.1 Large Language Model Reasoning

The emergence of sophisticated reasoning capabilities in large language models has fundamentally transformed how we approach complex problem-solving tasks. Chain-of-thought prompting demonstrated that generating intermediate reasoning steps significantly improves the ability of large language models to perform complex reasoning (Wei et al., 2022), opening new avenues for scientific reasoning. Recent advances have moved beyond manual prompt engineering, with researchers developing methods for chain-of-thought reasoning without explicit prompting (Wang and Zhou, 2024). The Tree of Thoughts framework (Yao et al., 2023) introduced deliberate decision-making by considering multiple reasoning paths simultaneously, while self-consistency decoding (Wang et al., 2022) improved reasoning reliability by sampling multiple reasoning paths. The development of specialized reasoning models like OpenAI's o1 has further pushed the boundaries, using reinforcement learning to refine problem-solving strategies, with recent surveys highlighting rapid progress in mathematical reasoning capabilities (Zhang et al., 2024).

2.2 Scientific Reasoning Benchmarks

The evaluation of language models on scientific content has become a critical research area, with several landmark datasets establishing the foundation for systematic assessment. Lu et al. introduced multimodal reasoning approaches for science question answering (Lu et al., 2022), while the MMLU benchmark (Hendrycks et al., 2020) provided comprehensive evaluation across multiple domains, including physics. Specialized benchmarks have emerged for mathematical problem-solving (Cobbe et al., 2021) and elementary science questions (Clark et al., 2018). The CMMLU benchmark (Li et al., 2023) extended multilingual evaluation to Chinese, while efforts to develop cross-lingual scientific reasoning benchmarks have highlighted persistent performance gaps in low-resource languages (Bang et al., 2023).

2.3 Bangla Reasoning and NLP Benchmarks

The development of reasoning capabilities for low-resource languages like Bangla has gained momentum through several key initiatives. Early work focused on foundational NLP tasks, with BanglaBERT (Bhattacharjee et al., 2021) establishing baselines for general language understanding. Recent work has specifically targeted reasoning capabilities, with Reveal-Bangla introducing a dataset for cross-lingual multi-step reasoning evaluation (Islam and Sarti, 2025). Additional contributions include SentNoB for sentiment analysis (Islam et al., 2021), BanglaT5 for text generation and translation (Bhattacharjee et al., 2022), BanglaBook for large-scale sentiment analysis (Kabir et al., 2023), and BanglaNLG for natural language generation (Bhattacharjee et al., 2022). The BanglaParaphrase dataset (Akil et al., 2022) has contributed to semantic understanding tasks, while IndicNLPSuite (Kakwani et al., 2020) provides broader South Asian language benchmarks. However, domain-specific reasoning benchmarks in scientific subjects like physics remain largely unexplored, creating the gap our work addresses.

3 P6Jiggasha Dataset

3.1 Data Collection

We systematically collected 1,500 multiple-choice physics questions from diverse Bangladeshi educational sources spanning various academic levels and institutional contexts. Figure 1 illustrates our complete data generation workflow consisting of four main stages: collection and OCR extraction, structure and preprocessing, filtering, and verification with categorization. We systematically collected 1,500 multiple-choice physics questions from diverse Bangladeshi educational sources including HSC Physics Textbooks, Physics Guides, Engineering Question Banks, and Test Papers. Each question follows a standard multiple-choice format with four options (A, B, C, D) and a single correct answer, covering easy-to-difficult MCQs with comprehensive physics concepts. All correct answers were rigorously re-checked directly against the original sources and retained only when the match was unambiguous.

Figure 2 shows the distribution of questions

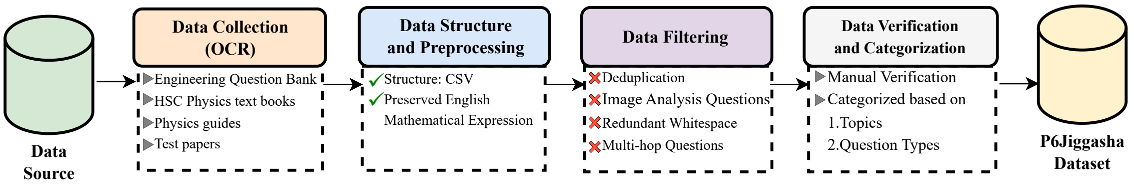


Figure 1: Data generation workflow for the P6Jiggasha dataset showing the pipeline from collection through OCR extraction, filtering, and verification to final categorization across physics topics and question types.

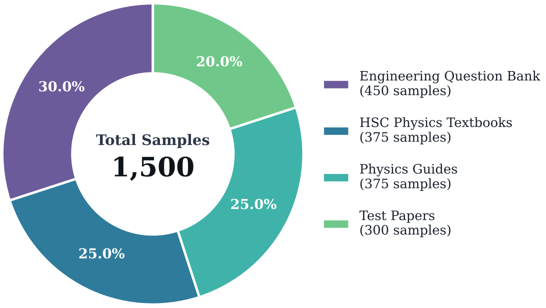


Figure 2: Source distribution showing Engineering Question Banks (30%), HSC Physics Textbooks (25%), Physics Guides (25%), and Test Papers (20%) representing diverse educational materials.

collected from various educational sources across Bangladesh's academic ecosystem, reflecting the diversity of materials used in physics education.

3.2 OCR and Data Extraction

The source materials were primarily available in PDF format and printed documents. We employed advanced language models including Grok 4 and GPT-5 for Optical Character Recognition (OCR), to extract questions from these diverse sources. The OCR process involved automated text extraction from PDF documents, image-to-text conversion for scanned materials, mathematical expression recognition and formatting, and rigorous quality verification of extracted content to ensure accuracy and completeness of physics terminology and mathematical notations.

3.3 Data Structure and Pre-processing

The dataset is structured in CSV format with preserved English mathematical expressions. During the pre-processing phase, we implemented comprehensive data filtering to remove duplications, graph/image-based questions, redundant whitespace, and multi-hop questions. This filtering process ensures consistency and maintains focus on text-based physics concepts while preserving the integrity of mathematical expressions essential for physics problem-solving.

Data Distribution	**Samples**
Total Dataset	1,500
Test Set (allocated from total)	500
Text Statistics	
Mean Word Count	23.2
Maximum Word Count	46
Minimum Word Count	4

Table 1: Dataset composition with 1,500 total questions (500 for testing) and text statistics showing mean word count of 23.2, reflecting concise question formulations.

3.4 Data Verification and Categorization

Following the data filtering phase, we enhanced the dataset through manual verification and systematic categorization. All mathematical expressions and physics terminology were manually verified for accuracy by domain experts. Using Gemini 2.5 Pro, we analyzed each question to assign appropriate topic and question type labels, ensuring comprehensive coverage across different physics domains and cognitive complexity levels.

Table 1 presents the dataset composition and statistical overview, with 500 questions allocated for testing purposes and comprehensive text statistics showing the linguistic characteristics of the dataset.

We categorized the dataset along two primary dimensions based on systematic analysis:

Topics (six categories): Electromagnetism, Mechanics, Thermodynamics, Wave Optics, Quantum Physics, and Modern Physics, reflecting comprehensive coverage of fundamental and advanced physics concepts.

Question Types (six categories): Based on the distribution analysis, we identified and categorized questions into Mathematical, Theoretical, Definition, Reasoning, Application, and Miscellaneous types, capturing different cognitive demands and problem-solving approaches.

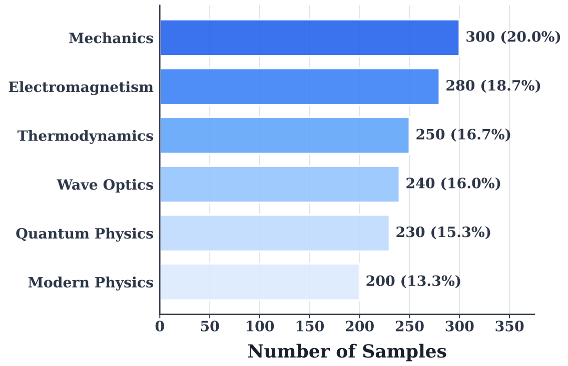


Figure 3: Physics topic distribution with Mechanics dominating (300 samples, 20%) followed by Electromagnetism (280, 18.7%), Thermodynamics (250, 16.7%), Wave Optics (240, 16%), Quantum Physics (230, 15.3%), and Modern Physics (200, 13.3%).

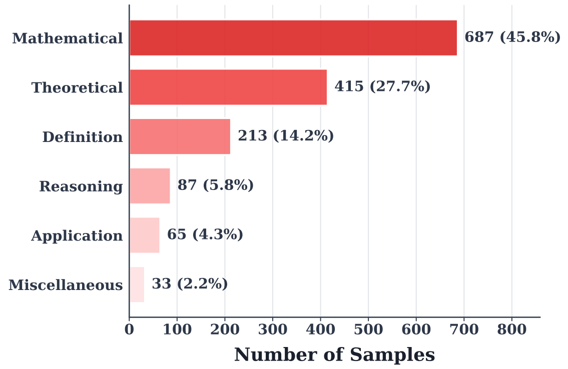


Figure 4: Question type distribution showing Mathematical questions as the largest category (687, 45.8%), followed by Theoretical (415, 27.7%), Definition (213, 14.2%), Reasoning (87, 5.8%), Application (65, 4.3%), and Miscellaneous (33, 2.2%).

Figure 3 presents the distribution across physics topics, showing that Mechanics dominates the topic distribution with 300 samples, followed by Electromagnetism (280 samples) and Thermodynamics (250 samples). Figure 4 shows the question type distribution, where Mathematical questions constitute the largest portion with 687 samples, reflecting the quantitative nature of physics education, while Miscellaneous questions represent the smallest category with 33 samples.

4 Experiment Setup

4.1 Bangla QA based Prompting

We employ *zero-shot prompting* to evaluate the reasoning capabilities of large language models (LLMs). For each question instance, the input to the model consists of three components: a system prompt P, the Bangla natural language question Q_{BAN}, and a set of candidate answer options $O_{\text{BAN}} = \{o_1, o_2, o_3, o_4\}$ where every o_i is given in Bangla. The model is tasked with selecting the optimal answer $o^* \in O_{\text{BAN}}$.

The system prompt P is designed to encourage deliberative reasoning by instructing the model to think step-by-step before committing to a final answer. This aligns with recent trends in prompt engineering where reasoning-based instructions can improve model performance, particularly in multi-step or ambiguous queries.

We evaluate three state-of-the-art LLM families in this setup: `GPT`, `Gemini`, and `DeepSeek`. All models are evaluated under identical prompting conditions to ensure fair comparison. The detailed prompt is given in the Appendix.

4.2 Translation-based Prompting

Most reasoning-capable large language models (LLMs) exhibit a strong bias toward English-language reasoning, primarily due to the predominance of English data in their pretraining corpora. Motivated by this, we also evaluate the performance of these models on an English-translated version of our dataset.

To create this version, we translated both the questions and the corresponding answer options into English. The prompt used for translation is provided in the Appendix. For this task, we employ `Gemini-2.5 Pro` and manually verify the generated translations, making corrections where necessary to ensure accuracy and consistency.

Following translation, we conduct zero-shot prompting experiments analogous to our original setup. Each input to the model consists of three components: a system prompt P, an English natural language question Q_{ENG}, and a set of answer options $O_{\text{ENG}} = \{o_1, o_2, o_3, o_4\}$, where each o_i is provided in Bangla. The model is tasked with selecting the optimal answer $o^* \in O_{\text{ENG}}$.

4.3 Evaluation Metrics

We evaluate model performance using a retry-based metric we term **Cumulative Accuracy@k** (Cumul Acc@k), where $k \in \{1, 2, 3, 4\}$. Unlike the standard Pass@k metric, which assumes access to k simultaneous guesses per sample, our approach allows the model up to k sequential attempts to answer each question correctly.

Formally, let S_1 be the original set of N questions. At each round $t \in \{1, 2, \ldots, k\}$, we define

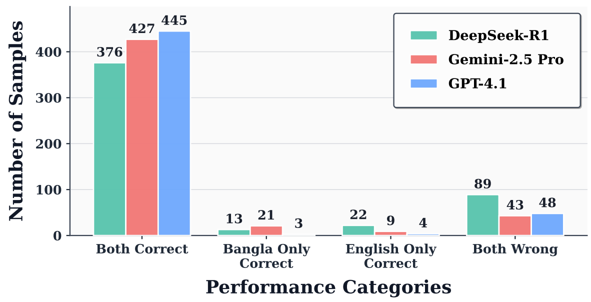


Figure 5: Cross-linguistic performance comparison across four categories: Both Correct, Bangla Only Correct, English Only Correct, and Both Wrong. GPT-4.1 achieves the highest consistency (445 both correct), while DeepSeek-R1 shows the most failures (89 both wrong).

S_t as the subset of questions not answered correctly in any of the previous $t-1$ rounds. In each round, the model is re-prompted with only the remaining incorrect samples from the prior round. Let $a_t(q)$ denote the model's prediction for question q in round t, and let o^*_q be the correct answer.

The Cumulative Accuracy@k is defined as:

$$\text{Cumul Acc@}k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\Big\{ \exists t \leq k \text{ such that } a_t(q_i) = o^*_{q_i} \Big\}$$

This metric captures the proportion of total questions that the model eventually answers correctly within k retries. It reflects a realistic, interactive usage scenario where a system is allowed to reattempt difficult questions rather than relying solely on a single response. We report Cumulative Accuracy@k for $k = 1, 2, 3, 4$.

5 Results and Analysis

Our evaluation of state-of-the-art LLMs on P6Jiggasha reveals substantial performance gaps and distinct behavioral patterns across models, languages, and question types, with resubmission strategies proving particularly effective for initially lower-performing models.

Cross-Linguistic Consistency Analysis. Figure 5 reveals multilingual robustness patterns across models. GPT-4.1 exhibits the strongest consistency with 445 questions correct in both languages, followed by Gemini-2.5 Pro (427) and DeepSeek-R1 (376). Language-specific patterns show GPT-4.1 with minimal Bangla-only successes (3) versus higher English-only successes (4), while DeepSeek-R1 and Gemini-2.5 Pro show balanced language-specific performance (13-21 and 22-9, respectively). The "Both Wrong" category highlights fundamental gaps, with DeepSeek-R1 (89) showing significantly more failures than GPT-4.1 (48) and Gemini-2.5 Pro (43).

Table 2 presents the comprehensive results across all evaluated models and languages.

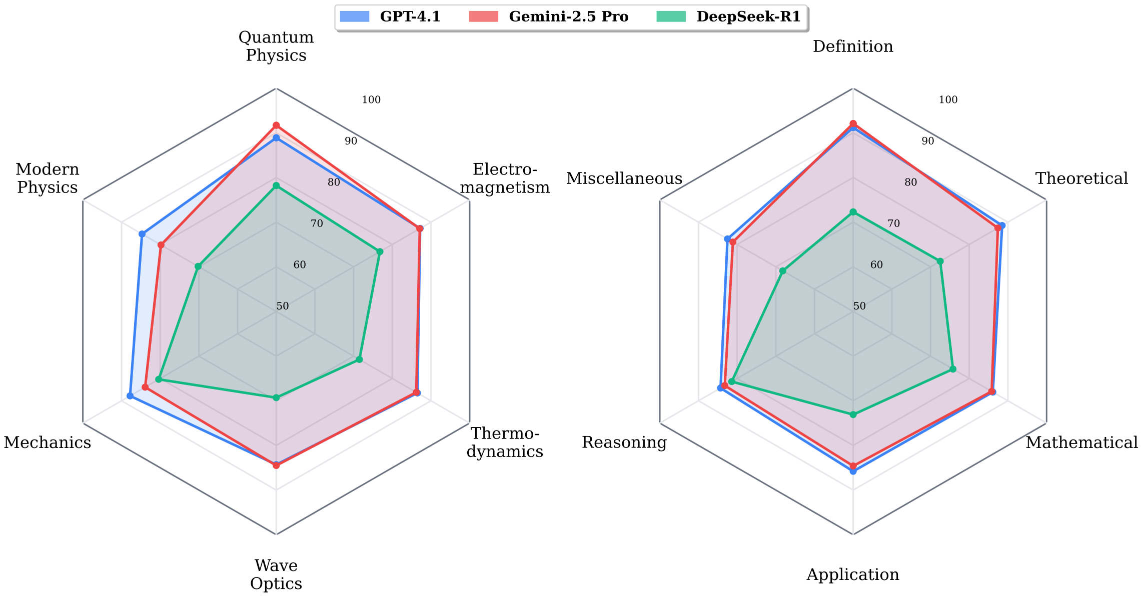
LLMs Comparison. Table 2 reveals significant performance variations across models in the initial inference round (*Cumul Acc@1*). GPT-4.1 demonstrates superior performance on Bangla questions with 86.67% accuracy, substantially outperforming Gemini-2.5 Pro (71.24%) and DeepSeek-R1 (48.00%). This 38.67 percentage point gap between the best and worst performing models highlights the varying capabilities of current LLMs in handling Bangla physics content. Notably, the performance hierarchy (GPT-4.1 $>$ Gemini-2.5 Pro $>$ DeepSeek-R1) remains consistent across both languages, though the magnitude of gaps varies, with DeepSeek-R1 showing particularly weak initial performance in Bangla compared to English.

Impact of Iterative Re-evaluation Strategy. The cumulative accuracy improvements across multiple inference iterations (*Cumul Acc@2–4*) demonstrate the effectiveness of the resubmission strategy. Gemini-2.5 Pro shows the most dramatic improvement, gaining 18.28 percentage points from *Cumul Acc@1* (71.24%) to *Cumul Acc@4* (89.52%), ultimately matching GPT-4.1's final performance. DeepSeek-R1 exhibits the largest absolute gain of 29.71 percentage points in Bangla, though still achieving the lowest final accuracy at 77.71%. GPT-4.1 shows the smallest improvement (2.85 percentage points), indicating high initial accuracy with limited room for enhancement through resubmission. The diminishing returns pattern across iterations (largest gains at *Cumul Acc@2*, smaller at subsequent attempts) suggests that most correctable errors are resolved within the first retry, with marginal benefits from additional resubmissions.

Cross-lingual Performance Analysis. English translation yields mixed results across models. GPT-4.1 shows minimal language dependency with similar performance in both languages (86.67% Bangla *vs.* 84.57% English in *Cumul Acc@1*), ultimately achieving slightly higher final accuracy in English (89.90% *vs.* 89.52%). Conversely, Gemini-2.5 Pro performs better with Bangla questions ini-

Model	**Language**	**Cumulative Accuracy (%)**			
		Cumul Acc@1	**Cumul Acc@2**	**Cumul Acc@3**	**Cumul Acc@4**
GPT-4.1	Bangla	86.67	88.57	89.00	89.52
GPT-4.1	English	84.57	88.38	89.14	89.90
Gemini-2.5 Pro	Bangla	71.24	78.86	84.95	89.52
Gemini-2.5 Pro	English	68.32	83.81	85.52	87.25
DeepSeek-R1	Bangla	48.00	72.00	72.19	77.71
DeepSeek-R1	English	62.67	65.52	78.67	79.50

Table 2: Overall model performance across languages and inference iterations, showing GPT-4.1's high initial accuracy with minimal improvement, Gemini-2.5 Pro's dramatic gains through resubmission (+18.28%), and DeepSeek-R1's substantial but insufficient improvement (+29.71% in Bangla).

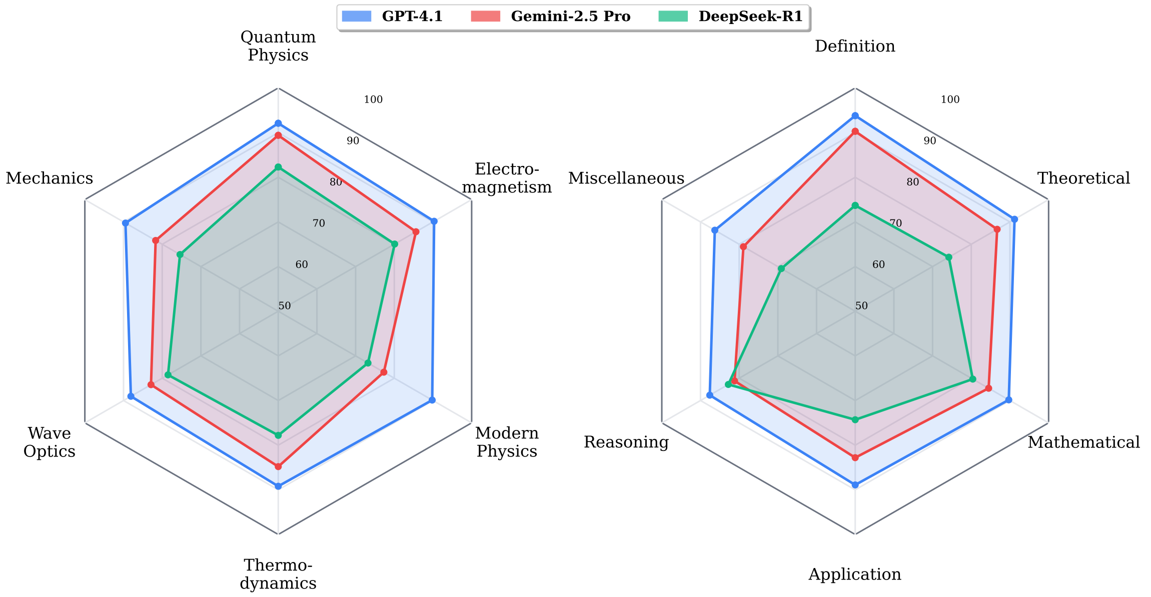


(a) (b)

Figure 6: Performance analysis for Bangla questions: (a) Topic-wise performance showing consistent strength in Quantum Physics and Electromagnetism versus challenges in Modern Physics and Mechanics, (b) Question type-wise performance revealing strong accuracy on Definition/Theoretical questions but progressive degradation for Application/Reasoning tasks.

tially (71.24% *vs.* 68.32%) but shows stronger improvement trajectories in English during resubmissions, reaching 87.25% final accuracy. DeepSeek-R1 demonstrates a notable preference for English at initial inference (62.67% *vs.* 48.00%, a 14.67 percentage point gap), though Bangla performance catches up through multiple iterations (77.71% *vs.* 79.50% final), suggesting language-specific reasoning patterns and imbalanced training data distribution that can be partially compensated through iterative refinement.

Topic-wise Performance Analysis. Figure 6(a) and Figure 7(a) show topic-wise performance on Bangla and English questions using *Cumul Acc@4*. All models perform better on *Quantum Physics* and *Electromagnetism*, with GPT-4.1 and Gemini-2.5 Pro achieving 88–95% accuracy across most domains. *Modern Physics* and *Mechanics* remain more challenging, especially for DeepSeek-R1, which trails by 10–20 percentage points. DeepSeek-R1 also struggles notably with *Wave Optics* on Bangla questions, showing the lowest accuracy among topics. Overall, radar charts indicate stable performance hierarchies, with GPT-4.1 leading consistently across all six domains. Cross-linguistically, models perform slightly better and more consistently on Bangla, while English results exhibit more compressed performance ranges.



(a) (b)

Figure 7: Performance analysis for English questions: (a) Topic-wise performance showing similar domain trends as Bangla with slightly compressed ranges, (b) Question type-wise performance confirming that cognitive complexity outweighs linguistic factors.

Question Type-wise Performance Analysis. Figure 6(b) and Figure 7(b) present type-wise performance on Bangla and English questions. All models excel on *Definition* questions, with GPT-4.1 and Gemini-2.5 Pro near 90% and DeepSeek-R1 around 85%. For *Mathematical* questions, GPT-4.1 leads at 95%, Gemini-2.5 Pro follows at 90%, and DeepSeek-R1 remains competitive at 82%, showing relative strength on reasoning-based queries. DeepSeek-R1 attains about 78% on *Mathematical* questions for both languages, reflecting moderate math solving ability. However, it performs worst on *Miscellaneous* questions (60%), marking the lowest score across all types and revealing challenges with unconventional formats. Consistent trends across both languages indicate that question-type complexity, not language, primarily drives performance differences.

Design Choice	**Cumul@k Acc**			**Pass@k Acc**	
	C@1	**C@2**	**C@3**	**Pass@1**	**Pass@1**
Zero Shot					
GPT-4.1	86.67	88.57	89.00	87.39	89.57
Gemini-2.5 Pro	71.24	78.86	84.95	73.64	76.32
DeepSeek-R1	48.00	72.00	72.19	55.28	67.92
CoT					
GPT-4.1	88.94	92.35	92.87	88.25	90.88
Gemini-2.5 Pro	76.58	84.02	89.72	77.48	82.14
DeepSeek-R1	56.57	78.93	83.83	68.27	76.39

Table 3: Comparison of different prompt in P6JIGGASHA dataset. We consider Bangla Question for this experiment. Here **C@k** means *Cumulative@k Acc*

LLM Prompting Results. Table 3 compares the performance of GPT-4.1, Gemini-2.5 Pro, and DeepSeek-R1 on Bangla questions from the P6JIGGASHA dataset under Zero-Shot and Chain-of-Thought (CoT) prompting. Across all metrics, GPT-4.1 consistently outperforms the other models, achieving high cumulative and pass@k accuracies even in the zero-shot setting, with further improvements under CoT prompting. CoT prompts benefit all models, with smaller models like DeepSeek-R1 showing the largest relative gains, indicating that reasoning prompts help weaker models more significantly. As expected, both cumulative@k and pass@k accuracies increase with k, reflecting the higher probability of obtaining a correct answer when multiple candidates are considered. Overall, the results highlight that larger models perform well even without reasoning prompts, while CoT prompts and multiple candidate outputs further enhance performance across models.

6 Error Analysis

We conducted a comprehensive error analysis examining representative failure cases across question types and physics topics. Detailed breakdowns are provided in Appendix A. Our findings reveal performance gaps as well as deeper structural weaknesses in cross-lingual physics reasoning, especially where linguistic variation interacts with domain-specific concepts.

Mathematical and Theoretical Challenges. Figure 8(a,e) shows universal success for straightforward calculations but reveals Gemini-2.5 Pro's inconsistent wave optics encoding. Figure 8(b,d) demonstrates cross-lingual inversions, indicating language-specific rather than unified formula retrieval (see Appendix A, subsection A.1–A.2). These errors suggest reliance on pattern-matching over conceptual understanding, with theoretical items exposing persistent fragility even under multiple inference attempts. Such weaknesses indicate that deeper symbolic or conceptual grounding remains limited across models.

Applied and Practical Reasoning. Figure 8(c,f) shows models performing well in Bangla but collapsing in English, particularly for thermodynamic comfort and engine efficiency (Appendix A, subsection A.3–A.4). These failures illustrate that real-world physics reasoning is highly sensitive to linguistic framing, where small contextual shifts produce divergent interpretations. This highlights that applied problems require robust world-knowledge alignment, which current LLMs struggle to maintain consistently across languages.

Cross-lingual Inconsistencies. Examples (c–f) in Figure 8 show inconsistent answers across languages for identical problems (Appendix A, subsection A.5). This suggests parallel but weakly aligned internal knowledge representations, where identical concepts trigger different reasoning paths depending solely on query language. Such instability poses concerns for bilingual learners who may switch between Bangla and English during study.

Iterative Reasoning Improvements. Figures 9 and 10 show DeepSeek-R1's +29.71% gain through cumulative accuracy, adding verification layers across attempts (Appendix A, subsection A.6). Although multi-round prompting stabilizes reasoning, improvements vary by domain, revealing that many initial errors stem from incomplete intermediate steps. Iterative refinement helps correct shallow mistakes but cannot reliably resolve deeper conceptual gaps.

Translation-Induced Failures. Figure 11 highlights semantic drift during translation that affects performance independently of physics ability (Appendix B). Even with manual verification, subtle phrasing changes can shift interpretation, adding instability—especially in applied or context-sensitive tasks. This indicates that translation artifacts compound existing reasoning vulnerabilities, disproportionately affecting cross-lingual evaluation.

7 Conclusion

This work presents the first comprehensive evaluation of Large Language Models on Bangla physics questions, introducing the `P6Jiggasha` dataset of 1,500 multiple-choice questions across six physics topics and question types. We evaluate three state-of-the-art models—GPT-4.1, Gemini-2.5 Pro, and DeepSeek-R1-Distill-Llama-70B—revealing significant performance variations across languages and complexity levels. We introduce the *Cumulative Accuracy@k* metric for sequential reasoning evaluation and identify systematic challenges in mathematical reasoning, theoretical understanding, and cross-lingual consistency. Our findings highlight that language modality significantly impacts model performance, emphasizing the need for balanced multilingual training data in scientific domains. Future work should explore open-ended formats, multimodal integration, and expansion to other scientific domains and low-resource languages.

Limitations

Our evaluation focuses exclusively on multiple-choice questions, which may not fully capture scientific reasoning required for open-ended explanations. The dataset currently excludes visual elements such as graphs and diagrams; we plan to incorporate multimodal capabilities in future work. Our study is confined to physics, and generalizability to other STEM subjects in Bangla remains uncertain. The observed cross-lingual inconsistencies raise reliability concerns for educational deployment. Furthermore, the English translations—although manually checked—may still introduce subtle semantic shifts that influence model performance, especially for context-dependent reasoning tasks. Our iterative evaluation strategy (CumulAcc@k) improves robustness but does not fully disentangle genuine reasoning improvements from repeated exposure effects.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Ajwad Akil, Najrin Sultana, Abhik Bhattacharjee, and Rifat Shahriyar. 2022. Banglaparaphrase: A high-quality bangla paraphrase dataset. *arXiv preprint arXiv:2210.05109*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *arXiv preprint arXiv:2101.00204*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2022. Banglanlg and banglat5: Benchmarks and resources for evaluating low-resource natural language generation in bangla. *arXiv preprint arXiv:2205.11081*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. Sentnob: A dataset for analysing sentiment on noisy bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271.

Khondoker Ittehadul Islam and Gabriele Sarti. 2025. Reveal-bangla: A dataset for cross-lingual multi-step reasoning evaluation. *arXiv preprint arXiv:2508.08933*.

Mohsinul Kabir, Obayed Bin Mahfuz, Syed Rifat Raiyan, Hasan Mahmud, and Md Kamrul Hasan. 2023. Banglabook: A large-scale bangla dataset for sentiment analysis from book reviews. *arXiv preprint arXiv:2305.06595*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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*.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. *Advances in Neural Information Processing Systems*, 37:66383–66409.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36:11809–11822.

Janice Ahn Zhang, Aidan Ning, Zhiyuan Jiao, Zijian Ma, Yikang Chen, Jiayi Zhang, et al. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

A Detailed Error Analysis with Examples

This section provides comprehensive analysis of model failure patterns observed in the P6Jiggasha dataset, with detailed examination of representative cases across different physics domains and question types.

A.1 Mathematical Reasoning Challenges

Figure 8 presents representative failure cases across different physics domains. Example (a) demonstrates universal success across all models (B, B, B in both languages) for a *Young's modulus* calculation, indicating strong capability when problems require straightforward formula application with clearly identified variables and direct substitution. Example (e) reveals model-specific challenges in wave optics interference calculations involving intensity superposition.

While GPT-4.1 and DeepSeek-R1 correctly identify option A in both languages, Gemini-2.5 Pro consistently fails, selecting B in Bangla and D in English. This language-specific error pattern suggests Gemini struggles with the conceptual understanding of intensity calculations in interference, incorrectly applying linear summation rather than accounting for phase relationships. The fact that Gemini selects different wrong answers across languages indicates the error mechanism itself is language-dependent, suggesting inconsistent encoding of wave optics principles across multilingual representations.

A.2 Theoretical Concept Understanding

In Figure 8(b), the test examines the relationship between diffraction and interference, where DeepSeek-R1 incorrectly selects refraction (D) in Bangla while correctly identifying interference (C) in English. GPT-4.1 and Gemini-2.5 Pro maintain consistency with correct answers in both languages. This suggests DeepSeek's Bangla physics vocabulary may conflate closely related optical phenomena, indicating weaker terminological precision in low-resource language understanding.

Figure 8(d) reveals dramatic cross-lingual inconsistencies in *moment of inertia* formula retrieval. In Bangla, GPT-4.1 fails (A) while Gemini-2.5 Pro and DeepSeek-R1 succeed (D, D). In English, this pattern completely inverts: GPT-4.1 succeeds (D) while Gemini-2.5 Pro and DeepSeek-R1 both fail (A, A). This systematic cross-lingual inversion—where models that succeed in one language fail in the other—provides strong evidence that physics formulas are not stored in language-agnostic representations but rather encoded separately for each language, leading to inconsistent retrieval depending on query language.

A.3 Applied Reasoning and Contextual Understanding

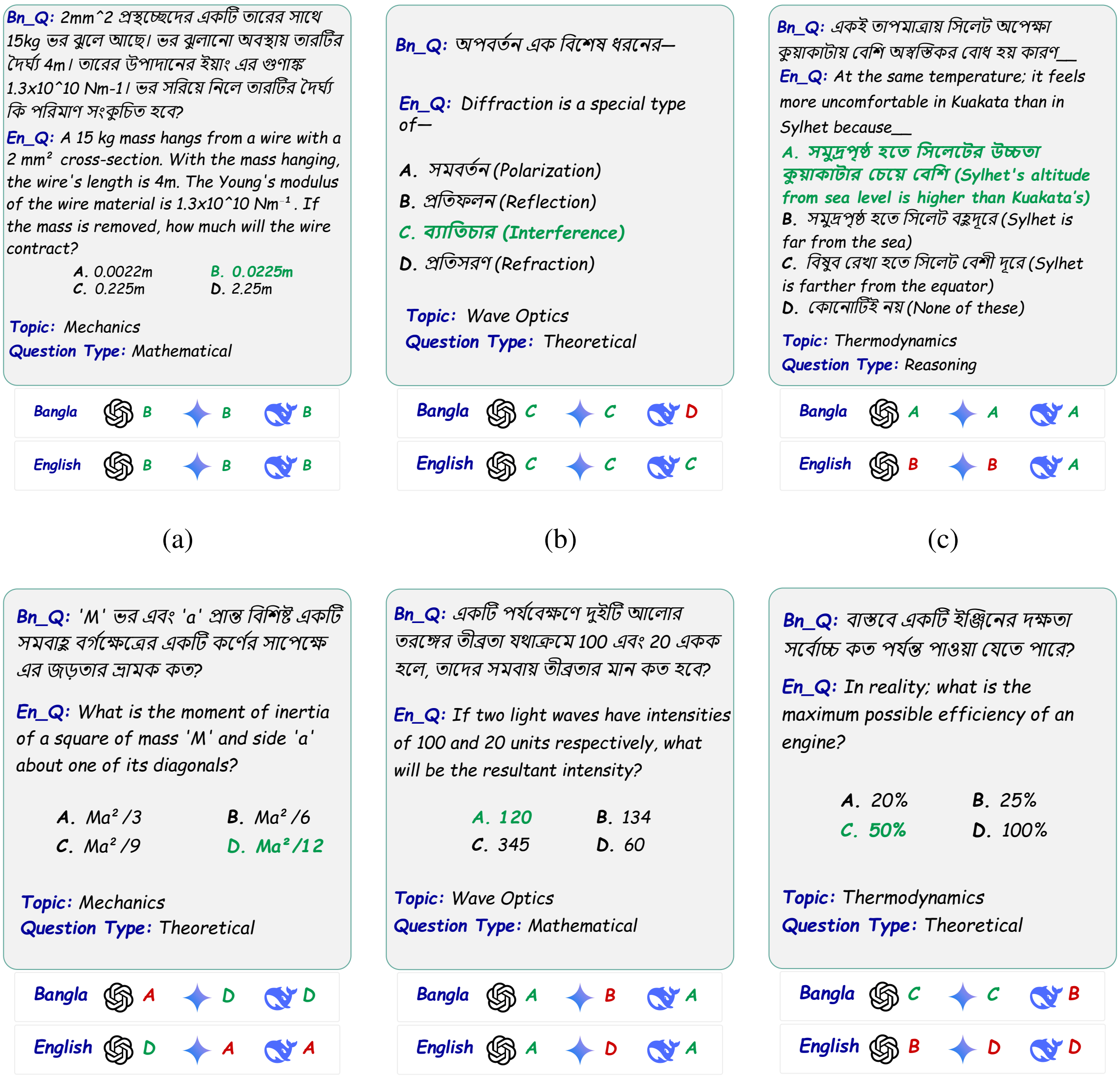
Figure 8(c) demonstrates interesting cross-lingual performance patterns in thermodynamic reasoning about altitude and thermal comfort. All models correctly identify option A in Bangla, but in English, GPT-4.1 and Gemini-2.5 Pro both incorrectly select B (distance from sea), while only DeepSeek-R1 maintains correct reasoning. This suggests that contextual real-world physics applications may be better encoded in Bangla training data for GPT and Gemini, or that English translation introduces semantic ambiguity that misleads these models. The fact that DeepSeek maintains consistency while larger models fail in English is particularly noteworthy, suggesting different architectural or training approaches to handling context-dependent reasoning.

A.4 Practical Physics Understanding

Figure 8(f) examines understanding of realistic engine efficiency limits, revealing significant challenges across all models. In Bangla, GPT-4.1 and Gemini-2.5 Pro both correctly identify C (50%) while DeepSeek-R1 fails (B, 25%). However, in English, *all three models fail*, with GPT-4.1 selecting B (25%), and both Gemini-2.5 Pro and DeepSeek-R1 selecting D (100%). This complete performance collapse in English—where even previously successful models fail—suggests that the conceptual bridge between idealized thermodynamic principles (Carnot efficiency) and practical engineering constraints is poorly represented in English training data. The scattered predictions across different percentage values indicate models confuse theoretical maximums, practical limits, and unrealistic ideal cases, highlighting systematic gaps in applied thermodynamics reasoning.

A.5 Cross-lingual Consistency Issues

Four examples in Figure 8—(c), (d), (e), and (f)—exhibit significant prediction inconsistencies between Bangla and English versions. Example (d) shows the most striking pattern: a complete performance inversion where successful models in



Bn_Q: 2mm^2 প্রস্থচ্ছেদের একটি তারের সাথে 15kg ভর ঝুলে আছে। ভর ঝুলানো অবস্থায় তারটির দৈর্ঘ্য 4m। তারের উপাদানের ইয়াং এর গুণাঙ্ক 1.3x10^10 Nm-1। ভর সরিয়ে নিলে তারটির দৈর্ঘ্য কি পরিমাণ সংকুচিত হবে?

En_Q: A 15 kg mass hangs from a wire with a 2 mm^2 cross-section. With the mass hanging, the wire's length is 4m. The Young's modulus of the wire material is 1.3x10^10 Nm^{-1} . If the mass is removed, how much will the wire contract?

A. 0.0022m
B. 0.0225m
C. 0.225m
D. 2.25m

Topic: Mechanics
Question Type: Mathematical

(a)

Bn_Q: অপবর্তন এক বিশেষ ধরনের—

En_Q: Diffraction is a special type of—

A. সমবর্তন (Polarization)
B. প্রতিফলন (Reflection)
C. ব্যাতিচার (Interference)
D. প্রতিসরণ (Refraction)

Topic: Wave Optics
Question Type: Theoretical

(b)

Bn_Q: একই তাপমাত্রায় সিলেট অপেক্ষা কুয়াকাটায় বেশি অস্বস্তিকর বোধ হয় কারণ__

En_Q: At the same temperature; it feels more uncomfortable in Kuakata than in Sylhet because__

A. সমুদ্রপৃষ্ঠ হতে সিলেটের উচ্চতা কুয়াকাটার চেয়ে বেশি (Sylhet's altitude from sea level is higher than Kuakata's)
B. সমুদ্রপৃষ্ঠ হতে সিলেট বহুদূরে (Sylhet is far from the sea)
C. বিষুব রেখা হতে সিলেট বেশী দূরে (Sylhet is farther from the equator)
D. কোনোটিই নয় (None of these)

Topic: Thermodynamics
Question Type: Reasoning

(c)

Bn_Q: 'M' ভর এবং 'a' প্রান্ত বিশিষ্ট একটি সমবাহু বর্গক্ষেত্রের একটি কর্ণের সাপেক্ষে এর জড়তার ভ্রামক কত?

En_Q: What is the moment of inertia of a square of mass 'M' and side 'a' about one of its diagonals?

A. $Ma^2/3$
B. $Ma^2/6$
C. $Ma^2/9$
D. $Ma^2/12$

Topic: Mechanics
Question Type: Theoretical

(d)

Bn_Q: একটি পর্যবেক্ষণে দুইটি আলোর তরঙ্গের তীব্রতা যথাক্রমে 100 এবং 20 একক হলে, তাদের সমবায় তীব্রতার মান কত হবে?

En_Q: If two light waves have intensities of 100 and 20 units respectively, what will be the resultant intensity?

A. 120
B. 134
C. 345
D. 60

Topic: Wave Optics
Question Type: Mathematical

(e)

Bn_Q: বাস্তবে একটি ইঞ্জিনের দক্ষতা সর্বোচ্চ কত পর্যন্ত পাওয়া যেতে পারে?

En_Q: In reality; what is the maximum possible efficiency of an engine?

A. 20%
B. 25%
C. 50%
D. 100%

Topic: Thermodynamics
Question Type: Theoretical

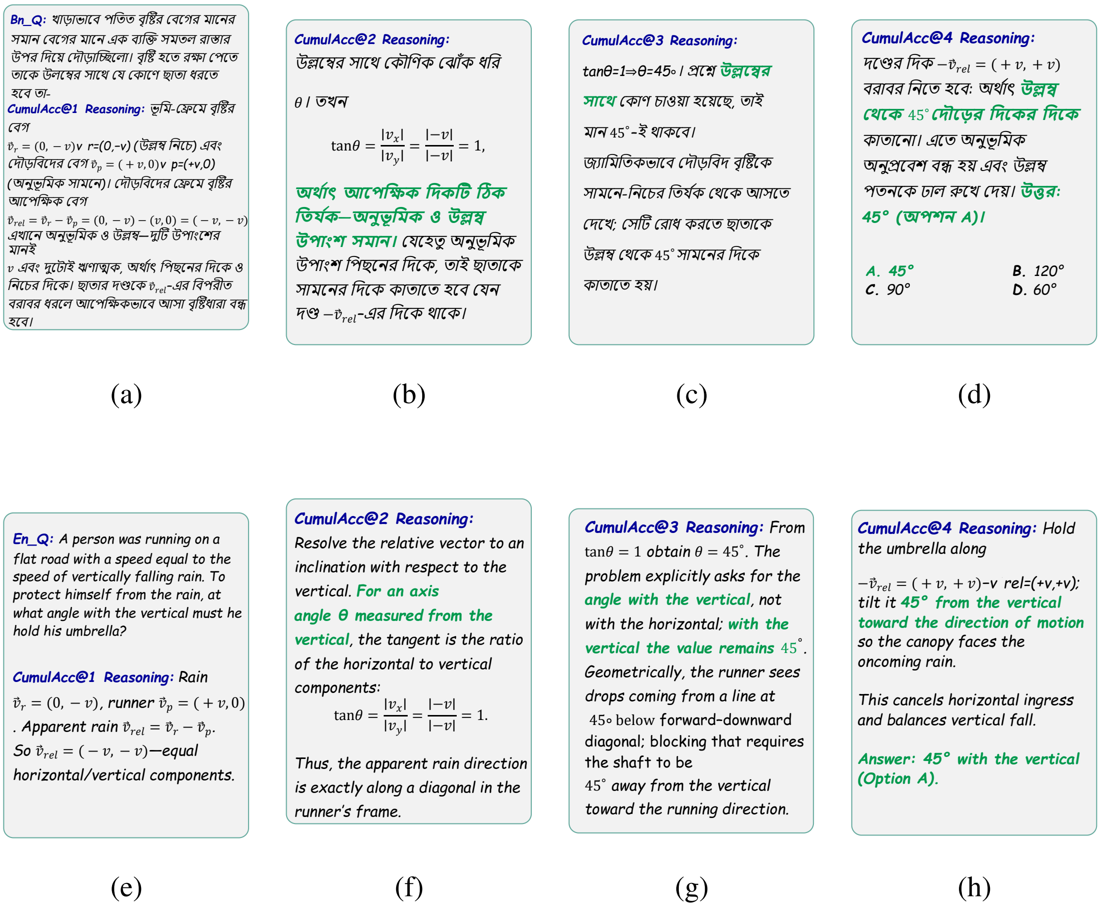
(f)

Figure 8: Representative error cases showing model performance on different question types. Each example displays the question in both Bangla and English, along with predictions from GPT-4.1, Gemini-2.5 Pro, and DeepSeek-R1 (left to right), with correct answers marked in green and incorrect ones marked in red.

Bangla fail in English and vice versa, strongly suggesting language-compartmentalized knowledge rather than unified multilingual understanding. Example (f) demonstrates Bangla superiority (2 of 3 correct) versus English failure (0 of 3 correct), indicating certain physics concepts are better preserved in low-resource language training despite smaller data volumes. These inconsistencies have critical implications for educational deployment: students querying the same problem in different languages may receive contradictory answers, undermining trust and pedagogical effectiveness. The patterns suggest current multilingual LLMs lack cross-lingual alignment mechanisms for domain-specific knowledge, necessitating explicit consistency constraints during training for reliable educational applications.

A.6 Cumulative Reasoning Improvements through Iterative Refinement

Beyond identifying failure patterns, our analysis reveals how DeepSeek-R1's cumulative accuracy mechanism systematically improves reasoning quality through iterative self-correction. Figures 9 and 10 demonstrate this progressive refinement across four iterations, where each cumulative pass adds verification layers that catch and correct errors from previous attempts. In the relative



Bn_Q: খাড়াভাবে পতিত বৃষ্টির বেগের মানের সমান বেগের মানে এক ব্যক্তি সমতল রাস্তার উপর দিয়ে দৌড়াচ্ছিলো। বৃষ্টি হতে রক্ষা পেতে তাকে উলম্বের সাথে যে কোণে ছাতা ধরতে হবে তা-

CumulAcc@1 Reasoning: ভূমি-ফ্রেমে বৃষ্টির বেগ

$\vec{v}_r = (0, -v)$v r=(0,-v) (উল্লম্ব নিচে) এবং দৌড়বিদের বেগ $\vec{v}_p = (+v, 0)$v p=(+v,0) (অনুভূমিক সামনে)। দৌড়বিদের ফ্রেমে বৃষ্টির আপেক্ষিক বেগ

$\vec{v}_{rel} = \vec{v}_r - \vec{v}_p = (0, -v) - (v, 0) = (-v, -v)$

এখানে অনুভূমিক ও উল্লম্ব—দুটি উপাংশের মানই

v এবং দুটোই ঋণাত্মক, অর্থাৎ পিছনের দিকে ও নিচের দিকে। ছাতার দণ্ডকে $\vec{v}_{rel}$-এর বিপরীত বরাবর ধরলে আপেক্ষিকভাবে আসা বৃষ্টিধারা বন্ধ হবে।

(a)

CumulAcc@2 Reasoning:

উল্লম্বের সাথে কৌণিক ঝোঁক ধরি

θ। তখন

$$\tan\theta = \frac{|v_x|}{|v_y|} = \frac{|-v|}{|-v|} = 1,$$

অর্থাৎ আপেক্ষিক দিকটি ঠিক তির্যক—অনুভূমিক ও উল্লম্ব উপাংশ সমান। যেহেতু অনুভূমিক উপাংশ পিছনের দিকে, তাই ছাতাকে সামনের দিকে কাতাতে হবে যেন দণ্ড $-\vec{v}_{rel}$-এর দিকে থাকে।

(b)

CumulAcc@3 Reasoning:

tanθ=1⇒θ=45∘। প্রশ্নে **উল্লম্বের সাথে** কোণ চাওয়া হয়েছে, তাই মান 45°-ই থাকবে।

জ্যামিতিকভাবে দৌড়বিদ বৃষ্টিকে সামনে-নিচের তির্যক থেকে আসতে দেখে; সেটি রোধ করতে ছাতাকে উল্লম্ব থেকে 45° সামনের দিকে কাতাতে হয়।

(c)

CumulAcc@4 Reasoning:

দণ্ডের দিক $-\vec{v}_{rel} = (+v, +v)$ বরাবর নিতে হবে: অর্থাৎ **উল্লম্ব থেকে 45° দৌড়ের দিকের দিকে** কাতানো। এতে অনুভূমিক অনুপ্রবেশ বন্ধ হয় এবং উল্লম্ব পতনকে ঢাল রুখে দেয়। **উত্তর: 45° (অপশন A)।**

A. 45° B. 120°
C. 90° D. 60°

(d)

En_Q: A person was running on a flat road with a speed equal to the speed of vertically falling rain. To protect himself from the rain, at what angle with the vertical must he hold his umbrella?

CumulAcc@1 Reasoning: Rain $\vec{v}_r = (0, -v)$, runner $\vec{v}_p = (+v, 0)$. Apparent rain $\vec{v}_{rel} = \vec{v}_r - \vec{v}_p$. So $\vec{v}_{rel} = (-v, -v)$—equal horizontal/vertical components.

(e)

CumulAcc@2 Reasoning:

Resolve the relative vector to an inclination with respect to the vertical. **For an axis angle θ measured from the vertical**, the tangent is the ratio of the horizontal to vertical components:

$$\tan\theta = \frac{|v_x|}{|v_y|} = \frac{|-v|}{|-v|} = 1.$$

Thus, the apparent rain direction is exactly along a diagonal in the runner's frame.

(f)

CumulAcc@3 Reasoning: From $\tan\theta = 1$ obtain $\theta = 45^\circ$. The problem explicitly asks for the **angle with the vertical**, not with the horizontal; **with the vertical the value remains** 45°. Geometrically, the runner sees drops coming from a line at 45∘ below forward-downward diagonal; blocking that requires the shaft to be 45° away from the vertical toward the running direction.

(g)

CumulAcc@4 Reasoning: Hold the umbrella along

$-\vec{v}_{rel} = (+v, +v)$–v rel=(+v,+v); tilt it **45° from the vertical toward the direction of motion** so the canopy faces the oncoming rain.

This cancels horizontal ingress and balances vertical fall.

Answer: 45° with the vertical (Option A).

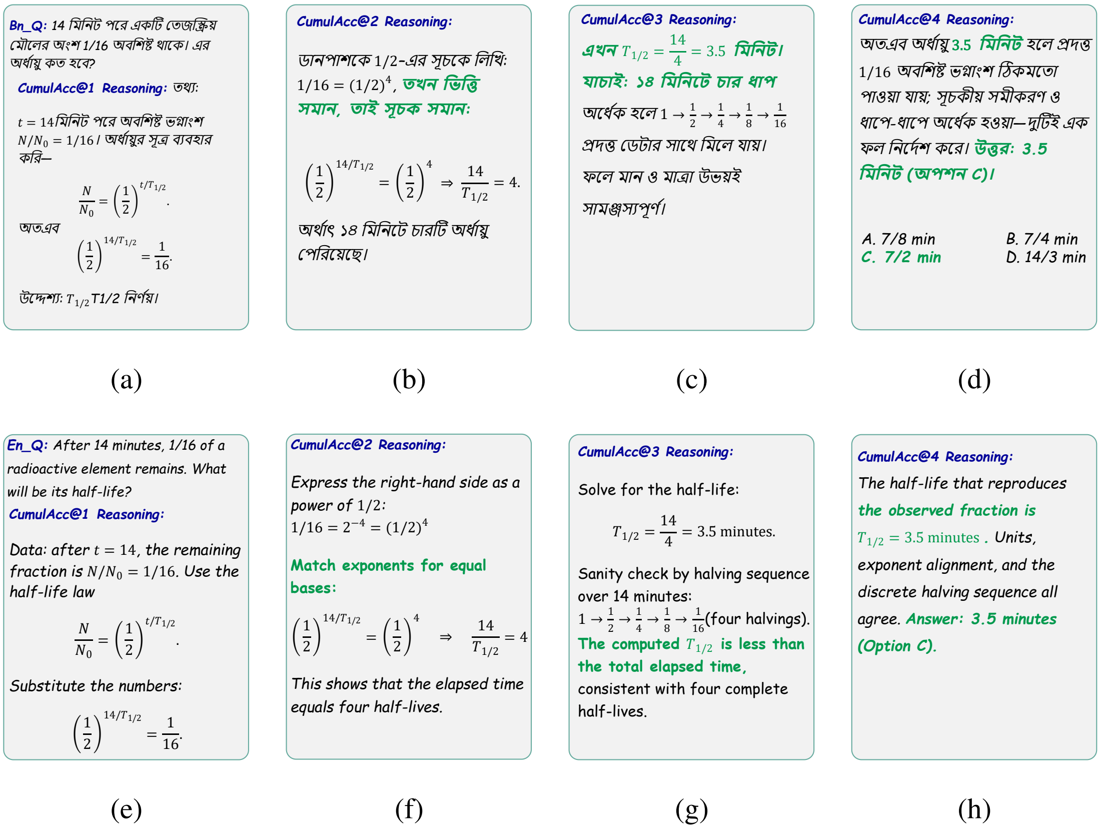
(h)

Figure 9: **DeepSeek-R1 iterative reasoning refinement for a relative motion problem.** Figure shows progressive reasoning across four cumulative accuracy steps for a problem on the umbrella angle needed to avoid rain at matching velocity (top: Bangla a–d, bottom: English e–h). Iterations (a,e) set velocity vectors, (b,f) compute relative velocity and tangent-based angles, (c,g) determine the 45° solution, and (d,h) verify via geometry (Option A). The parallel view highlights cross-lingual consistency in DeepSeek-R1's reasoning, with Bangla using detailed explanations and English favoring concise vector notation.

motion problem (Figure 9), CumulAcc@1 establishes the basic vector framework but lacks angular interpretation; CumulAcc@2 introduces tangent ratio calculations but doesn't fully resolve the geometric meaning; CumulAcc@3 explicitly derives $\theta = 45$ from $\tan\theta = 1$; and CumulAcc@4 provides complete geometric validation by confirming the umbrella orientation blocks the diagonal rain trajectory. This layered verification transforms incomplete initial reasoning into a robust, multi-perspective solution.

Similarly, the radioactive decay problem (Figure 10) showcases cumulative betterment: CumulAcc@1 sets up the exponential formula, CumulAcc@2 matches it to discrete half-life counts, CumulAcc@3 performs the arithmetic calculation with sanity checks, and CumulAcc@4 validates dimensional consistency across exponential and sequential interpretations. The progression from 48.00% (CumulAcc@1) to 77.71% (CumulAcc@4) in Bangla—a remarkable +29.71% absolute improvement—demonstrates that iterative refinement doesn't merely retry the same approach but systematically builds conceptual scaffolding, catching algebraic errors, verifying physical intuition, and ensuring cross-validation between mathematical formalism and discrete physical processes. This cumulative mechanism proves particularly effective for complex multi-step physics problems where single-pass reasoning often misses subtle conceptual connections or calculation errors.

The cross-lingual consistency observed across iterations (Figures 9e-h and 10e-h) further validates that these improvements stem from genuine reasoning refinement rather than language-specific pattern matching, as DeepSeek-R1 achieves similar



Bn_Q: 14 মিনিট পরে একটি তেজস্ক্রিয় মৌলের অংশ 1/16 অবশিষ্ট থাকে। এর অর্ধায়ু কত হবে?

CumulAcc@1 Reasoning: তথ্য:

$t = 14$মিনিট পরে অবশিষ্ট ভগ্নাংশ $N/N_0 = 1/16$। অর্ধায়ুর সূত্র ব্যবহার করি—

$$\frac{N}{N_0} = \left(\frac{1}{2}\right)^{t/T_{1/2}}.$$

অতএব

$$\left(\frac{1}{2}\right)^{14/T_{1/2}} = \frac{1}{16}.$$

উদ্দেশ্য: $T_{1/2}$T1/2 নির্ণয়।

(a)

CumulAcc@2 Reasoning:

ডানপাশকে $1/2$-এর সূচকে লিখি: $1/16 = (1/2)^4$, **তখন ভিত্তি সমান, তাই সূচক সমান:**

$$\left(\frac{1}{2}\right)^{14/T_{1/2}} = \left(\frac{1}{2}\right)^4 \Rightarrow \frac{14}{T_{1/2}} = 4.$$

অর্থাৎ ১৪ মিনিটে চারটি অর্ধায়ু পেরিয়েছে।

(b)

CumulAcc@3 Reasoning:

এখন $T_{1/2} = \frac{14}{4} = 3.5$ মিনিট।

যাচাই: ১৪ মিনিটে চার ধাপ

অর্ধেক হলে $1 \rightarrow \frac{1}{2} \rightarrow \frac{1}{4} \rightarrow \frac{1}{8} \rightarrow \frac{1}{16}$

প্রদত্ত ডেটার সাথে মিলে যায়।

ফলে মান ও মাত্রা উভয়ই

সামঞ্জস্যপূর্ণ।

(c)

CumulAcc@4 Reasoning:

অতএব অর্ধায়ু **3.5 মিনিট** হলে প্রদত্ত $1/16$ অবশিষ্ট ভগ্নাংশ ঠিকমতো পাওয়া যায়; সূচকীয় সমীকরণ ও ধাপে-ধাপে অর্ধেক হওয়া—দুটিই এক ফল নির্দেশ করে। **উত্তর: 3.5 মিনিট (অপশন C)।**

A. 7/8 min B. 7/4 min
C. 7/2 min D. 14/3 min

(d)

En_Q: After 14 minutes, 1/16 of a radioactive element remains. What will be its half-life?

CumulAcc@1 Reasoning:

Data: after $t = 14$, the remaining fraction is $N/N_0 = 1/16$. Use the half-life law

$$\frac{N}{N_0} = \left(\frac{1}{2}\right)^{t/T_{1/2}}.$$

Substitute the numbers:

$$\left(\frac{1}{2}\right)^{14/T_{1/2}} = \frac{1}{16}.$$

(e)

CumulAcc@2 Reasoning:

Express the right-hand side as a power of $1/2$:
$1/16 = 2^{-4} = (1/2)^4$

Match exponents for equal bases:

$$\left(\frac{1}{2}\right)^{14/T_{1/2}} = \left(\frac{1}{2}\right)^4 \Rightarrow \frac{14}{T_{1/2}} = 4$$

This shows that the elapsed time equals four half-lives.

(f)

CumulAcc@3 Reasoning:

Solve for the half-life:

$$T_{1/2} = \frac{14}{4} = 3.5 \text{ minutes}.$$

Sanity check by halving sequence over 14 minutes:
$1 \rightarrow \frac{1}{2} \rightarrow \frac{1}{4} \rightarrow \frac{1}{8} \rightarrow \frac{1}{16}$(four halvings). **The computed $T_{1/2}$ is less than the total elapsed time,** consistent with four complete half-lives.

(g)

CumulAcc@4 Reasoning:

The half-life that reproduces **the observed fraction is $T_{1/2} = 3.5$ minutes .** Units, exponent alignment, and the discrete halving sequence all agree. **Answer: 3.5 minutes (Option C).**

(h)

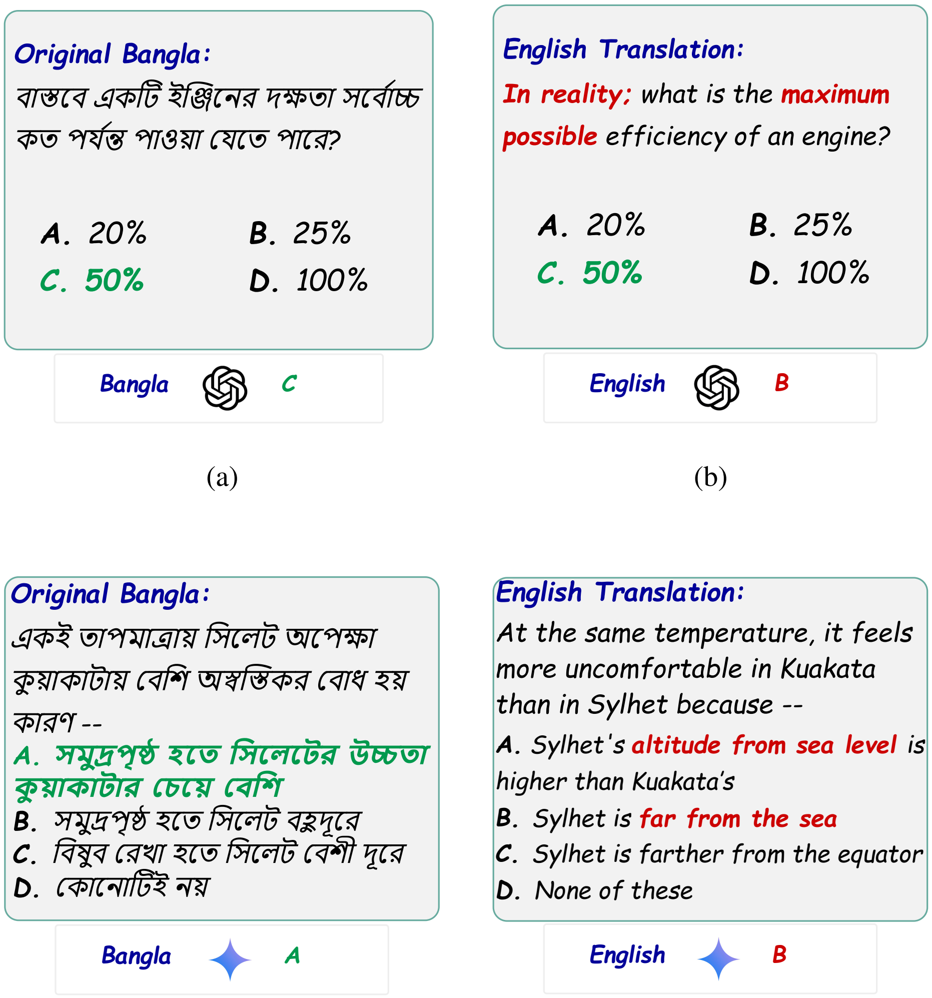
Figure 10: **DeepSeek-R1 iterative reasoning refinement for a radioactive decay problem.** Figure shows progressive reasoning across four cumulative accuracy steps for a half-life problem where 1/16 of a radioactive element remains after 14 minutes (top: Bangla a–d, bottom: English e–h). Iterations (a,e) set up the decay formula, (b,f) express 1/16 as $(1/2)^4$, (c,g) compute $T_{1/2} = 3.5$ min, and (d,h) verify dimensional consistency (Option C). This illustrates DeepSeek-R1's systematic decomposition of exponential decay problems into verifiable steps.

cumulative gains in English (62.67% to 79.50%, +16.83%). The success of this iterative approach suggests that educational AI systems should prioritize multi-pass reasoning architectures over single-shot prediction, enabling students to see how expert problem-solving progressively builds from foundations to verified solutions.

B Translation-Induced Evaluation Failures

Our analysis reveals systematic cross-lingual inconsistencies where translation artifacts alter model interpretations. Figure 11 illustrates two physics problems in which GPT-4.1 and Gemini-2.5 Pro provide different answers across Bangla and English versions due to semantic degradation during translation. In the first case, concerning maximum real engine efficiency (correct answer: 50%), GPT-4.1 correctly identifies the Bangla version's practical engineering framing and selects the right answer. However, in the English translation, the phrase “In reality” introduces semantic ambiguity, creating tension between practical and theoretical interpretations. As a result, GPT-4.1 shifts its reasoning toward typical operational ranges (25%) rather than engineering limits. In the second case, regarding why Kuakata feels more uncomfortable than Sylhet at equal temperatures (correct answer: A, based on altitude affecting atmospheric pressure and evaporative cooling), the Bangla version presents a truncated, obviously incorrect distractor, allowing Gemini-2.5 Pro to select the correct option. The English translation completes this distractor but alters its semantics, making it appear plausible. Consequently, Gemini-2.5 Pro chooses the wrong answer in English, despite understanding the physics correctly in Bangla.

These examples demonstrate that translation-induced failures can occur at multiple linguistic levels. Question-stem translations may introduce contextual ambiguity, as in the engine efficiency example, while option-level translation quality can



Original Bangla:
বাস্তবে একটি ইঞ্জিনের দক্ষতা সর্বোচ্চ কত পর্যন্ত পাওয়া যেতে পারে?

A. 20%
B. 25%
C. 50%
D. 100%

Bangla
C

(a)

English Translation:
In reality; what is the maximum possible efficiency of an engine?

A. 20%
B. 25%
C. 50%
D. 100%

English
B

(b)

Original Bangla:
একই তাপমাত্রায় সিলেট অপেক্ষা কুয়াকাটায় বেশি অস্বস্তিকর বোধ হয় কারণ --

A. সমুদ্রপৃষ্ঠ হতে সিলেটের উচ্চতা কুয়াকাটার চেয়ে বেশি
B. সমুদ্রপৃষ্ঠ হতে সিলেট বহুদূরে
C. বিষুব রেখা হতে সিলেট বেশী দূরে
D. কোনোটিই নয়

Bangla
A

(c)

English Translation:
At the same temperature, it feels more uncomfortable in Kuakata than in Sylhet because --

A. Sylhet's altitude from sea level is higher than Kuakata's
B. Sylhet is far from the sea
C. Sylhet is farther from the equator
D. None of these

(d)

Figure 11: Figure illustrates translation-induced evaluation failures in GPT-4.1 and Gemini-2.5 Pro. (a,b) Engine efficiency: GPT-4.1 answers correctly in Bangla (C, 50%) but selects B (25%) in English due to semantic ambiguity ("In reality" vs. "maximum possible"). (c,d) Thermodynamic comfort: Gemini-2.5 Pro selects A in Bangla but B in English because option translation alters meaning. Green marks correct answers; red highlights translation-induced faults.

change distractor plausibility, as seen in the thermodynamic comfort case. Importantly, aggregate performance metrics mask these failures: both models maintain high overall accuracy (GPT-4.1: 89.52% Bangla, 89.90% English; Gemini-2.5 Pro: 89.52% Bangla, 87.25% English), yet individual question analysis reveals systematic cross-lingual interpretation divergences. This has critical implications for multilingual educational AI, where students working in different languages may face substantively different reasoning challenges even on ostensibly identical problems.

Based on these findings, we recommend several practices for multilingual benchmark construction: preserving contextual precision across languages, validating semantic equivalence at the option level to avoid inadvertently changing distractor quality, analyzing cross-lingual response patterns to flag potential translation issues, and incorporating expert review protocols to ensure both question stems and options maintain the intended meaning. Figure 11 highlights that even without errors in answer key mapping, translation-altered semantics can undermine evaluation reliability, calling for careful consideration in the design of cross-lingual benchmarks.

C System Prompts

Single Inference (Cumul Acc@1). For initial evaluation, we used the following system prompt:

System Prompt for Bangla Physics MCQ (Single Inference)

You are a world-class Physics teacher specialized in solving MCQs in Bangla.
Your role:

- Read the question carefully. Questions may be in Bangla with English symbols, formulas, or expressions mixed in.
- Identify the topic (e.g., Mechanics, Thermodynamics, Wave Optics, Electromagnetism, Quantum Physics, Modern Physics).
- Identify the question type (Mathematical, Theoretical, Definition, Reasoning, Application, Miscellaneous).
- Work through the problem step by step, showing reasoning and calculations in Bangla where natural, but keep math notation in standard English symbols.
- Always check units, conversions, and scientific notation.
- Be concise but accurate in reasoning.
- At the end, give the correct option as a single letter: A, B, C, or D — nothing else on the final line.

Iterative Re-evaluation (Cumul Acc@2-4). For subsequent attempts on incorrectly answered questions, we modified the prompt to encourage deeper analysis:

System Prompt for Bangla Physics MCQ (Iterative Re-evaluation)

You are a world-class Physics teacher specialized in solving MCQs in Bangla.
Your role:

- Analyze each question deeply and answer carefully, rechecking the reasoning step by step, rethink and answer wisely.
- Read the question carefully. Questions may include English text mixed with formulas, symbols, or expressions.
- Identify the topic (e.g., Mechanics, Thermodynamics, Wave Optics, Electromagnetism, Quantum Physics, Modern Physics).
- Identify the question type (Mathematical, Theoretical, Definition, Reasoning, Application, Miscellaneous).
- Work through the problem step by step, showing reasoning and calculations.
- Always check units, conversions, and scientific notation.
- Be concise but accurate in reasoning.
- At the end, provide the correct option as a single letter: A, B, C, or D — nothing else on the final line.

C.1 Prompts for English Question Answering

Single Inference (Cumul Acc@1).

System Prompt for English Physics MCQ (Single Inference)

You are a world-class Physics teacher specialized in solving MCQs in English.
Your role:

- Read the question carefully.
- Identify the topic (e.g., Mechanics, Thermodynamics, Wave Optics, Electromagnetism, Quantum Physics, Modern Physics).
- Identify the question type (Mathematical, Theoretical, Definition, Reasoning, Application, Miscellaneous).
- Work through the problem step by step, showing reasoning and calculations clearly, keeping math notation in standard English symbols.
- Always check units, conversions, and scientific notation.
- Be concise but accurate in reasoning.
- At the end, give the correct option as a single letter: A, B, C, or D — nothing else on the final line.

Iterative Re-evaluation (Cumul Acc@2-4).

System Prompt for English Physics MCQ (Iterative Re-evaluation)

You are a world-class Physics teacher specialized in solving MCQs in English.
Your role:

- Analyze each question deeply and answer carefully, rechecking the reasoning step by step, rethink and answer wisely.
- Read the question carefully.
- Identify the topic (e.g., Mechanics, Thermodynamics, Wave Optics, Electromagnetism, Quantum Physics, Modern Physics).
- Identify the question type (Mathematical, Theoretical, Definition, Reasoning, Application, Miscellaneous).
- Work through the problem step by step, showing reasoning and calculations clearly.
- Always check units, conversions, and scientific notation.
- Be concise but accurate in reasoning.
- At the end, provide the correct option as a single letter: A, B, C, or D — nothing else on the final line.

C.2 Translation Prompt for Dataset Creation

For creating the English version of our dataset, we used Gemini-2.5 Pro with manual verification. The translation prompt was designed to preserve mathematical expressions and technical terminology:

Prompt for Bangla-to-English Translation

You are a professional translator.
Task: Translate ONLY Bangla text into English, keep English text unchanged.
Input format (comma-separated):
id,question,A,B,C,D,answer
Output rules:

- Output must use commas (,) as the column separator.
- Exactly 7 columns: id,question,A,B,C,D,answer
- In the question column: NEVER output commas. If the translation needs a comma, replace it with a semicolon (;) instead.
- Do not add extra columns.
- Do not add explanations, comments, or metadata.
- Keep numbers, math expressions, and symbols unchanged.
- Leave English text unchanged if it already exists.

C.3 Topic and Question Type Categorization Prompt

For systematic categorization of questions across physics domains and cognitive complexity levels, we used Gemini-2.5 Pro with the following prompt:

Prompt for Topic and Question Type Classification

You are a Physics exam classifier.
Task: Analyze each question (and optionally the options and correct answer) to determine:

1. The most relevant Physics topic from the allowed list.
2. The most appropriate Question Type from the allowed list.

Allowed Topics:
Electromagnetism, Mechanics, Thermodynamics, Wave Optics, Quantum Physics, Modern Physics
Allowed Question Types:
Mathematical, Theoretical, Definition, Reasoning, Application, Miscellaneous
Output Instructions:

- Input is CSV with columns: id,question,A,B,C,D,answer
- For each row, return ONLY: id,topic,question_type
- One line per question
- No extra text, commentary, or explanations
- If unsure, use Miscellaneous for both topic and question_type

Example Input:
1,Diffraction is a special type of—,Polarization,Reflection,Interference,Refraction,C
2,If the momentum of an object of mass 50 kg is 200 kgms-1; its kinetic energy will be-,200 J,300 J,400 J,500 J,C
Example Output:
1,Wave Optics,Theoretical
2,Mechanics,Mathematical