B-REASO: A Multi-Level Multi-Faceted Bengali Evaluation Suite for Foundation Models

Md Tanzib Hosain[1,3,], Md Kishor Morol[2,3]
[1]American International University-Bangladesh, [2]Cornell University
[3]ELITE Research Lab
20-42737-1@student.aiub.edu, mmorol@cornell.edu

Abstract

The fast growth of large language models (LLMs) necessitates the urgent need for new NLP benchmarks. We provide B-REASO, the first inclusive Bengali assessment suite created to evaluate advanced foundation model knowledge and reasoning skills in a Bengali language setup. The B-REASO includes multiple-choice questions with four different degrees of difficulty: professional, college, high school, and middle school. The questions cover 50 different fields, from science and engineering to the humanities. Alongside B-REASO, there is B-REASO HEAVY, a subset of extremely difficult B-REASO topics that need for sophisticated reasoning skills to answer. We do a thorough assessment of the most sophisticated LLMs on B-REASO, encompassing models with an English focus. Findings show that only Claude-3.5-Sonnet was able to get an average accuracy of more than 65%, indicating that contemporary LLMs still have a long way to go. We hope that B-REASO will support the creation and expansion of foundation models for Bengali users by assisting in the analysis of significant advantages and disadvantages of these models. We open-source our code and data at `https://github.com/kraritt/b-reaso`.

1 Introduction

Natural Language Processing (NLP) has seen a revolution with the advent of Large Language Models (LLMs). In the age of LLMs, assessment plays a critical function as the central force that shapes the direction of Artificial Intelligence (AI) progress. To assess LMs' natural language understanding skills, traditional NLP benchmarks (Wang et al., 2019; Wang, 2018) have been frequently used. However, since bigger models are performing at a level comparable to humans, these standards are no longer as applicable, leaving limited area for further study (Liu et al., 2023; Goyal et al., 2022; Hendrycks et al., 2020).

Recent LLM evaluation has focused mostly on assessing sophisticated reasoning skills and exploring advanced world knowledge in order to adequately benchmark LLMs (Wang et al., 2023; Cobbe et al., 2021; Hendrycks et al., 2021; Clark et al., 2018). Meanwhile, as multilingual and other LLMs tailored for various regional languages (Nguyen et al., 2023) have become more popular, evaluation benchmarks for languages other than English–have also been introduced (Huang et al., 2024; Li et al., 2023). Nevertheless, the English language remains the primary focus of the contemporary benchmarks, which leads to a restricted comprehension of LLMs' proficiency in other languages rapidly. As per our best knowledge, there is no comprehensive previous work that has been done on LLMs' understanding specifically in Bengali language setup. This study focuses on assessing foundation models' advanced capabilities in the context of Bengali–world's seventh[1] most commonly spoken languages.

We introduce B-REASO (Figure 1), the first complete Bengali evaluation suite designed to fully evaluate LLMs' advanced knowledge and reasoning skills in a Bengali environment, in an effort to close the gap between Bengali LLM development and assessment. As shown in Table 1, B-REASO comprises 13497 multiple-choice test questions covering 50 different fields, from science and engineering to the humanities. Middle school, high school, college, and professional exams are the four difficulty levels from which the questions are gathered (details in Appendix G). We also provide B-REASO HEAVY as an accompanying benchmark, which is a subset of B-REASO's most difficult tasks that need very sophisticated thinking skills to complete, such as, college physics and advanced mathematics. With an accuracy of 48.53%, B-REASO HEAVY is the first Bengali benchmark

---

[1] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9260–9274
November 4-9, 2025 ©2025 Association for Computational Linguistics

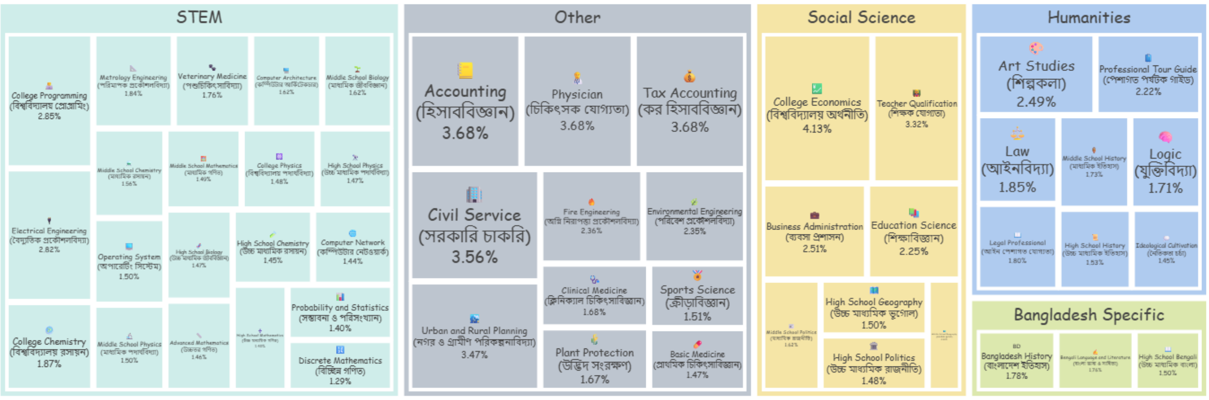

Figure 1: B-REASO overview diagram. The five fields—STEM, social science, humanities, Bangladesh specific and other—are represented by different colors for the subjects.

at this level and one of the few benchmarks for advanced reasoning where Gemini-1.5-Pro-002 still has trouble.

2 B-REASO Benchmark

2.1 Principles of Benchmark Design

Table 1: Statistics of B-REASO benchmark grouped by various fields of subjects and levels.

Group	#Fields	#Instances (%)
	grouped by field	
Social Science	8	2394 (17.74%)
STEM	20	4495 (33.30%)
Humanities	8	1997 (14.80%)
Others	11	3932 (29.13%)
Bangladesh Specific	3	679 (5.03%)
	grouped by level	
College	23	5798 (42.96%)
High School	8	1594 (11.81%)
Professional	12	4696 (34.79%)
Middle School	7	1409 (10.44%)
Total	50	13497 (100.00%)

B-REASO's goal is to assist developers in rapidly comprehending the capabilities of their models from a variety of angles so they can identify the models' flaws and make the necessary improvements. In order to do this, we concentrate on LLMs' more sophisticated skills, such reasoning and global knowledge, which are perhaps the most important ones for LLMs in the modern workforce. Complex tasks are often the primary differentiators between LLMs, even when multiple LLMs may perform equally in simple situations like informal talks (Achiam et al., 2023). As a result, we built B-REASO using actual, difficult human tests that are utilized in Bangladesh to evaluate people's skills in a variety of ways. Like (Ghahroodi et al., 2024; Hendrycks et al., 2021), we only choose multiple-choice questions because: (a) metrics are well-defined; and (b) multiple-choice questions are a straightforward but effective stand-in for assessing the potential of advanced abilities of foundation models, which we believe could be readily utilized and reflected in a variety of downstream applications through specialized instruction tuning (Chung et al., 2024; Wang et al., 2022). Only one of the four options (details in Appendix E) for each question is the right response. Through prompting, LLMs are meant to be used to answer these queries. 50 different disciplines are covered by the B-REASO questions, which we then group into more general categories like STEM, the Humanities, Social Science, Bangladesh Specific and Other fields.

2.2 Principles of Data Collection

2.2.1 Precis of Subject Selection

Middle school, high school, college, and professional are the four difficulty levels covered by B-REASO. With the exception of English, we cover the core subjects taught in Bengali middle and high schools. We choose 23 representative topics at the college level from the official undergraduate major categories specified by Bangladesh's University Grant Commission[2]. To ensure completeness, at least one subject from each category is included in B-REASO. We choose 12 sample professional-level qualifications, including those for doctors, lawyers, and civil servants, based on Bangladesh Civil Service[3] qualification directory. In addition, we group these topics into five topical categories: Social Science, Humanities, STEM

---

[2] https://ugc.gov.bd/

[3] https://bpsc.gov.bd/

(Science, Technology, Engineering, and Mathematics), Bangladesh Specific and Other fields. Table 1 shows each of the 50 topics together with the groups to which they belong.

2.2.2 Prevention of Data Contamination

National test questions, including those from Bangladesh's national professional examinations and college entrance exams are often made available online and disseminated extensively. As a result, these queries could unintentionally be stumbled upon and included to LLM pretraining, which might result in problems with data leaking. We get our data from either small-scale local tests, such those offered online by certain high schools, or from mock exams following the design principle of (Huang et al., 2024) in order to reduce this danger. Furthermore, rather than coming straight from structured questions or plain text, the majority of B-REASO examples are taken from online PDF or Microsoft Word publications. To create the final structured format, the writers analyze and meticulously annotate these papers; for certain topics, this procedure often entails complicated LaTeX equation translation. This reduces the possibility of data contamination even further.

2.2.3 Precis of Data Sources

Our data is mostly sourced from publicly accessible online mock tests. Furthermore, some of the college-level problems are publicly posted previous test questions from Bangladesh's leading colleges. A small percentage of college questions are national graduate entrance exam practice questions[4]. These questions are not publicly accessible, and we have got their permission to add around 1800 of them to B-REASO.

2.2.4 Precis of Data Processing

Most of the gathered data is in the form of Microsoft Word or PDF documents, with a small percentage also being web pages. The first step is to convert PDF documents into text. Following that, all questions are processed by the writers, either automatically where feasible or manually with the use of OCR toolkits[5]. For STEM subjects including significant symbols and equations, mathematical expressions are standardized into LaTeX(Figure 5) format in accordance with earlier studies (Tam et al., 2024; Hendrycks et al., 2021). Annotators compile and check the generated LaTeX formulae to make sure they are accurate. Furthermore, every instance is verified by a person for quality control in order to eliminate inappropriate instances. Annotations were made by knowledgeable participants (each holds at least a bachelor's degree, with someone having advanced degrees), all of whom are co-authors on this work, guaranteeing that each person gets full credit for their contributions. Additionally, explicit instructions were given on how to acquire participant permission, preserve data privacy, and honor local customs (see Appendix F).

2.2.5 Precis of Explanation Curation

LLMs have shown remarkable performance with reasoning-heavy tasks when using Chain-of-Thought (CoT) reasoning (Kojima et al., 2022; Wei et al., 2022), which demands them to provide a text sequence of the reasoning process along with the final response. Achieving state-of-the-art performance on a variety of tasks, the few-shot version of CoT is more widely utilized than the zero-shot version (Wang et al., 2022; Xie et al., 2024; Zhou et al., 2022; Gao et al., 2023). To enable the possible use of B-REASO in a few-shot CoT context, we provide high-quality explanation data for the development split by combining human annotation with automated creation. To be more precise, we first instruct Claude-3.5-Sonnet to provide a detailed explanation of the ground-truth response, and then we manually edit the explanations to produce the final ones. Details on prompting Claude-3.5-Sonnet are in Appendix A. A dev example with explanations is illustrated in Figure 2.

2.3 B-REASO HEAVY

From B-REASO, we choose eight difficult math, physics, and chemistry subjects to create a distinct, the B-REASO HEAVY benchmark encompasses probability, discrete mathematics, and advanced mathematics, high school mathematics, high school chemistry, college chemistry, college physics, and statistics, as well as high school physics. These topics often necessitate the use of intricate LaTeX equations and difficult-to-solve logical problems. Figure 3 illustrates a case from advanced physics. In line with current initiatives to provide challenging standards for evaluating sophisticated reasoning, B-REASO HEAVY aptitudes (Suzgun et al., 2022; Hendrycks et al., 2021), which are the primary variables that distinguish

---

[4] https://udvash.com/

[5] https://mathpix.com/

প্রশ্নঃ
একটি গাড়ি 10 সেকেন্ডে স্থির অবস্থা থেকে 20 m/s বেগে ত্বরান্বিত হয়। গাড়িটি স্থির ত্বরণে চলে। এই সময়ে গাড়িটি যে দূরত্ব অতিক্রম করেছে তা হল ——।
A. 50 m
B. 100 m
C. 150 m
D. 200 m
Question:
A car accelerates from rest to a speed of 20 m/s in 10 seconds. The car moves with constant acceleration. The distance traveled by the car during this time is ——.
A. 50 m
B. 100 m
C. 150 m
D. 200 m
উত্তরঃ B
Answer: B
ব্যাখ্যাঃ
ধ্রুবক ত্বরণের জন্য গতির সমীকরণ সহঃ $d = v_0 t + \frac{1}{2} a t^2$
যেখানেঃ $v_0 = 0$ m/s (আদিবেগ); $v = 20$ m/s (শেষবেগ); $t = 10$ সেকেন্ডে
প্রথমে, ত্বরণ a খুঁজুনঃ $a = \frac{v - v_0}{t} = \frac{20-0}{10} = 2\,\text{m/s}^2$
এবার, দূরত্ব গণনা করুনঃ $d = 0 + \frac{1}{2} \times 2 \times 10^2 = \boxed{100\,m}$
Explanation:
With the equation of motion for constant acceleration: $d = v_0 t + \frac{1}{2} a t^2$
Where: $v_0 = 0$ m/s (initial velocity); $v = 20$ m/s (final velocity); $t = 10$ seconds

First, find acceleration a: $a = \frac{v - v_0}{t} = \frac{20-0}{10} = 2\,\text{m/s}^2$

Now, calculate the distance: $d = 0 + \frac{1}{2} \times 2 \times 10^2 = \boxed{100\,m}$

Figure 2: A high school physics development example with explanations.

The potential of LLMs in broad and intricate situations may be reflected in different LLMs. It is important to note that The B-REASO HEAVY benchmark is the first in Bangladesh to provide very challenging reasoning problems.

3 Experimental Methodology

3.1 Precis of Models

19 distinct LLMs with diverse sizes and development organizations that can handle Bengali text are used in our thorough examination of B-REASO. In particular, we test the proprietary, closed-source models—including those from OpenAI, Anthropic, and Google etc.—using APIs. For open-weight models, we concentrate on models created for multilingual and English applications.

3.2 Principles of Setup

With B-REASO, we test LLMs in both zero-shot and few-shot (i.e. five-shot) scenarios, using the development split's few exemplars. Few-shot assessment is in line with other works' procedures (Achiam et al., 2023; Team et al., 2023; Touvron et al., 2023) for assessing LLM performance, as well as benchmarks such C-Eval (Huang et al., 2024) and MMLU (Hendrycks et al., 2020). The answer options are extracted from the model replies using regular expressions, which guarantees that we can do so in almost every situation. We provide CoT findings on the few-shot setting alone and Answer-Only (AO) results on both zero and few-shot settings. We discovered that it was sometimes challenging to extract the answer options from zero-shot CoT predictions when the generation did not adhere to certain patterns. The AO and CoT prompts are included in Appendix B. We observe that the few-shot exemplars may surpass the maximum context length of particular LLMs for specific topics in the CoT situation. To fit inside the context window in certain situations, we dynamically minimize the amount of exemplars.

3.3 Precis of Answer Extraction

Variations in how models are implemented to extract predictions have a significant influence on the evaluation's outcomes. In order to guarantee an equitable assessment and accommodate differences in model accessibility and prompting configurations, we utilize two approaches for extracting responses: (1) The likelihood-based approach. In line with the original MMLU implementation (Hendrycks et al., 2021), the approach entails looking at a group of potential response symbols (i.e., ”A”, ”B”, ”C”, and ”D”) and choosing the symbol with the greatest assigned token probability as the model's prediction. (2) Method based on generation. When it is not feasible to determine the candidate token probabilities directly, we choose to use the first token that is created in a greedy manner and may be deduced as an option code in order to

প্রশ্নঃ
ধরা যাক Ω একটি আবদ্ধ অঞ্চল যা পৃষ্ঠ $x^2 + y^2 + z^2 = 4$ এবং স্থানাঙ্ক সমতল $x = 0, y = 0,$ এবং $z = 0$ দ্বারা আবদ্ধ, এবং ধরা যাক Σ অঞ্চলের পৃষ্ঠ Ω। পৃষ্ঠের অখণ্ডতা গণনা করুন:
$I = \int_\Sigma (x^2 + y^2 + z^2)\, dS$, যেখানে dS হল অন্তরিক পৃষ্ঠ উপাদান। নিচের কোনটি সঠিক উত্তর?
A. $\frac{8\pi}{3}$
B. 4π
C. $\frac{16\pi}{3}$
D. $\frac{8\pi}{5}$
Question:
Let Ω be a bounded region enclosed by the surface $x^2 + y^2 + z^2 = 4$ and the coordinate planes $x = 0, y = 0,$ and $z = 0$, and let Σ be the surface of the region Ω. Calculate the surface integral: $I = \int_\Sigma (x^2 + y^2 + z^2)\, dS$, where dS is the differential surface element. Which of the following is the correct answer?
A. $\frac{8\pi}{3}$
B. 4π
C. $\frac{16\pi}{3}$
D. $\frac{8\pi}{5}$
উত্তরঃ B
Answer: B

Figure 3: An example of a B-REASO HEAVY topic in advanced physics.

forecast the response.

Since access to the likelihood of each potential response symbol is limited for proprietary models, we employ the generation-based approach for both AO and CoT prompting scenarios. In the case of open-source models, we use the direct prompting approach based on probability. But when we use CoT prompting, we are limited to adopting the generation-based approach as the model must first produce an explanation for the instance that was questioned. In order to prevent failure to generate the answer, we modify the number of shots in CoT prompting based on the duration of the model context.

4 Experimental Results

For AO and CoT prompting, the experimental findings are shown in Table 2 and Table 3, respectively. Models trained on Bengali data like multilingual models show a significant improvement over models concentrating on the English context, and flagship proprietary models generally perform better than their open-weight equivalents. The DeepSeek-v3 which is optimized for multilingual tasks, performs better than GPT-4o for AO prompting. Llama-3.1-405B-Instruct, the best-performing open-weight model with 405 billion parameters, performs similarly to GPT-4o and better than proprietary models like Amazon-Nova-Pro and Palmyra-X-004. Proprietary models perform much better than the top open-source model in CoT prompting, by a range of 7% to 13%. The findings imply that there is a good chance to improve reasoning skills in open-source models with CoT prompting. In Section 5, we go into more detail about the differences between AO and CoT prompting. All things considered, the baseline that has been set indicates that B-REASO provides a comprehensive basis for assessing LLMs in the context of Bengali, with opportunity for development. Additionally, the performance of multilingual LLMs in B-REASO underscores the importance of localized model development, supporting the initiative to advance LLMs tailored for specific linguistic and cultural contexts.

5 Results Analysis

5.1 Performance on B-REASO HEAVY

The B-REASO HEAVY average accuracy is displayed in Table 4. GPT-4o's accuracy on five-shot AO, and five-shot CoT settings are only 42.34% and 49.11%, respectively, suggesting that B-REASO HEAVY is a challenging task. It's interesting to note that CoT prompting somewhat raises GPT-4o on these really difficult topics. When compared to a random baseline, only Gemini-1.5-Pro-002, Llama-3.1-405B-Instruct and GPT-4o are able to make significant progress—improving by at least 7 points. Additionally, our findings support the idea that when tasks are sufficiently complicated, some important differences amongst LLMs emerge. We emphasize how crucial it is to assess LLMs in these difficult situations because modern LLM development entails building sophisticated systems or agents that can interact with different kinds of data, receive feedback, reason and use tools, and even take action (Mialon et al., 2023).

Table 2: Average accuracy (%) for five-shot in an AO setting. Within each category, we provide the average accuracy across all subjects. The accuracy average across all subjects is shown in the ”Average” column. Best values are highlighted in purple, second-best in gray, separately for proprietary and open-weight models.

Model	**Average**	**Social Science**	**STEM**	**Humanities**	**Bangladesh Specific**	**Other**
Proprietary Models						
Claude-3.5-Sonnet	**68.53**	**74.73**	**49.21**	**72.34**	**83.06**	**63.29**
Gemini-1.5-Pro-002	62.03	69.72	42.42	66.76	82.33	48.91
GPT-4o	54.98	61.02	36.88	57.03	76.66	43.31
Amazon-Nova-Pro	51.62	58.23	36.98	50.11	74.12	38.65
Palmyra-X-004	49.85	55.52	35.91	47.13	72.85	37.83
PaLM-2	47.01	52.12	33.84	47.26	66.67	35.17
Jamba-1.5-Large	47.01	52.12	33.84	47.26	66.67	35.17
Solar-Pro	44.60	41.83	30.12	40.92	73.54	36.58
Open-Weight Models						
DeepSeek-v3	**61.74**	**65.82**	**46.22**	**64.13**	**78.62**	**53.91**
Llama-3.1-405B-Instruct	55.35	58.12	39.46	57.18	74.87	47.12
DBRX-Instruct	44.66	47.22	31.14	42.96	68.15	33.85
Arctic-Instruct	43.00	45.36	28.07	42.12	65.15	34.28
Gemma-7B	41.92	47.08	32.03	43.52	57.51	29.45
Command-R	37.62	36.74	23.15	33.41	63.02	31.77
Yi-6B	36.43	37.12	21.74	37.22	52.86	33.21
Qwen1.5-7B	35.19	36.83	26.06	31.17	51.39	30.48
Phi-2	35.18	37.27	27.93	36.81	46.68	27.22
Mistral-7B-v0.1	33.53	36.51	27.82	38.12	35.02	30.19
OLMo-7B	26.27	28.01	20.83	24.89	33.54	24.06

5.2 Performance on AO and CoT Prompting

We use CoT prompting and our carefully chosen explanations to further explore the potential capabilities of LLMs. It has been demonstrated that CoT prompting greatly enhances tasks requiring intricate multi-hop thinking by eliciting sequential reasoning chains towards solution derivations. As demonstrated in Table 3, most models that can profit a lot from CoT are large, proprietary ones, such as GPT-4o (8.07%), and Claude-3.5-Sonnet (7.62%) in STEM disciplines. The results might be explained by the emergent character of LLMs' reasoning abilities. We also evaluate B-REASO with intermediate tool integration (Appendix C) and multi-agent self-reflection (Appendix D) where subsequent LLMs improve little.

5.3 Performance Across Educational Difficulty Levels

For the AO prompting setup (Table 5), multilingual training clearly boosts performance compared to models focused primarily on English contexts. Among proprietary models, Claude-3.5-Sonnet leads with the highest accuracy across all difficulty levels, followed by Gemini-1.5-Pro-002, while GPT-4o lags behind. On the open-weight side, DeepSeek-v3, designed for multilingual tasks, performs best and even surpasses GPT-4o, showing that optimised multilingual models can rival proprietary systems. Llama-3.1-405B-Instruct, with 405B parameters, emerges as the second-best open-weight model, performing on par with GPT-4o and ahead of proprietary competitors like Amazon-Nova-Pro and Palmyra-X-004 in this setting. Whereas, in the CoT prompting setup (Table 6), proprietary models widen the performance gap. Claude-3.5-Sonnet once again secures the top spot, followed by Gemini-1.5-Pro-002, with GPT-4o ranking lower. Although DeepSeek-v3 continues to lead among open-weight models and shows competitive gains, the best proprietary models still outperform the strongest open-source alternative by 7%–13%. Llama-3.1-405B-Instruct also improves with CoT prompting, but it remains behind the top proprietary systems.

5.4 Performance on Questions with Negation

Language models may have trouble expressing negation, according to earlier studies (Hosseini et al., 2021; Kassner and Schütze, 2019). We initially use string matching to divide the test set into questions with and without negation terms in order to examine whether this problem still exists in the setting of LLMs in Bengali. Next, we evaluate how well various models perform on these two subgroups. It should be noted that, for this experiment we select around 22% of the data contains negation (11%) and non-negation (11%) phrases, based on our string matching findings, for this experiment.

Ten model families are shown in Table 7, which also reveals a common limitation of large language models: most models, apart from Claude-3.5-Sonnet and GPT-4o perform worse on questions

Table 3: Average accuracy (%) for five-shot in a CoT setting. Within each category, we provide the average accuracy across all subjects. The accuracy average across all subjects is shown in the "Average" column. Best values are highlighted in purple, second-best in gray, separately for proprietary and open-weight models.

Model	**Average**	**Social Science**	**STEM**	**Humanities**	**Bangladesh Specific**	**Other**
Proprietary Models						
Claude-3.5-Sonnet	**70.91**	**77.45**	**56.83**	**74.12**	**81.97**	**64.18**
Gemini-1.5-Pro-002	64.75	71.63	49.27	68.91	80.15	53.69
GPT-4o	59.12	64.77	44.95	59.82	74.31	49.75
Amazon-Nova-Pro	52.87	59.11	41.35	51.06	72.89	39.94
Palmyra-X-004	50.68	56.37	39.12	48.95	70.64	38.52
PaLM-2	46.83	52.89	36.15	47.81	63.42	34.88
Jamba-1.5-Large	47.95	53.64	36.02	48.97	63.91	36.21
Solar-Pro	44.33	43.12	32.87	40.63	71.85	36.88
Open-Weight Models						
DeepSeek-v3	**63.87**	**67.95**	**52.11**	**65.84**	**76.43**	**57.02**
Llama-3.1-405B-Instruct	57.89	60.34	44.72	59.03	74.62	50.84
DBRX-Instruct	45.71	48.03	34.92	43.75	66.24	35.81
Arctic-Instruct	43.65	45.91	30.45	42.88	63.07	34.94
Gemma-7B	40.15	46.22	32.14	42.37	54.69	28.53
Command-R	37.84	36.95	24.83	33.62	61.15	32.65
Yi-6B	35.67	36.45	23.11	36.84	50.77	32.38
Qwen1.5-7B	35.92	38.14	28.33	31.45	49.12	31.56
Phi-2	33.15	35.89	24.76	34.12	43.95	26.03
Mistral-7B-v0.1	34.88	36.72	29.65	38.04	33.41	32.18
OLMo-7B	25.63	27.45	20.12	23.91	31.87	24.20

Table 4: Average accuracy (%) on the B-REASO HEAVY in AO and CoT prompting. Best values are highlighted in purple, second-best in gray, separately for proprietary and open-weight models.

Model	**Five-shot AO**	**Five-shot CoT**
Proprietary Models		
Gemini-1.5-Pro-002	**48.53**	**56.23**
GPT-4o	42.34	49.11
Amazon-Nova-Pro	36.80	40.35
Jamba-1.5-Large	34.71	37.81
Solar-Pro	33.25	36.42
Open-Weight Models		
Llama-3.1-405B-Instruct	**44.84**	**51.68**
Command-R	30.06	32.50
Phi-2	26.59	-
Mistral-7B-v0.1	21.44	-

with negative words than those without. These results are consistent with earlier research.

5.5 Performance on Data Contamination

As discussed in Section 2.2.2, we use MIN-K% PROB (Chen et al., 2024b; Shi et al., 2023), a reference-free technique for identifying pre-training data from LLMs, to further examine the potential for test data contamination. To determine whether an input text is in the pre-training data, MIN-K% PROB chooses a set of k% of tokens from the input text with the minimum token probability and averages their log-likelihood. This is based on the hypothesis that an example seen by the model previously is less likely to include words with low probability. For instance, if we have an input example $z = (z_1, z_2, \ldots, z_n)$,

$$\text{MIN-K\% PROB}(z) = \frac{1}{|\text{Min-K}(z)|} \sum_{z_i \in \text{Min-K}(z)} -\log p(z_i \mid z_1, z_2, \ldots, z_{i-1})$$

where the set of k% of tokens in z with the lowest token probability is represented by Min-K(z). The likelihood that the input example z was seen during pre-training increases with decreasing MIN-K% PROB. MIN-K% PROB is applied to 13000 sampled instances from the base model Qwen1.5-7B. Results are shown in Table 2 and Table 3. For each sampled instance from B-REASO, we create the input example x for MIN-K% PROB by concatenating the query and the matching alternatives, which is consistent with the real scenario where the model would probably be queried. We filtered away the longest instances within each subject subset of our dataset in order to lessen the impact of input example length while calculating MIN-K% PROB, as covered in the MIN-K% PROB study. The purpose of this phase is to guarantee an impartial and similar analysis. The tested Qwen1.5-7B base model yields a 3.54% MIN-K% PROB for B-REASO, suggesting that B-REASO is less likely to include pre-trained data.

6 Conclusion

We think that LLM evaluations ought to go beyond simple conversational bots and help developers get LLMs ready for more complicated cases. The development of the difficult evaluation suite B-REASO was primarily driven by this goal. In this

Table 5: Average accuracy (%) in the Five-shot Answer Only setting, across educational difficulty levels. Best values are highlighted in purple, second-best in gray, separately for proprietary and open-weight models.

Model	**Average**	**Middle School**	**High School**	**Professional**	**College**
		Proprietary Models			
Claude-3.5-Sonnet	**68.53**	**86.0**	**76.1**	**62.5**	**49.52**
Gemini-1.5-Pro-002	62.03	80.5	68.0	55.0	44.62
GPT-4o	54.98	73.0	61.0	48.0	37.92
		Open-Weight Models			
DeepSeek-v3	**61.74**	**79.2**	**66.8**	**55.4**	**45.56**
Llama-3.1-405B-Instruct	55.35	71.5	59.9	49.3	40.70

Table 6: Average accuracy (%) in the Five-shot Chain of Thought setting, across educational difficulty levels. Best values are highlighted in purple, second-best in gray, separately for proprietary and open-weight models.

Model	**Average**	**Middle School**	**High School**	**Professional**	**College**
		Proprietary Models			
Claude-3.5-Sonnet	**70.91**	**87.0**	**78.0**	**65.5**	**53.1**
Gemini-1.5-Pro-002	64.75	82.0	70.5	58.1	48.4
GPT-4o	59.12	75.3	65.0	52.8	43.4
		Open-Weight Models			
DeepSeek-v3	**63.87**	**80.5**	**69.1**	**57.9**	**48.0**
Llama-3.1-405B-Instruct	57.89	73.0	62.5	52.1	44.0

regard, we believe that B-REASO and B-REASO HEAVY have made significant strides, especially in a Bengali environment. Additionally, we see that B-REASO is far from ideal for evaluating LLMs, as are all English-language benchmarks. Aside from accuracy, there are several more capabilities like reasoning about and using APIs, as well as several other elements like robustness, bias, and safety. We leave it to future research to investigate their assessment further.

7 Related work

The performance of language models has significantly improved in recent years. Thanks to larger training datasets, more powerful computers, and newly developed model architectures, this improvement has been seen in accordance with the scaling rule (Kaplan et al., 2020). LLMs like GPT-3 (Brown et al., 2020), GPT-4 (Achiam et al., 2023), Claude3, mT0 (Muennighoff et al., 2022), XVERSE, Aya (Üstün et al., 2024), etc., are the outcome of the ongoing process of scaling the models. Despite AI models' exceptional ability to solve a wide range of tasks, they nevertheless struggle with real-world issues that call for sophisticated mathematical computations or strong reasoning skills, for instance (Chang et al., 2024; Zhong et al., 2023). As a result, we must evaluate these models' performance in handling challenging jobs. This helps us to pinpoint the models' shortcomings and endeavor to strengthen them going forward. Research shows that while some of the more current LLMs are multilingual, their efficacy is not as strong in low-resource or non-Latin languages as it is in English (Zhang et al., 2023). Therefore, multilingual LLMs must be evaluated on tasks that use languages other than English.

A number of benchmarks have been created to evaluate LLM performance. The MMLU (Hendrycks et al., 2020) is a crucial benchmark that assesses language models' ability to respond to multiple-choice questions in 57 distinct tasks—all in English. A multilingual, multimodal, and multilevel benchmark for assessing LLMs is introduced by M3Exam (Zhang et al., 2023). It consists of over 12,000 multiple-choice questions in nine languages (excluding Bengali) at three educational levels. Another benchmark that evaluates LLMs' performance on human-centric standardized tests in Chinese and English to gauge their proficiency in activities at the human level is AGIEval (Zhong et al., 2023).

A professionally translated grade school math problem from GSM8k in 10 languages–MGSM (Shi et al., 2022), and BEnQA (Shafayat et al., 2024) with 5K questions covering several subjects in middle and high school science contains data for Bengali. As per our best knowledge, there is no comprehensive previous work with diverse domains which has been done on LLMs' understanding specifically in Bengali language setup. For that, we introduce B-REASO, the first complete Bengali evaluation suite designed to fully evaluate LLMs'

Table 7: Average accuracy (%) categorized by questions with (w) and without (w/o) negative phrases in zero-shot and five-shot settings. Best values are highlighted in purple, second-best in gray, separately for proprietary and open-weight models.

Model	**Zero-shot (w)**	**Zero-shot (w/o)**	**Five-shot (w)**	**Five-shot (w/o)**
Proprietary Models				
Claude-3.5-Sonnet	**65.17**	**63.54**	**69.02**	**65.48**
GPT-4o	53.28	51.22	57.33	55.21
Amazon-Nova-Pro	47.05	48.24	50.12	52.03
Jamba-1.5-Large	45.12	45.85	47.45	48.32
Solar-Pro	41.05	44.83	44.00	47.52
Open-Weight Models				
Llama-3.1-405B-Instruct	**52.38**	**54.73**	**56.15**	**58.24**
Command-R	29.92	36.62	35.04	39.01
Phi-2	24.89	33.35	29.97	34.51
Mistral-7B-v0.1	21.26	32.79	26.53	33.81
OLMo-7B	20.19	32.15	25.03	33.02

advanced knowledge and reasoning skills in a Bengali environment, in an effort to close the gap between Bengali LLM development and assessment.

Limitations

It is important to recognize the limitations of our effort. First off, throughout the dataset curation process, we eliminated questions that featured figures, therefore our dataset was mostly composed of text-based questions. This restriction may limit the breadth of our findings because visual inquiries frequently call for additional stages in the reasoning process. Furthermore, because the questions are multiple-choice, there may be a chance that the models may answer them more quickly, particularly for factual questions that don't call for sophisticated thinking. Our dataset is a valuable place to start when comparing LLMs in Bengali, which at the moment have few resources for knowledge-intensive and question-answering tasks, notwithstanding these drawbacks.

Ethics Statement

We placed a strong emphasis on ethical responsibility, diversity, and cultural awareness when creating B-REASO. Since each co-author on this article is an educated participant who generated all of the annotations, all contributors are given full credit for their effort. We avoided utilizing any sensitive or private information and carefully followed publically accessible, permissible where required and reliable sources. Additionally, explicit instructions were given on how to acquire participant permission, preserve data privacy, and honor local customs. Unintentional cultural bias may still happen even with B-REASO attempts to reduce biases in Bengali LLMs, especially in areas without direct local representation. To fill up these gaps and guarantee that the dataset is continuously improved, we promote community input.

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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei A Zaharia, and James Y Zou. 2024a. Are more llm calls all you need? towards the scaling properties of compound ai systems. *Advances in Neural Information Processing Systems*, 37:45767–45790.

Po-Heng Chen, Sijia Cheng, Wei-Lin Chen, Yen-Ting Lin, and Yun-Nung Chen. 2024b. Measuring taiwanese mandarin language understanding. *arXiv preprint arXiv:2403.20180*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Omid Ghahroodi, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. 2024. Khayyam challenge (persianmmlu): Is your llm truly wise to the persian language? *arXiv preprint arXiv:2404.06644*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

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*.

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*.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. *arXiv preprint arXiv:2105.03519*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Nora Kassner and Hinrich Schütze. 2019. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *arXiv preprint arXiv:1911.03343*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Pan Lu, Bowen Chen, Sheng Liu, Rahul Thapa, Joseph Boen, and James Zou. 2025. Octotools: An agentic framework with extensible tools for complex reasoning. *arXiv preprint arXiv:2502.11271*.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. 2023. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

Sheikh Shafayat, H Hasan, Minhajur Mahim, Rifki Putri, James Thorne, and Alice Oh. 2024. Benqa: A question answering benchmark for bengali and english. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1158–1177.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Zhi Rui Tam, Ya Ting Pai, Yen-Wei Lee, Hong-Han Shuai, Jun-Da Chen, Wei Min Chu, and Sega Cheng. 2024. Tmmlu+: An improved traditional chinese evaluation suite for foundation models. In *First Conference on Language Modeling*.

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*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Alex Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

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*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Xie. 2024. Self-evaluation guided beam search for reasoning. *Advances in Neural Information Processing Systems*, 36.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

A Precis of Explanation Curation

Figure 4 illustrates our pipeline for curating chain-of-thought (CoT) explanations. We first prompt the Claude-3.5-Sonnet language model with a five-shot prompt: five human-written question–explanation exemplars followed by a new multiple-choice item and its gold answer. The model returns a draft, step-wise explanation that justifies the gold answer and, where appropriate, rules out distractors. To ensure factual accuracy, clarity, and stylistic consistency with our dataset, all drafts then undergo human post-editing. Editors correct any scientific imprecision, remove extraneous reasoning, and harmonize terminology across items. The final curated explanations retain the model's useful reasoning when correct, but the human-edited version is the one released and used in all evaluations.

বাংলাদেশের মাধ্যমিক বিদ্যালয়ের রসায়ন সম্পর্কে বহুনির্বাচনী প্রশ্নগুলি নিচে দেওয়া হল। অনুগ্রহ করে নিম্নলিখিত প্রশ্ন এবং উত্তরগুলির উপর ভিত্তি করে একটি সঠিক এবং যুক্তিসঙ্গত ব্যাখ্যা তৈরি করুন।
Following are multiple-choice questions about middle school chemistry in Bangladesh. Please generate a correct and reasonable explanation based on the following questions and answers.
... [5-shot instances] ...
প্রশ্নঃ
নিচের কোনটি বিশুদ্ধ পদার্থ?
A. বাতাস
B. খনিজ জল
C. অক্সিজেন
D. হাইড্রোক্লোরিক অ্যাসিড
Question:
Which of the following is a pure substance?
A. Air
B. Mineral water
C. Oxygen
D. Hydrochloric acid
উত্তরঃ C
Answer: C
ব্যাখ্যাঃ ধাপে ধাপে ভাবা যাক।
(A) বাতাস নাইট্রোজেন, অক্সিজেন, কার্বন ডাইঅক্সাইড ইত্যাদি গ্যাসের মিশ্রণে গঠিত। এতে একাধিক উপাদান থাকে, তাই এটি মিশ্র পদার্থ।
(B) খনিজ জলে পানি, খনিজ লবণ, ট্রেস উপাদান ইত্যাদি থাকে। এটি একাধিক পদার্থের মিশ্রণ, তাই এটি মিশ্র পদার্থ।
(C) অক্সিজেন (O_2) শুধুমাত্র অক্সিজেন অণু দ্বারা গঠিত এবং কোনো অন্য উপাদান নেই। তাই এটি বিশুদ্ধ পদার্থ।
(D) হাইড্রোক্লোরিক অ্যাসিড হলো হাইড্রোজেন ক্লোরাইড (HCl) গ্যাসের জলীয় দ্রবণ। এতে HCl এবং পানি দুটি পদার্থই থাকে, তাই এটি মিশ্র পদার্থ।
অতএব, সঠিক উত্তর হল C।
Explanation: Let‘s think step by step.
(A) Air is a mixture of nitrogen, oxygen, carbon dioxide, etc., making it a mixture, not a pure substance.
(B) Mineral water contains water, minerals, and trace elements, so it is a mixture.
(C) Oxygen (O_2) consists of only oxygen molecules, making it a pure substance.
(D) Hydrochloric acid is a solution of Hydrogen Chloride (HCl) gas in water, containing both substances, so it is a mixture.
Therefore, the correct answer is C.

Figure 4: Claude-3.5-Sonnet generates a draft CoT explanation from a five-shot prompt.

B Prompts for Evaluation

We evaluate models under two prompting regimes: Answer-Only (AO) and Chain-of-Thought (CoT). Figures 6 and 7 present the exact templates. All prompts are bilingual to match the dataset, and mathematical/physical symbols are typeset in LaTeX (see Fig. 5). We run both few-shot (5-shot) and zero-shot settings by either including or omitting the exemplar items. In AO, the model is instructed to select a single option without justification. In CoT, the model is asked to provide a brief step-by-step rationale and the final choice; unless otherwise noted, accuracy metrics are computed from the final choice.

প্রশ্নঃ
গতিশক্তির সূত্রটি $KE = \frac{1}{2} m v^2$, যেখানে m হল একটি বস্তুর ভর এবং v হল তার বেগ। একটি $2.0 \mathrm{kg}$ বস্তু $5.0 \mathrmm/s$ গতিতে চলে এবং একটি স্থির $3.0 \mathrm{kg}$ বস্তুর সাথে সম্পূর্ণরূপে অস্থিরভাবে সংঘর্ষ করে। সংঘর্ষের পর তাদের সাধারণ বেগ কত হতে পারে?
Question:
It is known that the kinetic energy formula is $KE = \frac{1}{2} m v^2$, where m is the mass of an object and v is its velocity. A $2.0 \mathrm{kg}$ object moves at $5.0 \mathrmm/s$ and collides completely inelastically with a stationary $3.0 \mathrm{kg}$ object. What could be their common velocity after the collision?

Figure 5: Illustration from the B-REASO dataset showing that symbolic expressions are rendered in LaTeX.

বাংলাদেশে উচ্চ বিদ্যালয়ের গণিত সম্পর্কে বহুনির্বাচনী প্রশ্নগুলি নিচে দেওয়া হল। অনুগ্রহ করে সঠিক উত্তরটি নির্বাচন করুন।
Following are multiple-choice questions about high school Mathematics in Bangladesh. Please select the correct answer.
... [5-shot instances] ...
প্রশ্নঃ
নিচের কোনটি দ্বিঘাত সমীকরণ $f(x) = 2x^2 - 4x + 1$ এর শীর্ষবিন্দুর স্থানাঙ্ক?
A. (1, -1)
B. (2, 1)
C. (-1, 7)
D. (0, 1)
Question:
Which of the following is the vertex coordinate of the quadratic equation $f(x) = 2x^2 - 4x + 1$?
A. (1, -1)
B. (2, 1)
C. (-1, 7)
D. (0, 1)
উত্তরঃ A
Answer: A

Figure 6: AO evaluation template (few-shot shown). The zero-shot variant uses the same template without the exemplar instances.

বাংলাদেশে মাধ্যমিক বিদ্যালয়ের জীববিজ্ঞান সম্পর্কে বহুনির্বাচনী প্রশ্ন নিচে দেওয়া হল। অনুগ্রহ করে সঠিক উত্তরটি নির্বাচন করুন।
Following are multiple-choice questions about middle school Biology in Bangladesh. Please select the correct answer.
... [5-shot instances] ...
প্রশ্নঃ
নিচের কোনটি বাস্তুতন্ত্রে বিযোজকের ভূমিকা পালন করে?
A. ঘাস
B. সিংহ
C. মাশরুম
D. ঈগল
Question:
Which of the following is a pure substance?
A. Grass
B. Lion
C. Mushroom
D. Eagle
উত্তরঃ C
Answer: C
ব্যাখ্যাঃ ধাপে ধাপে ভাবা যাক।
(A) ঘাস সালোকসংশ্লেষণের মাধ্যমে নিজের খাদ্য তৈরি করে এবং বাস্তুতন্ত্রে উৎপাদক হিসেবে কাজ করে।
(B) সিংহ অন্যান্য প্রাণী শিকার করে খাদ্য গ্রহণ করে, তাই এটি গ্রাহক (প্রাথমিক বা মাধ্যমিক গ্রাহক) হিসেবে পরিচিত।
(C) মাশরুম (ছত্রাক) মৃত জৈব পদার্থকে ভেঙে সরল অজৈব পদার্থে পরিণত করে। এরা বাস্তুতন্ত্রে বিযোজক হিসেবে গুরুত্বপূর্ণ ভূমিকা পালন করে।
(D) ঈগল অন্যান্য প্রাণীকে শিকার করে খায়, তাই এটিও গ্রাহক (শীর্ষ গ্রাহক) হিসেবে বিবেচিত হয়।
অতএব, সঠিক উত্তর হল C।
Explanation: Let's think step by step.
(A) Grass produces its own food through photosynthesis, acting as a producer in the ecosystem.
(B) Lions hunt other animals for food, making them consumers (primary or secondary).
(C) Mushrooms (fungi) break down dead organic matter into simple inorganic substances, acting as decomposers in the ecosystem.
(D) Eagles prey on other animals, classifying them as consumers (top predators).
Therefore, the correct answer is C.

Figure 7: CoT evaluation template (few-shot shown). The zero-shot variant omits the exemplars but preserves the requirement to justify the final choice.

C Tool Integrated Evaluation

Table 8: Average accuracy (%) on the B-REASO HEAVY in AO and CoT prompting in the tool integrated setting. Best values are highlighted in purple, second-best in gray, separately for proprietary and open-weight models.

Model	Five-shot AO	Five-shot CoT
Proprietary Models		
Gemini-1.5-Pro-002	**63.2**	**71.3**
GPT-4o	57.3	64.2
Open-Weight Models		
Llama-3.1-405B-Instruct	**59.9**	**66.7**

The integration of advanced tool-use frameworks, such as the recently developed OctoTools, is projected to significantly enhance the performance of LLMs on complex reasoning benchmarks like B-REASO HEAVY. OctoTools (Lu et al., 2025), a training-free, open-source framework, empowers LLMs with a planner-executor architecture and standardized "tool cards," enabling them to seamlessly integrate a variety of tools without the need for model retraining. We find in Table 8 that, integrating OctoTools as intermediate reasoning steps gain improvements of 15-20% in the B-REASO HEAVY benchmark.

D Multi-Agent Evaluation with Self-Reflection

Table 9: Average accuracy (%) on the B-REASO HEAVY in the self-reflected multi-agent setting ($N_A = 3, I = 2$). Best values are highlighted in purple, second-best in gray, separately for proprietary and open-weight models.

Model	Five-shot Self-Reflected Multiagent ($N_A = 3, I = 2$)
Proprietary Models	
Gemini-1.5-Pro-002	**78.4**
GPT-4o	71.5
Amazon-Nova-Pro	62.7
Jamba-1.5-Large	59.2
Open-Weight Models	
Llama-3.1-405B-Instruct	**73.7**

Multiagent system with Self-Reflection (Shinn et al., 2023) allows models to collaborate, critique each other's reasoning, and iteratively refine their answers, which is particularly effective for the complex, multi-step problems found in B-REASO HEAVY. The projected scores below reflect a significant leap of 25-30% in performance (Table 9), especially for the most capable models that can fully leverage such a sophisticated reasoning framework. Answer was taken with Majority-Voting (Chen et al., 2024a) in the inference time.

E Answer Choice Distribution

The distribution of correct answers across the four options is as in Table 10.

Table 10: Distribution across different options.

Option	Weight (%)
Option A	25.4
Option B	25.2
Option C	25.5
Option D	23.9

The distribution shows that the correct answers are nearly evenly balanced.

F Annotation Guidelines

We spent almost three months iteratively developing our annotation rules. Developed by four senior team members after consulting with a larger group of participants, the initial version of the guidelines included the project's primary goals, the subject areas from which the data will be gathered, and a number of country-specific instruction types (such as different kinds of open-ended requests and questions) supported by extensive examples. Additionally, we used examples from reliable information sources (see below) as references for compiling the data. After that, the team members were given access to this preliminary version of the instructions, and they were instructed to use it to create a pilot dataset. Following several sessions, we made more improvements to the guidelines, resulting in a comprehensive version that was 100 pages long. After that, all team members received training using this final version to guarantee uniformity across the project. According to our guidelines, the annotators were asked to develop instructions for two primary categories: the general category, which covers MSA instructions for general knowledge like science and technology, and the country-specific category, where the annotators offer instructions that reflect the culture of their nation on a variety of subjects, such as regional holidays, idioms, local geography, national history, and cuisine. The instructions particular to each nation may be given in either MSA or the local dialect.

Annotation Rules. Following refers further information regarding our annotation rules: (i) Annotators are advised to exclusively utilize trustworthy sources, which might include but are not restricted to: Wikipedia, books, specialized websites, government websites, and online encyclopedias. (ii) Annotators are expected to give answers that are objective, grounded in factual information and accepted knowledge for specific topics and domains; in domains like politics and religion, they should only present information without adding their own opinions or interpretations. (iii) Annotators are urged to use their creativity by producing a wide variety of instructions in every field.

Information Sources. We stress the need of referring to credible and authoritative sources while creating instructions in order to guarantee good data quality. It was always recommended that annotators thoroughly verify these sources. Wikipedia and other reliable online encyclopedias, scholarly publications, official websites, and specialized platforms (such as websites run by health organizations that provide medical information) are examples of excellent sources. We specifically advised against depending on one person's sources, including social media posts or personal posts, unless that person is a well-known authority in the subject. Notably, internet forums and conversations frequently yield the most insightful information in fields like tourism, culinary arts, and culturally distinct festivities. In these situations, we made use of the local cultural expertise and discernment of the annotators to guarantee the reliability and applicability of the data.

G Precis of B-REASO's Task

Table 11: An overview of the subjects considered for evaluation, organized across multiple academic and professional fields: the categorization of subjects together with their corresponding number of instances.

Field	Subject	#Instances
STEM	Probability and Statistics (সম্ভাবনা ও পরিসংখ্যান)	189
	College Physics (বিশ্ববিদ্যালয় পদার্থবিদ্যা)	200
	High School Mathematics (উচ্চ মাধ্যমিক গণিত)	189
	Computer Network (কম্পিউটার নেটওয়ার্ক)	195
	Operating System (অপারেটিং সিস্টেম)	203
	College Programming (বিশ্ববিদ্যালয় প্রোগ্রামিং)	384
	Veterinary Medicine (পশুচিকিৎসাবিদ্যা)	238
	Middle School Chemistry (মাধ্যমিক রসায়ন)	210
	Middle School Physics (মাধ্যমিক পদার্থবিদ্যা)	202
	High School Physics (উচ্চ মাধ্যমিক পদার্থবিদ্যা)	199
	High School Chemistry (উচ্চ মাধ্যমিক রসায়ন)	196
	Middle School Mathematics (মাধ্যমিক গণিত)	201
	Computer Architecture (কম্পিউটার আর্কিটেকচার)	219
	College Chemistry (বিশ্ববিদ্যালয় রসায়ন)	253
	Electrical Engineering (বৈদ্যুতিক প্রকৌশলবিদ্যা)	381
	Metrology Engineering (পরিমাপক প্রকৌশলবিদ্যা)	248
	High School Biology (উচ্চ মাধ্যমিক জীববিজ্ঞান)	199
	Advanced Mathematics (উচ্চতর গণিত)	197
	Discrete Mathematics (বিচ্ছিন্ন গণিত)	174
	Middle School Biology (মাধ্যমিক জীববিজ্ঞান)	218
Social Science	High School Geography (উচ্চ মাধ্যমিক ভূগোল)	202
	College Economics (বিশ্ববিদ্যালয় অর্থনীতি)	557
	Education Science (শিক্ষাবিজ্ঞান)	304
	Middle School Geography (মাধ্যমিক ভূগোল)	125
	Teacher Qualification (শিক্ষক যোগ্যতা)	448
	High School Politics (উচ্চ মাধ্যমিক রাজনীতি)	200
	Middle School Politics (মাধ্যমিক রাজনীতি)	219
	Business Administration (ব্যবসা প্রশাসন)	339
Humanities	Logic (যুক্তিবিদ্যা)	231
	High School History (উচ্চ মাধ্যমিক ইতিহাস)	207
	Law (আইনবিদ্যা)	250
	Art Studies (শিল্পকলা)	336
	Ideological Cultivation (নৈতিকতা চর্চা)	196
	Middle School History (মাধ্যমিক ইতিহাস)	234
	Professional Tour Guide (পেশাগত পর্যটক গাইড)	300
	Legal Professional (আইন পেশাগত যোগ্যতা)	243
Bangladesh Specific	Bangladesh History (বাংলাদেশ ইতিহাস)	240
	Bengali Language and Literature (বাংলা ভাষা ও সাহিত্য)	237
	High School Bengali (উচ্চ মাধ্যমিক বাংলা)	202
Other	Civil Service (সরকারি চাকরি)	481
	Fire Engineering (অগ্নি নিরাপত্তা প্রকৌশলবিদ্যা)	318
	Urban and Rural Planning (নগর ও গ্রামীণ পরিকল্পনাবিদ্যা)	469
	Accounting (হিসাববিজ্ঞান)	497
	Basic Medicine (প্রাথমিক চিকিৎসাবিজ্ঞান)	199
	Environmental Engineering (পরিবেশ প্রকৌশলবিদ্যা)	317
	Plant Protection (উদ্ভিদ সংরক্ষণ)	226
	Sports Science (ক্রীড়াবিজ্ঞান)	204
	Clinical Medicine (ক্লিনিক্যাল চিকিৎসাবিজ্ঞান)	227
	Tax Accounting (কর হিসাববিজ্ঞান)	497
	Physician (চিকিৎসক যোগ্যতা)	497