BEnQA: A Question Answering Benchmark for Bengali and English

Sheikh Shafayat, H M Quamran Hasan,
Minhajur Rahman Chowdhury Mahim, Rifki Afina Putri, James Thorne, Alice Oh
KAIST, Republic of Korea
{sheikh.shafayat, quamranhasan, minhaj, rifkiaputri, thorne}@kaist.ac.kr,
alice.oh@kaist.edu

Abstract

In this study, we introduce BEnQA[1], a dataset comprising parallel Bengali and English exam questions for middle and high school levels in Bangladesh. Our dataset consists of approximately 5K questions covering several subjects in science with different types of questions, including factual, application, and reasoning-based questions. We benchmark several Large Language Models (LLMs) with our parallel dataset and observe a notable performance disparity between the models in Bengali and English. We also investigate some prompting methods, and find that Chain-of-Thought prompting is beneficial mostly on reasoning questions, but not so much on factual ones. We also find that appending English translation helps to answer questions in Bengali. Our findings point to promising future research directions for improving the performance of LLMs in Bengali and more generally in low-resource languages.

1 Introduction

Large language models (LLMs) like GPT-4 have shown impressive performance in many complex natural language processing tasks including reasoning and question answering, all of which have been subject to intense research in recent years (Bang et al., 2023; Liu et al., 2023). However, most of this research has focused on English and other high-resource languages, often neglecting medium- to low-resource languages. To make matters more difficult, most of the benchmarks designed for measuring progress in LLM research are only available in English and a handful of other languages, making it very hard to compare different LLM skills in low-resource languages.

This issue is particularly critical in light of the growing use of LLMs like ChatGPT in the context of education (Khan Academy, 2023). The absence of LLMs that perform equally well in non-English languages would lead to further divide in access to education and technology, highlighting a critical need for more inclusive language model development.

Addressing this gap, our work specifically targets Bengali, a language spoken by 272 million speakers worldwide (Zeidan, 2023), yet considered a low-resource language. We make the following contributions:

- We introduce BEnQA, a science question-answering dataset in English and Bengali, consisting of 5,161 questions taken from the Bangladeshi national curriculum for grade school exams. This dataset covers a wide range of subjects and question types, including those requiring multi-step reasoning.
- Our dataset is parallel in English and Bengali, allowing us to benchmark existing LLMs and measure the performance gap between them. Additionally, the parallel nature of our dataset ensures a fairer comparison between the two languages.
- Through our evaluation of various LLMs, we indeed observe a significant performance discrepancy between questions in English and Bengali. Additionally, we investigate the impact of different prompting techniques on LLMs' performance in Bengali, such as Chain-of-Thought (CoT) prompting and appending English translation to the prompt.

2 Related Work

Multilingual Reasoning Benchmark

There have been many English benchmarks to evaluate the reasoning capabilities of LLMs, such as COPA (Roemmele et al., 2011), HellaSwag

[1] We release the dataset at: https://github.com/sheikhshafayat/BEnQA

Findings of the Association for Computational Linguistics: ACL 2024, pages 1158–1177
August 11-16, 2024 ©2024 Association for Computational Linguistics

(Zellers et al., 2019), CosmosQA (Huang et al., 2019), and CommonsenseQA (Talmor et al., 2019). For non-English languages, one of the widely used benchmarks is X-COPA (Ponti et al., 2020). A recent work (Doddapaneni et al., 2023) provides a human translation of this dataset in several Indic languages, including Bengali. Besides these, BIG-Bench Hard (Suzgun et al., 2022) is a dataset consisting of a collection of 23 challenging BIG-Bench tasks (Srivastava et al., 2022) that measure a model's reasoning ability across various tasks such as logical deduction, multi-step arithmetic, and more.

Recent works develop datasets using academic exam questions to benchmark the ability of current LLMs. The majority of academic exam question datasets are in English, with a few exceptions, such as IndoMMLU (Koto et al., 2023). Notable examples of English datasets include MMLU (Hendrycks et al., 2020), which covers multi-task problems from grade school to college level; MATH (Hendrycks et al., 2021), consisting of competitive math problems; GSM8k (Cobbe et al., 2021), which consists of grade school mathematics problems; and the ARC dataset (Clark et al., 2018) that proposes grade school level multiple-choice science questions. Recent GPT-4 technical report (OpenAI, 2023) also included benchmark results of many exams such as the GRE, bar exam, and AP exam, as did other works such as Kung et al. (2023) and Choi et al. (2023). For Bengali, this lack of resources is especially evident. One existing data for Bengali is MGSM (Shi et al., 2022), a professionally translated grade school math problem from GSM8k in 10 languages, including Bengali. Our work aims to address this gap by introducing BEnQA, covering questions from various exam subjects collected from the official Bangladeshi school exams. BEnQA stands out as the only academic exam benchmark dataset available in Bengali.

Multilingual Prompting

Chain-of-Thought (CoT) prompting (Wei et al., 2022, 2023) has ushered a new wave of work in eliciting reasoning behavior in large language models. Shi et al. (2022) demonstrated the effectiveness of CoT and reported an increase in performance in mathematical reasoning while doing step-by-step reasoning in non-English languages. The same work also reports that translating the questions using Google Translate and then doing step-by-step reasoning in English often works even better. Ahuja et al. (2023) also used a similar approach. More recently, Huang et al. (2023) introduced cross-lingual thought prompting that first translates the original query into English and then does CoT reasoning in English.

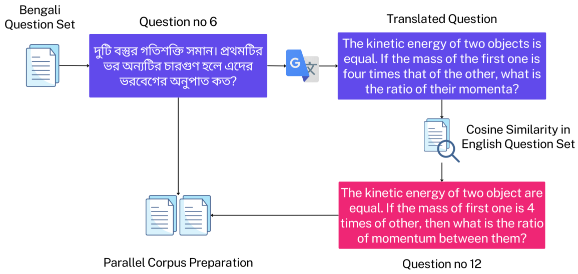

Bengali
Question Set
Question no 6
দুটি বস্তুর গতিশক্তি সমান। প্রথমটির ভর অন্যটির চারগুণ হলে এদের ভরবেগের অনুপাত কত?
Translated Question
The kinetic energy of two objects is equal. If the mass of the first one is four times that of the other, what is the ratio of their momenta?
Cosine Similarity in English Question Set
The kinetic energy of two object are equal. If the mass of first one is 4 times of other, then what is the ratio of momentum between them?
Parallel Corpus Preparation
Question no 12

Figure 1: Our pipeline for creating parallel corpus in BEnQA.

However, one obvious issue with doing CoT reasoning in English is that it renders the LLM output less useful for non-English users. Especially, when LLMs are used in educational contexts, CoT often provides helpful hints for learners (Han et al., 2023). In our work, we demonstrate that it is possible to do CoT reasoning in the target language and improve performance, as long as we have access to the English translation of the questions.

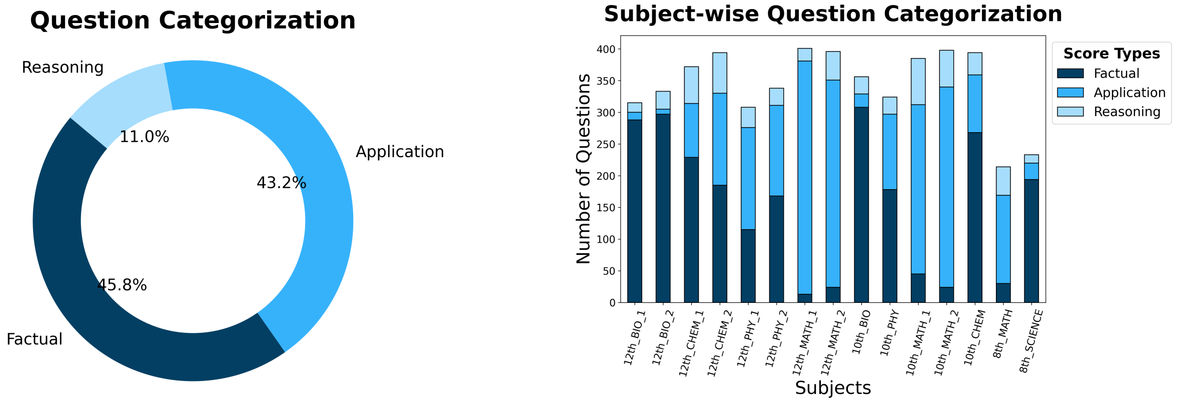
3 BEnQA Dataset

Our dataset, BEnQA, is a parallel corpus of English-Bengali science questions in a multiple-choice format from 8th, 10th, and 12th grade exams. These questions are sourced from nationwide board exams in Bangladesh and are officially available in both English and Bengali, giving us a high-quality parallel corpus.

3.1 Dataset Curation

Examination questions in Bangladesh are predominantly available in print rather than digital format. Therefore, we collect examination papers and use readily accessible solution books to create the ground truth for the questions. We then employ four typists, proficient in both Bengali and English, to digitize the questions and their corresponding answers. This process enables the conversion of questions and their answers into a digital format with minimal errors.

All questions are of the multiple-choice format, each presenting four options. The typists are instructed to format mathematical equations and chemical formulas using LaTeX. Questions involv-



(a) Statistics by categories.

(b) Statistics by grade and subject.

Figure 2: Statistics of our BenQA dataset.

ing figures are excluded. Additionally, we apply specific heuristics to filter out some questions that are not parallel, annotation errors, and poor-quality questions, including those with inconsistencies between the ground truths in English and Bengali, and non-parallel questions. The heuristics for filtering non-parallel questions are explained in detail in the next section.

3.2 Parallel Corpus Creation

Typically, questions in Bangladeshi national board exams are initially formulated in Bengali and then translated into English. This process allows us to create a parallel dataset. However, the question order in each version does not always align (i.e., the first question in the Bengali version can be the eleventh question in the English version). To address this, we develop a very simple yet effective algorithm for aligning the English and Bengali questions, illustrated in Figure 1.

First, we translate each Bengali question to English via Google Translate API. Within each subject, we then match the Google-translated version with the actual English translation of the question, using the cosine similarity of their embeddings. We use OpenAI `text-embedding-ada-002` embedding to achieve this. After the initial matching, two native Bengali speakers who are also proficient in English, manually verify each question to ensure that each Bengali question corresponds to its English counterpart. We also filter out a few questions for which the ground truth is different in English and Bengali. Manual inspection reveals that such cases typically result from annotator errors or issues with the questions themselves.

Since these translations are often carried out by school teachers with varying levels of English proficiency, some English questions may contain subtle grammatical errors or sound unnatural to native English speakers. We investigate the impact of these grammatical inconsistencies on model performance by creating a grammar-corrected version of the English questions using GPT-4 with human supervision in the loop. Our findings indicate that grammatical errors generally do not significantly affect model performance (see Appendix A).

3.3 Dataset Properties

In our proposed BEnQA corpus, there is a total of 5,161 questions available in both English and Bengali. This dataset comprises 55% (2,857) questions from the 12th grade, 36% (1,857) from the 10th grade, and 9% (447) from the 8th grade. The 12th-grade section encompasses various subjects like Mathematics, Physics, Chemistry, and Biology, further divided into part I and part II based on subtopics. For the 10th grade, it includes Mathematics I, Mathematics II, Physics, Chemistry, and Biology. In the case of the 8th grade, the dataset includes Mathematics and Science, with the latter being a comprehensive subject covering 8th-grade science. Figure 2b summarizes the subject and grade-wise question statistics.

3.4 Dataset Categorization

Our dataset encompasses various question types, each demanding specific skills, irrespective of the subject or grade. Based on the skills necessary for solving them, we categorize the questions into three distinct types:

- **Factual Knowledge:** These are questions that solely rely on knowledge of basic facts, events, concepts, dates, etc. Answering such questions does not involve any form of analysis or reasoning.

- **Procedural & Application:** This category comprises questions that require the ability to apply a procedure, utilize a familiar concept, or employ a formula for solving.
- **Reasoning:** Questions falling into this category demand multiple steps of analysis or reasoning to arrive at a solution.

We utilize GPT-4 for the purpose of categorizing the questions into these groups. We use the prompt given in Table 3 to categorize them. The categorization results are depicted in Figure 2a for question-type distribution and Figure 2b for subject-wise breakdown. The majority of Biology and Chemistry questions lean towards factual knowledge. Conversely, Mathematics questions are primarily centered around procedural & application skills, with some instances demanding reasoning. In the case of Physics, the number of factual and procedural questions is nearly equal.

4 Experiment Setup

To assess the performance of existing LLMs on our newly created dataset, we evaluate them using the BEnQA dataset across several open-source and proprietary models. Following the recommendation from prior work on ChatGPT (Lai et al., 2023), we keep the system prompt in English for both Bengali and English datasets. Most of the benchmark results presented in the main paper are conducted in a zero-shot setting to save tokenization costs in proprietary models (Petrov et al., 2023). Additionally, experiments in 3-shot and 5-shot settings were performed as supplementary analyses, details of which are provided in Appendix C.

4.1 Models

To evaluate the current LLMs capability in our dataset, we use the following models:

English

- **Proprietary LLMs**: GPT-4 (OpenAI, 2023)[2], GPT-3.5[3], Claude 2.1[4]
- **Open-source LLMs**: LLaMA-2 7B and 13B (Touvron et al., 2023)[5], Mistral 7B[6]

Bengali

- We utilize the proprietary models mentioned above for our Bengali experiments.

It is worth noting that most open-source models do not perform well on Bengali. For this reason, we conduct all open-source model experiments on only the English version of the dataset.

Most of the other experiments (i.e., ablations, exploration of prompting techniques, and others) in this paper are conducted using GPT-3.5 Turbo in order to avoid the high cost associated with more advanced proprietary models, particularly for Bengali (Ahia et al., 2023; Petrov et al., 2023). Opting for benchmarking on GPT-3.5 ensures a balanced trade-off between value and cost.

4.2 Prompts

In our experiments, we explored whether certain prompting methods could enhance the performance of LLMs on our dataset. We utilized the Chain-of-Thought (CoT) prompt, as recent work has noted its efficacy in improving reasoning tasks (Wei et al., 2023). Moreover, we also experimented without the CoT prompt for comparison. Detailed descriptions of the specific prompts used in our experiments are provided in Appendix F. Throughout all the Bengali experiments, the model was explicitly instructed to perform CoT reasoning in Bengali and avoid using English, aligning with our original goal of making the model's output useful to Bengali users, especially if LLMs are used in an educational context. Another potential approach to get CoT steps in Bengali is by utilizing a translation method, by first generating CoT steps in English then translating them to Bengali. However, we chose not to rely on this method to avoid "translationese" problem, where the translated sentences are awkward or grammatically incorrect, which might result in a failure to capture the true intent or meaning of the original sentence.

4.3 Evaluation Metric

The evaluation process involves a manual assessment of model outputs to determine if the final answer aligns with the ground truth. It is important to note that our evaluation does not scrutinize the

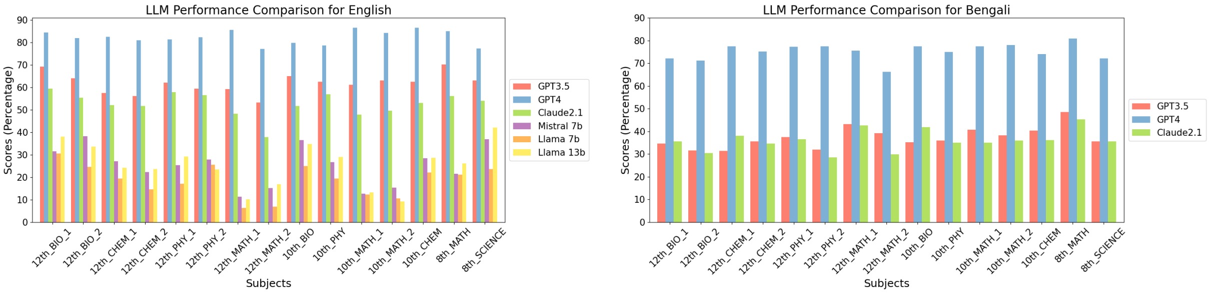
[2] We utilized `gpt-4-1106-preview` API endpoint (https://www.openai.com/research/gpt-4)

[3] We utilized `gpt-3.5-turbo-1106` API endpoint (https://platform.openai.com/docs/models/gpt-3-5-turbo)

[4] https://www.anthropic.com/index/claude-2

[5] https://llama.meta.com/llama2/

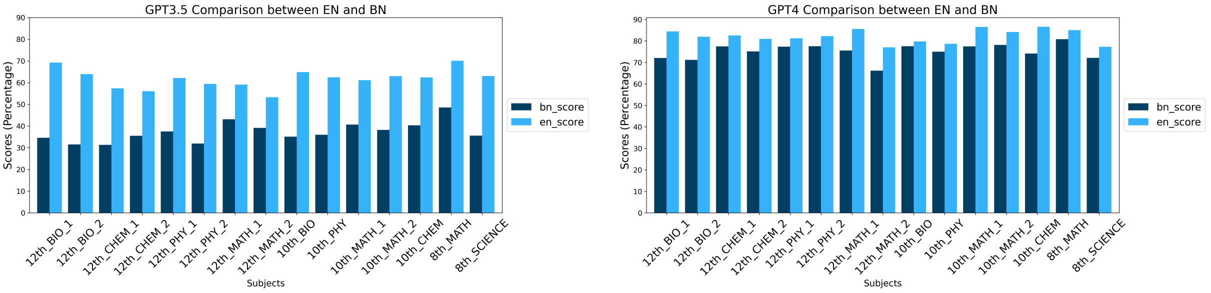
[6] https://mistral.ai/news/announcing-mistral-7b/



(a) LLMs results in English questions

(b) LLMs results in Bengali questions

Figure 3: Zero-shot performance of LLMs on BEnQA dataset



(a) GPT-3.5 results

(b) GPT-4 results

Figure 4: Performance comparison between Bengali and English questions

validity of intermediate reasoning steps.[7] In other words, we consider the answer correct if the final result is accurate, without delving into the examination of the validity of each intermediate reasoning step.

5 Results

5.1 How do LLMs perform on BEnQA?

Performance across all subjects In this experiment, we investigate the performance of various existing LLMs across all subjects in the BEnQA dataset. The results of all models are presented in Figure 3. For English questions (Figure 3a), our result shows that proprietary LLMs are far ahead of open-source LLMs such as LLaMA-2 and Mistral. Among proprietary LLMs, GPT-4 is significantly ahead, followed by GPT-3.5. We also observe that the LLMs' performance across subjects is relatively uniform, with dips in some subjects, such as 12th-grade Math II and 10th-grade Biology, indicating that these might be more challenging subjects for LLMs. The trends are also similar to Bengali questions (Figure 3b), where GPT-4 maintains a significant score gap. Interestingly, Claude2.1 shows a much closer performance to GPT-3.5 in Bengali, a contrast to its performance in English. The full benchmark table can be found in Appendix G.

English vs. Bengali performance To investigate the performance disparity between English and Bengali questions, we focus on two of the highest-performing models for comparison: GPT-3.5 and GPT-4. As depicted in Figure 4a, a substantial performance gap exists between Bengali and English in GPT-3.5. However, in GPT-4 (Figure 4b), the gap is much smaller, and in fact, performance in both English and Bengali has improved by a large margin across all subjects.

5.2 Does Chain-of-Thought help?

Chain-of-Thought (CoT) prompting is typically used to enhance the performance of LLMs in tasks that require reasoning. In order to evaluate its effectiveness on our dataset, we conducted several experiments based on subject and question categories.

Performance by Subject As shown in Figure 5, Chain-of-Thought reasoning boosts GPT-3.5 performances across the dataset in English. We also observe a similar trend with Bengali, with more details available in Appendix C. Interestingly, when

[7] We compared the intermediate CoT evaluation vs. final answer evaluation, and the result shows that the instances where the model gets the final answer right when the CoT steps are wrong are low. The details can be seen in Appendix B.

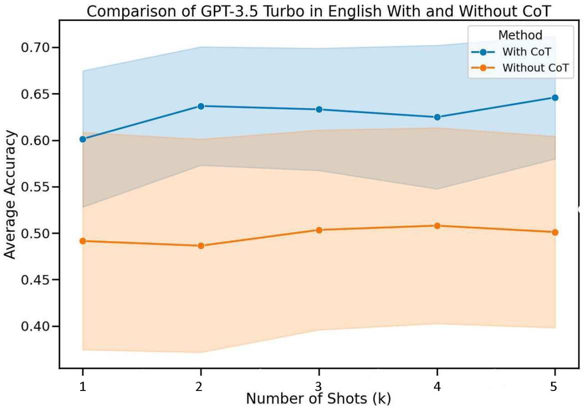


Figure 5: Accuracy of Chain-of-Thought (CoT) reasoning in few-shot settings for GPT-3.5 Turbo in English.

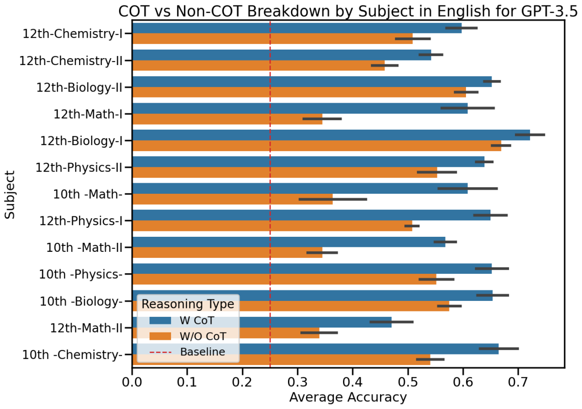


Figure 6: Chain-of-Thought (CoT) performance broken down by subject in English. Note that some subjects benefit more from CoT than others.

we decompose the result by subjects, CoT does not help all subjects uniformly as shown in Figure 6. CoT seems to help Math the most and Biology the least; this is indeed correlated with the different portion of question categories in each subject (Figure 2b), where we see that biology has more factual questions while math has more reasoning and application-based.

Performance by Question Category To validate our subject-based performance findings, we further analyze the GPT-3.5 performance by question category. As depicted in Figure 7, we indeed see that reasoning and application questions particularly benefit from CoT prompting, with accuracy improvements ranging by 10–20%. In contrast, we observed a minimal improvement in factual questions, highlighting the need for alternative methods to enhance the factual question performance. It is also worth noting that in English, the gains in application and reasoning questions through CoT still cannot surpass the performance of factual questions. For Bengali, the improvement is mostly seen in application questions, but not so much in factual and reasoning, as elaborated in Appendix C.2.

6 Additional Experiments

In this section, we conduct additional experiments to explore whether we can improve the performance in Bengali using better prompting. All the experiments described in this section were conducted on GPT-3.5, which provides a good value of cost-efficiency vs. capability in Bengali.

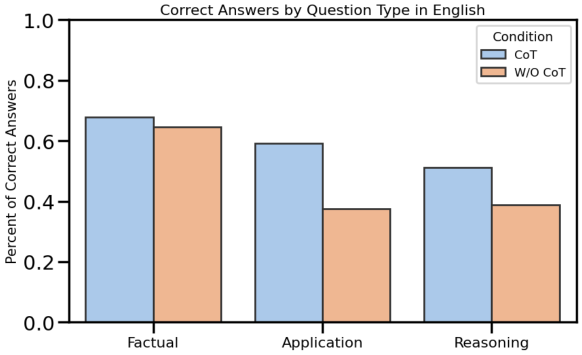
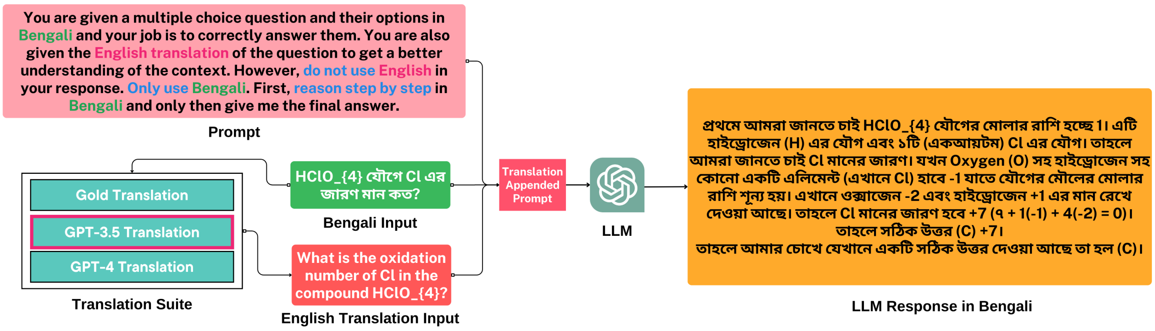


Figure 7: Performance breakdown by question category in English.

6.1 Can the performance be improved by appending English translation?

Our hypothesis for the models' subpar performance in Bengali might be attributed to two main factors: their unfamiliarity with Bengali scientific terminology and potential difficulty in understanding non-latin scripts (Lai et al., 2023). We expect that having access to a translation of the question would likely help the model to understand the context better. To confirm our hypothesis, we experiment with a randomly selected subset of 115 data points from each subject within the 10th-grade exam questions. Questions from 10th-grade exams were chosen because they provide a middle ground in terms of difficulty compared to 8th and 12th grade.

Prompt Details As illustrated in Figure 8, in addition to providing the model with the Bengali question, we append the English translation to the prompt. This results in a prompt that consists of both the original data and its translation.



You are given a multiple choice question and their options in Bengali and your job is to correctly answer them. You are also given the English translation of the question to get a better understanding of the context. However, do not use English in your response. Only use Bengali. First, reason step by step in Bengali and only then give me the final answer.

HClO_{4} যৌগে Cl এর জারণ মান কত?

What is the oxidation number of Cl in the compound HClO_{4}?

প্রথমে আমরা জানতে চাই HClO_{4} যৌগের মোলার রাশি হচ্ছে 1। এটি হাইড্রোজেন (H) এর যৌগ এবং ১টি (একআয়টম) Cl এর যৌগ। তাহলে আমরা জানতে চাই Cl মানের জারণ। যখন Oxygen (O) সহ হাইড্রোজেন সহ কোনো একটি এলিমেন্ট (এখানে Cl) হাবে -1 যাতে যৌগের মৌলের মোলার রাশি শূন্য হয়। এখানে ওক্সাজেন -2 এবং হাইড্রোজেন +1 এর মান রেখে দেওয়া আছে। তাহলে Cl মানের জারণ হবে +7 (৭ + 1(-1) + 4(-2) = 0)।
তাহলে সঠিক উত্তর (C) +7।
তাহলে আমার চোখে যেখানে একটি সঠিক উত্তর দেওয়া আছে তা হল (C)।

Figure 8: Illustration of our translated appended prompting method. This is an actual example taken from our experiments. The LLM in question, GPT-3.5, answers the question correctly; however, the answer contains spelling and grammar mistakes.

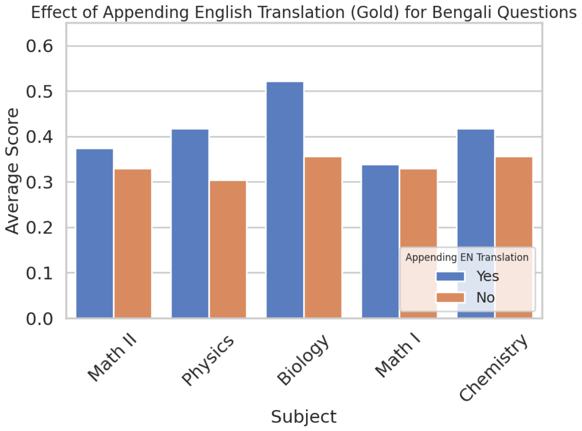


Figure 9: Appending English translation helps to answer questions in Bengali. The model was asked to do CoT in Bengali

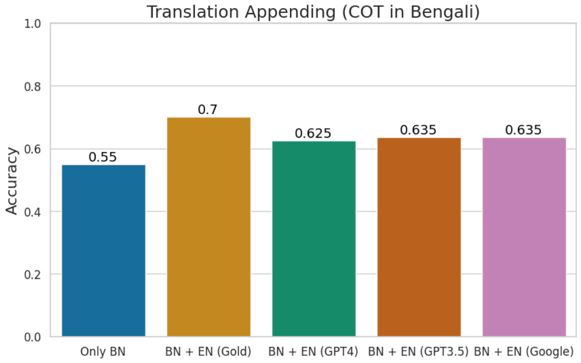


Figure 10: Effects of our appended translation prompting method on the Bengali portion of the IndicCOPA dataset.

Experiment Results As shown from the results in Figure 9, appending English translation appears to have a positive impact in all subjects. Throughout our work, all system prompts have been consistently in English. In this experiment, the questions were provided in Bengali along with the English translations. Hence the performance gain can be concluded to be coming from the appending of the English translations. The most improvement is shown in Biology, where scientific terminology can be found in the majority of questions. In contrast, for Math, the improvement is not as significant as in the other subjects. We also experimented with the case where we did not have access to gold English translations and replaced them with LLM-generated translations. Our initial experiments suggest that appending LLM-generated translations might work just as well as appending human translation. More details can be found in Appendix C.3.

6.2 Does the translation appended prompting method generalize to other datasets?

We extend our experiments to other datasets, such as COPA and Big-Bench-Hard, to evaluate the applicability of our prompting strategy across various datasets. For COPA, we ask the model to answer Bengali COPA problems taken from IndicCOPA (Doddapaneni et al., 2023).

COPA As shown in Figure 10, a noticeable improvement is observed in Bengali COPA when English translations are appended alongside the Bengali question. The improvement is more pronounced when we use the actual English version of the COPA dataset (marked as Gold). However, even using GPT-generated translation, accuracy improves the performance by 7 to 8%.

We also try to see if translation appended prompting can be helpful for the other languages included in the original X-COPA (Ponti et al., 2020) languages. The results are shown in the Figure 11. Appending English translation increases accuracy in all but one language, with the effects most pro-

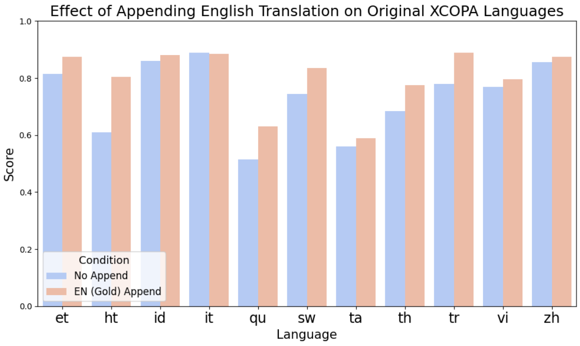


Figure 11: Effects of our appended translation prompting method in original X-COPA dataset languages. In all cases, the model was asked to do CoT in the corresponding languages.

nounced in low-resource languages. Further research in utilizing this prompting technique is an interesting direction left for future work.

Big-Bench-Hard Bengali For this experiment, we curate Big-Bench Hard Bengali (BBH-BN) dataset using GPT-4. Since only a small proportion of tasks require reasoning skills in BEnQA (11% of the total questions), we chose Big-Bench Hard as an additional dataset alternative. We selected some tasks in Big-Bench Hard that require reasoning and also have equivalence to Bengali to add more questions in this category. When we say equivalence to Bengali, we mean ignoring some tasks, like word sorting, because the alphabetical order may not be the same upon translation to Bengali. In this experiment, two native Bengali speakers generated prompts for each task iteratively and then generated the prompt.

Our experiment shows that on average, there is a 5.8% increase in BBH-BN performance in Bengali when we append gold English translation, while appending GPT-4-generated English translation provides an average of 4.2% improvement. Details about Big-Bench-Hard task selections and results can be found in Appendix E.

7 Discussion

Unlike most other bilingual or multilingual NLP datasets, which are usually first curated in English and then translated into other languages, questions in our BEnQA datasets are curated in Bengali first and then translated into English by school teachers. This approach could also provide an interesting outlook on how L2 English speakers might use English and whether LLMs are robust to such linguistic variations.

Besides, our benchmark results highlight that LLMs, such as ChatGPT, still have some way to go in terms of catching up with performances in low-resource languages like Bengali. One interesting aspect we noticed is that, in Bengali, it is much harder to make the model output in a predefined format for automated evaluation. Future research should focus on improving the instruction-following ability of multilingual models.

Another key finding is the substantial performance gap between open-source and proprietary language models. The open-source models, which are more accessible in the context of developing countries, must stride ahead to catch up with the proprietary models to make sure the benefits of AI, particularly LLMs, are not limited to only some demographics. A recently published open multilingual LLM, Aya (Üstün et al., 2024), represents a promising step in this direction.

Furthermore, we demonstrated the feasibility of having LLMs respond in the target language, while leveraging the advantage of high-resource languages, such as English, in the inference pipeline by utilizing translation of the query. The benefit of this approach is that we do not need to have access to the manually-created gold translation (as we had in this case): translation using the same model or more powerful/specialized models would work too. However, the choice of translation method, whether it is another more proficient LLM or a domain-specific fine-tuned translation model, can influence the performance. This finding necessitates further research to optimize and refine the use of LLMs, especially in the context of low-resource languages.

8 Conclusion and Future Work

In this paper, we introduced BEnQA, a locally sourced dataset from Bangladesh, comprising middle-school and high-school level exam questions in both English and Bengali. Our dataset is provided in parallel format, allowing us to benchmark the performance discrepancy between both languages in a fairer way. Upon benchmarking several Large Language Models (LLMs) with this dataset, we found that performance in Bengali lags behind English, even for the best-performing models; and current open-source models significantly lag behind proprietary models. We also explored whether the performance in Bengali questions can be improved by appending English translations to the prompts. This approach resulted in performance enhancements across most subjects in the BEnQA dataset on GPT-3.5 model. Notably, this improvement also extends to other datasets, such as COPA and Big-Bench-Hard. These findings pave a promising direction for future research in improving the LLMs performance in low-resource languages, specifically in Bengali.

Our findings open up several promising directions for future research. Our results have highlighted a possible gap in the current models' understanding of Bengali terminology, leading us to develop a very simple prompting method that helps the performance in Bengali data significantly. We also show the efficacy of this prompting approach on Bengali COPA, X-COPA, and Bengali Big-Bench-Hard dataset. Future research should also focus on to see to what extent such prompting methods utilizing high-resource languages generalize to other datasets and languages.

Limitation

Our works have several limitations that should be acknowledged. Firstly, our dataset primarily focused on text-based questions, as we discarded figure-based questions during the dataset curation. This limitation might restrict the scope of our findings, as questions with visual elements often require more reasoning steps. Additionally, since the questions are multiple-choice, there might be a possibility that the models use a shortcut to answer the questions, especially for questions that do not require advanced reasoning skills like factual questions. Despite these limitations, our dataset serves as an important starting point for benchmarking LLMs in Bengali, which currently have limited resources available for question-answering and knowledge intensive tasks.

Ethical Considerations

The exam questions in BEnQA dataset are freely available and have been manually curated and reviewed to minimize the presence of harmful content. This dataset is publicly accessible and distributed under the CC BY-SA 4.0 license, and the evaluation code is available under the MIT License. Our work has been reviewed and received approval from the Institutional Review Board (IRB) at our institution. All annotators involved in this project were compensated above the minimum wage.

Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW, supervised by the IITP(Institute of Information & Communications Technology Planning&Evaluation) in 2023 (2022-0-01092).

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 9904–9923, Singapore. Association for Computational Linguistics.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. arXiv preprint arXiv:2303.12528.

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.

Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. 2023. Chatgpt goes to law school. Available at SSRN.

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.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Jieun Han, Haneul Yoo, Yoonsu Kim, Junho Myung, Minsun Kim, Hyunseung Lim, Juho Kim, Tak Yeon Lee, Hwajung Hong, So-Yeon Ahn, and Alice Oh.

2023. Recipe: How to integrate chatgpt into efl writing education. In Proceedings of the Tenth ACM Conference on Learning @ Scale, L@S '23, page 416–420, New York, NY, USA. Association for Computing Machinery.

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.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. arXiv preprint arXiv:2305.07004.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Khan Academy. 2023. Harnessing ai so that all students benefit: A nonprofit approach for equal access. https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/. Accessed: 2023-12-15.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in indonesia: A comprehensive test on indommlu.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. PLoS digital health, 2(2):e0000198.

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.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. arXiv preprint arXiv:2304.03439.

OpenAI. 2023. Gpt-4 technical report.

Aleksandar Petrov, Emanuele La Malfa, Philip HS Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. arXiv preprint arXiv:2305.15425.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2362–2376, Online. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In 2011 AAAI Spring Symposium Series.

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.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.

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.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Adam Zeidan. 2023. Languages by total number of speakers.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model.

A Effect of Grammar Mistakes in BEnQA

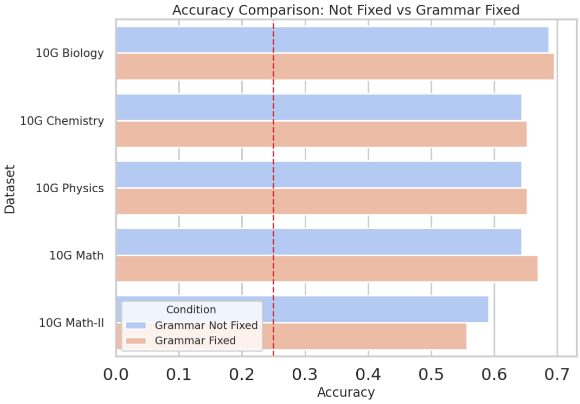


Figure 12: Effects of grammar mistakes on GPT-3.5 performance.

To see the effect of subtle grammar mistakes and unnatural English translations, we selected 120 questions from each subject of the 10th grade. We prompt GPT-4 to fix grammatical issues and unnatural translations if there are any. A native Bengali speaker proficient in English then went through the results and made necessary adjustments to the GPT-4 corrections.[8] The performance of GPT-3.5 on this subset of the dataset is presented in Figure 12. The difference between a grammar mistake-free version and the version with grammar mistakes is minimal, and a manual inspection revealed the slight discrepancy that is shown in Figure 12 emerged from the stochastic nature of GPT-3.5 rather than any grammar mistakes.

B Intermediate CoT Results

	Final Ans. Correct	Final Ans. Wrong
English translation Appended		
CoT Correct	30	1
CoT Wrong	4	26
English Translation Non-appended		
CoT Correct	22	0
CoT Wrong	6	29

Table 1: Examination of correctness of CoT steps and the final answer. In all cases, the intermediate reasoning steps were done in Bengali.

We did human evaluations on a subset of the responses, and the results can be seen in Table 1. The primary observations are: when the CoT is wrong, the model's final answer is wrong, and when the CoT is right, the final answer is correct. The instances where the model accidentally gets the final answer right with the wrong CoT steps are pretty low.

Evaluating the CoT steps for all samples using human evaluation is impractical, and based on the analysis of this sample, we decided to evaluate the final answer to make the evaluation more scalable.

C Additional BEnQA Results

C.1 Few-Shot Results

We adhere to zero-shot prompting throughout the paper as the preliminary experiments revealed that the difference between zero-shot and few-shot settings is not very significant (Figure 13) and conducting extensive few-shot experiments is costly

[8] Oftentimes, GPT-4 alters the original meaning of the question when doing grammar fixing, which is why we did not use this protocol to fix grammar for the whole dataset.

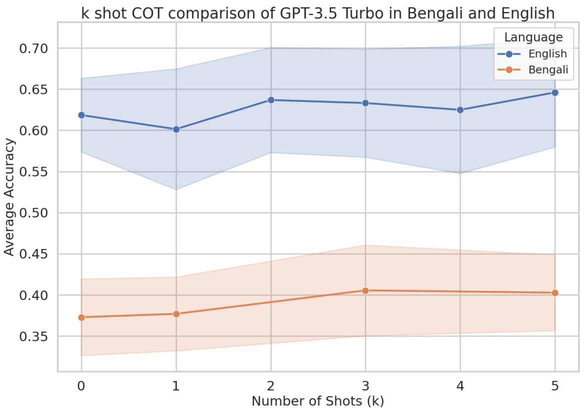


Figure 13: k-shot prompting in Bengali and English on GPT-3.5. Note that zero shot prompting had more detailed system prompt than for few shot prompting (See Appendix F)

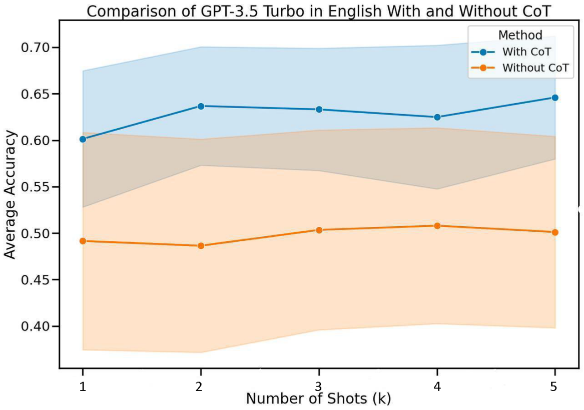


Figure 14: Effect of Chain of Thought reasoning across k shot on GPT-3.5 for BEnQA English

for proprietary models, and even more so in Bengali due to infertile tokenization.

Few-Shot Example Preparation: When preparing the few shot examples, it was made sure that at least the first three examples cover our categorization (factual, procedure and knowledge, reasoning) as described in Section 3.3. The five examples also covered all types of questions that (multiple choice and multi-selection type) usually appear in the exams. Finally, the examples in each topic were topic-wise diverse for each subject.

C.2 Effect of Chain-of-Thought Prompting

Chain-of-Thought (CoT) performance for English in 1 to 5-shot settings is illustrated in Figure 14. In all cases, employing CoT improves performance.

Similar observations can be made for GPT-3.5 in Bengali as well (Figure 15). We only chose k = 1, 3, and 5 to prevent high API costs associated with low-resource languages.

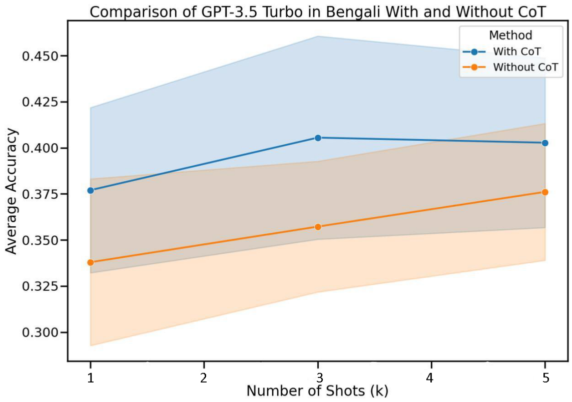


Figure 15: Effect of CoT reasoning across k shot on GPT-3.5. Note that in this Figure k is 1, 3, and 5

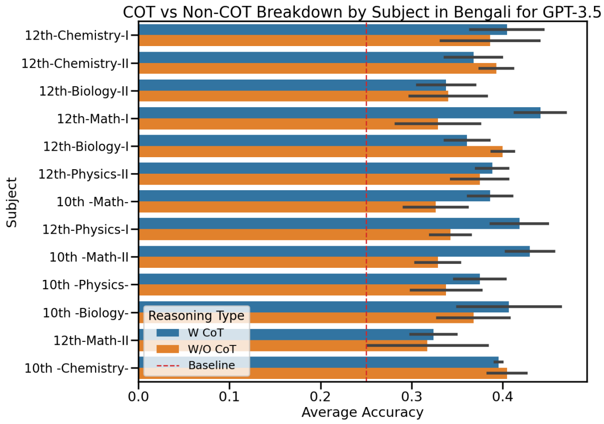


Figure 16: CoT performance breakdown by subject in Bengali.

CoT Subjectwise Breakdown: Similar to what we have observed in Section 5.2 We also note that doing CoT improves performance for Bengali. Figure 16 shows the subjectwise CoT performance breakdown for GPT-3.5.

We can also see the CoT performance breakdown by question type for Bengali in Figure 17. Unlike English, the performance in reasoning questions turns out to be on par with Bengali. Upon analysis, we conjecture that this discrepancy is due to the smaller size of data for the reasoning questions in some subjects, such as Physics. Moreover, the model by default also finds these questions harder to solve in Bengali compared to English as we can see that the base accuracy is much lower for Bengali than for English.

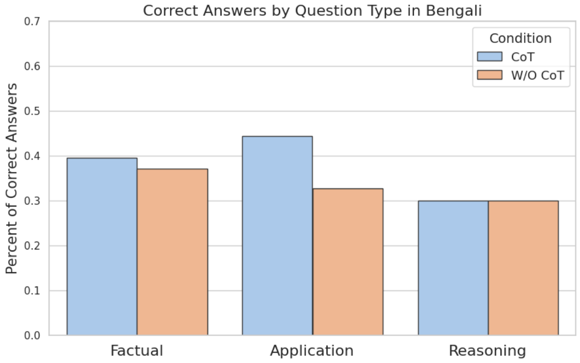


Figure 17: CoT performance breakdown by question type in Bengali.

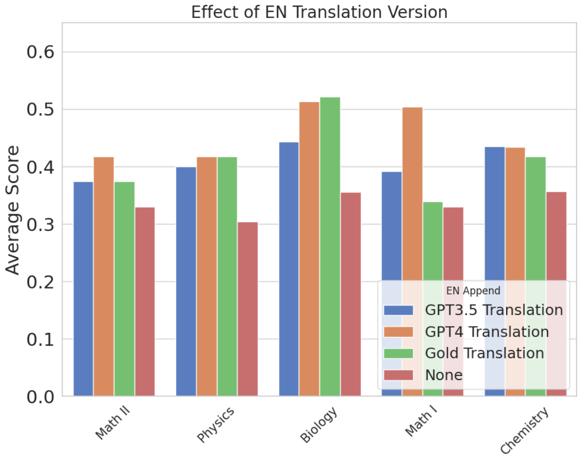


Figure 18: Effect of appending LLM generated English translation in appending experiment. None refers to the case where no English translation was appended.

C.3 Does LLM-generated translation work as well for appending?

One obvious question raised by our translation appending prompt method is whether machine translation can work as a replacement in cases where we do not have access to human-generated translations.

Figure 18 shows that appending LLM-generated English translation works just as well, and in fact, outperformed the original translation for some of the subjects. We only used GPT-3.5 and GPT-4 translations in this experiment and did not incorporate Google Translate; because, in our observations, equations written in LaTeX format tend to get broken by Google Translate.

The result of this experiment leaves open the possibility of incorporating such prompting techniques for a wide range of multilingual tasks where obtaining manual English translation would be quite difficult.

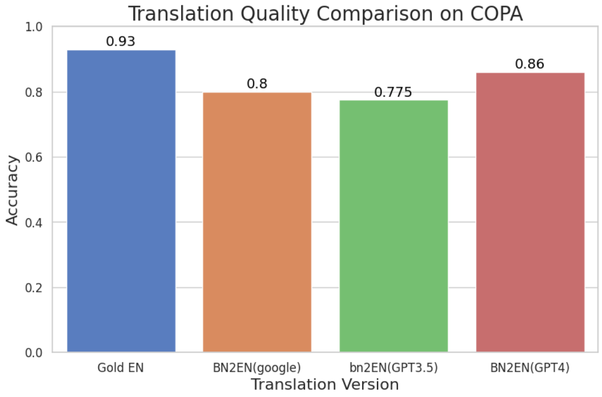


Figure 19: Backtranslating COPA from Bengali to English and seeing their performance. Gold EN refers to the original English COPA dataset

D Evaluation of Translation Quality Using Backtranslation

We use the COPA dataset to validate the translation quality of different models. We first take the Bengali translation of COPA dataset from (Doddapaneni et al., 2023) and backtranslate the Bengali translation of COPA dataset from Bengali to English using GPT-4, GPT-3.5 and Google Translate and then compare it with the performance on original English COPA. The results are in the Figure 19.

As we can see, GPT-4 translation (BN to EN) provides the closest results to the original English translation, while GPT-3.5 and Google Translate are slightly lower.

E Additional Experiments on COPA and BIG-Bench

Prompt Variation: We explore four possible variations of the prompting method:

- **BN:** Asking only in Bengali.
- **BN + EN (Gold):** Asking in Bengali and appending the English version from the original data.
- **BN + EN (GPT-3.5):** Asking in Bengali and appending the English translation done by GPT-3.5.
- **BN + EN (GPT-4):** Asking in Bengali and appending the English translation done by GPT-4.

Big-Bench Hard Task Selection: From the 23 datasets officially provided by BIG-Bench Hard, we selected 14 tasks that can have equivalence in Bengali[9]. We chose Causal Judgement, Date Understanding, Disambiguation QA, Formal Fallacies, Logical Deductions Five, Logical Deductions Seven, Logical Deductions Three, Multistep Arithmetic, Navigate, Object Counting, Penguin In a Table, Reasoning About Colored Objects, Temporal Sequences, and Web of Lies tasks for our work.

These 14 tasks were machine-translated into Bengali using GPT-4 with three human-annotated examples as prompts for each task. The prompts were made iteratively by two Bengali speakers to reflect the nature of each task.

Experiment Result: Figure 20 shows the results for appending English translations in Big-Bench-Hard experiments. In 7 out of 14 cases, appending English translation helped, while in three cases, it slightly hurt the performance. For the remaining four cases, performance remained largely unaffected. The detailed numbers can be found in Table 11.

F Prompts Used for Experiments

This section contains all the original prompts used for all the experiments.

Grammar-Corrected BEnQA Question Preparation: We use the prompt given in Table 2 to fix the grammatical issues of the original questions in English with GPT-4.

BEnQA Dataset Categorization: We mentioned in 3.3 that we classified our dataset questions into these three categories: Factual Knowledge, Procedural & Application, and Reasoning using GPT-4. We use the prompt given in Table 3 to categorize them.

For each question in the dataset, we take zero-shot approach with the prompt for categorization.

Table 7 represents the subject-wise question category in tabular visualization.

Chain-of-Thought Prompting for BEnQA Zero-shot benchmark: Table 4 shows the prompt for zero-shot benchmarking with the Proprietary models.

Chain-of-Thought Prompting for BEnQA Few-shot benchmark: The prompt we use for few-shot benchmarking with the Proprietary models is shown in Table 5.

Prompt for Translation Appended Benchmark: As described in E, we use the prompt shown in Table 6 to restrict the model to only reason in English in the translation-append experiment.

G Benchmark Statistics

This section contains benchmark results by types and datasets used for our experiments

BEnQA Zero-shot Benchmark Table 8 shows the zero-shot benchmark results on BEnQA.

BEnQA Few-shot Benchmark without Chain-of-Thought Reasoning Table 9 shows the few-shot benchmark results on BEnQA without Chain-of-Thought Reasoning.

BEnQA Few-shot Benchmark with Chain-of-Thought Reasoning Table 10 shows the few-shot benchmark results on BEnQA with the use of Chain-of-Thought.

BIG-Bench-Hard Zero-shot Benchmark Table 11 shows the zero-shot benchmark results on selected reasoning-based BIG-Bench-Hard datasets.

H BEnQA Questions Examples

We present some examples from the 10th-grade subjects by the categories consisting of both English and Bengali versions in Figures 21 & 22. We kept the questions as they are; that is, we did not correct the subtle grammatical awkwardness in the English version of some of the questions.

[9] Some BBH tasks like manipulating English alphabet letters, which do not have a direct 'translation' in Bengali.

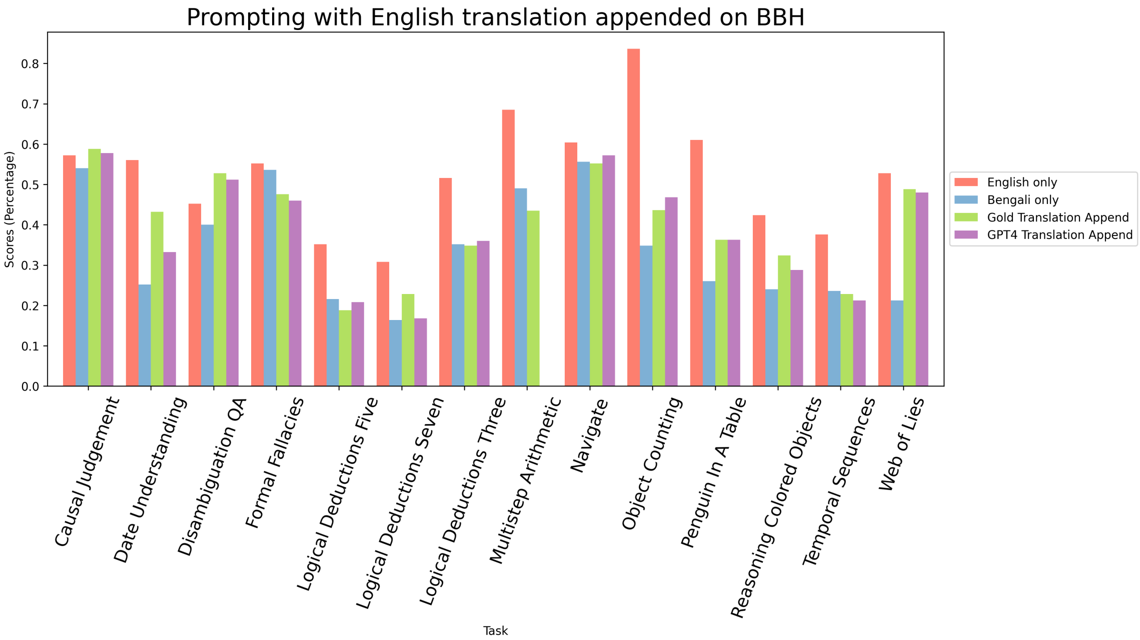


Figure 20: Effects of our translation appended prompting method on BIG-Bench Hard dataset. Here Gold Translation Append refers to the case where the question was asked in Bengali and the original English translation was appended, while GPT-4 Translation Append refers to the case where the English translation was generated using GPT-4

You are given a multiple choice exam question in English (along with their choices).

The question is mostly good, but sometimes contains minor grammatical mistakes and non-standard vocabulary. Your job is to make it sound natural and fluent.
You are also given the choices of the question to get a better understanding of the context.
However, do not return those choices in your response. Only return the question. Make sure you do the full question, not just the part of it.
If there is any equation or formula in the question, do not modify them.

The question to be fixed is given below:

Table 2: Prompt for Making Grammar-Corrected Questions by GPT-4

You are given a multiple choice question and your job is to tell what kind of reasoning is required to solve the problem.

Choose from the following options:
1. Factual Knowledge: The question requires only the knowledge of basic facts, dates, events, concepts, etc.
2. Procedural and Application: The question requires the ability to apply a procedure or a formula to solve the problem.
3. Reasoning: The question requires the ability to do multistep reasoning to solve the problem.

The question is given below:

Table 3: Prompt used to categorize questions in BEnQA dataset.

You are given a multiple choice question and their options in English/Bengali and your job is to correctly answer the question. First reason step by step in English/Bengali and only then give me the final answer as "a", "b", "c" or "d".

Keep these in mind:
1. Only include the letter a, b, c, d as your final answer. Do not include the option text.
2. Every question will have an answer in the given options. So, DO NOT say that none of the answers are correct.
3. ONLY ONE of the given options will have the answer. So DO NOT provide multiple options as answers.
4. The questions contain enough information to solve the problem, so DO NOT say that you need additional information.

The question is given below:

Table 4: Prompt for Proprietary Models for BEnQA Zero-Shot Benchmark

You are given a multiple choice question in English/Bengali. Your job is to answer it correctly. First reason step by step and then answer the question.

{Question 1}
{Answer 1}

{Question 2}
{Answer 2}
- -
{Question}

Table 5: Prompt for Proprietary Models for BEnQA Few-Shot Benchmark

You are given a situation and its possible reason effect/answer in Bengali and your job is to correctly identify the reason effect/answer of the situation.

For your better understanding, English translation is also given. However, you must answer in Bengali only.
Reason step by step in Bengali and only then give me the final answer.

The question is given below:

Table 6: Prompt for Zero-shot Append Experiment

Subject	Factual Knowledge	Procedural & Application	Reasoning	Total by Dataset
12th Bio I	288	12	15	315
12th Bio II	297	8	28	333
12th Chem I	229	85	58	372
12th Chem II	185	145	64	394
12th Phy I	115	161	32	308
12th Phy II	168	143	27	338
12th Math I	13	368	20	401
12th Math II	24	327	45	396
10th Bio	308	21	27	356
10th Phy	178	119	27	324
10th Math I	45	267	73	385
10th Math II	24	316	58	398
10th Chem	268	91	35	394
8th Math	30	139	45	214
8th Sci	194	26	13	233
Total	**2366**	**2228**	**567**	**5161**

Table 7: BEnQA Dataset Question Count by Subjects and Categories

Language	Dataset	GPT 4	GPT 3.5	Claude 2.1	LLaMA2 (13b)	LLaMA2 (7b)	Mistral 7b
English	12th Bio I	84.44	69.21	59.37	38.10	30.48	31.43
	12th Bio II	81.98	63.96	55.26	33.63	24.62	38.14
	12th Chem I	82.57	57.37	52.06	24.13	19.35	27.08
	12th Chem II	80.96	56.09	51.75	23.60	14.47	22.34
	12th Phy I	81.23	62.14	48.25	29.13	17.15	25.24
	12th Phy II	82.25	59.47	37.78	23.37	25.44	27.81
	12th Math I	85.54	59.10	57.93	10.22	6.23	11.22
	12th Math II	77.02	53.28	56.51	16.92	6.82	15.15
	10th Bio	79.78	64.89	51.75	34.83	25.00	36.52
	10th Phy	78.70	62.46	53.02	28.92	19.38	26.77
	10th Math I	86.53	61.14	47.94	13.21	12.18	12.69
	10th Math II	84.17	63.00	49.52	9.25	10.50	15.25
	10th Chem	86.55	62.44	56.83	28.68	22.08	28.43
	8th Math	85.05	70.09	56.07	26.17	21.03	21.50
	8th Sci	77.25	63.09	54.08	42.06	23.61	36.91
Bengali	12th Bio I	72.06	34.60	35.56			
	12th Bio II	71.17	31.53	30.48			
	12th Chem I	77.48	31.37	38.10			
	12th Chem II	75.13	35.53	34.60			
	12th Phy I	77.35	37.54	42.54			
	12th Phy II	77.51	31.95	29.84			
	12th Math I	75.56	43.14	36.57			
	12th Math II	66.16	39.14	28.57			
	10th Bio	77.53	35.11	41.85			
	10th Phy	75.00	36.00	36.19			
	10th Math I	77.46	40.67	34.92			
	10th Math II	78.14	38.25	35.87			
	10th Chem	74.11	40.36	34.92			
	8th Math	80.84	48.60	45.33			
	8th Sci	72.10	35.62	35.62			

Table 8: BEnQA Zero-shot Benchmark

Language	**Dataset**	**GPT 3.5**		**LLaMA2 (7b)**		**LLaMA2 (13b)**	
		5-shot	**3-shot**	**5-shot**	**3-shot**	**5-shot**	**3-shot**
English	10th Bio	58.26	59.13	48.70	43.48	43.48	39.13
	10th Phy	56.52	60.87	37.39	39.13	29.57	27.83
	10th Math I	41.74	43.48	30.43	25.22	33.91	28.70
	10th Math II	46.09	37.39	21.74	21.74	40.00	42.61
	10th Chem	58.26	58.26	36.52	37.39	26.96	25.22
Bengali	10th Bio	33.04	32.17				
	10th Phy	41.74	35.65				
	10th Math I	26.96	32.17				
	10th Math II	33.04	33.91				
	10th Chem	41.74	41.74				

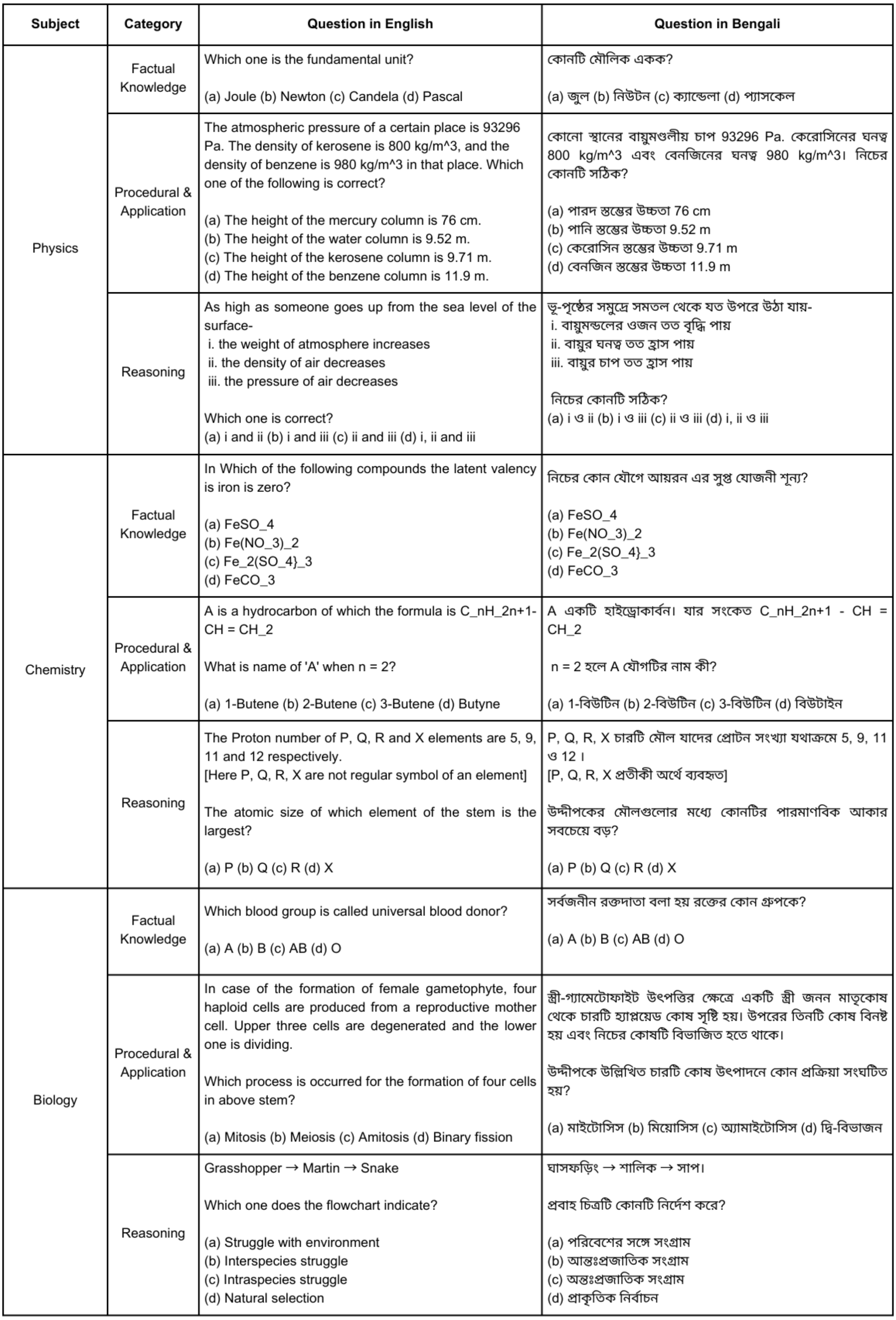
Table 9: BEnQA 10th Grade Few-shot Benchmark without Chain-of-Thought Reasoning

Language	**Dataset**	**GPT 3.5**		**LLaMA2 (7b)**		**LLaMA2 (13b)**	
		5-shot	**3-shot**	**5-shot**	**3-shot**	**5-shot**	**3-shot**
English	10th Bio	67.83	69.57	49.57	49.57	48.7	43.48
	10th Phy	66.09	70.43	40.00	38.26	38.26	43.48
	10th Math I	66.09	69.57	28.70	26.09	37.39	33.04
	10th Math II	62.61	58.26	29.57	27.83	25.22	30.43
	10th Chem	68.70	63.48	44.35	42.61	43.48	42.61
Bengali	10th Bio	36.52	33.91				
	10th Phy	46.09	39.13				
	10th Math I	34.84	41.74				
	10th Math II	40.87	45.22				
	10th Chem	40.87	47.83				

Table 10: BEnQA 10th Grade Few-shot Benchmark with Chain-of-Thought Reasoning

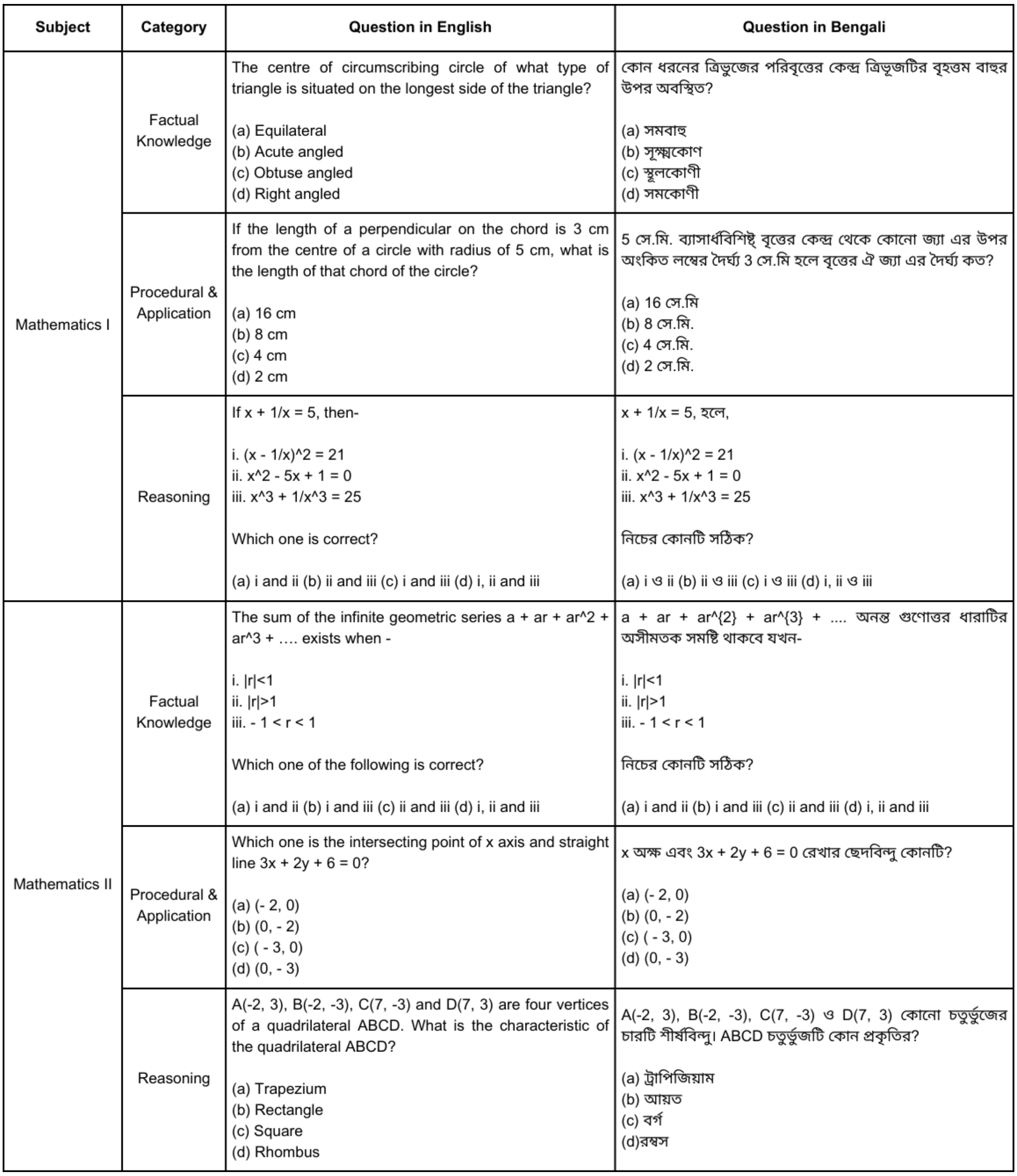
Dataset	**English only**	**Bengali only**	**Gold Translation Append**	**GPT Translation Append**
Causal Judgement	57.2	54.0	**58.8**	57.8
Date Understanding	56.0	25.2	**43.2**	33.2
Disambiguation QA	45.2	40.0	**52.8**	51.2
Formal Fallacies	55.2	**53.6**	47.6	46.0
Logical Deductions Five	35.2	**21.6**	18.8	20.8
Logical Deductions Seven	30.8	16.4	**22.8**	16.8
Logical Deductions Three	51.6	35.2	34.8	**36.0**
Multistep Arithmetic	68.5	**49.0**	43.5	
Navigate	60.4	55.6	55.2	**57.2**
Object Counting	83.6	34.8	43.6	**46.8**
Penguin In A Table	61.0	26.0	**36.3**	**36.3**
Reasoning About Colored Objects	42.4	24.0	**32.4**	28.8
Temporal Sequences	37.6	**23.6**	22.8	21.2
Web of Lies	52.8	21.2	**48.8**	48.0

Table 11: Big-Bench Hard Zero-shot Benchmark



Subject	Category	Question in English	Question in Bengali
Physics	Factual Knowledge	Which one is the fundamental unit? (a) Joule (b) Newton (c) Candela (d) Pascal	কোনটি মৌলিক একক? (a) জুল (b) নিউটন (c) ক্যান্ডেলা (d) প্যাসকেল
	Procedural & Application	The atmospheric pressure of a certain place is 93296 Pa. The density of kerosene is 800 kg/m^3, and the density of benzene is 980 kg/m^3 in that place. Which one of the following is correct? (a) The height of the mercury column is 76 cm. (b) The height of the water column is 9.52 m. (c) The height of the kerosene column is 9.71 m. (d) The height of the benzene column is 11.9 m.	কোনো স্থানের বায়ুমণ্ডলীয় চাপ 93296 Pa. কেরোসিনের ঘনত্ব 800 kg/m^3 এবং বেনজিনের ঘনত্ব 980 kg/m^3। নিচের কোনটি সঠিক? (a) পারদ স্তম্ভের উচ্চতা 76 cm (b) পানি স্তম্ভের উচ্চতা 9.52 m (c) কেরোসিন স্তম্ভের উচ্চতা 9.71 m (d) বেনজিন স্তম্ভের উচ্চতা 11.9 m
	Reasoning	As high as someone goes up from the sea level of the surface- i. the weight of atmosphere increases ii. the density of air decreases iii. the pressure of air decreases Which one is correct? (a) i and ii (b) i and iii (c) ii and iii (d) i, ii and iii	ভূ-পৃষ্ঠের সমুদ্রে সমতল থেকে যত উপরে উঠা যায়- i. বায়ুমন্ডলের ওজন তত বৃদ্ধি পায় ii. বায়ুর ঘনত্ব তত হ্রাস পায় iii. বায়ুর চাপ তত হ্রাস পায় নিচের কোনটি সঠিক? (a) i ও ii (b) i ও iii (c) ii ও iii (d) i, ii ও iii
Chemistry	Factual Knowledge	In Which of the following compounds the latent valency is iron is zero? (a) FeSO_4 (b) Fe(NO_3)_2 (c) Fe_2(SO_4}_3 (d) FeCO_3	নিচের কোন যৌগে আয়রন এর সুপ্ত যোজনী শূন্য? (a) FeSO_4 (b) Fe(NO_3)_2 (c) Fe_2(SO_4}_3 (d) FeCO_3
	Procedural & Application	A is a hydrocarbon of which the formula is C_nH_2n+1-CH = CH_2 What is name of 'A' when n = 2? (a) 1-Butene (b) 2-Butene (c) 3-Butene (d) Butyne	A একটি হাইড্রোকার্বন। যার সংকেত C_nH_2n+1 - CH = CH_2 n = 2 হলে A যৌগটির নাম কী? (a) 1-বিউটিন (b) 2-বিউটিন (c) 3-বিউটিন (d) বিউটাইন
	Reasoning	The Proton number of P, Q, R and X elements are 5, 9, 11 and 12 respectively. [Here P, Q, R, X are not regular symbol of an element] The atomic size of which element of the stem is the largest? (a) P (b) Q (c) R (d) X	P, Q, R, X চারটি মৌল যাদের প্রোটন সংখ্যা যথাক্রমে 5, 9, 11 ও 12। [P, Q, R, X প্রতীকী অর্থে ব্যবহৃত] উদ্দীপকের মৌলগুলোর মধ্যে কোনটির পারমাণবিক আকার সবচেয়ে বড়? (a) P (b) Q (c) R (d) X
Biology	Factual Knowledge	Which blood group is called universal blood donor? (a) A (b) B (c) AB (d) O	সর্বজনীন রক্তদাতা বলা হয় রক্তের কোন গ্রুপকে? (a) A (b) B (c) AB (d) O
	Procedural & Application	In case of the formation of female gametophyte, four haploid cells are produced from a reproductive mother cell. Upper three cells are degenerated and the lower one is dividing. Which process is occurred for the formation of four cells in above stem? (a) Mitosis (b) Meiosis (c) Amitosis (d) Binary fission	স্ত্রী-গ্যামেটোফাইট উৎপত্তির ক্ষেত্রে একটি স্ত্রী জনন মাতৃকোষ থেকে চারটি হ্যাপ্লয়েড কোষ সৃষ্টি হয়। উপরের তিনটি কোষ বিনষ্ট হয় এবং নিচের কোষটি বিভাজিত হতে থাকে। উদ্দীপকে উল্লিখিত চারটি কোষ উৎপাদনে কোন প্রক্রিয়া সংঘটিত হয়? (a) মাইটোসিস (b) মিয়োসিস (c) অ্যামাইটোসিস (d) দ্বি-বিভাজন
	Reasoning	Grasshopper → Martin → Snake Which one does the flowchart indicate? (a) Struggle with environment (b) Interspecies struggle (c) Intraspecies struggle (d) Natural selection	ঘাসফড়িং → শালিক → সাপ। প্রবাহ চিত্রটি কোনটি নির্দেশ করে? (a) পরিবেশের সঙ্গে সংগ্রাম (b) আন্তঃপ্রজাতিক সংগ্রাম (c) অন্তঃপ্রজাতিক সংগ্রাম (d) প্রাকৃতিক নির্বাচন

Figure 21: BEnQA 10th Grade Sample Questions by Subject and Category



Subject	Category	Question in English	Question in Bengali
Mathematics I	Factual Knowledge	The centre of circumscribing circle of what type of triangle is situated on the longest side of the triangle? (a) Equilateral (b) Acute angled (c) Obtuse angled (d) Right angled	কোন ধরনের ত্রিভুজের পরিবৃত্তের কেন্দ্র ত্রিভূজটির বৃহত্তম বাহুর উপর অবস্থিত? (a) সমবাহু (b) সূক্ষ্মকোণ (c) স্থূলকোণী (d) সমকোণী
	Procedural & Application	If the length of a perpendicular on the chord is 3 cm from the centre of a circle with radius of 5 cm, what is the length of that chord of the circle? (a) 16 cm (b) 8 cm (c) 4 cm (d) 2 cm	5 সে.মি. ব্যাসার্ধবিশিষ্ট্ বৃত্তের কেন্দ্র থেকে কোনো জ্যা এর উপর অংকিত লম্বের দৈর্ঘ্য 3 সে.মি হলে বৃত্তের ঐ জ্যা এর দৈর্ঘ্য কত? (a) 16 সে.মি (b) 8 সে.মি. (c) 4 সে.মি. (d) 2 সে.মি.
	Reasoning	If x + 1/x = 5, then- i. (x - 1/x)^2 = 21 ii. x^2 - 5x + 1 = 0 iii. x^3 + 1/x^3 = 25 Which one is correct? (a) i and ii (b) ii and iii (c) i and iii (d) i, ii and iii	x + 1/x = 5, হলে, i. (x - 1/x)^2 = 21 ii. x^2 - 5x + 1 = 0 iii. x^3 + 1/x^3 = 25 নিচের কোনটি সঠিক? (a) i ও ii (b) ii ও iii (c) i ও iii (d) i, ii ও iii
Mathematics II	Factual Knowledge	The sum of the infinite geometric series a + ar + ar^2 + ar^3 + …. exists when - i. |r|<1 ii. |r|>1 iii. - 1 < r < 1 Which one of the following is correct? (a) i and ii (b) i and iii (c) ii and iii (d) i, ii and iii	a + ar + ar^{2} + ar^{3} + ... অনন্ত গুণোত্তর ধারাটির অসীমতক সমষ্টি থাকবে যখন- i. |r|<1 ii. |r|>1 iii. - 1 < r < 1 নিচের কোনটি সঠিক? (a) i and ii (b) i and iii (c) ii and iii (d) i, ii and iii
	Procedural & Application	Which one is the intersecting point of x axis and straight line 3x + 2y + 6 = 0? (a) (- 2, 0) (b) (0, - 2) (c) (- 3, 0) (d) (0, - 3)	x অক্ষ এবং 3x + 2y + 6 = 0 রেখার ছেদবিন্দু কোনটি? (a) (- 2, 0) (b) (0, - 2) (c) (- 3, 0) (d) (0, - 3)
	Reasoning	A(-2, 3), B(-2, -3), C(7, -3) and D(7, 3) are four vertices of a quadrilateral ABCD. What is the characteristic of the quadrilateral ABCD? (a) Trapezium (b) Rectangle (c) Square (d) Rhombus	A(-2, 3), B(-2, -3), C(7, -3) ও D(7, 3) কোনো চতুর্ভুজের চারটি শীর্ষবিন্দু। ABCD চতুর্ভুজটি কোন প্রকৃতির? (a) ট্রাপিজিয়াম (b) আয়ত (c) বর্গ (d)রম্বস

Figure 22: BEnQA 10th Grade Sample Questions by Subject and Category