SOMAJGYAAN: A Dataset for Evaluating LLMs on Bangla Culture, Social Knowledge, and Low-Resource Language Adaptation

Fariha Anjum Shifa[2,*], Muhtasim Ibteda Shochcho[1,*], Abdullah Ibne Hanif Arean[2,*], Mohammad Ashfaq Ur Rahman[1], AKM Moshiur Rahman [1], Ahaj Mahhin Faiak[1], Md Fahim[1,3,†], M Ashraful Amin[1], Amin Ahsan Ali[1], AKM Mahbubur Rahman[1]

[1]*Center for Computational & Data Sciences* [2]*University of Dhaka* [3]*Penta Global Limited*

[*]Equal Contribution [†]Project Lead

Correspondence: fahimcse381@gmail.com

Abstract

Despite significant progress in large language models (LLMs), their knowledge and evaluation continue to be centered around high-resource languages, leaving critical gaps in low-resource settings. This raises questions about how effectively LLMs handle subjects that require locally relevant knowledge. To address this challenge, we need a robust dataset that reflects the knowledge of underrepresented regions such as Bangladesh. In this paper, we present SOMAJGYAAN, a Bangla multiple-choice dataset consisting of 4,234 questions, annotated across five levels of difficulty. The questions are drawn from Bangladesh's National Curriculum and Global Studies textbooks, covering a wide range of domains including History, Geography, Economics, Social Studies, Politics and Law, and Miscellaneous topics. Difficulty levels were assigned by four expert annotators to minimize annotation bias. The experiments reveal that closed-source LLMs perform better than open-source LLMs. While fine-tuning open-source models on SOMAJGYAAN improves their performance, they still fall short of matching closed-source LLMs. Our findings highlight the importance of culturally grounded evaluation datasets and task-specific adaptation to improve LLM performance in low-resource language settings. Our dataset is available at https://github.com/farihanjum/SOMAJGYAAN.

1 Introduction

LLMs have demonstrated remarkable abilities in language understanding, generation, and reasoning, achieving state-of-the-art performance across a wide range of natural language processing (NLP) tasks. Their widespread adoption spans applications from information retrieval to decision support systems, positioning them as critical intermediaries in the global dissemination of knowledge. However, as these models increasingly shape how information is accessed and interpreted, concerns have surfaced regarding their ability to equitably represent cultural and regional knowledge, particularly for communities underrepresented in their training data. Previous research (Yao et al., 2024) has shown that LLMs exhibit cultural and geographic biases, often underperforming on tasks involving low-resource languages. LLMs exhibit biases against locations with lower socioeconomic conditions, such as most of Africa, on subjective topics like attractiveness and morality, with Spearman's ρ up to 0.70 (Manvi et al., 2024). As language and culture are deeply intertwined, the marginalization of regional knowledge in LLMs can lead to misinformation, stereotyping, and loss of cultural nuance (Shafayat et al., 2024b), particularly for communities like Bangladesh.

Advances in multilingual NLP have allowed LLM accessibility for non-English contexts, yet significant gaps persist in South Asian language coverage and region-specific knowledge modeling. Many of these advancements primarily cater to global contexts but often overlook the uniqueness of local cultures. This has resulted in a significant gap in the evaluation of LLM capabilities in low-resource and culturally distinct settings (Myung et al., 2024; Etxaniz et al., 2024).

Despite Bangladesh's significant cultural, historical, and geopolitical relevance, its representation in large-scale datasets commonly used to train LLMs is minimal. Even though roughly 173 million people speak Bangla, Bangladeshi content constitutes around 0.1 percent of Common Crawl according to the "Statistics of Common Crawl Monthly Archives" (Crawl, 2025). This data scarcity creates a knowledge lacuna, leaving unanswered critical questions about LLMs' capacity to handle Bangladeshi history, legal systems, geography, and sociocultural practices. As a result, there is little empirical understanding of how these models perform when providing accurate, nuanced, and con-

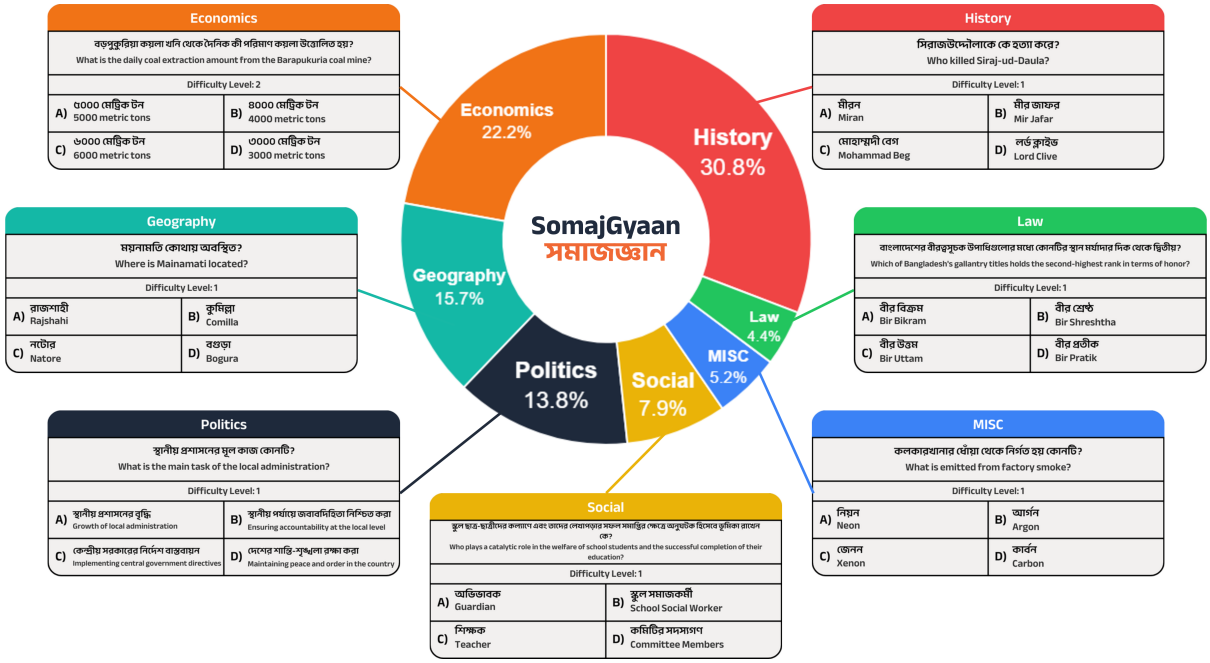


Figure 1: SOMAJGYAAN Dataset Overview

textually relevant information about Bangladesh (Poli et al., 2024). To assess how LLMs perform on local topics from an underrepresented culture, we introduce SOMAJGYAAN, a multiple-choice question-answering dataset with 4234 samples divided into single-hop and multi-hop questions. The question spans a wide range of subjects related to Bangladesh's history, geography, economics, legal systems, social systems, and other miscellaneous areas.

A broader comparison with open-source and proprietary models reveals a performance gap, with leading multilingual models such as DeepSeek-R1-Distill-14B and Phi 4. Meanwhile, top-performing closed-source models such as Gemini 2.0 Flash (0.73), GPT-4o (0.69), and Claude 3.7 (0.64) continue to outperform open models, particularly in knowledge-intensive domains such as Law and Economics. Notably, smaller architectures like Phi 4 achieve a competitive 0.52, challenging the notion that larger models always perform better. Our experiment demonstrated that Fine-tuning with Low-Rank Adaptation (LoRA) (Ge, 2024) further enhances the performance of the open-source performance but still lags behind the closed-source LLMs. Our contribution can be summarized as follows:

- We release SOMAJGYAAN, a novel multiple-choice question-answering dataset designed to evaluate the knowledge of LLMs about Bangladesh across seven categories. The dataset contains single-hop and multi-hop questions, comprising 4,234 samples spanning five difficulty levels. To the best of our knowledge, we are the first to propose a Bangla Cultural and History-based dataset.
- We analyze the performance of SOTA open- and closed-source LLMs on our dataset through both prompting and fine-tuning through an extensive benchmarking.

2 Related Work

Research in NLP primarily focuses on the English language due to extensive resource availability, often resulting in multilingual datasets initially created or translated from English, such as MLQA (Lewis et al., 2020) and XNLI (Conneau et al., 2018). This translation-based approach introduces biases and frequently overlooks cultural nuances, especially in morphologically diverse languages (Anik et al., 2025). To address the scarcity of reliable resources for languages like Bangla, several manually curated datasets have been developed. These include BanglaRQA (Ekram et al., 2022), BenQA (Shafayat et al., 2024a), and BanglaQUAD (Rony et al., 2024), which provide extensive question-answer pairs across various categories. Additionally, NOIR-BETTIK (Aurpa et al., 2025) specifically targets

reading comprehension tasks with multiple-choice questions sourced from authentic Bangla texts, enhancing both NLP research and educational applications. Further efforts have led to the development of multilingual benchmarks that assess the knowledge and reasoning capabilities of LLMs. Prominent examples include CommonsenseQA (Talmor et al., 2019), CosmosQA (Huang et al., 2019), and X-COPA (Ponti et al., 2020), which particularly support Indic languages through human translation. Studies such as BertaQA (Etxaniz et al., 2024; Sengupta et al., 2024) also evaluate LLM generalization abilities in low-resource QA contexts.

Dataset	Domains	Type	Difficulty
BanglaRQA	STEM	Extractive QA	✗
BenQA	STEM	Open-domain QA	✗
BanglaQUAD	STEM	Open-domain QA	✗
Noirbettik	STEM	MCQ (Single Answer)	✗
SOMAJGYAAN	Non-STEM	MCQ (Single & Multi-hop)	✓(5 levels)

Table 1: Comparison of existing Bangla QA datasets with SOMAJGYAAN .

It is worth noting that the aforementioned datasets primarily emphasize STEM-related tasks, whereas our work focuses on ***non-STEM*** domains to assess the pretrained knowledge of LLMs in Bangla. Despite these advances, LLMs still struggle with language-specific challenges, particularly idiomatic expressions and culturally nuanced content in low-resource languages such as Bangla. This limitation is primarily due to the English-centric nature of training datasets, as demonstrated by (Khoshtab et al., 2025) and (Liu et al., 2024). Consequently, current multilingual models often default to Western cultural norms, lacking depth in local cultural understanding unless explicitly fine-tuned (Hasan et al., 2025; Sadhu et al., 2025; Sultana et al., 2025). However, targeted fine-tuning on culturally specific datasets significantly improves performance, as shown by TigerLLM (Raihan and Zampieri, 2025) and CultureLLM (LI et al., 2024).

3 SOMAJGYAAN : Dataset Creation

In this section, we discuss details about our dataset collection and annotation process. The whole pipeline is depicted in Figure 2.

Data Collection. Our dataset is primarily sourced from Bangladesh and Global Studies books of the National Curriculum and Panjeree guide books for classes VI to X. These books are the standard source of information about Bangladesh, ranging in multiple categories. As the books were available in a PDF version and lacked any structured version of the questions, we created the dataset from scratch. Prior to data collection, formal permission was obtained from the relevant authorities to use these materials. In this dataset, we aimed to develop a comprehensive multiple-choice question-answering dataset that covers an extensive range of essential subjects about Bangladesh. These include history, geography, politics, economics, social studies, law, and miscellaneous subjects. The dataset contains factual and causal-type questions about Bangladesh. Furthermore, we added both single-hop and multi-hop questions to increase the difficulty and depth of the dataset. In the final dataset, we got 4234 samples containing seven categories and classified across five levels of difficulty.

Dataset Filtering. We primarily filtered out invalid or inconsistent Bangla sentences. During preprocessing, we discarded any questions or answer choices that were incomplete, grammatically incorrect, or linguistically inconsistent. The dataset was further cleaned by removing code-mixed sentences containing non-Bangla words, which were identified using regular expressions. Additionally, we excluded any questions that required interpreting external content, such as tables, charts, or images, since our focus was on text-based multiple-choice question answering.

Data Annotation. The source data was originally available in PDF format, comprising the National Curriculum and Textbook Board (NCTB) books and supplementary guidebooks. Due to the low-resource language status of Bangla, existing tools for extracting Bangla text from PDFs lack the accuracy and precision necessary for high-quality dataset creation. Consequently, we developed the dataset from the ground up. We employed two expert typist undergraduate students from the Department of Linguistics, who were fluent in both reading and writing Bangla. They were also adept at tools such as Microsoft Excel and collaborative platforms like Google Sheets. A detailed set of guidelines (Appendix B) was provided to ensure consistency. The PDF versions of the books included multiple-choice questions at the end of each chapter, each with answer options and correct answers. The typists manually extracted and transcribed these questions to create the dataset. They formatted the content into CSV files, placing each question, its options, and the correct answer

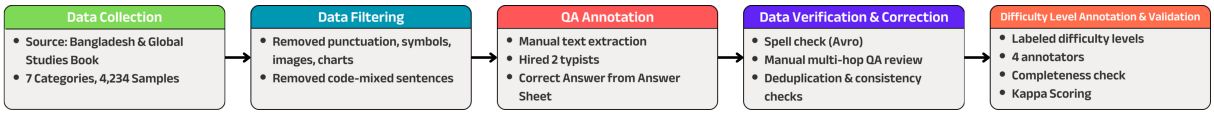


Figure 2: SOMAJGYAAN : Dataset creation pipeline

in separate columns. This process was completed within a one-month timeframe. Throughout this period, any ambiguities or issues encountered were resolved immediately to prevent workflow disruptions. Given the potential unfamiliarity of the typists with specialized software, we did not provide any annotation tools. Each typist received fair compensation for their contributions at a rate of 2TK BDT per question. After the annotation process, we applied a normalization step. Only clean and consistent Bangla texts were retained in the final dataset for training and evaluation purposes.

Data Verification and Correction. The dataset was verified using the Avro Keyboard and Bangla Spell Checker[1] to ensure accurate spelling and the use of valid Bangla words. In addition, a manual review was conducted to ensure that each multi-hop question was contextually consistent with its corresponding passage. To further improve quality, duplicate entries were removed to eliminate redundancy, and question-answer pairs were carefully examined for logical consistency and correctness.

Difficulty Level Annotation. We hired four annotators to label the difficulty level of the data. They were graduate students from the Department of Linguistics and the Department of Computer Science and Engineering. All annotators were native Bangla speakers with experience in reading, writing, and understanding Bangla content. Each annotator received the dataset along with a detailed annotation guideline (Appendix C) to help maintain consistency. Every data sample was annotated independently by four annotators using a difficulty scale from 1 to 5. If any confusion or disagreement came up during the process, it was quickly resolved by the domain expert to ensure consistency and quality. The annotators received fair compensation for their work at a rate of 0.5 TK BDT per question.

Inter-annotators' agreement. As mentioned earlier, for the difficulty score, each data sample was annotated by four annotators to ensure consistency and capture a shared perspective. To validate the data and assess the quality of the annotations, we calculated the inter-annotator agreement score. We used Fleiss's Kappa score (Fleiss, 1971) as our measure of inter-annotator agreement. The agreement scores for each label in the categories are reported in Table 2. From the table, we observed that no agreement score was below 0.62. According to (Islam et al., 2021; Fleiss, 1971), an inter-annotator agreement score of 0.62 or higher, an average agreement score of 0.71, when considering four annotators, indicates strong agreement across the dataset.

[1]https://www.omicronlab.com/avro-keyboard.html

Category	Kappa(κ) Score
History	0.70
Economics	0.62
Geography	0.86
Politics	0.68
Social	0.72
Law	0.64
Misc	0.78
Avg	0.71

Table 2: Inter-Annotator Agreement Score

4 Data Analysis

In Table 3, we present a statistical summary of SOMAJGYAAN. The dataset contains 4234 samples, spanning over 12k unique options. We split the dataset into a training and test dataset with a ratio of 85:15. To ensure balanced representation across categories, we applied category-wise group K-fold techniques for dataset splitting.

Figure 1 shows the category-wise distribution of samples, with History, Economics, and Geography comprising the majority. Figure 12a presents the average lengths of questions, answers, and options for each category. From this, it is evident that questions related to the Law category tend to be longer, while Geography-related questions are generally shorter compared to other categories. Table 3 illustrates the distribution of difficulty scores. Most samples fall within the 1-3 difficulty range. Interestingly, questions with a difficulty score of 1 are fewer than those with difficulty scores of 2 or 3 in most cases. Samples with higher difficulty scores are less common, as there is a correlation

General Statistics		
Samples		4234
Questions		4234
Unique Options		12640
Splits		
Train		3600
Test		634
Difficulty Level		
Level 1		554
Level 2		2592
Level 3		369
Level 4		686
Level 5		33
WH-words		5
Categories/Types		7
Q&A Statistics	**Q**	**A**
Mean word count	19.25	2.15
Max word count	83	10
Min word count	6	1

Table 3: Dataset statistics of SOMAJGYAAN Dataset.

between multi-hop questions and higher difficulty levels. Typically, multi-hop questions are assigned higher difficulty scores.

Figure 12b displays the category-wise count of multi-hop and normal questions, with the Law category having the fewest multi-hop questions. Lastly, Figure 13 shows the distribution of WH-word-based questions, indicating the count of samples in each category relative to WH-words.

5 Experiment Setup

In this study, we design two types of experiments: i) LLM Prompting, which includes zero-shot, few-shot, and CoT prompting, and ii) LoRA Fine-tuning

5.1 LLM Prompting

Prior research on Bangla and multilingual settings (Haider et al., 2025; Fahim et al., 2024; Ahmed et al., 2024) reports that LLM performance varies with prompting strategies. Building on these insights, we explore multiple prompt variations in our dataset to examine their impact on model behavior. In the prompting setting, we evaluate a selection of open-source models along with two closed-source models. The open-source models can be categorized into two main types: i) Monolingual (English-centric) and ii) Multilingual models. For the monolingual open-source models, we consider Vicuna-7B (Chiang et al., 2023) RWKV6-7B (Peng et al., 2024). We also consider TituLLMs (Nahin et al., 2025), which is trained on a huge Bangla corpus. For the multilingual open-source models, we evaluate Llama3.1-8B (Grattafiori et al., 2024) , Qwen2.5-7B(Yang et al., 2024) , Mistral-7B(Jiang, 2024) , DeepSeek-R1-Distill-Qwen-14B (Guo et al., 2025) , Phi-4 (Abdin et al., 2024) . In addition to the open-source models, we also consider the proprietary Gemini-2.0-Flash (Team et al., 2023), Claude 3.7 Sonnet, Grok, GPT-4o, and GPT-4o-mini models (Hurst et al., 2024). For the GPT family, we specifically chose GPT-4o-mini due to its cost-effectiveness compared to other versions. For prompting, we mainly consider the zero-shot prompting. To investigate the effect of different prompt techniques, we also consider Few-shot prompting and Chain-of-Thought (CoT). The details about those prompting techniques are detailed in Appendix D.

5.2 LoRA Fine-Tuning

Wwe aim to explore the performance of LLMs after fine-tuning with context. We apply Low-Rank Adaptation (LoRA) (Hu et al., 2021) to fine-tuning. Instead of directly modifying all the pre-trained parameters, LoRA utilizes a low-rank matrix. This matrix requires significantly fewer parameters to represent the task-specific adaptations. If W represents the frozen parameters of the pre-trained model, LoRA introduces a low-rank update $\Delta W = A \times B^T$, where A $(d \times r)$ and B $(r \times d)$ are trainable matrices with a much smaller rank r compared to d. These matrices capture task-specific adjustments with fewer parameters. The updated weights W' are the sum of the original weights W and the fine-tuned update ΔW: $W' = W + \Delta W = W + AB^T$

Why do we consider LoRA Finetuning? Our proposed SOMAJGYAAN dataset leverages the context and memorization capabilities of LLMs. One might question why we consider LoRA fine-tuning. Firstly, we aim to investigate the performance improvements that fine-tuning can bring. Additionally, LoRA fine-tuning avoids the issues of catastrophic forgetting while incorporating the necessary downstream knowledge, all while preserving the original model's knowledge. In our dataset, answering questions from certain categories requires establishing relationships between the questions and their options, which is why we explore LoRA fine-tuning. However, we found that to achieve better performance with LLMs, having contextual understanding is more crucial than relying on dataset-

Models	Categories								Difficulty Levels				
	Hist	Eco	Geo	Poli	Social	Law	Misc	Avg	L1	L2	L3	L4	L5
Zero Shot Prompt													
Bangla LLM													
TituLLM-1B	0.15	0.16	0.13	0.16	0.12	0.36	0.12	0.17	0.20	0.18	0.19	0.16	0.14
TituLLM-3B	0.21	0.21	0.25	0.19	0.22	0.21	0.24	0.22	0.20	0.24	0.26	0.23	0.19
Open LLM [Monolingual]													
Vicuna 7B	0.23	0.34	0.25	0.31	0.30	0.46	0.21	0.30	0.34	0.31	0.28	0.29	0.26
RWKV6-7B	0.21	0.26	0.17	0.28	0.16	0.32	0.42	0.26	0.28	0.30	0.24	0.28	0.23
Open LLM [Multilingual]													
Llama-3.1-8B	0.42	0.52	0.38	0.53	0.38	0.43	0.48	0.43	0.49	0.38	0.44	0.46	0.35
Qwen2.5-7B	0.45	0.48	0.32	0.57	0.38	0.46	0.67	0.47	0.46	0.46	0.44	0.52	0.49
Mistral-7B	0.30	0.33	0.31	0.30	0.24	0.25	0.27	0.29	0.27	0.28	0.35	0.29	0.28
DeepSeek-R1-14B	0.51	0.61	0.47	0.62	0.54	0.61	0.59	0.56	0.62	0.60	0.54	0.56	0.47
Phi 4	0.41	0.53	0.42	0.60	0.52	0.61	0.58	0.52	0.48	0.53	0.49	0.49	0.59
Close LLM													
Gemini2.0 Flash	0.59	**0.75**	**0.72**	**0.78**	**0.73**	**0.82**	**0.74**	**0.73**	0.79	0.76	0.71	0.72	0.65
Claude 3.7	0.48	0.70	0.48	0.67	**0.73**	0.77	0.62	0.64	0.68	0.64	0.65	0.62	0.61
Grok	0.58	0.66	0.61	0.65	0.70	0.76	0.68	0.66	0.70	0.62	0.66	0.61	0.71
GPT-4o-mini	0.51	0.66	0.50	0.53	0.59	0.61	0.71	0.59	0.64	0.59	0.61	0.58	0.51
GPT-4o	**0.63**	0.73	0.67	0.64	0.72	0.76	0.67	0.69	0.73	0.65	0.69	0.70	0.66
LoRA Fine Tuning Result													
TituLLM-1B	0.29	0.26	0.31	0.33	0.42	0.29	0.21	0.30	0.40	0.34	0.25	0.28	0.22
TituLLM-3B	0.38	0.33	0.31	0.42	0.44	0.18	0.39	0.35	0.46	0.35	0.37	0.29	0.26
Llama-3.1-8B	0.45	0.57	0.34	0.59	0.32	0.46	0.58	0.47	0.43	0.43	0.43	0.40	0.59
Qwen2.5-7B	0.45	0.51	0.33	0.62	0.38	0.61	0.55	0.49	0.51	0.52	0.47	0.49	0.43
Mistral-7B	0.32	0.33	0.23	0.34	0.26	0.39	0.27	0.31	0.38	0.31	0.27	0.32	0.25
DeepSeek-R1-14B	0.45	0.46	0.44	0.50	0.36	0.54	0.61	0.48	0.52	0.52	0.41	0.42	0.54
Phi 4	0.49	0.57	0.44	0.69	0.52	0.57	0.64	0.56	0.62	0.56	0.51	0.59	0.53

Table 4: The model benchmarking on the test split of the SOMAJGYAAN dataset is reported in terms of accuracy percentage. Here, Hist, Eco, Geo, and Polit stand for History, Economics, Geography, and Politics, respectively.

specific patterns.

The LoRA configuration is set with α and r = 64, a dropout rate of 0.05, a learning rate of $2e^{-4}$, and a batch size of 32. We train all the models for 1 epoch. LLaMA-Factory (Zheng et al., 2024) is used for LoRA fine-tuning. To ensure reproducibility, greedy decoding, with the temperature set to 0 without any sampling mechanism, is used during evaluation.

6 Result Analysis

Bangla Open Source LLMs. TituLLM is a continually pretrained version of the meta-llama/Llama-3.2 architecture with extended about 42K Bangla tokens, fine-tuned on extensive Bangla datasets. TituLLM-1B achieves its best performance in the Law dataset with an accuracy of 0.36, but struggles in other domains, resulting in a lower overall average accuracy of 0.17. On the other hand, TituLLM-3B generally outperforms TituLLM-1B across most categories, yielding a higher average accuracy. However, in the Law dataset specifically, zero-shot prompting results show that TituLLM-1B performs better than TituLLM-3B.

Open Source LLMs. Open-source LLMs are crucial for democratizing AI and enabling transparent, customizable solutions for underrepresented languages like Bangla. Despite these advantages, our benchmarking reveals a persistent performance gap compared to proprietary models. Among open models, *DeepSeek-R1-Distill-14B* achieves the highest average accuracy (0.56), followed by *Phi 4* (0.52), yet both lag behind top closed-source models (see Table 4). Performance tends to be stronger in relatively well-represented domains such as Politics and Economics, while Law and Social categories remain more challenging. Again, open models show higher accuracy on difficulty Level 1 and Level 2 questions but drop on higher levels, indicating limitations in handling complex, multi-step reasoning.

Closed-source LLMs. Closed-source LLMs maintain a clear lead in overall accuracy and robustness, particularly on knowledge-intensive and culturally specific categories. Models such as *Gemini 2.0 Flash* (0.73 avg) and *GPT-4o* (0.69 avg) set the benchmark, outperforming all open-source alternatives across all evaluated domains. Their strength is

most pronounced in Law, Politics, and Economics, where access to richer and more structured training data likely plays a role. These models also perform consistently well across difficulty levels, including Level 4 and Level 5, showing a stronger capacity for handling complex or context-heavy reasoning tasks.

Monolingual vs Multilingual LLMs. This experiment highlights the consistent advantage of multilingual LLMs over monolingual ones when applied to Bangla cultural and general knowledge tasks (see Table 4). Top-performing monolingual models, such as *Vicuna 7B* (0.30) and *RWKV6-7B* (0.26), fall behind their multilingual counterparts across all evaluated metrics. In contrast, models like *DeepSeek-R1-Distill-14B* (0.56), *Phi 4* (0.52), and *Qwen2.5-7B* (0.47) demonstrate stronger generalization across diverse knowledge categories. This trend is particularly visible in high-context domains like Law and Politics, where multilingual models exhibit higher category-wise accuracy. Furthermore, they show greater resilience on Level 3 and above questions, suggesting that their broader training corpus better equips them to handle increased complexity and depth of reasoning.

Does Larger Model Size Do Better? Increasing model size generally correlates with better performance. For example, *DeepSeek-R1-Distill-14B* (0.56 avg) outperforms *Mistral-7B* (0.29 avg), and *GPT-4o* (0.69 avg) surpasses smaller closed models. However, smaller well-trained multilingual models like *Phi 4* (0.52 avg) can achieve competitive accuracy. Larger models excel in complex areas like Law and Politics, where deeper factual recall is necessary, and they show greater stability at higher difficulty levels (L4–L5). In contrast, smaller models often experience significant performance drops in these categories.

Effects of Different Prompting Techniques. This experiment compares zero-shot, few-shot, and Chain-of-Thought (CoT) prompting for three models: *Gemini 2.0 Flash*, *Grok*, and *GPT-4o* on the SOMAJGYAAN dataset (Table 5). On average, CoT prompting provides the highest accuracy for Grok and matches zero-shot for Gemini Flash, while few-shot slightly underperforms across all models. It shows that Gemini Flash remains the most consistent and highest-performing model regardless of prompting strategy, achieving up to 0.73 accuracy. CoT improves performance in Law and Misc categories but shows inconsistent gains in History and Social, and it provides moderate benefits on mid-range difficulty levels (L3–L4) without major improvements at the easier or harder ends.

LoRA Finetuning Result. LoRA fine-tuning proves to be an effective strategy for improving LLM performance, particularly in resource-constrained settings. Fine-tuned models consistently outperform their base versions, with improvements ranging from 5–10% in accuracy across different categories. The most significant improvements are observed in *History*, *Geography*, and *Social* categories, where nuanced, context-heavy questions require refined knowledge representations. Notably, *TituLLM-3B* sees substantial performance gains post-fine-tuning, indicating the effectiveness of LoRA in enhancing knowledge recall for low-resource languages. However, fine-tuning does not fully bridge the gap between open and closed models, particularly in categories requiring extensive factual recall like *Law* and *Politics*. This suggests that while LoRA enhances performance, the underlying pretraining data remains a key determinant of LLM capabilities.

LoRA fine-tuning also improves the handling of *multihop* questions, which require reasoning across multiple knowledge points. Baseline models often struggle with multihop QA due to their reliance on surface-level pattern matching rather than deeper contextual reasoning. After fine-tuning, models show noticeable improvements in answering such questions, particularly in *History* and *Geography*, where linking multiple facts is essential. However, performance in complex multihop legal and policy-related questions remains relatively low, suggesting that additional specialized fine-tuning or retrieval-augmented generation techniques might be required. Despite these limitations, the observed improvements reinforce LoRA's effectiveness in refining an LLM's ability to engage with structured reasoning tasks.

Multihop vs Normal QA. We compare model performance on multihop and normal QA tasks across seven categories (Figure 3). On average, normal QA yields slightly better performance for most models; however, certain models like *GPT-4o* and *Qwen2.5-7B* show competitive or even better results in multihop settings. Interestingly, models tend to perform better on multihop questions in categories such as *Economics* and *Misc*, while *Law*

Models	Categories								Difficulty Levels				
	Hist	Eco	Geo	Poli	Social	Law	Misc	Avg	L1	L2	L3	L4	L5
Zero Shot Prompt													
Gemini2.0 Flash	0.59	0.75	0.72	0.78	0.73	0.82	0.74	0.73	0.79	0.76	0.71	0.72	0.65
Grok	0.58	0.66	0.61	0.65	0.70	0.76	0.68	0.66	0.70	0.62	0.66	0.61	0.71
GPT-4o	0.63	0.73	0.67	0.64	0.72	0.76	0.67	0.69	0.73	0.65	0.69	0.70	0.66
Few Shot Prompt													
Gemini2.0 Flash	0.58	0.73	0.74	0.76	0.71	0.80	0.75	0.72	0.81	0.75	0.71	0.72	0.59
Grok	0.55	0.62	0.57	0.66	0.72	0.71	0.70	0.65	0.71	0.57	0.63	0.68	0.68
GPT-4o	0.65	0.67	0.66	0.64	0.73	0.74	0.69	0.68	0.74	0.63	0.70	0.68	0.66
Chain of Thought Prompt													
Gemini2.0 Flash	0.57	0.74	0.72	0.75	0.68	0.86	0.76	0.73	0.78	0.76	0.74	0.72	0.63
Grok	0.55	0.67	0.58	0.64	0.66	0.73	0.74	0.65	0.72	0.63	0.62	0.67	0.62
GPT-4o	0.62	0.71	0.72	0.64	0.67	0.79	0.65	0.68	0.69	0.63	0.71	0.71	0.67

Table 5: The effect of different prompt techniques on the test split of the SOMAJGYAAN dataset is reported in terms of accuracy percentage. We consider Gemini, Grok, and GPT-4o for this experimentation. Here, Hist, Eco, Geo, and Polit stand for History, Economics, Geography, and Politics, respectively.

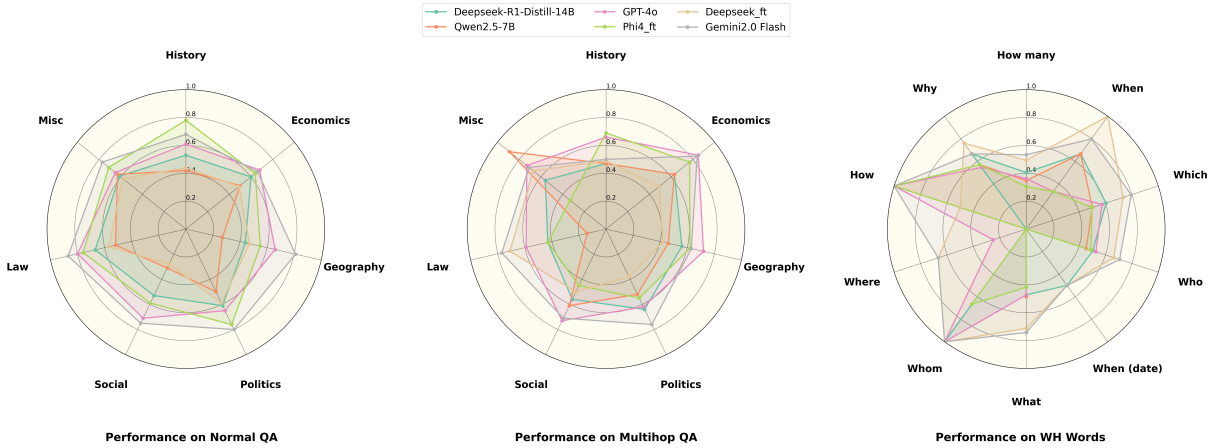


Figure 3: Error Analysis on SOMAJGYAAN Dataset.

remains consistently difficult across both formats. Among different difficulty levels, multihop QA reveals greater performance drops at higher levels (L4–L5), reflecting the added complexity in reasoning. Among open models, *DeepSeek-R1-Distill-14B* is most consistent across both QA types, while *GPT-4o* performs best overall among closed models. For fine-tuned models, *Phi 4* shows balanced performance, though gains in multihop QA are less pronounced. For detailed analysis, refer to Appendix Table 6

Performance on Question Types. We analyze model behavior across different WH-word question types to understand their strengths and weaknesses (Figure 3). Most models perform reliably on questions beginning with "*Whom*," "*When*," and "*How*," with *Gemini 2.0 Flash*, *GPT4o*, and *Phi 4 ft* achieving perfect or near-perfect accuracy. In contrast, "*When (date)*" and "*Where*" questions are more challenging, especially for multilingual open models like *Qwen2.5 7B*, which scored zero on both. While *Gemini 2.0 Flash* consistently performs best across nearly all WH types, *GPT4o* also shows strong generalization across factual (e.g., *What*) and reasoning-based (e.g., *Why*) forms.

Error Analysis. From the (Figure 14), we observe three distinct examples. In the first case, all open-source models interpret the question literally in the zero-shot scenario, and even after fine-tuning, they continue to predict the literal meaning, failing to capture the intended context. In the second example, none of the models predict the correct answer in the zero-shot setting. After fine-tuning, some models change their responses, but they still fail to arrive at the correct answer due to a lack of domain knowledge. In the final example, all models, ex-

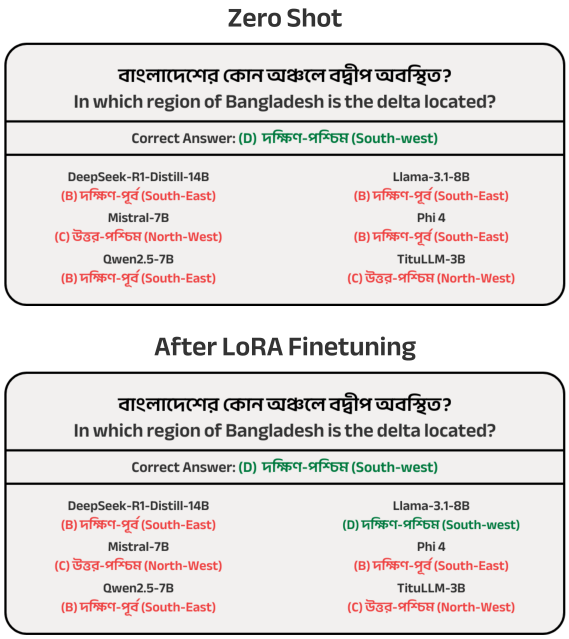


Figure 4: Model predictions before and after fine-tuning. In this example, Llama-3.1-8B correctly answered after fine-tuning.

cept Deepseek-R1-Distill-14B, predict the correct answer initially. The Deepseek model does not provide any prediction in this case. After fine-tuning, some models that previously answered correctly begin to produce incorrect responses.

7 Conclusion

We introduce SOMAJGYAAN, a multiple-choice question-answering dataset designed to evaluate LLMs on their understanding of Bangla culture, social knowledge, and adaptability to low-resource languages. The dataset spans diverse categories and is meticulously annotated for fairness and consistency. Experiments reveal that while fine-tuning LLMs on Bangla data improves their performance, closed-source models still outperform open-source ones due to limited Bangla knowledge in pre-training corpora. This underscores the need for more high-quality datasets and better adaptation techniques for low-resource languages.

Limitations

The findings from this study reveal several key limitations. First, open-source Bangla models, such as TituLLM-1B and TituLLM-3B, continue to lag behind larger multilingual and proprietary models in terms of overall accuracy, especially in complex domains like Law and Social Studies. While fine-tuning helps in improving responses for simpler questions, it has a limited impact on higher-difficulty, multi-step reasoning tasks. In some cases, like DeepSeek-R1-Distill-14B, fine-tuning even leads to a decline in performance, suggesting potential overfitting or loss of general reasoning capability. Furthermore, models consistently struggle in categories that require richer domain knowledge, indicating gaps in training data coverage.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Fahim Ahmed, Md Fahim, Md Ashraful Amin, Amin Ahsan Ali, and AKM Rahman. 2024. Improving the performance of transformer-based models over classical baselines in multiple transliterated languages. In *ECAI 2024*, pages 4043–4050. IOS Press.

Mahfuz Ahmed Anik, Abdur Rahman, Azmine Toushik Wasi, and Md Manjurul Ahsan. 2025. Preserving Cultural Identity with Context-Aware Translation Through Multi-Agent AI Systems. In *NAACL 2025 Workshop on Language Models for Underserved Communities*.

Tanjim Taharat Aurpa, Md. Shahriar Hossain Apu, Farzana Akter, Richita Khandakar Rifat, and Md. Ahsan Habib. 2025. Noirbettik: A reading comprehension based multiple choice question answering dataset in bangla language. *Data in Brief*, 59:111395.

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.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Common Crawl. 2025. Common crawl language statistics. Accessed March 29, 2025.

Syed Mohammed Sartaj Ekram, Adham Arik Rahman, Md. Sajid Altaf, Mohammed Saidul Islam, Mehrab Mustafy Rahman, Md Mezbaur Rahman, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2022. BanglaRQA: A Benchmark Dataset for Under-resourced Bangla Language Reading Comprehension-based Question Answering with Diverse Question-Answer Types. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2518–2532. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Bertaqa: How much do language models know about local culture? *Advances in Neural Information Processing Systems*, 37:34077–34097.

Md Fahim. 2023. Aambela at blp-2023 task 2: Enhancing banglabert performance for bangla sentiment analysis task with in task pretraining and adversarial weight perturbation. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 317–323.

Md Fahim, Fariha Tanjim Shifat, Fabiha Haider, Deeparghya Dutta Barua, MD Sakib Ul Rahman Sourove, Md Farhan Ishmam, and Md Farhad Alam Bhuiyan. 2024. Banglatlit: A benchmark dataset for back-transliteration of romanized bangla. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14656–14672.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Yuhang Ge. 2024. A survey on lora of large language models. *arXiv preprint arXiv:2407.11046*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Md Sakib Ul Rahman Sourove, Deeparghya Dutta Barua, Md Fahim, and Md Farhad Alam Bhuiyan. 2025. Banth: A multi-label hate speech detection dataset for transliterated bangla. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7217–7236.

Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025. NativQA: Multilingual culturally-aligned natural queries for LLMs.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

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. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Fengqing Jiang. 2024. Identifying and mitigating vulnerabilities in llm-integrated applications. Master's thesis, University of Washington.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. Comparative study of multilingual idioms and similes in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8680–8698, Abu Dhabi, UAE. Association for Computational Linguistics.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. *arXiv preprint*. ArXiv:1910.07475 [cs].

CHENG LI, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. CultureLLM: Incorporating cultural differences into large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems*, 36:43136–43155.

Rohin Manvi, Nikhil Madan, Wanrong Zhu, Amanpreet Singh, and Diyi Yang. 2024. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Shahriar Kabir Nahin, Rabindra Nath Nandi, Sagor Sarker, Quazi Sarwar Muhtaseem, Md Kowsher, Apu Chandraw Shill, Md Ibrahim, Mehadi Hasan Menon, Tareq Al Muntasir, and Firoj Alam. 2025. Titullms: A family of bangla llms with comprehensive benchmarking. *arXiv preprint arXiv:2502.11187*.

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Xingjian Du, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, Kranthi Kiran GV, Jan Kocoń, Bartłomiej Koptyra, Satyapriya Krishna, Ronald McClelland Jr., Jiaju Lin, Niklas Muennighoff, Fares Obeid, Atsushi Saito, Guangyu Song, Haoqin Tu, Cahya Wirawan, Stanisław Woźniak, Ruichong Zhang, Bingchen Zhao, Qihang Zhao, Peng Zhou, Jian Zhu, and Rui-Jie Zhu. 2024. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *Preprint*, arXiv:2404.05892.

Alberto Poli, Ilya Sergey, and Greg Morrisett. 2024. Scalable verification of mitigations for memory safety. *arXiv preprint arXiv:2404.12464*.

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. Association for Computational Linguistics.

Nishat Raihan and Marcos Zampieri. 2025. Tigerllm – a family of bangla large language models. *Preprint*, arXiv:2503.10995.

Md Rashad Al Hasan Rony, Sudipto Kumar Shaha, Rakib Al Hasan, Sumon Kanti Dey, Amzad Hossain Rafi, Amzad Hossain Rafi, Ashraf Hasan Sirajee, and Jens Lehmann. 2024. BanglaQuAD: A Bengali Open-domain Question Answering Dataset. *arXiv preprint*. ArXiv:2410.10229 [cs].

Jayanta Sadhu, Maneesha Rani Saha, and Rifat Shahriyar. 2025. Social bias in large language models for Bangla: An empirical study on gender and religious bias. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 204–218, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Saptarshi Sengupta, Wenpeng Yin, Preslav Nakov, Shreya Ghosh, and Suhang Wang. 2024. Exploring language model generalization in low-resource extractive qa. *Preprint*, arXiv:2409.18446.

Sheikh Shafayat, H Hasan, Minhajur Mahim, Rifki Putri, James Thorne, and Alice Oh. 2024a. BEnQA: A Question Answering Benchmark for Bengali and English. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1158–1177. Association for Computational Linguistics.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. 2024b. Multi-FAct: Assessing factuality of multilingual llms using factscore. *arXiv preprint*, abs/2402.18045. [Preprint].

Nusrat Sultana, Rumana Yasmin, Bijon Mallik, and Mohammad Shorif Uddin. 2025. Onubad: A comprehensive dataset for automated conversion of bangla regional dialects into standard bengali dialect. *Data in Brief*, 58:111276.

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. Association for Computational Linguistics.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

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*.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd*

ACL, Bangkok, Thailand. Association for Computational Linguistics.

A Category Wise Examples of Questions and Answers from Dataset

The whole dataset is divided into 7 categories: **History, Geography, Politics, Economics, Social, Law, and Miscellaneous**. Each of these categories contains multiple-choice questions (MCQs) covering factual and reasoning-based knowledge about Bangladesh. All the 7 categories of QA are also divided into two parts: **Normal QA and Multihop QA**.

- **Normal QA** consists of a **question with four options**, where only one of the four options is correct. These questions primarily test direct factual knowledge and basic understanding of the subject matter.
- **Multihop QA** is structured differently to evaluate the model's ability to perform **multi-step reasoning**. Each multihop question consists of **a question and three options**, where any one option can be correct, any two options can be correct, or all three options can be correct.

The **Normal QA** and **Multihop QA** examples are organized by category and displayed side by side below for easy comparison.

History. This category focuses on historical events, figures, dates, and terminology, particularly related to Bangladesh (e.g., Liberation War, Mughal Empire) and global history (e.g., World Wars, ancient civilizations). Includes questions on battles, treaties, and cultural milestones. Figure 5 shows a side-by-side comparison of History-related Normal QA and Multihop QA.

Economics. This category contains questions related to economic principles, systems, and Bangladesh's economy. Topics include GDP, inflation, banking, agriculture, trade policies, and developmental challenges like poverty and unemployment. Figure 6 shows an example of Normal QA and a Multihop QA under the economics category.

Geography. This category deals with questions related to physical and human geography, including landforms (rivers, mountains), climate, and regional specifics of Bangladesh (e.g., Sundarbans, Padma River) and global phenomena (e.g., ocean currents, tectonic plates). Figure 7 shows examples under the Geography category.

Politics. This category covers the questions of centers on governance, political systems, elections, and government functions in Bangladesh and globally. This category includes questions on parliamentary procedures, political parties, and international relations (e.g., UN, SAARC). One example of Normal QA and one of Multihop is shown in Figure 8.

Social. Figure 9 shows examples of this category. This category addresses questions on societal structures, cultural practices, family dynamics, and community issues. Topics include social welfare, gender roles, education, and challenges like child labor or urbanization effects.

Law. This category includes questions on legal frameworks, constitutional provisions, judicial processes, and rights in Bangladesh. Examples include the Constitution, election laws, and landmark court cases. One example of Normal QA and One from Multihop under this category are shown in Figure 10

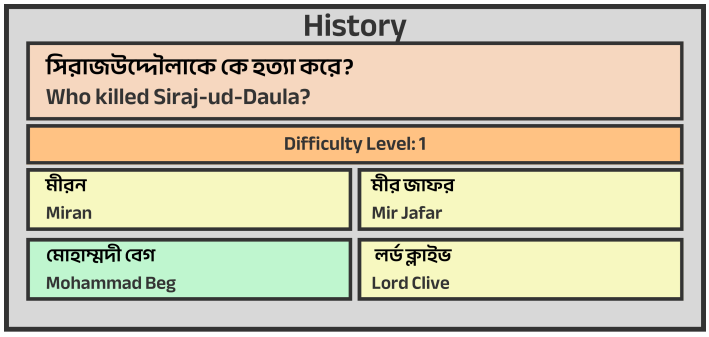
Miscellaneous. This category covers diverse general knowledge topics that are not tied to specific categories. Examples include science (e.g., gas composition, environmental issues), arts (e.g., famous paintings), and notable personalities. Examples are shown in Figure 11

B Guidelines for Typists

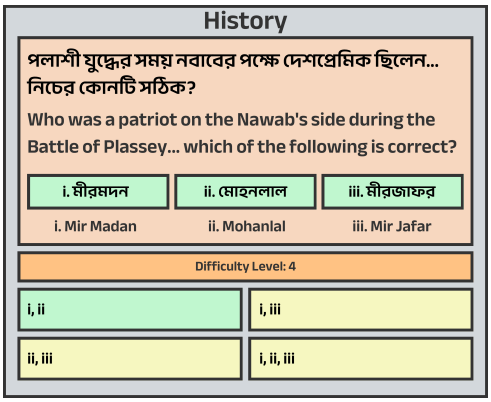
Document Structure. The dataset should consist of seven distinct columns, each serving a specific purpose. The **Question** column will contain the multiple-choice question written in Bangla script. The four subsequent columns, labeled **Option A**, **Option B**, **Option C**, and **Option D**, will present the possible answer choices, also in Bangla. The **Answer** column should explicitly indicate the correct answer by referencing the corresponding option (e.g., A, B, C, or D). Lastly, the **Category** column will classify the question under an appropriate subject or domain, ensuring clear organization of information.

Question Formatting. All questions must be typed in Bangla script with proper spelling and grammar. It is important to maintain accurate punctuation to ensure clarity. Additionally, non-textual questions, such as those requiring images, tables, or charts, should be excluded from the dataset.

Answer Choices Formatting. Each multiple-choice question should have four answer options, labeled as **Option A**, **Option B**, **Option C**, and

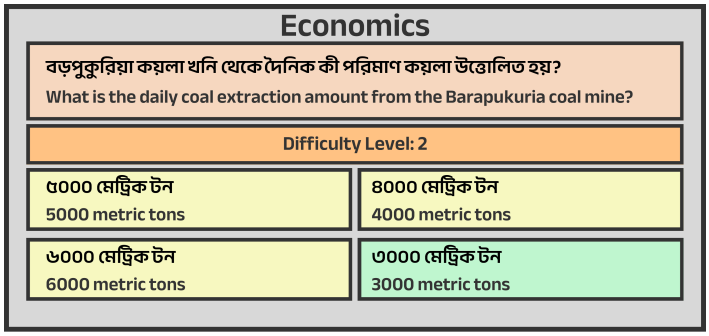


(a) Normal QA - History Question

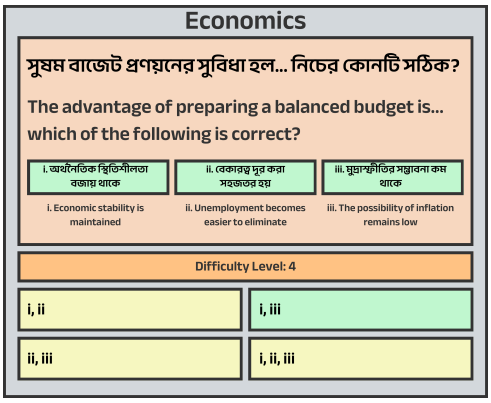


(b) Multihop QA - History Question

Figure 5: Comparison of Normal QA and Multihop QA for History Questions



(a) Normal QA - Economy Question



(b) Multihop QA - Economy Question

Figure 6: Comparison of Normal QA and Multihop QA for Economy Questions

Option D. The options must be consistently formatted to ensure clarity and alignment. The correct answer should be clearly indicated in the **Answer** column using its corresponding option label (A, B, C, or D). Proper spacing between each option is recommended.

Numerical and Symbolic Representation. All numerical values should be represented using Bangla numerals instead of English numerals. Similarly, Bangla punctuation marks should be used, such as the Bangla full stop instead of the English period.

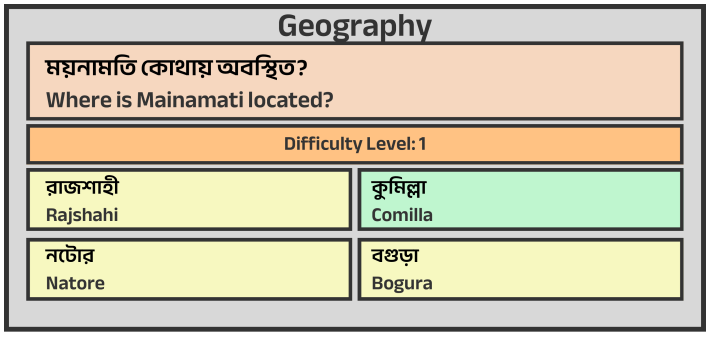
Spacing and Formatting Consistency. The dataset should use a consistent font such as **SolaimanLipi**, **Siyam Rupali**, or any Unicode-supported Bangla font. Uniform font size should be maintained throughout the document to ensure visual coherence. Proper spacing must be applied between questions, options, and answers, promoting clear readability and a professional appearance.

Data Verification and Proofreading. After typing, all questions, options, and answers should be cross-checked for spelling, grammar, and formatting accuracy. Proofreading is essential to identify and correct any errors. Additionally, it is important to confirm that no question or answer is missing. Ensure that the category information is appropriately assigned.

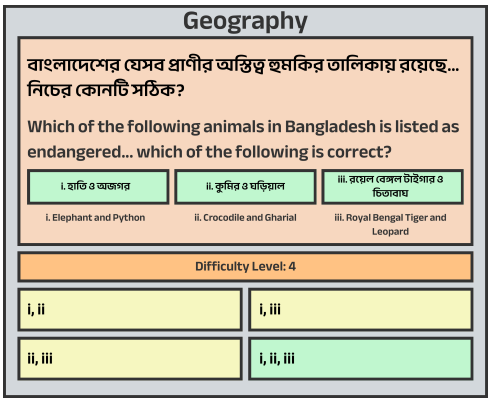
File Format and Naming Convention. The completed dataset should be saved in CSV or Excel format. The file name should follow a clear and structured format. For instance, a suitable filename could be `Bangla_Question_Dataset_2025.csv`, which specifies the dataset's content, language, and creation year.

C Guidelines for Difficulty Level Annotation

This section provides clear guidelines for annotators to follow when assigning difficulty levels to sentences. Each level is defined based on the complexity of reasoning and contextual understanding required to derive the correct answer. Annotators should carefully evaluate the sentence using these

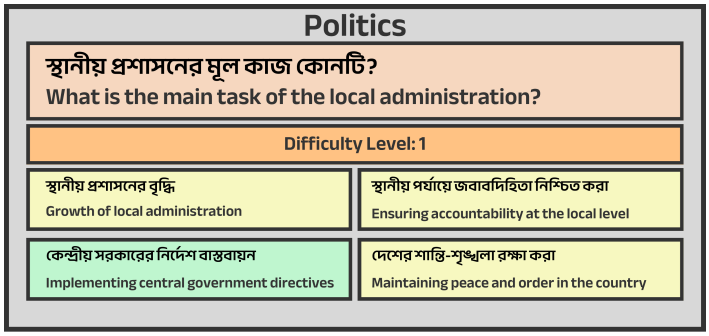


(a) Normal QA - Geography Question

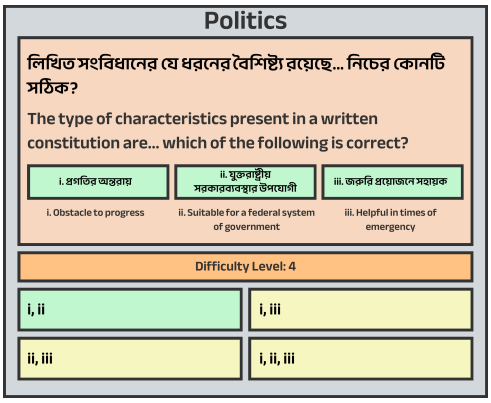


(b) Multihop QA - Geography Question

Figure 7: Comparison of Normal QA and Multihop QA for Geography Questions



(a) Normal QA - Politics Question



(b) Multihop QA - Politics Question

Figure 8: Comparison of Normal QA and Multihop QA for Politics Questions

criteria.

Level 1: Factual (Easy)

- Requires a direct fact-based answer.
- No reasoning or contextual understanding is needed.
- The correct answer is explicitly present in the sentence.

Level 2: Reasoning-Based (Moderate)

- Requires logical or deductive reasoning.
- Information is known, but the answer is not immediately obvious.
- Annotators must apply simple reasoning to reach the answer.

Level 3: Fact + Reasoning (Moderate-High)

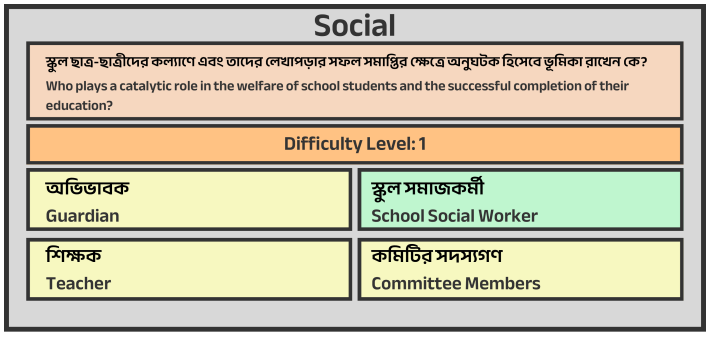
- Requires both factual knowledge and reasoning.
- Involves combining known facts with logical thinking.

Level 4: Context-Based Single-Hop (High)

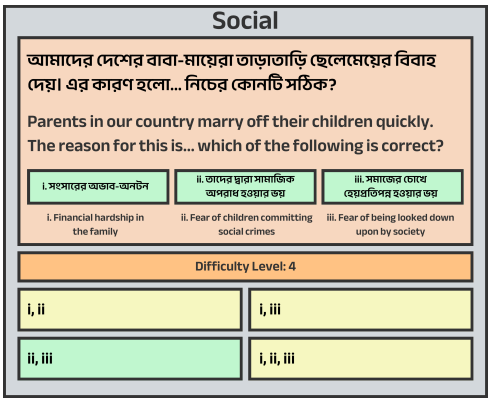
- Requires understanding the context before answering.
- Involves retrieving information using a single step of reasoning.
- The sentence may reference external information, but a direct inference can be made.

Level 5: Context-Based Multi-Hop (Very High)

- Requires understanding multiple pieces of context and connecting them.
- Involves multiple reasoning steps and deeper knowledge synthesis.
- The sentence may demand cross-referencing different facts to find the correct answer.

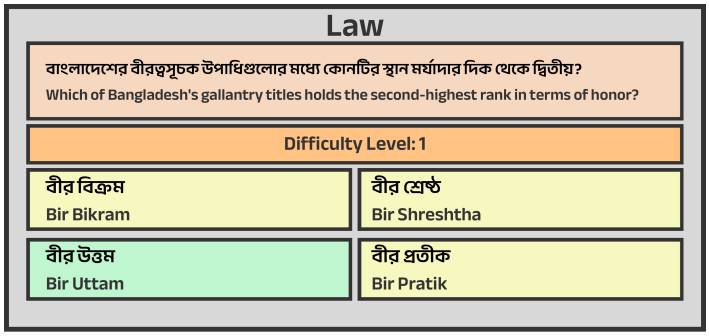


(a) Normal QA - Social Question

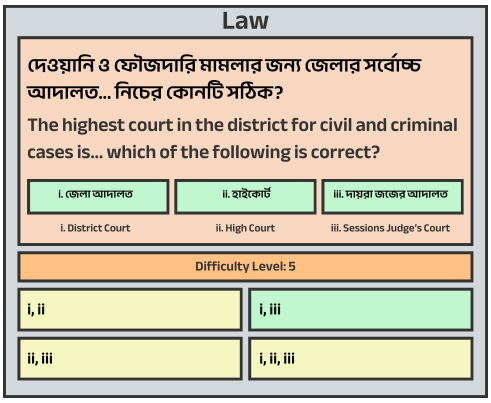


(b) Multihop QA - Social Question

Figure 9: Comparison of Normal QA and Multihop QA for Social Questions



(a) Normal QA - Law Question



(b) Multihop QA - Law Question

Figure 10: Comparison of Normal QA and Multihop QA for Law Questions

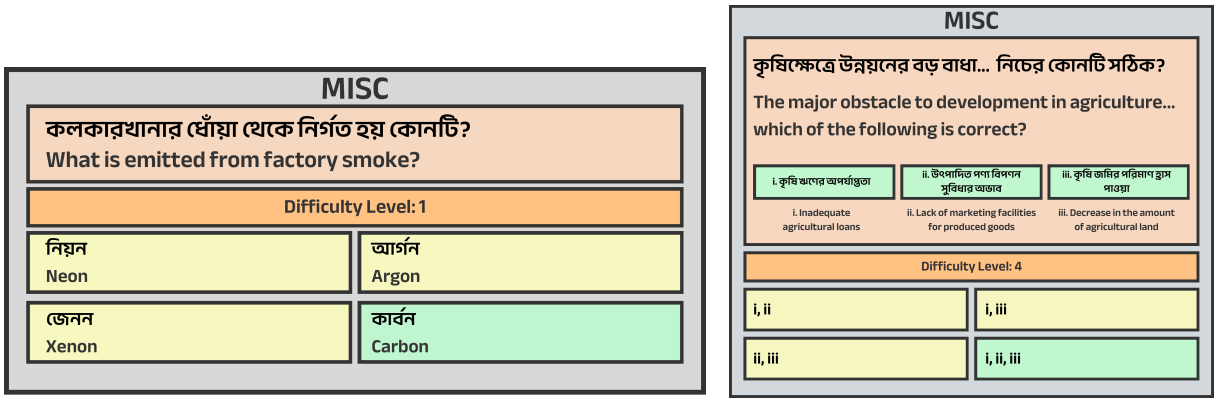
D Experimentation on Different Prompt Style.

Few-shot Prompting. This prompting technique involves providing a language model with a limited number of examples to guide its responses to specific tasks (Brown et al., 2020). This technique helps models understand task structure and expected output, enhancing performance, especially when extensive labeled data is unavailable (Ma et al., 2023). In our evaluation, we incorporated two standard multiple-choice questions (MCQ) and two multi-hop MCQ into the prompts for the Grok and GPT-4o models. This approach aimed to assess the models' ability to handle complex reasoning tasks by leveraging the provided examples. (Fahim, 2023)

CoT Prompting. In our evaluation, we also used Chain-of-Thought (CoT) prompting, a technique that guides Large Language Models (LLMs) to generate intermediate reasoning steps (Wei et al., 2022), thereby enhancing their performance on complex tasks. This approach has been shown to improve reasoning abilities in LLMs by encouraging step-by-step thought processes before arriving at a final answer. To tailor the CoT prompting to our focus on Bangla culture, history, geography, law, and social knowledge, we designed a structured prompt with the following instructions:

- Analyze the question.
- Recall relevant Bangla cultural, historical, geographical, legal, and social knowledge.
- Eliminate incorrect options.
- Choose the most suitable answer.

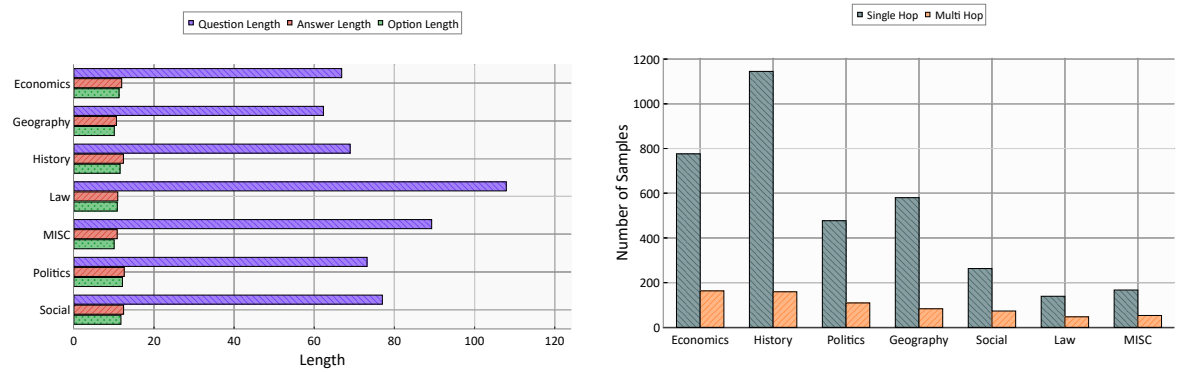
This culturally informed CoT prompting aims to mitigate the inherent biases of LLMs, which often favor Western perspectives due to their predominantly English training data. By explicitly directing the models to consider Bangla-specific contexts, we enhance their ability to provide accurate and culturally relevant responses (Tao et al.,



(a) Normal QA - Miscellaneous Question

(b) Multihop QA - Miscellaneous Question

Figure 11: Comparison of Normal QA and Multihop QA for Miscellaneous Questions



(a) Category-wise Mean Lengths

(b) Multi-hop and Normal Questions per Category

Figure 12: Data Distribution by Category, Difficulty Level, and Question Type

2024). In practice, GPT-4o and Grok model received a multiple-choice question with the above CoT prompt. This methodology allows us to assess the effectiveness of CoT prompting in eliciting culturally aware reasoning from LLMs.

E Result Analysis and Findings

E.1 Difficulty vs Performance

Understanding how different models perform across varying difficulty levels is crucial for assessing their robustness and reliability. Table 4 presents a comparative analysis of multiple language models across five levels of difficulty, using different prompting strategies, including Zero Shot and LoRA fine-tuning. The evaluation is conducted with and without additional context to analyze the effectiveness of contextual information in improving model performance.

From Table 5, we observe that larger models, such as Gemini2.0 Flash, GPT-4o, Grok, and Claude 3.7, tend to exhibit higher performance across all difficulty levels. Among the zero-shot results, GPT-4o achieves an average accuracy of 0.69, making it one of the top-performing models. On the other hand, smaller models like TituLLM-3B struggle with consistently lower scores, indicating the limitations of smaller parameter sizes in handling complex questions.

When comparing performance across different difficulty levels, we note that model accuracy generally decreases as difficulty increases. This trend is expected, as more challenging questions often require multi-hop reasoning and deeper knowledge retrieval. For instance, DeepSeek-R1-Distill-14B performs well in lower difficulty levels but shows a significant drop in performance at Level 5.

E.2 Multihop vs Normal QA

To evaluate how well different models handle Normal QA versus Multihop QA, we analyzed their accuracy across various categories. The results indicate distinct patterns in performance:

Overall Performance Difference. Most models exhibit higher accuracy in **Normal QA** compared to **Multihop**, suggesting that they struggle with

multi-step reasoning tasks. This trend is expected, as reasoning over multiple pieces of information requires stronger contextual understanding and logical inference capabilities.

Figure 13: WH terms distribution over Seven Categories

Model-Specific Trends. Some models, such as GPT-4o and Deepseek-R1-Distill-14B, demonstrate relatively strong performance in Multihop QA, indicating better reasoning abilities compared to smaller models like TituLLM-3B and Mistral-7B-v0.3.

Category-Wise Observations:

- *History and Geography:* Performance in these categories tends to be lower for **Multihop**, likely due to the complexity of linking historical events or geographic relationships.
- *Politics and Law:* These domains also show a significant performance gap, reflecting the challenge of understanding legal frameworks and political dynamics that require nuanced reasoning.
- *Miscellaneous:* This category shows a mixed performance, with some models performing better in **Multihop** than in **Normal QA**, possibly due to the presence of pattern-recognizable questions.

E.3 Multihop vs CoT

The table 7 compares the performance of various models on multi-hop questions using two prompting techniques: Zero Shot Prompting and Chain of Thought (CoT) Prompting, across seven categories (History, Economics, Geography, Politics, Social Sciences, Law, and Miscellaneous). Models like Gemini2.0 Flash and GPT-4o perform well in both prompting techniques, with GPT-4o achieving the highest average score in both cases (0.70 for Zero Shot and 0.72 for CoT). On the other hand, Qwen 2.5 and Llama 3.1 show relatively lower and more inconsistent performance, particularly in the Zero Shot setting. Here, GPT-4o performs best overall, showing slight improvement with CoT, especially in Law, while Grok declines, suggesting CoT effectiveness depends on the model's reasoning strength. Economics is the easiest category, whereas History and Geography show variability. Stronger models like GPT-4o benefit from CoT, while weaker ones (e.g., Grok) may not, highlighting that CoT is model-dependent.

E.4 Data Analysis

The distribution of WH terms across the seven categories reveals interesting patterns in the types of questions asked. WH-terms, such as "what," "where," "when," "why," and "how much," are commonly used in questions that require specific information retrieval or clarification. In some categories, like History and Geography, WH-terms are more frequently associated with questions that ask for factual details or explanations. On the other hand, categories like Law may feature a different pattern, with WH-terms being used less often due to the more complex, interpretative nature of questions in this domain. The chart in Figure 13 provides a clear visual representation of how WH terms are distributed across these categories, highlighting the differences in question types and their reliance on such terms.

Models	Categories							
	Hist	**Eco**	**Geo**	**Poli**	**Social**	**Law**	**Misc**	**Avg**
Zero Shot Prompt								
Meta-Llama-3.1-8B	0.43/0.38	0.54/0.54	0.50/0.39	0.52/0.57	0.33/0.28	0.29/0.48	0.44/0.61	0.46/0.45
Mistral-7B-v0,.3	0.37/0.33	0.23/0.25	0.35/0.18	0.38/0.36	0.22/0.25	-/0.43	0.33/0.38	0.31/0.29
Qwen2.5-7B	0.47/0.42	0.63/0.50	0.46/0.27	0.52/0.50	0.61/0.31	0.14/0.52	0.89/0.63	0.53/0.45
Deepseek-R1-Distill-14B	0.47/0.53	0.63/0.60	0.56/0.44	0.64/0.61	0.56/0.53	0.43/0.67	0.56/0.61	0.56/0.55
Microsoft-Phi-4	0.44/0.42	0.49/0.57	0.42/0.47	0.52/0.68	0.44/0.44	0.43/0.57	0.89/0.58	0.49/0.51
TituLLM-3B	0.20/0.30	0.26/0.26	0.20/0.29	0.21/0.24	0.35/0.20	0.29/0.22	0.33/0.35	0.24/0.27
Claude 3.7	0.38/0.53	0.85/0.63	0.40/0.51	0.69/0.67	0.78/0.70	0.68/0.79	0.58/0.64	0.62/0.64
Grok	0.52/0.61	0.78/0.62	0.68/0.58	0.62/0.67	0.61/0.76	0.68/0.78	0.77/0.65	0.67/0.67
GPT 4o	0.66/0.61	0.85/0.68	0.72/0.66	0.62/0.65	0.73/0.71	0.59/0.80	0.73/0.65	0.70/0.68
GPT 4o Mini	0.59/0.47	0.78/0.62	0.48/0.51	0.48/0.56	0.51/0.63	0.32/0.68	0.77/0.69	0.56/0.59
LoRA Fine Tuning								
Meta-Llama-3.1-8B	0.33/0.39	0.54/0.40	0.35/0.43	0.59/0.54	0.44/0.31	0.43/0.48	0.56/0.42	0.45/0.42
Mistral-7B-v0.3	0.55/0.72	0.66/0.47	0.65/0.56	0.62/0.61	0.44/0.50	0.29/0.52	0.22/0.67	0.56/0.60
Qwen2.5-7B	0.31/0.47	0.49/0.53	0.31/0.34	0.52/0.69	0.50/0.38	0.43/0.62	0.44/0.58	0.41/0.50
Deepseek-R1-Distill-14B	0.49/0.44	0.49/0.49	0.42/0.47	0.38/0.59	0.50/0.28	0.71/0.57	0.67/0.58	0.48/0.48
Microsoft-Phi-4	0.69/0.78	0.77/0.64	0.62/0.55	0.55/0.76	0.45/0.59	0.43/0.76	0.33/0.71	0.62/0.69
TituLLM-3B	0.20/0.39	0.23/0.30	0.38/0.29	0.31/0.37	0.17/0.28	0.43/0.19	0.11/0.33	0.25/0.33

Table 6: MultiHop vs Normal QA Performance Analysis. The cell format is *Multihop Result/Normal Result*.

Models	Categories							
	Hist	**Eco**	**Geo**	**Poli**	**Social**	**Law**	**Misc**	**Avg**
Zero Shot Prompt								
Gemini2.0 Flash	0.50	0.84	0.63	0.76	0.71	0.77	0.71	0.70
Grok	0.52	0.78	0.68	0.62	0.61	0.68	0.77	0.67
Claude 3.7	0.38	0.85	0.40	0.69	0.78	0.68	0.58	0.62
GPT-4o	0.66	0.85	0.72	0.62	0.73	0.59	0.73	0.70
GPT-4o Mini	0.59	0.78	0.48	0.48	0.51	0.32	0.77	0.56
Llama 3.1	0.41	0.51	0.63	0.55	0.44	0.29	0.40	0.46
Phi 4	0.51	0.60	0.52	0.62	0.67	0.57	0.80	0.61
Qwen 2.5	0.49	0.40	0.33	0.31	0.33	0.57	0.60	0.43
Chain of Thought Prompt								
Grok	0.52	0.80	0.52	0.59	0.55	0.68	0.75	0.63
GPT-4o	0.66	0.81	0.72	0.62	0.76	0.73	0.77	0.72

Table 7: Performance of different models on multi-hop question based on Zero Shot Prompting and Chain of Thought Prompting

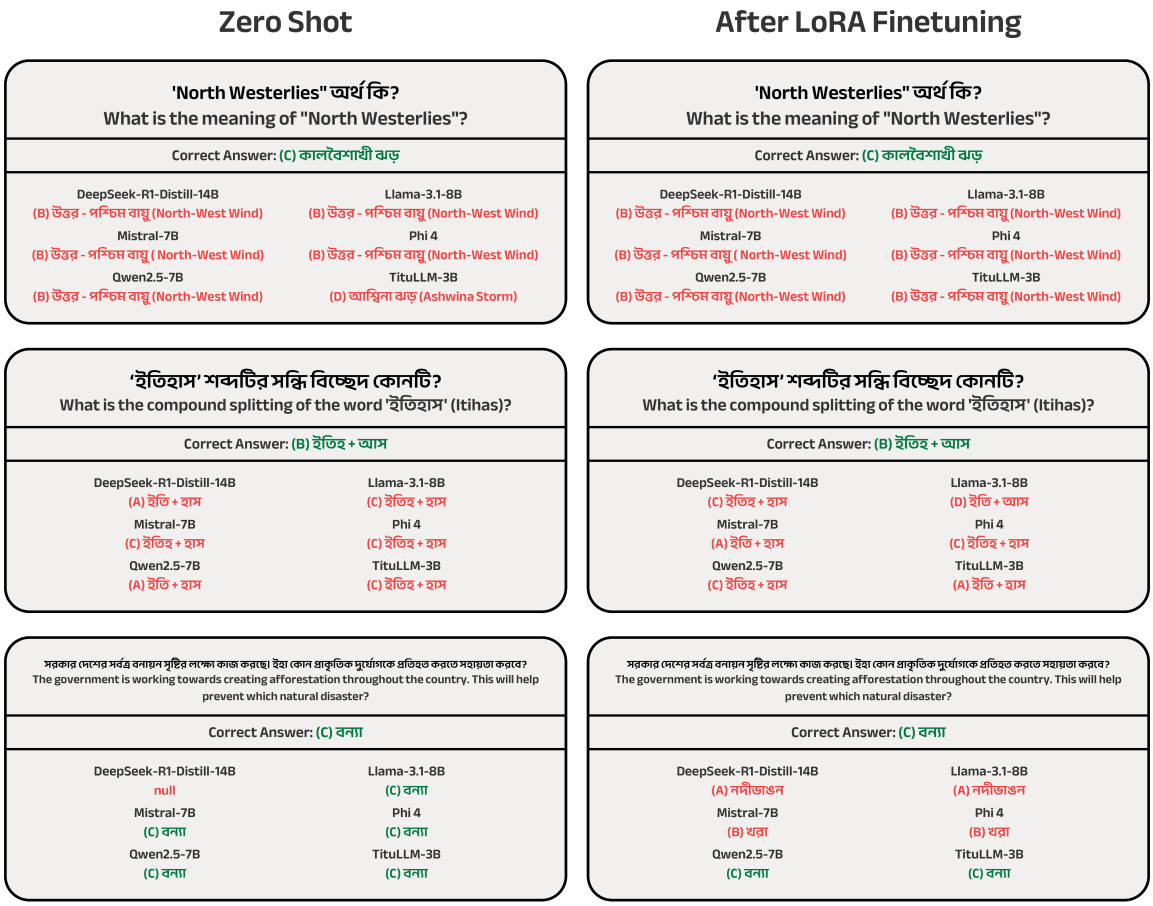


Figure 14: Illustration of model prediction behaviors across three examples, comparing zero-shot and fine-tuned outputs. The first example shows models consistently providing literal interpretations despite fine-tuning. The second highlights the influence of domain knowledge, where fine-tuning leads to varied but still incorrect responses. The third example demonstrates how fine-tuning can sometimes degrade performance, with certain models losing previously correct predictions, while one model (Deepseek-R1-Distill-14B) fails to produce any response in the zero-shot setting.

F Prompting

All the prompts are detailed in this section.

Base Prompt.

Base Prompt Used for Prediction

You are a multilingual AI expert assistant. You will be provided with a question and four options, which may pertain to topics such as Bangla culture, social life, law, and more. The question will fall under one of the following categories: History, Economics, Geography, Politics, Society, Law, or Miscellaneous. The options will be formatted as A, B, C, and D. Your task is to answer the question by selecting the corresponding letter of the correct option: A, B, C, or D.
Question: [Insert text-based question here]
Options:
A) [Option A]
B) [Option B]
C) [Option C]
D) [Option D]

Select the correct option from A, B, C, or D

Few-Shot Prompt.

Few Prompt Used for Prediction

You are a multilingual AI expert assistant. You will be provided with a question and four options, which may pertain to topics such as Bangla culture, social life, law, and more. The question will fall under one of the following categories: History, Economics, Geography, Politics, Society, Law, or Miscellaneous. You will be given either multiple-choice questions or plain questions. You have to answer the question by providing the letter of the option. For example, if you want to answer the question by selecting option 'A', you have to provide 'A' as the answer.
Normal Questions Example:
Example Question 1: [Insert text-based question here]
Options:
A) [Option A]
B) [Option B]
C) [Option C]
D) [Option D]
Answer: C
Multihop Questions Example:
Example Question 2: [Insert text-based question here]
i. [Multihop Options]
ii. [Multihop Options]
iii. [Multihop Options]

Options:
A: i & ii B: i & iii C: ii & iii D: i, ii & iii Answer: D

Chain-of-Thought

Chain of Thought Prompt Used for Prediction

You are a multilingual AI expert assistant. You will be provided with a question and four options, which may pertain to topics such as Bangla culture, social life, law, and more. The question will fall under one of the following categories: History, Economics, Geography, Politics, Society, Law, or Miscellaneous. The options will be formatted as A, B, C, and D. Your task is to answer the question by selecting the corresponding letter of the correct option: A, B, C, or D.
Question: [Insert text-based question here]
Options:
A) [Option A]
B) [Option B]
C) [Option C]
D) [Option D]

Think carefully step-by-step before selecting the final answer:
1. Analyze the question.
2. Recall relevant Bangla Cultural, History, Geography, Law, and Social knowledge.
3. Eliminate incorrect options.
4. Choose the most suitable answer.

Answer the following question by providing only the letters of options A, B, C, or D.