BnMMLU: Measuring Massive Multitask Language Understanding in Bengali

Anonymous ACL submission

Abstract

The Massive Multitask Language Understanding (MMLU) benchmark has been widely used to evaluate language models across various domains. However, existing MMLU datasets primarily focus on high-resource languages such as English, which leaves low-resource languages like Bengali underrepresented. In this paper, we introduce BnMMLU, a benchmark to evaluate the multitask language understanding capabilities of Bengali in language models. The dataset spans 23 domains, including science, humanities, mathematics and general knowledge and is structured in a multiple-choice format to assess factual knowledge, application-based problem-solving and reasoning abilities of language models. It consists of 138,949 question-option pairs. We benchmark several proprietary and open-source large language models (LLMs) on the BnMMLU test set. Additionally, we annotate the test set with three cognitive categories—factual knowledge, procedural application and reasoning—to gain deeper insights into model strengths and weaknesses across various cognitive tasks. The results reveal significant performance gaps, highlighting the need for improved pre-training and fine-tuning strategies tailored to Bengali data. We release the dataset and benchmark results to facilitate further research in this area.

1 Introduction

The advancement of natural language processing (NLP) has been significantly driven by large-scale benchmarks that assess the capabilities of language models across various domains. Among these, the Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021) benchmark has emerged as a widely recognized evaluation framework. MMLU covers 57 diverse subjects, spanning disciplines such as mathematics, science, humanities, history, law, medicine and general knowledge. It is designed to measure a model's ability to generalize across multiple domains. While MMLU has significantly contributed to evaluating models in high-resource languages like English, it provides little to no coverage for low-resource languages like Bengali.

Although Bengali is the seventh most spoken language globally[1], Bengali remains underrepresented in NLP research, with limited high-quality datasets, pre-trained models and benchmarks. The absence of a standardized knowledge-driven evaluation data set for Bengali language models restricts their ability to generalize across real-world tasks. While some multilingual benchmarks include Bengali (Kakwani et al., 2020), their coverage is sparse and does not adequately test subject-specific knowledge or reasoning skills in Bengali.

To address this gap, we introduce BnMMLU, a multidisciplinary benchmark for evaluating the multitask language understanding of Bengali in language models. This dataset spans multiple disciplines and is structured in a multiple-choice format to assess factual, application and reasoning ability.

Our contributions in this work are:

- **Dataset Creation:** We construct a domain-specific Bengali knowledge benchmark, named BnMMLU, sourced from academic textbooks, competitive exams, educational resources and multiple educational websites.
- **Model Evaluation:** We benchmark several proprietary LLMs and open-source LLMs on the BnMMLU test set in a zero-shot setting.
- **Analysis and Insights:** We conduct a detailed analysis of model performance across various subjects and annotate the test set with three cognitive categories to gain deeper insights into the strengths and weaknesses of the models in different knowledge areas.

[1] https://www.dhakatribune.com/world/201648/bangla-ranked-at-7th-among-100-most-spoken

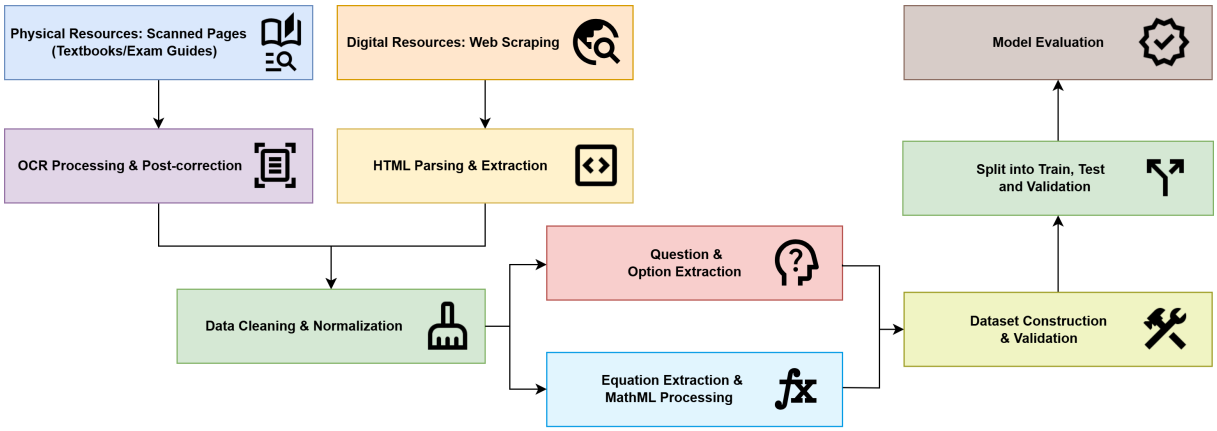


Figure 1: An overview of the data collection, extraction, cleaning and processing pipeline for constructing the BnMMLU benchmark, leading to dataset validation and evaluation.

2 Related Works

The Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021) established a rigorous standard by evaluating models across diverse disciplines, including mathematics, science, humanities and law. However, they fail to capture the linguistic, cultural and syntactic nuances of non-English languages. As a result, models trained and evaluated primarily on English data often exhibit suboptimal performance when applied to different linguistic contexts.

To address this, researchers have developed language-specific benchmarks tailored to different linguistic and cultural settings. The Korean MMLU (KMMLU) dataset (Son et al., 2024), derived from native Korean examinations, captures the linguistic and contextual intricacies unique to Korean. Similarly, the Chinese MMLU (CMMLU) (Li et al., 2024) provides a comprehensive Chinese benchmark and reveals that most existing LLMs struggle to achieve an average accuracy of 50%, emphasizing the need for improved pretraining and fine-tuning strategies for non-English languages. Another paper, M3KE (Liu et al., 2023) also focuses on Chinese language and their benchmark dataset reports that GPT-3.5 achieves an accuracy of 48%. ArabicMMLU (Koto et al., 2024) constitutes the first multi-task language understanding benchmark made for Modern Standard Arabic. The top-performing Arabic-centric model reported an overall accuracy 62.3%.

In the context of multilingual benchmarks, IndicGLUE (Kakwani et al., 2020) evaluates NLP models across multiple Indian languages, including Bengali, focusing on classification, sentiment analysis and named entity recognition (NER). However, IndicGLUE lacks the extensive multitask evaluation seen in MMLU. XGLUE (Liang et al., 2020), another multilingual benchmark covering 19 languages, primarily focuses on text classification and question answering but does not provide a comprehensive assessment of reasoning and domain-specific knowledge in Bengali.

The BEnQA benchmark (Shafayat et al., 2024) presents a dataset of parallel Bengali and English exam questions for middle and high school levels in Bangladesh, covering approximately 5000 science questions. The authors observe a performance disparity between LLMs on Bengali and English, finding that Chain-of-Thought prompting helps with reasoning but not factual questions. They also show that adding English translations improves answers in Bengali.

3 The BnMMLU Benchmark

3.1 Task Design

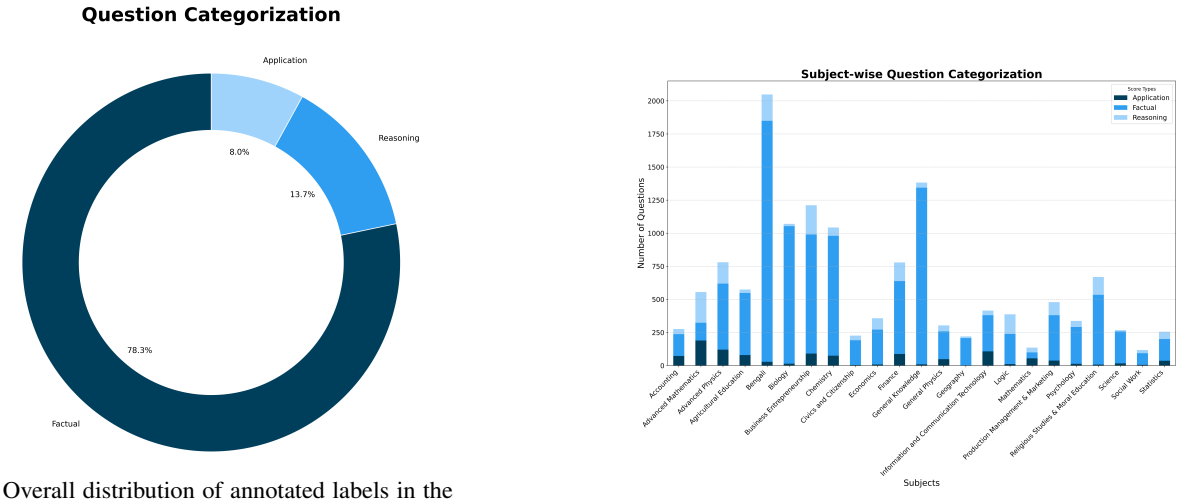
We create a multitask benchmark comprising 23 multiple-choice tasks spanning STEM, humanities, social sciences and other domains. Supercategories are detailed in Table 4.

3.2 Dataset Construction

The full workflow is shown in Figure 1. Questions were sourced from Bangladeshi educational and professional materials through two channels.

- **Physical Resources**: Scanned pages from NCTB-approved textbooks[2] and competitive

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



(a) Overall distribution of annotated labels in the test set.

(b) Subject-based distribution of annotated labels.

Figure 2: Distribution of annotated labels: (a) overall distribution, (b) subject-based distribution.

exam guides, processed using OCR tools[3] with post-correction for script accuracy.

- **Digital Sources**: Web-scraped questions from Bangladeshi educational portals. The web scraping was performed using Selenium[4] and BeautifulSoup[5].

3.3 Equation Storage and Representation

To ensure the structured preservation of mathematical expressions within the dataset, we stored all equations in MathML[6], an XML-based markup language for representing mathematics. However, to maintain consistency and simplify processing, we remove the outer <*math*>...</*math*> tags while retaining the inner MathML content. By doing this, the equations remain machine-readable preserving the data.

3.4 Task Categories

The task types includes a broad range of topics, each addressing a specific domain of expertise and practice.

3.4.1 Humanities

This includes studies of social structures, political systems and economics. It covers subjects such as Accounting (financial accounting, auditing and cost accounting), Economics (economic theories and markets), Finance (investment strategies), Business (entrepreneurship and business strategy) and Civics and Citizenship (the study of governance, rights and law). Additionally, psychology tasks explore human behavior, cognition, and mental health, while geography includes topics like earth sciences and geopolitics.

3.4.2 STEM (Science, Technology, Engineering, Mathematics)

Tasks in STEM emphasize scientific reasoning, mathematics and technology. Analytical subjects such as advanced mathematics, algebra and calculus (Advanced Mathematics) challenge problem-solving skills, while biology covers areas like cellular biology, genetics and ecology. Chemistry tasks focus on organic and inorganic chemistry and physics explores mechanics, thermodynamics and electromagnetism. Information and communication technology assesses skills in programming, networking and databases, while statistics tasks focus on data analysis, probability and inference.

3.4.3 Social Sciences

This domain studies human culture, philosophy and ethics. Bengali literature, poetry and syntax (Bengali) are explored through the lens of language and culture, while logical reasoning and argument analysis (Logic) test critical thinking. Religion and moral education (Religion and Moral Education) are also included.

3.4.4 Others

General knowledge and current affairs (General Knowledge) are assessed.

[3] https://github.com/JaidedAI/EasyOCR
[4] https://www.selenium.dev
[5] https://pypi.org/project/beautifulsoup4
[6] https://www.w3.org/TR/MathML

3.5 Dataset Splitting and Test Set Annotation

To ensure robust model development and evaluation, the dataset is partitioned into three non-overlapping subsets following an 80:10:10 ratio. The test set is reserved for the final performance evaluation of the models.

Like BEnQA paper (Shafayat et al., 2024), in addition to this partitioning, the test set has been meticulously annotated with three distinct cognitive categories to facilitate a more fine-grained assessment of model performance.

- **Factual Knowledge**: Includes questions that require the retrieval of specific facts or information.
- **Procedural and Application**: It assesses the ability to apply known procedures, methods or algorithms to solve problems.
- **Reasoning**: Tasks that require higher-order thinking skills, such as analysis, synthesis and logical deduction to derive conclusions.

3.5.1 Annotation Process

The annotation is conducted by three undergraduate students from diverse academic backgrounds. Annotators are provided only with the questions from the test set, without additional context ensuring an unbiased classification process. The annotators assign each question to one of the three categories. The inter-annotator agreement is measured using Fleiss' kappa (κ) (Fleiss, 1971) with a score of 0.6327, indicating substantial agreement (Landis and Koch, 1977). In cases of disagreement, the final label is determined by majority voting.

The overall distribution is shown in Figure 2a and the subject-based distribution is in Figure 2b.

4 Methodology

4.1 Model Selection

We evaluate several proprietary and open-source models on BnMMLU dataset. Following the recommendation from prior work (Lai et al., 2023), we keep the system prompt in English. The benchmark results presented in the paper are conducted in a zero-shot setting to save tokenization costs in proprietary models (Petrov et al., 2023).

We evaluate the performance of several language models on the BnMMLU dataset. These include:

- **Proprietary LLMs**: GPT-3.5-Turbo[7], GPT-4o[8], Claude 3.5-Haiku[9], Claude 3.5-Sonnet[10], Gemini 2.0 Flash[11], Gemini 2.0 Flash Lite[12].
- **Open-source LLMs**: Llama 3.1 - 8b and Llama 3.3 - 70b (Grattafiori et al., 2024), Gemma 2 - 9b and Gemma 2 - 27b (Gemma, 2024).

4.2 Evaluation Metrics

For evaluation, we employed accuracy as our primary metric, defined as the proportion of correctly predicted answers across all tasks.

5 Results

The evaluation of language models on the BnMMLU benchmark reveals distinct performance patterns across model architectures and knowledge domains. As shown in Table 1, proprietary models generally outperform open-source alternatives, with Gemini 2.0 Flash achieving the highest overall accuracy (75.80%), followed by GPT-4o (69.38%) and Claude 3.5-Sonnet (66.71%). Open-source models demonstrate a significant performance gap, with Llama 3.3-70b (59.30%) and Gemma 2-27b (53.45%) trailing behind their proprietary counterparts.

5.1 Performance Across Evaluation Metrics

Table 1 highlights critical disparities in model capabilities.

- **Factual Accuracy**: Gemini 2.0 Flash leads (76.53%), followed by GPT-4o (70.68%) and Claude 3.5-Sonnet (67.01%).
- **Reasoning Challenges**: All models show reduced performance in reasoning tasks, with the largest gap observed in open-source models (Llama 3.3-70b: 58.72% vs Gemini 2.0 Flash: 73.13%).
- **Procedural Application**: Performance declines across all models, particularly in GPT-4o (63.28% vs its 70.68% factual accuracy).

[7] https://platform.openai.com/docs/models#gpt-3-5-turbo
[8] https://platform.openai.com/docs/models#gpt-4o
[9] https://www.anthropic.com/claude/haiku
[10] https://www.anthropic.com/news/claude-3-5-sonnet
[11] https://deepmind.google/technologies/gemini/flash
[12] https://deepmind.google/technologies/gemini/flash-lite

Model	Overall Accuracy	Factual Accuracy	Application Accuracy	Reasoning Accuracy
gpt-3.5-turbo-0125	0.3830	0.3883	0.3848	0.3516
gpt-4o-2024-08-06	0.6938	0.7068	0.6328	0.6529
claude-3-5-haiku-20241022	0.5456	0.5349	0.6356	0.5564
claude-3-5-sonnet-20241022	0.6671	0.6701	0.6751	0.6455
gemini-2.0-flash	**0.7580**	**0.7653**	**0.7307**	**0.7313**
gemini-2.0-flash-lite	0.7199	0.7260	0.6761	0.7091
Llama 3.1 - 8b	0.3996	0.3989	0.4011	0.4033
Llama 3.3 - 70b	**0.5930**	**0.5969**	**0.5631**	**0.5872**
Gemma 2 - 9b	0.4835	0.4859	0.4539	0.4860
Gemma 2 - 27b	0.5345	0.5368	0.5075	0.5363

Table 1: Performance of Language Models on BnMMLU across different evaluation metrics

Model	STEM Accuracy	Humanities Accuracy	Social Sciences Accuracy	Others Accuracy
gpt-3.5-turbo-0125	0.3746	0.3671	0.3986	0.4067
gpt-4o-2024-08-06	0.7047	0.6066	0.7231	0.7627
claude-3-5-haiku-20241022	0.5688	0.4575	0.5712	0.5781
claude-3-5-sonnet-20241022	0.6728	0.5925	0.6854	0.7598
gemini-2.0-flash	**0.7893**	**0.6875**	**0.7529**	**0.8090**
gemini-2.0-flash-lite	0.7471	0.6440	0.7296	0.7562
Llama 3.1 - 8b	0.3895	0.3553	0.4373	0.4298
Llama 3.3 - 70b	**0.5926**	**0.5026**	**0.6409**	**0.6592**
Gemma 2 - 9b	0.4838	0.4088	0.5372	0.4942
Gemma 2 - 27b	0.5460	0.4533	0.5802	0.5391

Table 2: Domain-Specific Performance of Language Models on BnMMLU

5.2 Domain-Specific Competencies

Table 2 reveals substantial variations across knowledge domains.

- **STEM Superiority**: Gemini 2.0 Flash achieves 78.93% accuracy in STEM, outperforming GPT-4o (70.47%) by 8.46 percentage.
- **Humanities Disparity**: Performance gaps narrow in humanities, with Gemini 2.0 Flash (68.75%) leading Claude 3.5-Sonnet (59.25%) by 9.5 percentage.
- **Social Science Divide**: Proprietary models maintain strong performance (Gemini 2.0 Flash: 75.29%) while open-source models struggle (Llama 3.3-70b: 64.09%)
- **General Knowledge**: Gemini 2.0 Flash achieves peak performance (80.9%) in others domains.

5.3 Architectural Comparisons

- **Proprietary Models**: Gemini 2.0 Flash leads across all metrics, showing particular strength in STEM (+8.46% over GPT-4o) and reasoning tasks (+7.84% over Claude 3.5-Sonnet).
- **Open-Source Models**: Performance scales with parameter count - Llama 3.3-70b (59.30%) outperforms its 8B version by 19.34 percentage, while Gemma 2-27b (53.45%) surpasses its 9B counterpart by 5.1 percentage.

5.4 Critical Limitations

The evaluation reveals four fundamental constraints in current model capabilities.

- **Reasoning Deficits**: All models show significantly lower reasoning accuracy compared to factual recall, with proprietary models maintaining a 7.84-17.97% advantage over open-source counterparts (Gemini 2.0 Flash: 73.13% vs Llama 3.3-70b: 58.72% reasoning accuracy).
- **Proprietary-Open Source Divide**: A consistent 16.25-19.85% performance gap exists between top proprietary and open-source models

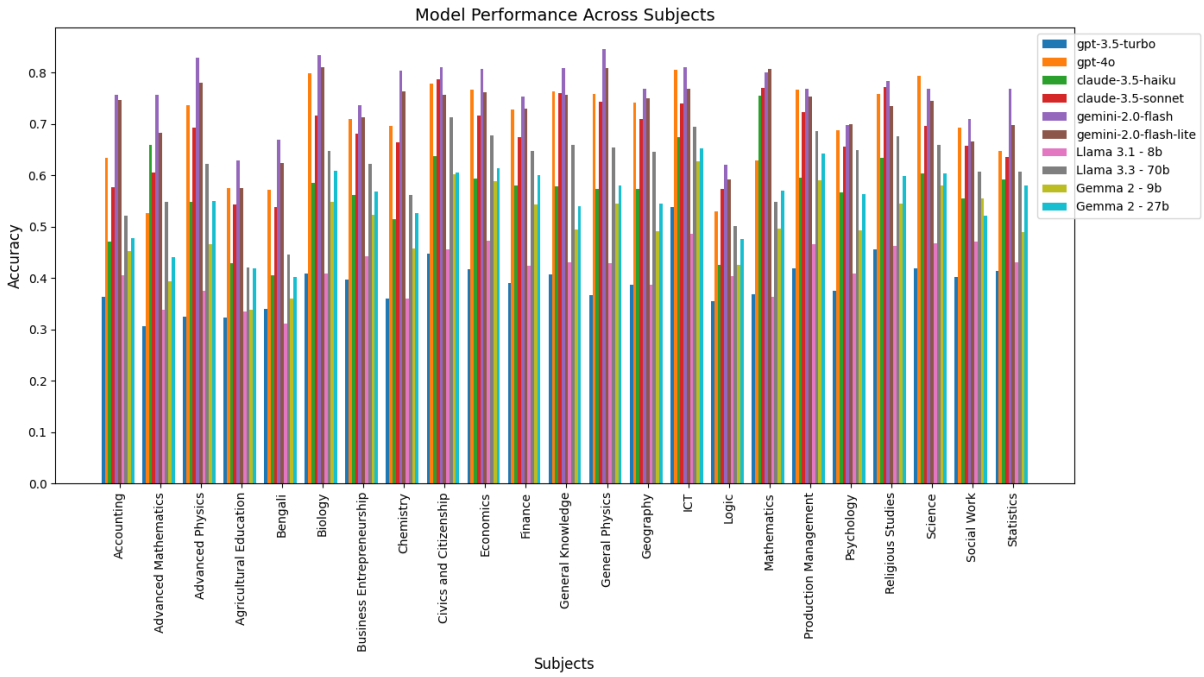


Figure 3: Subject-wise performance distribution across evaluated language models

across all metrics (Gemini 2.0 Flash: 75.80% vs Llama 3.3-70b: 59.30% overall accuracy).

- **Parameter Efficiency Issues**: Scaling effects remain sublinear - Gemma 2-27b (53.45%) only achieves 5.1% higher accuracy than its 9B version despite 3× parameters, while Gemini 2.0 Flash Lite (71.99%) shows minimal degradation from its full version (75.80%).

- **Procedural Application Challenges**: All models struggle with application tasks, showing 3-7% accuracy drops compared to factual performance (GPT-4o: 63.28% vs 70.68%, Claude 3.5-Sonnet: 67.51% vs 67.01%).

6 Conclusion

In this work, we introduce BnMMLU, a comprehensive multitask benchmark designed to evaluate the language understanding capabilities of Bengali language models. By covering 23 diverse domains and annotating the test set with cognitive categories, we provide a fine-grained assessment of model performance in factual knowledge retrieval, procedural application and reasoning. Our benchmarking results highlight significant performance gaps between proprietary and open-source models, with Gemini 2.0 Flash and GPT-4o achieving the highest overall accuracy, while open-source models like Llama 3.3 - 70b and Gemma 2 - 27b show moderate performance. A key insight from our findings is that reasoning and procedural application tasks remain challenging for Bengali language models, suggesting the need for improved pretraining and fine-tuning strategies. Additionally, the disparity in subject-wise performance underscores the limitations of existing Bengali training corpora, particularly in STEM and applied sciences. These results shows the urgent need for more diverse and high-quality Bengali datasets to enhance the generalization ability of LLMs in real-world applications. To facilitate further research, we release the BnMMLU dataset and benchmark results, aiming to drive advancements in Bengali NLP and multilingual model development. Future work could explore fine-tuning techniques, domain-specific data augmentation and more effective reasoning-based training methodologies to bridge the performance gap in Bengali multitask language understanding.

7 Limitations

In this section, we discuss the key limitations of our study, including constraints in dataset utilization, computational resources and model selection. These limitations impact the scope and generalizability of our findings, highlighting areas for future improvements and expansions.

7.1 Limited Dataset Utilization

Due to resource limitations, we could not use the full dataset for benchmarking. Instead, we evaluated models on a subset of the dataset (10%), which, while representative, does not fully capture the breadth of knowledge and reasoning capabilities across all subject areas.

7.2 Reasoning Models

While we benchmarked several proprietary and open-source models, we could not include reasoning models such as DeepSeek R1[13], OpenAI o1[14] due to computational limitations. These models are designed to excel in reasoning and logical inference tasks and their inclusion would provide valuable insights into the current gap in Bengali language understanding. Future work should benchmark these advanced models to assess their effectiveness on complex Bengali reasoning tasks.

7.3 General Knowledge Ambiguity

General knowledge and current affairs rely on real-time information, which may not be universally accessible or relevant.

7.4 Shift to Multimodal Models

While the dataset in question is not inherently multimodal, recent advancements in AI and machine learning have led to the rise of multimodal models that process and integrate text images and other data types simultaneously. This shift may create challenges in adapting traditional datasets to such models, as they are optimized for a more complex, multimodal approach that may not be fully compatible with purely textual or single-modality datasets.

8 Future Works

In this section, we discuss potential future directions to enhance BnMMLU and improve the performance of Bengali language models.

8.1 Expansion of Benchmark Coverage

Future iterations of BnMMLU could incorporate additional subject areas, particularly in technical fields such as medicine, law and engineering. Expanding the dataset to cover these domains would allow for a more comprehensive evaluation of domain-specific knowledge in Bengali. Additionally, including emerging disciplines such as artificial intelligence, data science and climate studies would ensure that Bengali language models remain relevant for evolving knowledge domains.

8.2 Fine-tuning Strategies for Bengali LLMs

The performance gaps observed in reasoning and procedural application tasks highlight the need for improved fine-tuning methodologies. Future research could explore instruction tuning, retrieval-augmented generation and Chain-of-Thought (CoT) (Wei et al., 2023) prompting tailored for Bengali language models. These techniques have shown promise in enhancing logical reasoning and problem-solving abilities in large language models (Shafayat et al., 2024) and could significantly improve performance on complex cognitive tasks in Bengali.

8.3 Development of Open-Source Bengali LLMs

The benchmark results indicate a significant performance disparity between proprietary and open-source models. Investing in the development and fine-tuning of large-scale, open-source Bengali LLMs trained on high-quality, domain-specific datasets could help bridge this gap. Creating publicly available, well-documented Bengali language models would enable researchers and developers to build robust NLP applications while reducing reliance on proprietary models.

8.4 Evaluation of Multilingual and Code-Switched Models

Many Bengali speakers frequently switch between Bengali and English in real-world communication. Investigating how multilingual and code-switched models perform on BnMMLU could provide insights into their practical usability in bilingual environments. Future research could analyze whether multilingual pretraining and translation-based augmentation improve the accuracy and fluency of Bengali-English code-switched text processing.

9 Ethical Considerations

BnMMLU dataset's test portion's annotators were fairly compensated for their work. The dataset will be publicly available under the CC BY-SA 4.0 license, ensuring free accessibility.

[13] https://www.deepseek.com
[14] https://openai.com/o1

References

JL Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378—382.

Gemma. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, and Andy Zou. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., and Avik Bhattacharyya. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. *Preprint*, arXiv:2402.12840.

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. *Preprint*, arXiv:2304.05613.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Haonan Li, Yixuan Zhang, and Fajri Koto. 2024. Cmmlu: Measuring massive multitask language understanding in chinese. *Preprint*, arXiv:2306.09212.

Yaobo Liang, Nan Duan, Yeyun Gong, and Ning Wu. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Chuang Liu, Renren Jin, Yuqi Ren, Linhao Yu, and Tianyu Dong. 2023. M3ke: A massive multi-level multi-subject knowledge evaluation benchmark for chinese large language models. *Preprint*, arXiv:2305.10263.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *Preprint*, arXiv:2305.15425.

Sheikh Shafayat, H Hasan, Minhajur Mahim, and Rifki Putri. 2024. BEnQA: A question answering benchmark for Bengali and English. In *ACL 2024*, pages 1158–1177, Bangkok, Thailand.

Guijin Son, Hanwool Lee, Sungdong Kim, and Seungone Kim. 2024. Kmmlu: Measuring massive multitask language understanding in korean. *Preprint*, arXiv:2402.11548.

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. *Preprint*, arXiv:2201.11903.

A Appendix

A.1 Dataset Statistics

Table 3 provides a detailed breakdown of the dataset, including the number of questions per domain and their data splits. Table 4 gives us overview of all subject domains tested concepts in BnMMLU.

B Model Details

B.1 Prompting Strategies for LLMs

LLMs were evaluated using zero-shot.

Prompt:
You are an AI trained to answer multiple-choice questions accurately.
Read the following question carefully and provide the correct answer choice.
Respond only with the letter corresponding to the correct option (e.g., A, B, C, or D).

```
{row[question]}
A) {row[option A]}
B) {row[option B]}
C) {row[option C]}
D) {row[option D]}
```

C Experimental Setup

All experiments were conducted using NVIDIA RTX A6000 GPU with PyTorch 2.0.

Domain	Subdomain (Subjects)	Total Questions	Train / Test / Validation
STEM	Advanced Mathematics	5,556	4,444 / 555 / 557
	Advanced Physics	7,807	6,245 / 780 / 782
	Agricultural Education	5,745	4,596 / 574 / 575
	Biology	10,706	8,564 / 1,070 / 1,072
	Chemistry	10,434	8,347 / 1,043 / 1,044
	General Mathematics	1,359	1,087 / 135 / 137
	General Physics	3,037	2,429 / 303 / 305
	Information and Communication Technology	4,151	3,320 / 415 / 416
	Science	2,136	2,046 / 267 / 268
	Statistics	2,558	2,046 / 255 / 257
Humanities	Bengali	20,482	16,385 / 2,048 / 2,049
	Logic	3,870	3,096 / 387 / 387
	Religious Studies	6,691	5,352 / 669 / 670
Social Sciences	Accounting	2,768	2,214 / 276 / 278
	Business Entrepreneurship	12,112	9,689 / 1,211 / 1,212
	Civics and Citizenship	2,262	1,809 / 226 / 227
	Economics	3,572	2,857 / 357 / 358
	Finance	7,799	6,239 / 779 / 781
	Geography	2,208	1,766 / 220 / 222
	Production Management	4,796	3,836 / 479 / 481
	Psychology	3,371	2,696 / 337 / 338
	Social Work	1,174	939 / 117 / 118
Others	General Knowledge	13,820	11,056 / 1,382 / 1,382

Table 3: Dataset distribution across domains, subjects and Train/Test/Validation splits.

SL	Subject Name	Tested Concepts	Supercategory
1	Accounting	Financial Accounting, Auditing, Cost Accounting...	Social Sciences
2	Advanced Mathematics	Calculus, Algebra, Advanced Topics...	STEM
3	Advanced Physics	Mechanics, Thermodynamics, Electromagnetism...	STEM
4	Agricultural Studies	Agronomy, Soil Science, Agro-business...	STEM
5	Bengali	Literature, Poetry, Syntax...	Humanities
6	Biology	Cellular Biology, Genetics, Ecology...	STEM
7	Business Entrepreneurship	Entrepreneurship, Business Strategy...	Social Sciences
8	Chemistry	Organic Chemistry, Inorganic Chemistry...	STEM
9	Civics and Citizenship	Governance, Rights, Law...	Social Sciences
10	Economics	Economic Theories, Markets, Finance...	Social Sciences
11	Finance	Investment, Corporate Finance, Economics...	Social Sciences
12	General Knowledge	General Knowledge, Current Affairs...	Others
13	General Mathematics	Algebra, Geometry, Arithmetic...	STEM
14	General Physics	Mechanics, Thermodynamics, Electromagnetism...	STEM
15	Geography	Earth Sciences, Geopolitics...	Social Sciences
16	Information and Communication Technology	Programming, Networking, Databases...	STEM
17	Logic	Logical Reasoning, Argument Analysis...	Humanities
18	Psychology	Behavior, Cognition, Mental Health...	Social Sciences
19	Production Management	Operations, Marketing, Supply Chain...	Social Sciences
20	Religion and Moral Education	Ethics, Philosophy, Different Religions...	Humanities
21	Science	General Science, Scientific Method...	STEM
22	Social Work	Community Development, Welfare Policies...	Social Sciences
23	Statistics	Data Analysis, Probability, Inference...	STEM

Table 4: Overview of Subject Domains Tested Concepts in BnMMLU.