BnMMLU: Measuring Massive Multitask Language Understanding in Bengali

Saman Sarker Joy[1, 2] and **Swakkhar Shatabda**[2]
[1]University of Malaya, Malaysia
[2]BRAC University, Bangladesh
saman.sarker.joy@gmail.com, swakkhar.shatabda@bracu.ac.bd

Abstract

Large-scale multitask benchmarks have driven rapid progress in language modeling, yet most emphasize high-resource languages such as English, leaving Bengali underrepresented. We present BnMMLU, a comprehensive benchmark for measuring massive multitask language understanding in Bengali. BnMMLU spans 41 domains across STEM, humanities, social sciences, and general knowledge, and contains 134,375 multiple-choice question–option pairs-the most extensive Bengali evaluation suite to date. The dataset preserves mathematical content via MathML, and includes BnMMLU-HARD, a compact subset constructed from questions most frequently missed by top systems to stress difficult cases. We benchmark 24 model variants across 11 LLM families, spanning open-weights general/multilingual, Bengali-centric open-weights, and proprietary models, covering multiple parameter scales and instruction-tuned settings. We evaluate models under standardized protocols covering two prompting styles (Direct vs. Chain-of-Thought) and two context regimes (0-shot vs. 5-shot), reporting accuracy consistently across families. Our analysis highlights persistent gaps in reasoning and application skills and indicates sublinear returns to scale across model sizes. We release the dataset and evaluation templates to support rigorous, reproducible assessment of Bengali language understanding and to catalyze progress in multilingual NLP.

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.

Although Bengali[1] is the seventh most spoken language globally (Eberhard et al., 2025), 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.

In the absence of such a benchmark, researchers lack the means to assess whether a model's responses in Bengali reflect genuine understanding, memorization of bilingual cues or hallucination. Our study is guided by the following:

(RQ1) How far do multilingual vs. Bengali-centric models transfer to native Bengali tasks across various domains?

(RQ2) What are the returns to scale under standardized prompting/context regimes?

(RQ3) When does elicited reasoning help (or hurt), especially on difficult items?

(RQ4) Which subject areas are systematically hard vs. easy across different LLMs?

[1]We use *Bengali* and *Bangla* interchangeably to denote the same language (ISO 639-1: bn; ISO 639-3: ben). The IANA Language Subtag Registry entry for bn lists both (https://www.iana.org/assignments/language-subtag-registry).

Findings of the Association for Computational Linguistics: ACL 2026, pages 12211–12230
July 2-7, 2026 ©2026 Association for Computational Linguistics

Dataset	Format	# Items	# Subjects	Math	S:H:SS:O
BanglaQuAD (Rony et al., 2024)	Extractive	30,808	14	✗	3:4:5:2
BanglaRQA (Ekram et al., 2022)	Extractive	14,889	20	✗	1:2:2:3
BEnQA (Shafayat et al., 2024)	MCQ	5,161	5	✓	1:0:0:0
BLUCK (Kabir et al., 2025)	MCQ	2,366	23	✗	0:1637:729:0
NOIRBETTIK (Aurpa et al., 2025)	MCQ	5,215	8	✗	2:8:1:4
TituLM-Bangla MMLU (Nahin et al., 2025a)	MCQ	87,869	11	✗	98:19:17:1
UDDIPOK (Aurpa et al., 2023)	Extractive	3,636	–	✗	–
BnMMLU	**MCQ**	**134,375**	**41**	✓	**4:2:3:1**

Table 1: Comparison of prominent Bengali QA datasets. The table lists format (extractive vs. multiple choice), size (items and subjects), preservation of mathematical content (MathML), and proportional distribution across STEM, Humanities, Social Sciences and Others. S:H:SS:O denotes STEM:Humanities:Social Sciences:Others.

To address these questions, we introduce BnMMLU, a benchmark to evaluate the multitask language understanding of Bengali in language models. Our contributions in this work are:

- A 41-domain MCQ suite with 134,375 spanning STEM, humanities, social science and general knowledge.
- Introduced BnMMLU-HARD, formed by ranking questions most frequently missed by models while preserving subdomain balance for stress testing.
- Evaluated 24 model variants, spanning open-weights general/multilingual, Bengali-centric open-weights and proprietary models.
- Comparable reporting across Direct vs. CoT and 0-shot vs. 5-shot settings and Reasoning and Non-reasoning comparisons with consistent prompts and accuracy metrics.

2 Related Work

The Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021) set a standard for evaluating language models on broad domain knowledge (e.g., mathematics, science, humanities, law), but is essentially English-centric and does not capture the linguistic, cultural and syntactic nuances of other languages.

Language-specific MMLU-style benchmarks extend this paradigm to local exams: KMMLU for Korean (Son et al., 2024), CMMLU for Chinese (Li et al., 2024), and ArabicMMLU for Modern Standard Arabic (Koto et al., 2024), all reporting that non-English models still lag behind their English counterparts.

In the multilingual setting, IndicGLUE (Kakwani et al., 2020) and XGLUE (Liang et al., 2020) include Bengali among many languages and cover tasks such as classification, sentiment analysis, NER and QA, but they are not broad multitask knowledge benchmarks in the MMLU sense.

For Bengali specifically, existing resources such as BanglaQuAD (Rony et al., 2024), BanglaRQA (Ekram et al., 2022), BEnQA (Shafayat et al., 2024), and related datasets (e.g., NOIRBETTIK, BLUCK) provide task-specific QA and reading-comprehension style evaluations, often focusing on span extraction or short-answer questions rather than structured, curriculum-style subject coverage.

TituLM-Bangla MMLU (Nahin et al., 2025a) adapts MMLU-style diagnostics to Bengali multiple-choice questions across various topics, but with narrower subject breadth and less fine-grained coverage than our BnMMLU, which targets a wider set of Bengali academic and professional domains for multitask knowledge and reasoning evaluation.

3 The BnMMLU Benchmark

We create BnMMLU, a multitask benchmark composed of multiple-choice question–answer pairs across 41 subjects spanning STEM, humanities, social sciences and other domains. We refer to this complete benchmark as BnMMLU-FULL throughout the remainder of the paper. The overview of the full pipeline is shown in Figure 1.

3.1 Dataset Construction

The questions were sourced from Bangladeshi educational and professional materials through two channels.

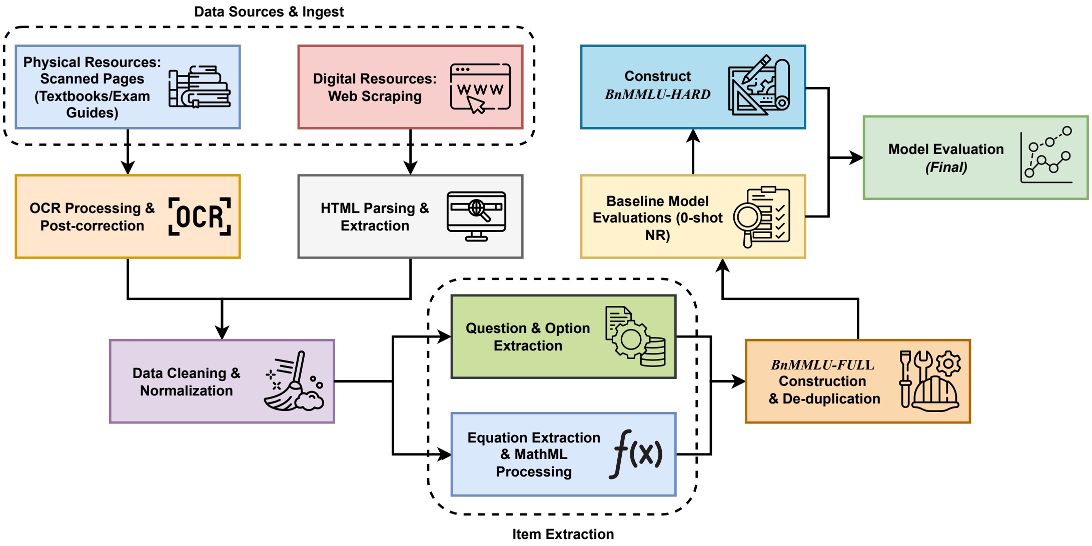

Figure 1: An overview of the pipeline for constructing the BnMMLU benchmark.

Physical Resources. Scanned pages from NCTB-approved textbooks and competitive exam guides, processed using OCR tool with post-correction for script accuracy. Due to the unstructured formatting of many print materials, 20% of the data came from these sources, and they did not contain properly formatted multiple-choice questions and answers. Examples of these books are shown in Figure 5.

Digital Sources. Web-scraped questions from Bangladeshi educational portals that host structured, exam-style multiple-choice questions. The web scraping was performed using Selenium[2] and BeautifulSoup[3]. The majority of the dataset, around 80% of the data came from these digital sources.

3.2 Optical Character Recognition (OCR) & Post-Correction

We scan printed book pages, apply a standard pre-processing pipeline (grayscale conversion, adaptive binarisation, and deskewing) and then run OCR system followed by LLM-based copy-editing to clean the text while preserving math and answer keys. Full implementation details and the exact copy-editing prompt are provided in Appendix B.

Post-correction reduced formatting issues and spelling errors. Additionally, approximately 10% of the question-option pairs were manually reviewed by the authors to ensure that the OCR text, math expressions and answer keys matched the original source pages.

We stored all equations in MathML[4], an XML-based markup language for representing mathematics.

3.3 Duplicate-Question De-duplication

We embed each question–option string using *text-embedding-3-small*[5] and run approximate nearest-neighbor search with the angular metric. For each question q_i, we retrieve top-k neighbors and convert angular distances $d(q_i, q_j)$ to similarities $s(q_i, q_j) = 1 - d(q_i, q_j)/2$; pairs with $s \geq 0.90$ define edges in an undirected graph whose connected components form duplicate clusters. We keep a single canonical item per cluster to obtain a de-duplicated benchmark. Full details are in Appendix C.

3.4 Task Categories

The benchmark covers 41 subjects across STEM, Humanities, Social Sciences and Other domains; a full list of subjects and tested concepts is provided in Appendix A.

3.5 Training-test decontamination

Because roughly 80% of BnMMLU is sourced from web-based question banks (section 3.1), we explicitly quantify potential train-test overlap on

[2] https://www.selenium.dev
[3] https://pypi.org/project/beautifulsoup4
[4] https://www.w3.org/TR/MathML
[5] https://platform.openai.com/docs/models/text-embedding-3-small

Model (Best per Family)	STEM	Humanities	Social Sciences	Others	Overall (Δ)
English-Centric / Bilingual Instruction-Tuned Models (Best per Family)					
LLAMA-3.3-70B-INSTRUCT	62.53	52.47	<u>65.81</u>	<u>68.99</u>	61.87 (+36.87)
QWEN3-32B	<u>72.03</u>	<u>53.01</u>	65.57	66.44	<u>65.34 (+40.34)</u>
GEMMA-3-27B-IT	63.61	51.43	63.90	67.44	61.27 (+36.27)
Bengali Pretrained / Instruction-Tuned Models (Best per Family)					
TITULLM-1B	27.15	27.53	28.42	28.19	27.72 (+2.72)
TIGERLLM-9B-IT	<u>56.02</u>	<u>47.85</u>	<u>59.48</u>	<u>61.29</u>	<u>55.70 (+30.70)</u>
BANGLALLAMA-3.1-8B-INSTRUCT	26.10	27.56	27.58	26.52	26.95 (+1.95)
Proprietary Models					
GPT-5-MINI	48.25	43.96	55.00	55.78	50.09 (+25.09)
GROK 4 FAST	61.98	51.63	64.02	67.60	60.82 (+35.82)
GEMINI 2.5 FLASH	72.38	**62.32**	**71.08**	**73.85**	**69.85 (+44.85)**
DEEPSEEK-V3.2-EXP	72.72	58.62	70.06	73.84	68.82 (+43.82)
QWEN-PLUS	**73.49**	56.15	66.89	70.17	67.29 (+42.29)

Table 2: Average accuracy (%) of models on the BNMMLU-FULL benchmark under 0-shot Direct (Non-Reasoning) evaluation. We report only the best-performing checkpoint per model family. Bold marks the highest overall score; underlines denote the best model within each category. (Δ) in overall is compared with random baseline (25%).

LLMs via an n-gram decontamination analysis. Following the GPT-3 contamination protocol and subsequent work, a shared 13-token span is treated as a conservative signal of near-verbatim memorization rather than chance overlap (Brown et al., 2020; Ravaut et al., 2025).

Overall contamination is low: for most corpora, fewer than 0.1% of questions exhibit any overlapping 13-gram. Full preprocessing details, per-corpus breakdowns are provided in Appendix D.

3.6 BnMMLU-HARD

We construct BNMMLU-HARD as a compact subset focused on the questions most frequently missed by the *top-10* models on BNMMLU-FULL, using their 0-shot (Direct) scores. Questions are ranked by aggregate error across these models, and we select the highest-error set while preserving a proportional subdomain balance. The distribution for both of them is shown in Figure 13, with subject counts in Table 8.

4 Experimental Evaluation

Following the recommendation from prior work (Lai et al., 2023), we keep the system prompt in English unless stated otherwise.

4.1 Model Selection

We evaluate a diverse set of language models on the **BnMMLU** dataset. Our selection is designed to cover both proprietary and open-weight families, multiple parameter scales, instruction-tuned checkpoints where available and a balance between Bengali-centric and English-centric models. Detailed access and setup information is provided in Table 7.

4.2 Evaluation Protocol

We evaluate each model under two prompting styles (Direct and Chain-of-Thought, CoT), two context regimes (0-shot and 5-shot) and two reasoning configurations (Reasoning-On and Non-Reasoning). The exact 0-shot/5-shot and Direct/CoT templates are shown in Figure 6.

The exact 0-shot/5-shot and Direct/CoT templates are shown in Figure 6.

Exemplar Construction for 5-shot. We selected five questions from each domain and used GPT-5-MINI WebUI[6] to make reasoning traces (CoT) the prompt in Figure 7. Then we manually screened the exemplars for correctness and style consistency. These were used as in-context demonstrations in the 5-shot setting (Direct uses the same exemplars but with the reasoning text removed).

[6] https://chatgpt.com/

Model	0-shot Direct (Non-Reasoning)	0-shot CoT (Δ) (Non-Reasoning)	5-shot Direct (Non-Reasoning)	5-shot CoT (Δ) (Non-Reasoning)
English-Centric / Bilingual Instruction-Tuned Models				
LLAMA-3.2-3B-INSTRUCT	19.95	18.33 (-1.62)	22.16	23.25 (+1.09)
LLAMA-3.3-70B-INSTRUCT	23.78	35.17 (+11.39)	31.15	<u>37.50 (+6.35)</u>
QWEN3-14B	14.67	14.32 (-0.35)	18.35	16.88 (-1.47)
QWEN3-32B	<u>25.52</u>	28.63 (+3.11)	34.63	31.19 (-3.44)
GEMMA-3-12B-IT	10.54	14.55 (+4.01)	18.50	23.52 (+5.02)
GEMMA-3-27B-IT	14.72	<u>37.59 (+22.87)</u>	<u>35.65</u>	34.65 (-1.00)
Bengali Pretrained / Instruction-Tuned Models				
TIGERLLM-9B-IT	<u>11.01</u>	<u>16.78 (+5.77)</u>	<u>18.44</u>	<u>23.32 (+4.88)</u>
Proprietary Models				
GPT-5-MINI	14.13	19.12 (+4.99)	19.66	18.63 (-1.03)
GROK 4 FAST	20.94	20.89 (-0.05)	44.06	51.12 (+7.06)
GEMINI 2.5 FLASH	**34.46**	45.38 (+10.92)	51.62	61.48 (+9.86)
DEEPSEEK-V3.2-EXP	29.89	**59.04 (+29.15)**	**58.83**	**64.53 (+5.70)**
QWEN-PLUS	32.47	58.74 (+26.27)	57.40	55.09 (-2.31)

Table 3: Accuracy (%) on BNMMLU-HARD for a reduced set of representative models. Δ is computed as CoT − Direct at the *same shot* (0-shot or 5-shot). **Bold** marks the global best per column; <u>underline</u> marks the best *within each category* per column.

4.3 Evaluation Metrics

For evaluating performance on BNMMLU-FULL & BNMMLU-HARD, we use accuracy as the primary metric. Accuracy is defined as the proportion of correctly predicted answers out of the total questions attempted.

5 Discussion

Table 2 summarizes 0-shot Direct (Non-Reasoning) accuracy on BNMMLU-FULL and detailed summary is shown in Table 9. Proprietary models lead overall: GEMINI 2.5 FLASH tops the chart (69.85) with best or near-best scores across Humanities, Social Sciences, and Others, while QWEN-PLUS holds the STEM peak (73.49) and strong overall (67.29). Among open-weights, QWEN3-32B (65.34) and LLAMA-3.3-70B-INSTRUCT (61.87) are the strongest, followed closely by GEMMA-3-27B-IT (61.27).

Bengali-centric models show competitive mid-tier performance led by TIGERLLM-9B-IT (55.70; best in its group), while small Bengali models cluster near the high-20s. Domain-wise, STEM tends to be the highest-scoring slice for top systems, with Humanities relatively lower for open-weights. Net: proprietary models currently set the frontier, large open-weights close much of the gap, and targeted Bengali pretraining helps at moderate scale but has not yet matched the largest bilingual/global families.

So, scale helps but with diminishing returns; consistent ladders imply healthy training pipelines; and matched-compute gaps highlight the outsized role of data and recipe design, especially beyond the mid-compute regime.

5.1 Prompting & Context Regimes

As shown in Table 3, adding reasoning and shots generally boosts accuracy, with the largest gains typically from *5-shot CoT*. Standout jumps include DEEPSEEK-V3.2-EXP (29.89→64.53, +34.64), GEMINI 2.5 FLASH (34.46→61.48, +27.02), and GROK 4 FAST (20.94→51.12, +30.18). Among open weights, GEMMA-3-27B-IT benefits markedly (+22.87 with 0-shot CoT; +20.93 with 5-shot Direct), and LLAMA-3.3-70B-INSTRUCT rises to 37.50 (+13.72). Bengali-centric TIGERLLM-9B-IT starts low (11.01) but more than doubles under 5-shot CoT (23.32; +12.31), indicating prompting can partly offset limited scale.

Gains are heterogeneous and sometimes negative at small–mid scales: LLAMA-3.2-3B-INSTRUCT (0-shot CoT: -1.62), QWEN3-8B (-0.68), and QWEN3-14B (-0.35); moreover, 5-shot CoT can underperform 5-shot Direct in some cases (e.g., QWEN3-32B: +5.67 vs. +9.11).

Model	STEM		Humanities		Social Sciences		Others		Overall	
	NR	R	NR	R	NR	R	NR	R	NR	R
Qwen3-32B	35.41	68.76	13.07	27.82	20.24	37.12	20.02	41.57	25.52	49.41
GPT-5-Mini	14.88	69.51	10.57	33.29	14.73	47.15	15.04	59.20	14.13	55.25
Grok 4 Fast	22.12	77.34	15.68	44.79	21.10	57.61	25.38	68.01	20.94	64.64
Gemini 2.5 Flash	37.62	73.75	26.86	47.45	**33.14**	57.18	**39.56**	67.91	**34.46**	63.39
DeepSeek-V3.2-Exp	35.17	**80.65**	**28.11**	**51.83**	27.00	**63.90**	29.00	**74.43**	29.89	**69.79**
Qwen-Plus	**43.65**	77.93	19.74	46.24	25.47	56.89	28.16	67.15	32.47	64.83

Table 4: 0-shot Direct evaluation accuracy (%) of reasoning-capable models on the **BnMMLU-HARD** subset. **NR** denotes Non-Reasoning and **R** denotes Reasoning.

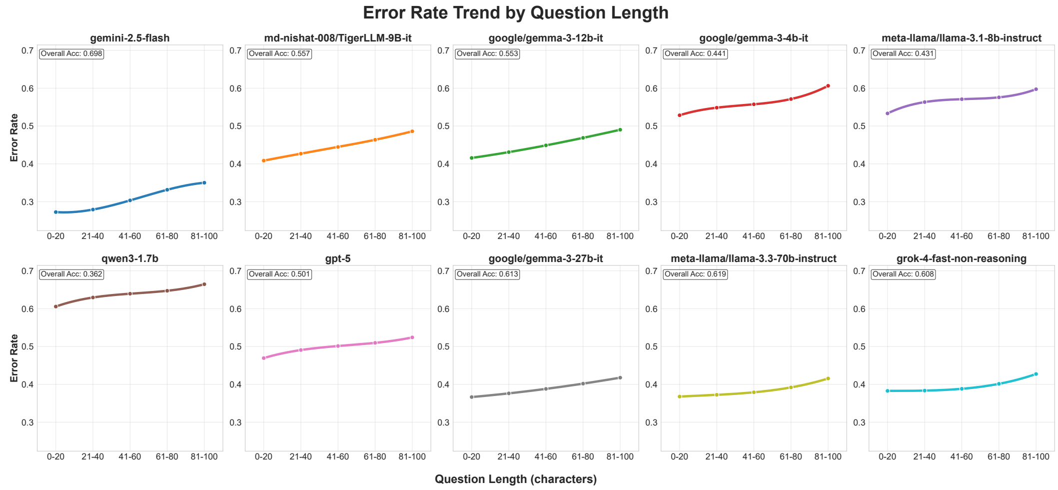

Figure 2: Error rate trends by question length (in characters) across ten evaluated models on the BnMMLU-FULL benchmark. Each subplot represents an individual model, with the x-axis indicating question length bins and the y-axis showing corresponding error rates. Overall accuracy for each model is annotated in its respective panel for reference.

5.2 Reasoning Effects

Across all reasoning-capable models on BnMMLU-HARD, enabling reasoning consistently lifts accuracy in every domain and for every model. Overall gains range from Qwen3-1.7B (+14.75; 14.53→29.28) to Grok 4 Fast (+43.70; 20.94→64.64), with substantial jumps also for GPT-5-Mini (+41.12) and DeepSeek-V3.2-Exp (+39.90). Under the reasoning setting, DeepSeek-V3.2-Exp attains the top scores across all domains-STEM 80.65, Humanities 51.83, Social Sciences 63.90, Others 74.43-and the highest overall (69.79). By contrast, under non-reasoning, the strongest baselines are split: Qwen-Plus leads STEM (43.65), DeepSeek-V3.2-Exp leads Humanities (28.11), and Gemini 2.5 Flash leads Social Sciences (33.14), Others (39.56), and Overall (34.46). These patterns indicate that reasoning particularly amplifies STEM and “Others” performance for mid/large models (e.g., Qwen3-14B STEM 18.42→65.36; Grok 4 Fast Others 25.38→68.01), while still yielding reliable improvements in Humanities and Social Sciences. All figures are on Table 4.

5.3 Sequence-Length Robustness

Across models, error rates increase monotonically with question length, with the sharpest degradation typically occurring between the 0–20 and 81–100 character bins. The strongest systems maintain the lowest error curves throughout: gemini-2.5-flash, llama-3.3-70b-instruct and gemma-3-27b-it show relatively shallow slopes as length grows. Mid-tier models such as gemma-3-12b-it, TigerLLM-9B-it, GPT-5-Mini exhibit a clearer length penalty past 60 characters. Smaller/earlier-generation instruction models like llama-3.1-8b-instruct, gemma-3-4b-it and qwen3-1.7b have the highest error rates

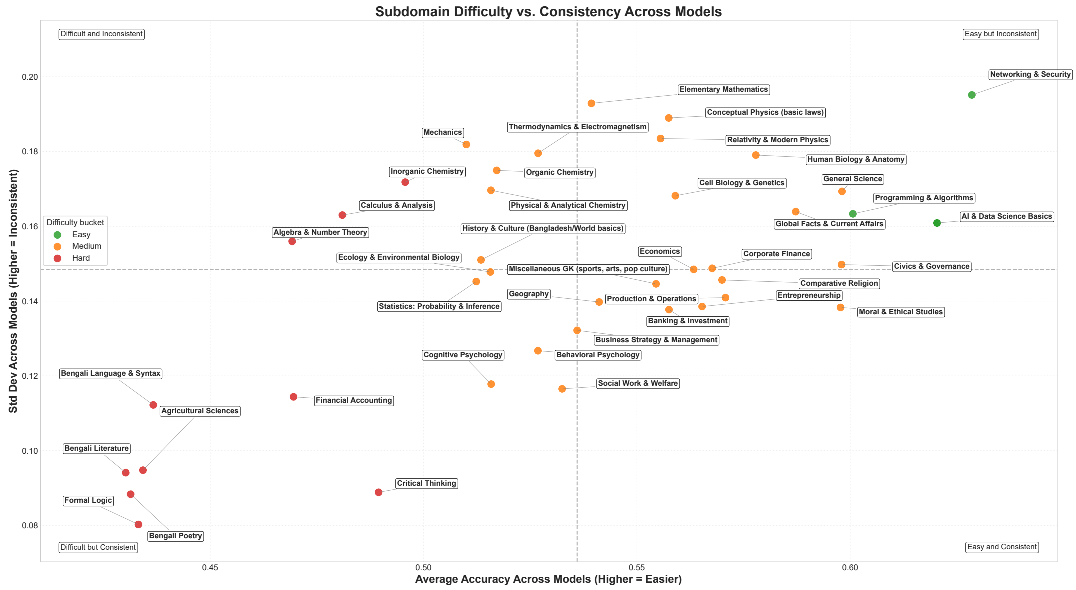

Figure 3: Subdomain difficulty versus cross-model consistency on the **BnMMLU-FULL** benchmark under 0-shot Direct prompting. The x-axis shows mean accuracy across models (higher = easier), and the y-axis shows standard deviation (higher = more inconsistent); each point is a subdomain color-coded by difficulty bucket (*Easy*, *Medium*, *Hard*). The four quadrants (*Easy & Consistent*, *Easy but Inconsistent*, *Difficult and Inconsistent*, *Difficult but Consistent*) summarize how subdomain complexity and variability interact in assessing LLM robustness.

and the steepest length-dependent drop-offs. Consequently, the performance gap between top and weaker models widens in the longest bin, indicating reduced robustness to longer, likely more compositionally complex, prompts. The per-model length-specific error profiles are visualised in Figure 2.

5.4 Subject-Specific Failure Modes

Analysis. Using median-based splits (dashed lines), we see four bands: Difficult & Inconsistent-advanced STEM (e.g., algebra/analysis, inorganic chemistry, mechanics) with low accuracy and wide spread; Easy & Inconsistent-computing/tech survey areas (networking, AI/data basics, programming) that score high but vary by model; Difficult & Consistent-Bengali/logic plus applied topics (accounting, agriculture) that are uniformly hard; Easy & Consistent-management/psych/finance/geography that most models handle reliably. Figure 3 shows the domain difficulty versus consistency.

5.5 Error Taxonomy & Case Studies

Across all prompting regimes, we observe a small set of recurring slips that surface with different facades.

Instruction-Following vs. Heuristic Shortcuts. A first class of errors stems from the model seizing a plausible heuristic instead of following the full instruction. When asked to expand an acronym, for instance, the model often latches onto the most *frequent* completion rather than the *domain-correct* one. In one item ("e-GP stands for: …"), a 0-shot direct answer defaulted to electronic government purchasing, likely because "purchasing" is a frequent neighbor of "e-government" in pretraining text. Chain-of-thought (CoT) prompting nudged the model to reason about procurement systems and public-sector terminology, which shifted the answer to electronic government procurement - the intended domain term. CoT slows the jump to a high-frequency collocation and creates space to align to the task's governing instruction (disambiguate by function, not by frequency).

Ambiguity at the Interface: Formatting, Scripts, and Mixed Notation. A second cluster originates upstream of reasoning: mixed scripts (Bengali + Roman), MathML-like tokens (`<msup>`, `<msqrt>`), and lookalike glyphs ("ln" vs. "1n") can be partially misparsed, leading the model to answer a *nearby* question.

Calibration and the Amount of Examples. More examples is not always better. We see overfitting to few-shot context where long Chain-of-Thought (CoT) rationales import the wrong frame (e.g., shown in Figure 9). Conversely, right-sized scaffolds 2-4 concise checks tied to the item's gating cues (time unit, regime, exponent, unit) - deliver the largest flips from wrong to right.

Design Implications (Across Domains). Three prompt-level nudges generalize: (i) *normalize then solve* for mixed markup/symbols; (ii) *scaffold lightly* to surface intermediate commitments without inducing spurious patterns; and (iii) *option-calibrate* by matching the derived condition (unit/exponent/scope) to the exact wording of the alternatives.

5.6 Bengali-Specific Error Patterns

Bengali items add characteristic frictions that interact with the taxonomy above. Below are several case studies illustrating these patterns.

Orthography & Mixed Markup at the Math/Language Boundary. Bengali prose often co-occurs with inline MathML-style tags in the options. Under 0-shot direct prompting, models sometimes select the most salient-looking option (e.g., a tidy fraction or exponent) without fully parsing the markup. For instance, in questions involving calculus or optics, performance improves once the expression is restated in standard math and only then compared against candidates.

Anglicized Cue Phrases Inside Bengali Questions. Embedded English slogans or titles can bias frequency-driven guesses in direct mode. A single reasoning step that maps the phrase to world knowledge before selecting the option reliably corrects this.

Bengali Numerals, Currency Tokens and the Danda. Arithmetic questions mixing Bengali numerals with the word for currency ("Taka") and closing with the Bengali danda tend to elicit rounded, visually salient choices in direct mode. Light reasoning that ties numerals to operations and checks the unit phrase flips such items to the correct answer.

5.7 CoT vs. Reasoning-On

This section examines cases where *5-shot CoT (non-reasoning)* answered incorrectly but *0-shot Reasoning-On* answered correctly. We preserve the provided snippets and bold the decisive cues.

Why Reasoning-On Helps. Across slices where 5-shot CoT (non-reasoning) fails but 0-shot Reasoning-On succeeds, the dominant pattern is regime selection vs. heuristic lock-in. CoT often stabilizes on a salient rule and never revisits it; e.g., from Figure 10 we can see, applying the inverse-square law outside Earth even though the query's span is center $\rightarrow$ surface, where $g \propto r$ and thus increases proportionally (Mechanics; B). By contrast, Reasoning-On explicitly enumerates alternatives, chooses the inside-sphere regime, and then maps the wording to the option. A second failure mode is granularity misread: CoT carries over exemplar priors shown on Figure 11, about monthly limits and answers "4," whereas Reasoning-On reparses the temporal cue (per week) and validates against the weekly constraint before selecting "2" (Business Strategy & Management; A). More generally, Reasoning-On performs lightweight option-checking after resolving the operative cue (temporal/categorical/physical), which prevents near-miss mappings and salience/round-number bias. The net effect is not longer chains, but earlier branching to the correct regime and a final consistency check with the provided options.

6 Conclusion

We introduced BNMMLU, a 41-subject Bengali benchmark of 134,375 multiple-choice questions spanning STEM, Humanities, Social Sciences, and Others, and BNMMLU-HARD, a stress-test subset constructed from items that strong systems most often miss. To support Bengali evaluation, we preserve mathematical content, normalize OCR-derived text, apply de-duplication and training-test decontamination analyses. We benchmark proprietary and open-weight LLMs under a controlled protocol covering prompting style (Direct vs. CoT), context regime (0-shot vs. 5-shot), and explicit reasoning configurations. Results show that proprietary systems lead & the best open-weight models narrow the gap; gains are largest when reasoning is enabled, especially on BNMMLU-HARD. Scaling trends improve accuracy but exhibit diminishing returns and meaningful cross-family differences, suggesting that data and post-training recipe quality matter beyond parameter count. We also analyze robustness to question length and subject-specific failure modes to highlight where current models remain brittle.

Limitations

We evaluate text-only capabilities and do not cover multimodal settings (vision-aided reasoning), so the results may not reflect performance in real-world multimodal use cases. While we tested a broad set of models, we were constrained by compute and access costs; therefore, some newer, larger or more expensive frontier models (and larger-scale tuning/inference setups) were not included, which could shift absolute performance levels. Broader human evaluation across all 41 subjects would require substantial expert time and financial resources, so we leave it for future work.

Ethical Statement

The dataset is publicly available under the CC BY-SA 4.0 license, ensuring free accessibility. We release the dataset at https://huggingface.co/datasets/samanjoy2/BnMMLU and the codes at https://github.com/samanjoy2/bnmmlu.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Acknowledgment

No external funding was received.

References

Tanjim Taharat Aurpa, Md Shoaib Ahmed, Richita Khandakar Rifat, Md. Musfique Anwar, and A.B.M. Shawkat Ali. 2023. Uddipok: A reading comprehension based question answering dataset in bangla language. *Data in Brief*, 47:108933.

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. ECollection 2025 Apr.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

DeepSeek-AI. 2025. Deepseek-v3.2-exp: Boosting long-context efficiency with deepseek sparse attention.

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

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. *Ethnologue: Languages of the World*, 28th edition. SIL International, Dallas, Texas. Online reference work.

Syed Mohammed Sartaj Ekram, Adham Arik Rahman, Md. Sajid Altaf, and Mohammed Saidul Islam. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Taisei Enomoto, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. 2025. A fair comparison without translationese: English vs. target-language instructions for multilingual LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 649–670, Albuquerque, New Mexico. Association for Computational Linguistics.

Google. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

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*.

Daeen Kabir, Minhajur Rahman Chowdhury Mahim, Sheikh Shafayat, Adnan Sadik, Arian Ahmed, Eunsu Kim, and Alice Oh. 2025. Bluck: A benchmark dataset for bengali linguistic understanding and cultural knowledge. *Preprint*, arXiv:2505.21092.

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.

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

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.

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.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025. Is translation all you need? a study on solving multilingual tasks with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9594–9614, Albuquerque, New Mexico. Association for Computational Linguistics.

Shahriar Kabir Nahin, Rabindra Nath Nandi, and Sagor Sarker. 2025a. Titullms: A family of bangla llms with comprehensive benchmarking. *Preprint*, arXiv:2502.11187.

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. 2025b. TituLLMs: A family of Bangla LLMs with comprehensive benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24922–24940, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2025. Gpt-5-mini system card.

Pedro Javier Ortiz Su'arez, Laurent Romary, and Benoit Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Nishat Raihan and Marcos Zampieri. 2025. TigerLLM - a family of Bangla large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 887–896, Vienna, Austria. Association for Computational Linguistics.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. 2025. A comprehensive survey of contamination detection methods in large language models. *Preprint*, arXiv:2404.00699.

Md Rashad Al Hasan Rony, Sudipto Kumar Shaha, and Rakib Al Hasan. 2024. Banglaquad: A bengali open-domain question answering dataset. *Preprint*, arXiv:2410.10229.

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.

Sanjay Suryanarayanan, Haiyue Song, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, and Raj Dabre. 2025. Pralekha: Cross-lingual document alignment for indic languages. *Preprint*, arXiv:2411.19096.

Gemma Team. 2025a. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzm'an, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Abdullah Khan Zehady, Shubhashis Roy Dipta, Naymul Islam, Safi Al Mamun, and Santu Karmaker. 2025. Banglallama: Llama for bangla language. *Preprint*, arXiv:2410.21200.

A Task Categories

The task types include a broad range of academic and professional topics, each addressing a specific domain of expertise and practice. The subject list and its tested concepts are in Table 12.

Humanities. Focuses on language, literature, philosophy, and ethics. Core areas include Bengali language & syntax, Bengali literature and poetry, formal logic and critical thinking, comparative religion, and moral & ethical studies. Coverage balances textual analysis with argumentation and value-oriented topics.

STEM (Science, Technology, Engineering, Mathematics). Emphasizes quantitative reasoning, natural sciences, and computing. Mathematics spans elementary topics, algebra & number theory, calculus & analysis, and statistics: probability & inference; physics includes mechanics, thermodynamics & electromagnetism, conceptual physics (basic laws), and relativity & modern physics; chemistry covers physical & analytical, inorganic, and organic subfields; life sciences include cell biology & genetics, human biology & anatomy, and ecology & environmental biology. Computing tracks programming & algorithms, networking & security, AI & data science basics, plus general science integration.

Social Sciences. Covers institutions, markets, and human behavior. Economics, banking & investment, financial accounting, and corporate finance sit alongside business strategy & management, production & operations, and entrepreneurship. The domain also includes civics & governance, geography, history & culture, cognitive and behavioral psychology, and social work & welfare.

Others. Includes general knowledge and global/current affairs, ranging from sports, arts, and media to international organizations, events, and world politics. Coverage reflects publicly available sources up to September 2024.

B OCR & Post-Correction Details

Printed book pages were scanned at 300 dpi into lossless TIFF images. Example scanned pages are shown in Figure 5. Then these images were pre-processed via (i) grayscale conversion, (ii) Sauvola adaptive thresholding, and (iii) Hough-transform deskewing before text extraction. We then employed EASYOCR (V1.7.1)[7] with its Bengali language model to obtain raw transcriptions. OCR output was cleaned and formatted using GPT-3.5-TURBO-0125[8] via the OpenAI API, with the Bengali copy-editing prompt shown in Figure 4. Post-correction reduced formatting issues and spelling errors; additionally, approximately 10% of question–option pairs were manually reviewed for quality assurance.

Bengali Copy-Editing Prompt

You are a Bengali copy-editor. Correct spelling and grammar only. Preserve MathML math, numerals and answer keys unchanged. Return just the corrected line.

Figure 4: Prompt used for Bengali copy-editing, formatted consistently with our evaluation prompt boxes.

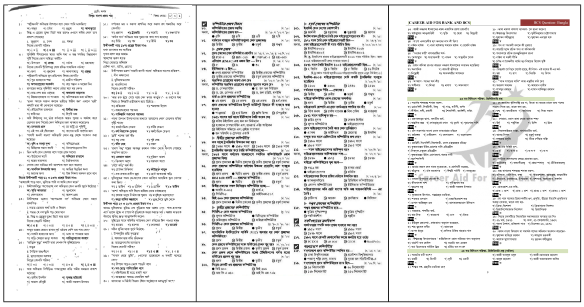

Figure 5: Sample scanned pages of Bengali multiple-choice questions collected from academic and preparatory guidebooks.

C Duplicate-Question Detection and De-duplication

Each question–option pair was embedded into a 1536-dimensional semantic space using the *text-embedding-3-small*[9] model and approximate nearest-neighbor (ANN) search with the angular metric was used to identify semantically similar items. For each question q_i, the top-k neighbors $\{q_j\}$ were retrieved and similarity was computed as Equation 1.

$$s(q_i, q_j) = 1 - \frac{d(q_i, q_j)}{2} \tag{1}$$

In the Equation 1, $d(\cdot, \cdot)$ is the ANN angular distance. Pairs with $s(q_i, q_j) \geq 0.90$ were flagged

[7] https://github.com/JaidedAI/EasyOCR

[8] https://platform.openai.com/docs/models/GPT-3.5-Turbo

[9] https://platform.openai.com/docs/models/text-embedding-3-small

as duplicates. These pairs formed an undirected graph $G = (V, E)$, whose connected components defined duplicate clusters. One canonical item per cluster was retained according to a deterministic rule, yielding a de-duplicated and semantically balanced benchmark. The algorithm is shown in Algorithm 1.

Algorithm 1 Duplicate-Question Detection and De-duplication

Require: Dataset $\mathcal{Q} = \{q_1, \dots, q_N\}$; neighbors k; similarity threshold τ=0.90
Ensure: Deduplicated set $\mathcal{Q}'$
Initialize graph $G \leftarrow \varnothing$
for each $q_i \in \mathcal{Q}$ **do**
    $v_i \leftarrow \textsc{Embed}(q_i) \in \mathbb{R}^{1536}$
end for
Build ANN index over $\{v_i\}_{i=1}^{N}$
for each $q_i \in \mathcal{Q}$ **do**
    $\mathcal{N} \leftarrow \textsc{TopKNeighbors}(v_i, k)$
    for each $q_j \in \mathcal{N}$ **do**
        $s \leftarrow 1 - d(v_i, v_j)/2$
        if $i \neq j$ **and** $s \geq \tau$ **then**
            Add edge (i, j) to G
        end if
    end for
end for
Find connected components $\{C_1, \dots, C_m\}$ of G
For each component C_ℓ, retain one canonical question and discard the rest
return $\mathcal{Q}'$

D Training-test Decontamination Details

To more precisely quantify possible training contamination on LLMs, we perform an n-gram decontamination analysis between our multiple-choice test set (questions including answer options) and a broad collection of Bengali corpora and pre-training datasets that are publicly documented or known to be used in at least some of the evaluated models. Because around 80% of BnMMLU is sourced from web-based question banks, this analysis is critical for ruling out benchmark inflation due to memorization.

Preprocessing and n-gram extraction. For each test question, we apply Unicode NFKC normalization and collapse consecutive whitespace. We then concatenate the question stem with all answer options into a single sequence, tokenize via simple whitespace splitting and extract all contiguous 13-grams (sequences of 13 tokens). We adopt 13-grams following the GPT-3 contamination protocol and subsequent studies, which treat a shared 13-token span between training and evaluation text as a conservative indicator of near-verbatim reuse rather than incidental overlap (Brown et al., 2020; Ravaut et al., 2025).

Corpora and contamination criterion. For each candidate corpus, we stream through the training split and compute the set of 13-grams for every document. A test question is marked as contaminated if any of its 13-grams appear in any training document from at least one corpus. This yields both per-corpus contamination rates and an overall contamination flag per question.

Results. Table 6 reports the per-corpus contamination statistics. For Pralekha, Bangla-Instruct, Bangla-TextBook, IndicCorp, OSCAR, CC100, and TituLM, fewer than 0.1% of test questions contain any overlapping 13-gram.

E Prompting Styles

Direct (No-CoT). Models are prompted *without* any instruction to explain or "think step by step." The prompt states the task and requests only the final answer. No intermediate reasoning cues or scaffolded hints are provided.

Chain-of-Thought (CoT). Models are explicitly invited to reason before giving the final answer. Prompts include a short instruction to first provide reasoning and then the answer. Answer must be clearly marked at the end.

F Context Regimes

Zero-shot (0-shot). No exemplars are given; the model receives only the task instruction and the test item (plus CoT cue when applicable). 0-shot Direct and CoT example prompts are given in Figure 6(a) and Figure 6(b). Bengali 0-shot Direct and CoT prompts are shown in Figure 12(a) and Figure 12(b).

Five-shot (5-shot). We supply five worked exemplars per *subdomain*. Each exemplar contains the question, a correct answer, and (for CoT) a concise reasoning trace. The same exemplars are reused for all test items within that subdomain. 5-shot Direct and CoT example prompts are given in Figure 6(c) and Figure 6(d). Bengali 5-shot Direct and CoT prompts are shown in Figure 12(c) and Figure 12(d).

G Reasoning Configurations

Reasoning-On (internal). For models that has an internal "reasoning" or "thinking" mode, we additionally evaluate a *Reasoning-On* configuration in the 0-shot setting. Instead of injecting explicit CoT exemplars into the prompt, we enable the provider's reasoning controls so that model generates and uses its internal reasoning traces.

Non-Reasoning-On (internal). We run reasoning-capable models with their explicit reasoning or "thinking" features disabled, using each provider's control parameter (e.g., `reasoning_effort`, `thinking_budget`) to suppress chain-of-thought tokens and approximate a standard non-reasoning chat setting. For GPT-5-MINI specifically, we set `reasoning_effort = minimal` and `verbosity = low`; according to OpenAI's documentation, this configuration greatly reduces visible reasoning tokens.

0-shot Direct Eval Prompt

Question: {Question}
A. {A} B. {B}
C. {C} D. {D}
Answer. Output in JSON: `{"answer":"A/B/C/D"}`

(a)

0-shot CoT Eval Prompt

Question: {Question}
A. {A} B. {B}
C. {C} D. {D}
Think step by step. After thinking, provide the final answer in JSON: `{"answer":"A/B/C/D"}`

(b)

5-shot Direct Eval Prompt

Example 1
Question: {Q1}
A. {A1} B. {B1}
C. {C1} D. {D1}
Answer: {A/B/C/D}
Example 2
...
Now solve:
Question: {Question}
A. {A} B. {B}
C. {C} D. {D}
Answer:
`{"answer":"A/B/C/D"}`

(c)

5-shot CoT Eval Prompt

Example 1
Question: {Q1}
A. {A1} B. {B1}
C. {C1} D. {D1}
Reasoning: {Reasoning}
Answer: {A/B/C/D}
Example 2
...
Now solve:
Question: {Question}
A. {A} B. {B}
C. {C} D. {D}
Think step by step. Answer:
`{"answer":"A/B/C/D"}`

(d)

Figure 6: Prompts used in our evaluation: 0-shot (Direct, CoT) and 5-shot (Direct, CoT).

CoT Elicitation Prompt

The following is a MCQ question {subject}. Produce a detailed step-by-step explanation in English. Answer and reasoning:
`{"answer":"A/B/C/D", "reasoning":"..."}`
Question: {Question}
A. {A} B. {B} C. {C} D. {D}

Figure 7: Prompt used to create the CoT selected questions' reasonings for CoT evaluation.

H Scaling & Family Effects

Across families, the scaling plot in Figure 8 (ExaFLOP vs. average accuracy) shows mostly monotonic "family ladders": larger, higher-compute checkpoints outperform smaller ones, but gains taper as compute rises. At comparable compute, noticeable cross-family gaps persist-pointing to differences in data curation, pretraining mix, and instruction-tuning rather than scale alone. Bengali-centric families are competitive in the low–mid compute band yet appear to plateau earlier than the largest bilingual/global families.

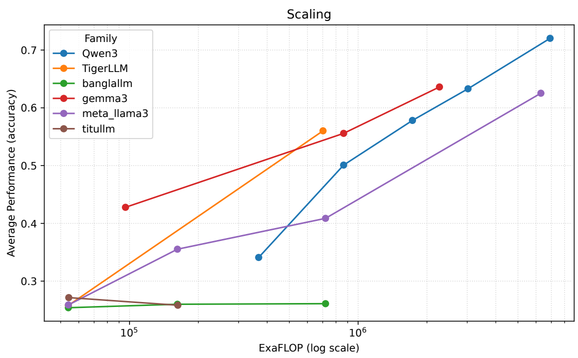

Figure 8: Average accuracy in 0-shot Direct (Non-Reasoning) versus estimated training compute (ExaFLOP; log scale). ExaFLOP is estimated as $6 \times \text{params}_\text{B} \times \text{train\_tokens}_\text{B}$ (both in billions), following Scaling Laws (Kaplan et al., 2020).

I Sequence-Length Robustness

Setup. To quantify how reliably each model handles longer contexts, we measure error rates as a function of question length. Let q denote a question, m a model and $|q|$ the number of characters in q. The procedure is formalised in Equations 2–7.

$$\text{Length}(q) = |q| \tag{2}$$

$$\text{Bin}(|q|) = \text{bin}_i \quad \text{if} \quad \text{bin}_{i-1} < |q| \leq \text{bin}_i \tag{3}$$

$$E(q,m) = \begin{cases} 0, & \text{if } m \text{ answers } q \text{ correctly} \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

$$A(m, \text{bin}_i) = 1 - \frac{\sum_{q:\, \text{Bin}(|q|)=\text{bin}_i} E(q,m)}{n_i} \tag{5}$$

$$ER(m, \text{bin}_i) = 1 - A(m, \text{bin}_i) \tag{6}$$

$$ER(m) = \frac{\sum_q E(q,m)}{N} \tag{7}$$

The length bins are fixed at $\{0, 20, 40, 60, 80, 100\}$, n_i is the number of questions falling in bin_i and N is the total number of questions. yields the length-specific error rate.

J Subject-Specific Failure Modes

Setup. To better understand how language models perform across different subjects, we analyze their subject-wise accuracy and variability. This analysis identifies which subjects are consistently easy or difficult for most models and which ones reveal significant disagreement. We assign subjects to the four regions using the criteria in Table 5.

Quadrant	Condition
Difficult & Inconsistent	$\text{Avg}_s < \mu_x \wedge \text{SD}_s > \mu_y$
Easy & Inconsistent	$\text{Avg}_s > \mu_x \wedge \text{SD}_s > \mu_y$
Difficult & Consistent	$\text{Avg}_s < \mu_x \wedge \text{SD}_s < \mu_y$
Easy & Consistent	$\text{Avg}_s > \mu_x \wedge \text{SD}_s < \mu_y$

Table 5: Quadrant definitions for subject difficulty versus consistency based on average accuracy (Avg_s) and standard deviation (SD_s) thresholds μ_x and μ_y.

Here, $\text{Accuracy}_{s,i}$ denote the accuracy of model i on subject s and N the number of evaluated models.

$$\text{Avg}_s = \frac{1}{N} \sum_{i=1}^{N} \text{Accuracy}_{s,i} \tag{8}$$

$$\text{SD}_s = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\text{Accuracy}_{s,i} - \text{Avg}_s\right)^2} \tag{9}$$

Avg_s serves as the X-coordinate and SD_s as the Y-coordinate in the *Subject Difficulty vs. Consistency* plot in Figure 3.

K Compute Resources

Open-weight models were evaluated on an internal compute node with *1 × NVIDIA RTX A6000* (48GB) for smaller models and *2 × NVIDIA RTX PRO 6000* (192GB) for larger models. Open-weight models were executed at their highest native precision; typically `bfloat16/float16` with no quantization. Proprietary models were accessed via their official APIs using identical prompts and decoding parameters to ensure comparability across systems.

Bengali (Literature)

Question (Bengali). 'টিকিয়া থাকাই চরম সার্থকতা নয়.'- মন্তব্যটি কোন্ রচনা থেকে নেওয়া?

Translation: *'Merely surviving is not the ultimate success.' - From which composition is this comment taken?*

Options. A. হৈমন্তী *(Haimanti)*
B. বিলাসী *(Bilasi)*
C. অর্ধাঙ্গী *(Ardhangi)*
D. যৌবনের গান. *(The Song of Youth)*

RIGHT - 0-shot-cot-nonreasoning → B
Model: gemini-2.5-flash

WRONG - 5-shot-cot-nonreasoning → D
Model: gemini-2.5-flash

Figure 9: Exemplar-induced overthinking: 5-shot CoT gravitates to a salient title (D), while 0-shot selects the correct source (B).

Mechanics

Question (Bengali). পৃথিবীর কেন্দ্র হতে ভূপৃষ্ঠ পর্যন্ত দূরত্ব বৃদ্ধির সাথে সাথে 'g'-এর মান-

Translation: *From the center of the Earth to the surface, with increasing distance, the value of 'g' -*

Options. A. সমানুপাতে হ্রাস পায় *(decreases proportionally)*
B. সমানুপাতে বৃদ্ধি পায় *(increases proportionally)*
C. ব্যস্তানুপাতে হ্রাস পায় *(decreases inversely)*
D. ব্যস্তানুপাতে বৃদ্ধি পায় *(increases inversely)*

WRONG - 5-shot CoT (non-reasoning) → C
MODEL REASONING: …$g \propto 1/r^2$ …as r increases, g decreases …"ব্যস্তানুপাতে হ্রাস পায়" matches inverse square …`{"answer":"C"}` …
Context: CoT assumes **outside-Earth regime** and never switches to the inside-Earth model.

RIGHT - 0-shot Reasoning-On → B
MODEL REASONING: …**inside Earth** ($r < R$), $g \propto r$ …from center → surface, g **increases proportionally** …**Therefore B** …
Context: Reasoning-on selects the **correct regime** (interior of a uniform sphere) and **maps wording → option.**

Figure 10: Physics: Reasoning-On toggles to the inside-sphere linear model ($g \propto r$), correcting CoT's **inverse-square overgeneralization**.

Business Strategy & Management

Question (Bengali). একটি বাণিজ্যিক ব্যাংক সঞ্চয়ী হিসাব খুলে তুমি সপ্তাহের সর্বাধিক কতবার টাকার তুলতে পারো?

Translation: *By opening a savings account in a commercial bank, how many times at most can you withdraw money* ***per week****?*

Options. A. ২ বার *(2 times)*
B. ৩ বার *(3 times)*
C. ৪ বার *(4 times)*
D. যতবার ইচ্ছা *(As many times as desired)*

WRONG - 5-shot CoT (non-reasoning) → C
MODEL REASONING: …savings accounts have a limit …up to **4 withdrawals** …Therefore **4** …`{"answer":"C"}`…
Context: CoT defaults to **monthly limits** learned from exemplars; misreads the **"per week"** cue.

RIGHT - 0-shot Reasoning-On → A
MODEL REASONING: …interpret as **"per week"** …savings accounts limited withdrawals …**commonly 2 per week** …**Therefore A** …
Context: Reasoning-on **re-parses the Bengali phrase**, aligns to the **weekly rule**, then **option-checks**.

Figure 11: Strategy: Reasoning-On corrects **temporal granularity** (weekly vs. monthly), avoiding CoT's **exemplar-driven heuristic**.

L Prompt Language Choice and Potential Bias

We adopt English system prompts because cross-lingual prompting studies find that English instructions can be more effective than target-language instructions in multilingual settings, and that direct comparisons can be confounded by translation artifacts ("translationese") (Enomoto et al., 2025). This effect is not universal: performance can vary by task type and by whether the model is English-centric or non-English-centric, so prompt language should be treated as an empirical variable rather than an assumption (Liu et al., 2025). Selected Bengali-versus-English prompt comparisons on BNMMLU-HARD are reported in Table 11.

This design choice also matches the training reality of several Bengali-capable checkpoints: many are produced via continual pretraining from strong base models that were originally trained heavily on English and multilingual corpora. TigerLLM is built by starting from Gemma-2 (9B) (and LLaMA 3.2 for smaller scale) and then continuing pretraining on Bangla resources before instruction tuning (Raihan and Zampieri, 2025).

0-shot Direct Eval Prompt

প্রশ্ন: {Question}
বিকল্পসমূহ::
A) {A} B) {B}
C) {C} D) {D}
সরাসরি সঠিক উত্তর দিন। শুধু A, B, C, বা D লিখুন। অন্য কিছু লিখবেন না।

(a)

0-shot CoT Eval Prompt

প্রশ্ন: {Question}
বিকল্পসমূহ::
A) {A} B) {B}
C) {C} D) {D}
উত্তর দেওয়ার আগে ধাপে ধাপে ভাবুন (সংক্ষিপ্ত ও নির্ভুলভাবে)। বিশ্লেষণ শেষ হলে শেষ লাইনে ঠিক একটি JSON অবজেক্ট লিখুন: `{"উত্তর":"A"}` যেখানে A/B/C/D হবে আপনার পছন্দ। অন্য কিছু লিখবেন না।

(b)

5-shot Direct Eval Prompt

উদাহরণ প্রশ্ন: {Question}
বিকল্পসমূহ::
A) {A1} B) {B1}
C) {C1} D) {D1}
উত্তর: {Answer}
…
এখন নিচের প্রশ্নটির উত্তর দিন::
প্রশ্ন: {question}
বিকল্পসমূহ::
A) {A2} B) {B2}
C) {C2} D) {D2}
সরাসরি সঠিক উত্তর দিন। শুধু A, B, C, বা D লিখুন। অন্য কিছু লিখবেন না।

(c)

5-shot CoT Eval Prompt

উদাহরণ প্রশ্ন: {Question}
বিকল্পসমূহ::
A) {A1} B) {B1}
C) {C1} D) {D1}
সংক্ষিপ্ত ব্যাখ্যা: {Reasoning}
উত্তর: {Answer}
…
এখন নিচের প্রশ্নটির উত্তর দিন::
প্রশ্ন: {question}
বিকল্পসমূহ::
A) {A2} B) {B2}
C) {C2} D) {D2}
উত্তর দেওয়ার আগে ধাপে ধাপে ভাবুন (সংক্ষিপ্ত ও নির্ভুলভাবে)। বিশ্লেষণ শেষ হলে শেষ লাইনে ঠিক একটি JSON অবজেক্ট লিখুন: `{"উত্তর":"A"}` যেখানে A/B/C/D হবে আপনার পছন্দ। অন্য কিছু লিখবেন না।

(d)

Figure 12: Bengali Prompts used in our evaluation: 0-shot (Direct, CoT) and 5-shot (Direct, CoT).

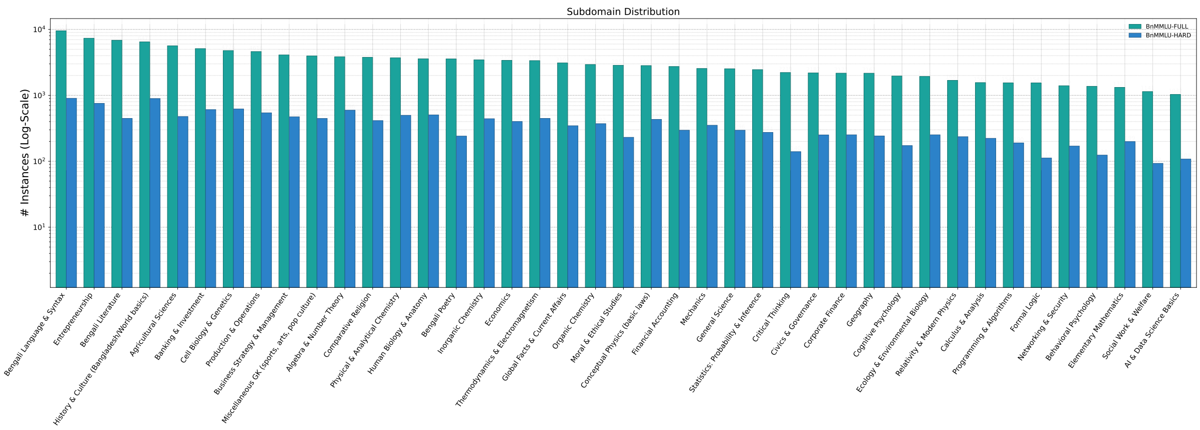

Figure 13: Subject-wise counts for BnMMLU-FULL and BnMMLU-HARD.

Corpus (dataset / split)	# Ref. Exp.	# Cont. Qs.	# Cont. Qs. (%)	# Unq. 13-g
Pralekha (ben) (Suryanarayanan et al., 2025)	95,813	0	0.00	0
Pralekha (eng-ben) (Suryanarayanan et al., 2025)	86,815	0	0.00	0
Pralekha (unal / ben) (Suryanarayanan et al., 2025)	47,906	1	0.00	2
TituLM Corpus (Nahin et al., 2025a)	31,225,356	122	0.09	239
IndicCorpV2 (asm–Beng) (Doddapaneni et al., 2023)	1,256,513	0	0.00	0
IndicCorpV2 (Beng) (Doddapaneni et al., 2023)	13,553,516	30	0.02	80
OSCAR (bn) (Ortiz Su'arez et al., 2020)	14,346,126	34	0.02	79
Bangla-Instruct (Instruction) (Raihan and Zampieri, 2025)	268,145	4	0.00	3
Bangla-Instruct (Response) (Raihan and Zampieri, 2025)	329,872	49	0.04	219
Bangla-TextBook (Raihan and Zampieri, 2025)	87,105	48	0.03	209
CC100 (bn) (Wenzek et al., 2020)	12,427,522	72	0.05	141

Table 6: 13-gram decontamination statistics. A question is marked as contaminated if its normalized text (question plus answer options) shares at least one contiguous 13-gram with any example in the corresponding training corpus.

Model	# Params	Access	Language
English-Centric / Bilingual Instruction-Tuned Models			
Llama-3.x-Instruct (Grattafiori et al., 2024)	1B, 3B, 8B, 70B	Weights Available	En / Multilingual
Qwen3 (Team, 2025b)	1.7B, 4B, 8B, 14B, 32B	Weights Available	En / Zh
Gemma-3-IT (Team, 2025a)	4B, 12B, 27B	Weights Available	En / Multilingual
Bengali Pretrained / Instruction-Tuned Models			
TituLLM (Nahin et al., 2025b)	1B, 3B	Weights Available	Bn / En
TigerLLM-IT (Raihan and Zampieri, 2025)	1B, 9B	Weights Available	Bn
BanglaLlama Instruct (Zehady et al., 2025)	1B, 3B, 8B	Weights Available	Bn
Proprietary Models			
GPT-5-Mini[*](OpenAI, 2025)	undisclosed	API	–
Grok 4 Fast[a]	undisclosed	API	–
Gemini 2.5 Flash (Google, 2025)	undisclosed	API	–
DeepSeek-V3.2-Exp (DeepSeek-AI, 2025)	685B	API / Weights Available	–
Qwen-Plus[b]	undisclosed	API	–

[a] https://x.ai/news/grok-4-fast
[b] https://www.alibabacloud.com/help/en/model-studio/models
[*] It has reasoning capability but cannot be fully disabled and thus, we use minimal reasoning when mentioning no-reasoning.

Table 7: Overview of evaluated models, grouped by family.

SL	Subject Name	BnMMLU-FULL	BnMMLU-HARD
1	Elementary Mathematics	1321	199
2	Algebra & Number Theory	3849	596
3	Calculus & Analysis	1561	223
4	Statistics: Probability & Inference	2456	274
5	Mechanics	2555	352
6	Conceptual Physics (basic laws)	2827	432
7	Thermodynamics & Electromagnetism	3365	447
8	Relativity & Modern Physics	1687	236
9	Physical & Analytical Chemistry	3713	498
10	Inorganic Chemistry	3458	442
11	Organic Chemistry	2934	372
12	Cell Biology & Genetics	4776	621
13	Human Biology & Anatomy	3585	505
14	Ecology & Environmental Biology	1936	252
15	Agri Sciences	5646	477
16	Networking & Security	1399	170
17	Programming & Algorithms	1548	190
18	AI & Data Science Basics	1032	108
19	General Science	2531	296
20	Bengali Language & Syntax	9528	902
21	Bengali Literature	6849	447
22	Bengali Poetry	3570	241
23	Comparative Religion	3776	414
24	Moral & Ethical Studies	2854	231
25	Formal Logic	1545	112
26	Critical Thinking	2219	140
27	Economics	3394	401
28	Banking & Investment	5118	608
29	Financial Accounting	2738	296
30	Corporate Finance	2171	252
31	Business Strategy & Management	4102	473
32	Production & Operations	4616	543
33	Entrepreneurship	7350	755
34	Cognitive Psychology	1963	173
35	Behavioral Psychology	1371	124
36	Civics & Governance	2187	251
37	Geography	2167	242
38	History & Culture	6473	894
39	Social Work & Welfare	1141	93
40	Miscellaneous GK	3961	446
41	Global Facts & Current Affairs	3103	346
Total		**134,375**	**15,074**

Table 8: Subject sizes in BnMMLU-FULL and BnMMLU-HARD.

Model	STEM	Humanities	Social Sciences	Others	Overall (Δ)
English-Centric / Bilingual Instruction-Tuned Models					
LLAMA-3.2-1B-INSTRUCT	25.88	27.22	27.26	26.71	26.69 (+1.69)
LLAMA-3.2-3B-INSTRUCT	35.53	34.91	41.34	40.30	37.68 (+12.68)
LLAMA-3.1-8B-INSTRUCT	40.87	39.49	47.14	46.89	43.07 (+18.07)
LLAMA-3.3-70B-INSTRUCT	62.53	52.47	65.81	68.99	61.87 (+36.87)
QWEN3-1.7B	34.10	34.58	39.70	36.00	36.25 (+11.25)
QWEN3-4B	50.09	40.60	48.46	47.21	47.27 (+22.27)
QWEN3-8B	57.82	43.96	51.87	53.83	52.51 (+27.51)
QWEN3-14B	63.30	49.37	59.23	61.13	58.76 (+33.76)
QWEN3-32B	72.03	53.01	65.57	66.44	65.34 (+40.34)
GEMMA-3-4B-IT	42.78	39.46	47.82	47.95	44.07 (+19.07)
GEMMA-3-12B-IT	55.58	47.50	59.13	59.82	55.26 (+30.26)
GEMMA-3-27B-IT	63.61	51.43	63.90	67.44	61.27 (+36.27)
Bengali Pretrained / Instruction-Tuned Models					
TITULLM-1B	27.15	27.53	28.42	28.19	27.72 (+2.72)
TITULLM-3B	25.83	27.59	27.70	26.56	26.87 (+1.87)
TIGERLLM-1B-IT	25.80	27.47	27.42	26.11	26.73 (+1.73)
TIGERLLM-9B-IT	56.02	47.85	59.48	61.29	55.70 (+30.70)
BANGLALLAMA-3.2-1B-INSTRUCT	25.40	27.40	27.07	25.93	26.40 (+1.40)
BANGLALLAMA-3.2-3B-INSTRUCT	26.00	27.56	27.44	26.26	26.82 (+1.82)
BANGLALLAMA-3.1-8B-INSTRUCT	26.10	27.56	27.58	26.52	26.95 (+1.95)
Proprietary Models					
GPT-5-MINI	48.25	43.96	55.00	55.78	50.09 (+25.09)
GROK 4 FAST	61.98	51.63	64.02	67.60	60.82 (+35.82)
GEMINI 2.5 FLASH	72.38	**62.32**	**71.08**	**73.85**	**69.85 (+44.85)**
DEEPSEEK-V3.2-EXP	72.72	58.62	70.06	73.84	68.82 (+43.82)
QWEN-PLUS	**73.49**	56.15	66.89	70.17	67.29 (+42.29)

Table 9: Average accuracy (%) of models on the BNMMLU-FULL benchmark under 0-shot Direct (Non-Reasoning) evaluation. Bold marks the highest overall score; underlines denote the best model within each category. (Δ) in overall is compared with random baseline (25%).

Model	0-shot Direct (Non-Reasoning)	0-shot CoT (Δ) (Non-Reasoning)	5-shot Direct (Non-Reasoning)	5-shot CoT (Δ) (Non-Reasoning)
English-Centric / Bilingual Instruction-Tuned Models				
LLAMA-3.2-3B-INSTRUCT	19.95	18.33 (-1.62)	22.16	23.25 (+1.09)
LLAMA-3.1-8B-INSTRUCT	19.14	22.91 (+3.77)	21.99	22.62 (+0.63)
LLAMA-3.3-70B-INSTRUCT	23.78	35.17 (+11.39)	31.15	<u>37.50 (+6.35)</u>
QWEN3-1.7B	14.53	21.67 (+7.14)	23.27	23.55 (+0.28)
QWEN3-4B	12.26	19.09 (+6.83)	26.46	29.74 (+3.28)
QWEN3-8B	21.59	20.91 (-0.68)	29.14	28.39 (-0.75)
QWEN3-14B	14.67	14.32 (-0.35)	18.35	16.88 (-1.47)
QWEN3-32B	<u>25.52</u>	28.63 (+3.11)	34.63	31.19 (-3.44)
GEMMA-3-4B-IT	14.85	15.72 (+0.87)	16.78	19.51 (+2.73)
GEMMA-3-12B-IT	10.54	14.55 (+4.01)	18.50	23.52 (+5.02)
GEMMA-3-27B-IT	14.72	<u>37.59 (+22.87)</u>	<u>35.65</u>	34.65 (-1.00)
Bengali Pretrained / Instruction-Tuned Models				
TIGERLLM-9B-IT	<u>11.01</u>	<u>16.78 (+5.77)</u>	<u>18.44</u>	<u>23.32 (+4.88)</u>
Proprietary Models				
GPT-5-MINI	14.13	19.12 (+4.99)	19.66	18.63 (-1.03)
GROK 4 FAST	20.94	20.89 (-0.05)	44.06	51.12 (+7.06)
GEMINI 2.5 FLASH	**34.46**	45.38 (+10.92)	51.62	61.48 (+9.86)
DEEPSEEK-V3.2-EXP	29.89	**59.04 (+29.15)**	**58.83**	**64.53 (+5.70)**
QWEN-PLUS	32.47	58.74 (+26.27)	57.40	55.09 (-2.31)

Table 10: Accuracy (%) on BNMMLU-HARD. Δ is computed as CoT − Direct at the *same shot* (0-shot or 5-shot). **Bold** marks the global best per column; <u>underline</u> marks the best *within each category* per column.

Model	Setting	English Prompt	Bengali Prompt	Δ (Bn−En)
TIGERLLM-9B-IT	0-shot Direct	11.01	11.40	+0.39
TIGERLLM-9B-IT	0-shot CoT	16.78	17.23	+0.45
TIGERLLM-9B-IT	5-shot Direct	18.44	18.19	-0.25
TIGERLLM-9B-IT	5-shot CoT	23.32	23.90	+0.58
QWEN3-14B	5-shot CoT	16.88	16.50	-0.38
GEMMA-3-12B-IT	5-shot CoT	23.52	23.11	-0.41

Table 11: Selected English-Bengali prompt comparisons on BNMMLU-HARD.

SL	Subject Name	Tested Concepts	Supercategory
1	Elementary Mathematics	Arithmetic, Fractions, Ratios, Basic Problem Solving...	STEM
2	Algebra & Number Theory	Equations, Functions, Prime Numbers, Theorems...	STEM
3	Calculus & Analysis	Differentiation, Integration, Sequences, Series...	STEM
4	Statistics: Probability & Inference	Descriptive Statistics, Probability, Hypothesis Testing...	STEM
5	Mechanics	Dynamics, Statics, Kinematics, Laws of Motion...	STEM
6	Conceptual Physics (basic laws)	Motion, Forces, Energy, Newtonian Principles...	STEM
7	Thermodynamics & Electromagnetism	Laws of Thermodynamics, Heat Transfer, Electricity...	STEM
8	Relativity & Modern Physics	Einstein's Theories, Quantum Concepts, Atomic Models...	STEM
9	Physical & Analytical Chemistry	Stoichiometry, Molecular Structure, Spectroscopy...	STEM
10	Inorganic Chemistry	Periodic Table, Coordination Compounds...	STEM
11	Organic Chemistry	Hydrocarbons, Functional Groups, Reactions...	STEM
12	Cell Biology & Genetics	Cell Structure, DNA/RNA, Inheritance, Evolution...	STEM
13	Human Biology & Anatomy	Organ Systems, Physiology, Human Genetics...	STEM
14	Ecology & Environmental Biology	Ecosystems, Biodiversity, Conservation, Sustainability...	STEM
15	Agri Sciences	Agronomy, Crop, Soil Management, Agribusiness...	STEM
16	Networking & Security	Internet Protocols, Cybersecurity, Encryption, Firewalls...	STEM
17	Programming & Algorithms	Python, Logic, Data Structures, Computational Thinking...	STEM
18	AI & Data Science Basics	Machine Learning, Neural Networks, Data Processing...	STEM
19	General Science	Scientific Method, Basic Physics, Chemistry, Biology...	STEM
20	Bengali Language & Syntax	Morphology, Grammar, Sentence Structure, Semantics...	Humanities
21	Bengali Literature	Prose, Poetry, Authors, Literary Devices...	Humanities
22	Bengali Poetry	Poetic Forms, Symbolism, Meter, Notable Poets...	Humanities
23	Comparative Religion	Theology, World Religions, Ethical Teachings...	Humanities
24	Moral & Ethical Studies	Ethics, Values, Philosophy, Social Responsibility...	Humanities
25	Formal Logic	Propositional Logic, Proofs, Logical Systems, Paradoxes...	Humanities
26	Critical Thinking	Logic, Reasoning, Argumentation, Analytical Skills...	Humanities
27	Economics	Microeconomics, Macroeconomics, Fiscal Policy, Trade...	Social Sciences
28	Banking & Investment	Financial Systems, Banking Principles, Securities...	Social Sciences
29	Financial Accounting	Balance Sheets, Cash Flow, Auditing, Cost Analysis...	Social Sciences
30	Corporate Finance	Capital Budgeting, Valuation, Risk Management...	Social Sciences
31	Business Strategy & Management	Strategic Planning, Leadership, Organizational Theory...	Social Sciences
32	Production & Operations	Process Design, Quality Control, Supply Chain...	Social Sciences
33	Entrepreneurship	Startup Models, Business Planning, Innovation...	Social Sciences
34	Cognitive Psychology	Memory, Perception, Decision-Making, Theories...	Social Sciences
35	Behavioral Psychology	Emotions, Behaviorism, Conditioning, Human Interaction...	Social Sciences
36	Civics & Governance	Constitution, Rights, Political Systems, Citizenship ...	Social Sciences
37	Geography	Physical Geography, Climate, Maps, Human Geography...	Social Sciences
38	History & Culture	Historical Events, Heritage, Civilization, Global Affairs...	Social Sciences
39	Social Work & Welfare	Social Policy, Community Engagement, Case Studies...	Social Sciences
40	Miscellaneous GK	Global Trivia, Sports Facts, Entertainment, Arts, Media...	Others
41	Global Facts & Current Affairs	International Organizations, Events, World Politics...	Others

Table 12: Overview of subject domains and tested concepts in BnMMLU.