BenNumEval: A Benchmark to Assess LLMs' Numerical Reasoning Capabilities in Bengali

Kawsar Ahmed[♣], **Md Osama**[♣], **Omar Sharif**[♠♣], **Eftekhar Hossain**[¥], **Mohammed Moshiul Hoque**[♣]

[♠]Department of Computer Science, Dartmouth College, USA
[¥]Department of Computer Science, University of Central Florida, USA
[♣]Department of Computer Science and Engineering,
[♣]Chittagong University of Engineering & Technology, Bangladesh
{u1804017@student.cuet.ac.bd, moshiul_240@cuet.ac.bd}

Abstract

Large Language Models (LLMs) demonstrate exceptional proficiency in general-purpose tasks but struggle with numerical reasoning, particularly in low-resource languages like Bengali. Despite advancements, limited research has explored their numerical reasoning capabilities in these languages. To address this gap, we present *BenNumEval (Bengali Numerical Evaluation)*, a benchmark designed to assess LLMs on numerical reasoning tasks in Bengali. It comprises six diverse tasks and a total of 3.2k samples curated from real-world problem-solving scenarios. Our extensive evaluations reveal that even with advanced prompting techniques such as *Cross-Lingual Prompting (XLP)* and *Cross-Lingual Chain-of-Thought Prompting (XCoT)*, LLMs fall notably short of human-level performance, particularly when using *Bengali Native Prompting (BNaP)*. These findings underscore the substantial gap between current LLM capabilities and human expertise in numerical reasoning, highlighting the need for more robust and linguistically inclusive AI models to advance Bengali Language Processing and equitable AI development. The source code for the system and evaluation pipeline is publicly available on GitHub[1].

1 Introduction

Numerical reasoning is an essential ability that is vital in numerous everyday situations. Given the prevalence of numerical data in textual content, the ability to perform numerical reasoning is essential for interpreting and solving problems encountered in everyday life. This process involves extracting critical details from text and transforming them into mathematical formats.

Recently, LLMs have demonstrated considerable potential in addressing general-purpose tasks within the AI community, showcasing their efficiency across a broad spectrum of applications (Zhao et al., 2023). However, despite these advancements, current state-of-the-art AI systems remain fragile and often need help when presented with mathematical reasoning problems that are posed in a slightly altered way (Bang et al., 2023). This constraint makes LLMs less effective for real-world problem-solving, particularly in handling numerical reasoning tasks requiring multiple reasoning steps (Ahn et al., 2024). While these models perform adequately with simple, single-step tasks, they frequently falter when confronted with more complex problems, such as math word problems, which require the models to comprehend the text, determine the appropriate mathematical operations, and arrive at the correct solution (Shi et al., 2022; Schwartz et al., 2024). This highlights the need for further advancements to enhance their robustness in handling such multifaceted challenges. To address these challenges, various benchmarks have been developed for numerical reasoning. For example, MGSM (Shi et al., 2022), a multilingual benchmark that includes ten languages with diverse linguistic structures. Mishra et al. (2022b) developed NumGLUE and LILA benchmark (Mishra et al., 2022a), which assess how effectively advanced language models can manage numerical reasoning tasks across various problems in English.

This work introduces *BenNumEval* (Bengali Numerical Evaluation), a novel benchmark dataset designed for numerical reasoning tasks in the Bengali language. Currently, there is a lack of dedicated datasets for evaluating State-of-the-Art (SoTA) models on these tasks in Bengali, making it difficult to assess their performance accurately. This limitation highlights a significant gap in evaluating the numerical reasoning capabilities of models in low-resource languages. *BenNumEval* seeks to address this challenge by providing a structured

[1] https://github.com/kawsar-pie/BenNumEval

Findings of the Association for Computational Linguistics: ACL 2025, pages 17782–17799

benchmark, fostering new research opportunities in Bengali Language Processing (BLP), and advancing AI capabilities in underrepresented languages. The main contributions of this work are as follows.

- We introduce *BenNumEval*, a benchmark dataset for evaluating numerical reasoning in Bengali, covering six diverse tasks with 3.2k examples sourced from multiple domains.
- We conduct a comprehensive evaluation of LLMs on *BenNumEval* using *Bengali Native Prompting (BNaP), Cross-Lingual Prompting (XLP),* and *Cross-Lingual Chain-of-Thought prompting(XCoT)*. Results indicate that while advanced prompting improves performance, LLMs still lag behind human-level reasoning, underscoring key challenges in Bengali numerical understanding.

2 Related Work

Numerical reasoning has grown from small-scale to large-scale datasets, supporting deep learning research. This section chronologically reviews vital datasets and models, showcasing their impact on advancing mathematical reasoning in NLP.

Benchmark Datasets. Early works in this domain focused on small-scale datasets aimed at understanding the quantitative aspects of natural language. For instance, Kushman et al. (2014) introduced a template-based dataset that solved questions with equations as parameters. This was followed by the addition-subtraction dataset by Hosseini et al. (2014), which focused on simple arithmetic operations, and a dataset for arithmetic problems developed by Koncel-Kedziorski et al. (2015). These datasets were foundational in addressing basic arithmetic reasoning tasks.

Over time, the complexity of arithmetic reasoning questions has steadily increased. Roy and Roth (2016) and Upadhyay et al. (2016) developed more challenging datasets, pushing the boundaries of arithmetic reasoning. As research progressed, larger datasets were introduced, notably by Ling et al. (2017) and Dua et al. (2019), to support deep learning models. To enhance diversity and explainability, Amini et al. (2019) and Miao et al. (2021) created datasets with varied problem types. Further advancements were made by Zhang et al. (2020) and Hendrycks et al. (2021), who focused on scale information in embeddings and increasing question difficulty. Mishra et al. (2022b) introduced two benchmarks: NumGLUE, with 100K problems across eight tasks to assess arithmetic reasoning, and LILA (Mishra et al., 2022a), which combines 20 datasets for mathematical reasoning, categorized into 23 tasks. Both benchmarks highlight the difficulty for advanced models, with GPT-3 (Brown, 2020) struggling on NumGLUE and RoBERTa (Liu, 2019) performing poorly on SVAMP, a 1,000-problem challenge set developed by Patel et al. (2021). Recent advancements in arithmetic reasoning benchmarks include GSM8K by (Cobbe et al., 2021), featuring 8.5K diverse grade-school math problems, and MATH (Hendrycks et al., 2021) dataset with 12,500 complex competition-level problems. Both highlight the difficulty models face in solving these tasks. Shi et al. (2022) introduced MGSM, testing multilingual reasoning with 250 GSM8K problems in 10 languages. Additionally, the Bengali Math Word Problem (BMWP) dataset by Mondal et al. (2023) contains 8,653-word problems, focusing on operator recognition rather than the mathematical reasoning addressed in this study.

Models. In addition to general-purpose LLMs, researchers have increasingly focused on enhancing these models' mathematical and logical reasoning capabilities (Ahn et al., 2024). This has led to a recent trend of developing LLMs specifically tailored for mathematical reasoning, such as DeepSeekMath (Shao et al., 2024), Mathstral-7B[2], Qwen-2.5-math (Yang et al., 2024), InternLM2-math (Ying et al., 2024), Llemma (Azerbayev et al., 2023), and WizardMath (Luo et al., 2023). These models are explicitly designed to improve performance on math reasoning tasks.

Most existing research focuses on high-resource languages like English, creating a major gap in resources for evaluating numerical reasoning in low-resource languages. To address this disparity, we introduce *BenNumEval*, a comprehensive benchmark dataset for numerical reasoning in Bengali. Besides this work aims to facilitate the evaluation of NLP models in a low-resource setting and contribute to the development of robust multilingual and cross-lingual model capabilities.

3 Overview of Tasks in *BenNumEval*

Evaluating LLMs on numerical reasoning is challenging, particularly when creating a benchmark

[2] https://mistral.ai/news/mathstral/

in Bengali due to the lack of resources. To address this, we introduce *BenNumEval*, a benchmark dataset with six tasks focused on numerical reasoning in Bengali (see Table 1).

Commonsense + Arithmetic (CA). Consider the problem: "আবুল বোর্ডে প্রতিটি বাংলা স্বরবর্ণ লিখেছিলেন। যদি প্রতিটি স্বরবর্ণ ৭ বার লেখা হয়। বোর্ডে মোট কতটি বর্ণমালা লেখা ছিল?" *(Abul wrote each Bengali vowel on the board. If each vowel was written 7 times, how many letters were written on the board in total?)* Solving this requires both arithmetic reasoning and commonsense knowledge, such as recognizing that Bengali has 11 vowels. These problems combine numerical commonsense facts (e.g., a hand has 5 fingers, a week has 7 days) with arithmetic challenges.

Domain-Specific (DS). "১২ রেডিয়ানকে ডিগ্রিতে প্রকাশ করলে কত হবে?" *(What is the value of 12 radians in degrees?)* This task requires numerical reasoning combined with domain-specific knowledge, such as applying the formula, $Degrees = \frac{180}{\pi} \times Radians$. The dataset covers various fields, including chemistry, physics, advanced mathematics, and computer science.

Commonsense + Quantitative (CQ). This category involves quantitative comparisons, such as determining if one number is larger or smaller than another. For example: "১২৩৪৫৬৭৮৯০ সংখ্যাটি কি ৯৮৭৬৫৪৩২১০ এর চেয়ে ছোট নাকি বড়? (A) ছোট (B) বড়" *(Is the number 1234567890 smaller or larger than 9876543210? (A) Smaller (B) Larger)*. These questions often require subtraction or other arithmetic operations to solve.

Fill-in-the-blanks Format (FiB). "৪টি কলমের মূল্য ৮০ টাকা। ১০টি কলমের মূল্য ___ টাকা।" *(The price of 4 pens is 80 Taka. The price of 10 pens is ___ Taka.)* This task, unlike others requiring external knowledge, is self-contained and focuses solely on numerical reasoning. It represents a variation of traditional math word problems in a fill-in-the-blank style.

Quantitative NLI (QNLI). "Premise: রহিম ৫০ টাকায় বই কিনল *(Rahim bought a book for 50 Taka.)* Hypothesis: রহিম টাকায় ৮০টি বই কিনলেন" *(Rahim bought 80 books for one Taka.)* Quantitative NLI involves evaluating the relationship between a premise and hypothesis using basic numerical reasoning, transforming word problems into a natural language inference format to assess quantitative reasoning.

Arithmetic Word Problems (AWP). Word problems involve interpreting a scenario and performing arithmetic calculations. For example: "৭৫ টাকায় ১৫ টি বলপেন কিনে ৯০ টাকায় বিক্রয় করলে শতকরা কত ভাগ লাভ হবে ?" *(If 15 ballpoint pens are bought for 75 Taka and sold for 90 Taka, what is the percentage of profit?)* This task tests the ability to apply arithmetic reasoning in real-world situations.

3.1 Dataset Development

Developing a benchmark dataset for numerical reasoning is inherently challenging due to its complexity and sensitivity. These challenges are further amplified when working in a low-resource language like Bengali. Key obstacles include the limited availability of high-quality, domain-specific data, the difficulty of curating diverse examples across different task formats, and the need for precise annotation to ensure mathematical accuracy. Furthermore, the process is time-consuming, requiring a thorough and consistent data collection and validation workflow. The detailed corpus development process for *BenNumEval* is illustrated in Figure 1.

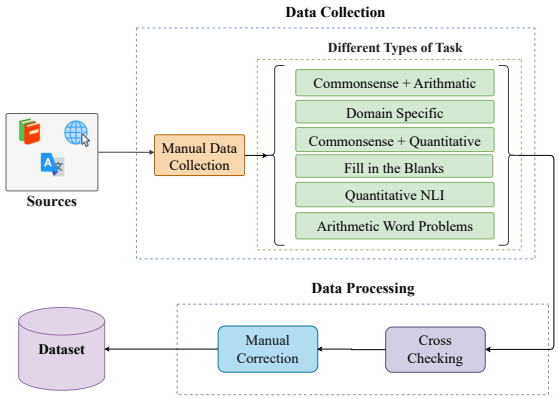

Figure 1: Development process of the *BenNumEval* dataset

3.1.1 Data Collection

The data collection process was conducted in multiple stages:

(i) Manual Data Collection: Arithmetic problems were manually sourced from school textbooks published by *NCTB*[3] and *WBBSE*[4]. For the Domain-Specific (DS) task, problems were collected from *Royal Publications*[5] books on Physics, Chemistry, and Higher Mathematics. Our primary focus was to gather data from school

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

[4] https://wbbse.wb.gov.in/

[5] https://the-royal-scientific-publications.com/

Tasks	Size	Examples
Commonsense + Arithmetic (CA)	410	**Question:** পাঁচ ডজন ডিমের দাম ৩০০ টাকা হলে ৭ টি ডিমের দাম কত টাকা? *(If the price of five dozen eggs is 300 taka, what is the price of 7 eggs?)* **Answer:** ৩৫ টাকা *(35 taka)*
Domain-Specific (DS)	705	**Question:** একটি বাইক ৭ সেকেন্ডে ২১ মিটার ভ্রমণ করেছে। বাইকের গড় গতি বের কর। *(A bike traveled 21 meters in 7 seconds. Calculate the average speed of the bike.)* **Answer:** ৩ মিটার/সেকেন্ড *(3 meters/second)*
Commonsense + Quantitative (CQ)	400	**Question:** গাছের ডাল ৬ ইঞ্চি পুরু যেখানে ফুলের ডাল ০.২৫ ইঞ্চি পুরু। তাই গাছের ডাল ছিল? (A) দুর্বল (B) শক্তিশালী *(The tree branch is 6 inches thick, while the flower stem is 0.25 inches thick. Therefore, the tree branch was? (A) Weak (B) Strong)* **Answer:** শক্তিশালী *(Strong)*
Fill-in-the-blanks (FiB)	665	**Question:** তিনটি সংখ্যার যোগফল ৫৪৫২। সংখ্যা তিনটি হল ২৫৬৩, ___ এবং ১২৪৫। *(The sum of three numbers is 5,452. The three numbers are 2,563, ___, and 1,245.)* **Answer:** ১৬৪৪ *(1644)*
Quantitative NLI (QNLI)	425	**Premise:** রাশেদ ৬৫ টি বই কিনলেন। *(Rashed bought 65 books.)* **Hypothesis:** রাশেদ ২৫ টির বেশি বই কিনলেন। *(Rashed bought more than 25 books.)* **Options:** Entailment, Contradiction or Neutral **Answer:** Entailment
Arithmetic Word Problems (AWP)	650	**Question:** যদি ৮ কেজি পোলাওয়ের চালের মূল্য ৯৬০ টাকা হয়, তবে ৪৮০০ টাকা দিয়ে কত কেজি চাল কেনা যাবে? *(If the price of 8 kilograms of polao rice is 960 taka, how many kilograms of rice can be bought with 4,800 taka?)* **Answer:** ৪০ কেজি *(40 kg)*
Total	**3255**	

Table 1: Types of various numerical reasoning tasks in Bengali, task sizes, and sample examples of each task. See Table 6 in Appendix B for additional examples.

books. However, we observed that the Quantitative NLI (QNLI) and Commonsense+Quantitative (CQ) tasks were unavailable in these books. To address this gap, we sourced these problems from the NumGLUE dataset (Mishra et al., 2022b) and translated them into Bengali using Google Translator via the `deep-translator`[6] package. We then manually reviewed the translations to correct errors, including English words, numerals, and any mistranslations.

Additionally, Commonsense + Arithmetic (CA) problems were curated using fundamental commonsense knowledge related to various everyday concepts, such as the number of fingers on a hand, days in a week, and counts in a dozen. Furthermore, we incorporated some native Bengali commonsense knowledge (e.g., "1 hali" meaning "4 pieces," analogous to "1 dozen" meaning "12 pieces") to enhance the cultural and contextual relevance of the dataset.

(ii) Task Formatting: The arithmetic problems collected from school textbooks were structured into diverse task-specific formats. For instance, arithmetic word problems were reformulated into a fill-in-the-blanks format to enhance alignment with different evaluation tasks.

[6] https://pypi.org/project/deep-translator/

3.1.2 Data Processing

To ensure the integrity and consistency of the dataset, we applied the following steps:

(i) Cross-Checking: The dataset was primarily derived from two different sources: school textbooks (for tasks CA, DS, FiB, and AWP) and translated questions from the NumGLUE dataset (for tasks CQ and QNLI). These sources had been prevalidated by experts, ensuring the correctness of the original question-answer pairs. Consequently, recalculating the answers was unnecessary. While minor numerical typographical errors may have occurred during data collection, we performed thorough cross-checking to ensure accuracy. Additionally, we manually verified that all problems in the dataset are unique and appear only once.

(ii) Manual Correction: During cross-checking, we identified minor annotation errors, such as incorrect question-answer pairings, in fewer than 1% of the instances. To ensure accuracy and reliability, these errors were carefully reviewed and corrected by an experienced high school science teacher, who volunteered for this task. This validation process further improves the dataset's overall quality, consistency, and robustness.

4 Dataset Distributions

The *BenNumEval* dataset is designed to cover a wide range of numerical reasoning tasks. This sec-

tion provides an overview of the distribution of the dataset, detailing the composition and structure across different task categories. We analyze the dataset's variety in terms of task type, sample size, and format, highlighting its balanced approach to challenging models in various aspects of numerical reasoning.

4.1 Task-wise Dataset Distribution

The task-wise distribution of *BenNumEval* reflects a strategic focus on diverse aspects of numerical reasoning. Domain-Specific (DS) tasks, making up 21.66% of the dataset, highlight the emphasis on contextual arithmetic problems, tailored for specific fields and real-world applications. In contrast, the Commonsense + Quantitative (CQ) category, which constitutes 12.29%, demonstrates the integration of general reasoning with numerical analysis, though less frequently explored. The Fill-in-the-Blanks (FiB) tasks, which form 20.43% of the dataset, underscore the importance of fundamental arithmetic skills in a range of contexts. This balanced distribution showcases a holistic approach to evaluate numerical reasoning, catering to both specialized applications and general arithmetic proficiency (see Table 1 and Figure 2).

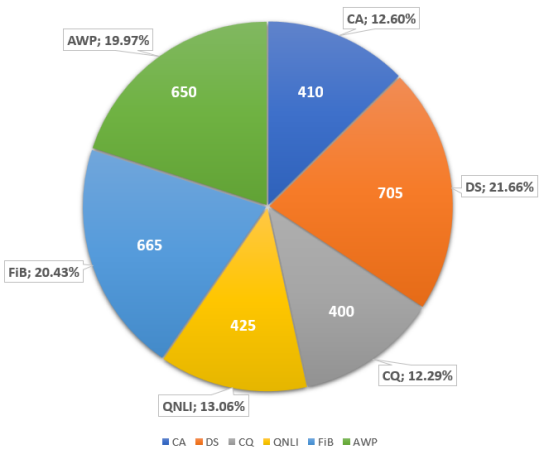

Figure 2: Task-wise dataset distribution

4.2 Source-wise Dataset Distribution

The *BenNumEval* dataset combines traditional educational resources with translated and manually curated problems, ensuring comprehensive coverage of numerical reasoning tasks across multiple domains. A significant 40.03% of the dataset comes from school mathematics textbooks, contributing heavily to CA, FiB, and AWP tasks, aligning it with formal educational resources. To enhance linguistic and contextual diversity, 32.96% of the dataset consists of translated problems. These questions primarily support CQ and QNLI tasks, while also contributing to CA. Domain-specific mathematical problems account for 7.65% of the dataset, catering to the Domain-Specific (DS) task. Additionally, subject-specific problems from Physics (6.51%) and Chemistry (6.42%) further enrich the dataset by incorporating real-world scientific applications that require numerical reasoning. A smaller subset (1.1%) of problems originates from Computer Science (CS), focusing on number systems and contributing to DS tasks. Moreover, 5.35% of the dataset comprises manually curated problems designed to test CA reasoning. These problems incorporate everyday numerical concepts, ensuring the dataset captures a broad spectrum of commonsense-based reasoning challenges. By combining diverse sources, BenNumEval achieves a balanced representation of traditional educational materials, domain-specific expertise, and linguistic diversity in Bengali, making it a robust benchmark for evaluating numerical reasoning models (see Figure 3 & Figure 4).

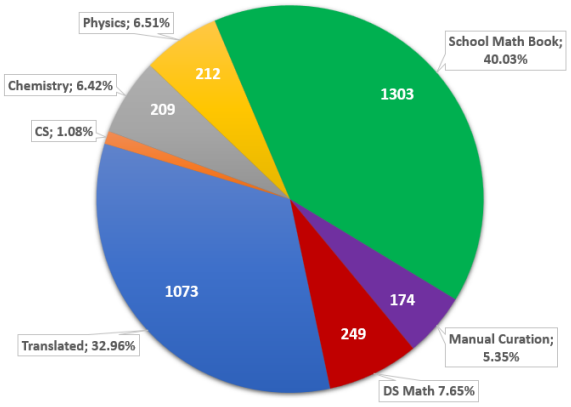

Figure 3: Source-wise dataset distribution

5 Experiments

Our experimental process evaluates the performance of LLMs on numerical reasoning tasks in Bengali. As depicted in Figure 5, the workflow initiates with a Bengali numerical reasoning question. This question is then processed using one of three prompting techniques: *BNaP*, *XLP* (Qin et al., 2023), or *XCoT* (Huang et al., 2023) and subsequently input into the LLM. The LLM's generated response undergoes answer extraction to produce the final solution, which is then evaluated for

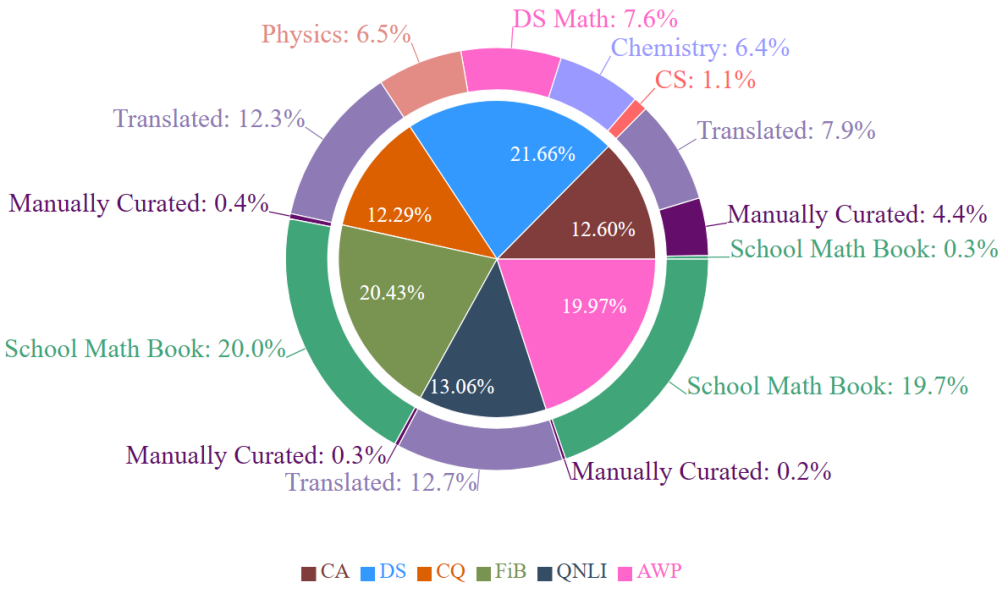

Figure 4: Distrubution of data in each task with sources

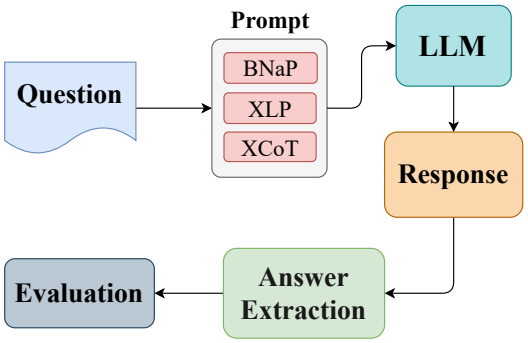

Figure 5: Overview of the experimental process using different prompts for numerical reasoning.

accuracy. Figure 6 provides a detailed example of this process.

5.1 LLMs

This study assessed the performance of a diverse set of Large Language Models (LLMs), encompassing specialized math models and multilingual general-purpose models. The evaluated models included: Gpt-4o (Achiam et al., 2023), Gemini-2.0-flash [7], Llama-3.3-70B [8], Mathstral-7B [9], and DeepSeek-R1-Distill-Llama-70B (Guo et al., 2025). We utilized instruction-tuned versions of all models to ensure a consistent evaluation of numerical reasoning tasks. Across all experiments, we set the model temperature and top-p value to 1.0. Additionally, we applied a standardized prompt template (see Tables 3, 4 and 5 in Appendix B) uniformly throughout the experiments.

5.2 Prompting Techniques

The design and structure of prompts play a critical role in optimizing the performance of LLMs, as they are highly sensitive to prompt formats (Zhao et al., 2021). However, instruction tuning has shown that LLMs can be trained to better follow natural language instructions, reducing the reliance on explicit prompt engineering. Given that LLMs exhibit zero-shot learning capabilities (Wei et al., 2021; Kojima et al., 2022), we utilized zero-shot *Cross-Lingual Prompting (XLP)* (Qin et al., 2023), *Cross-Lingual Chain-of-Thought Prompting (XCoT)* (Huang et al., 2023), and *Bengali Native Prompting (BNaP)* to assess and compare the performance of various LLMs on the *BenNumEval* dataset.

Cross-Lingual Chain-of-Thought Prompting (XCoT) extends chain-of-thought (*CoT*) reasoning (Wei et al., 2022) to multilingual contexts, improving the ability of LLMs to solve complex mathematical problems across different languages (Shi et al., 2022). *XCoT* ensures coherent reasoning, particularly in cross-lingual numerical and logical computations (Qin et al., 2023). We implemented *XCoT* by designing a template

[7] https://deepmind.google/technologies/gemini/flash/

[8] https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md

[9] https://mistral.ai/news/mathstral/

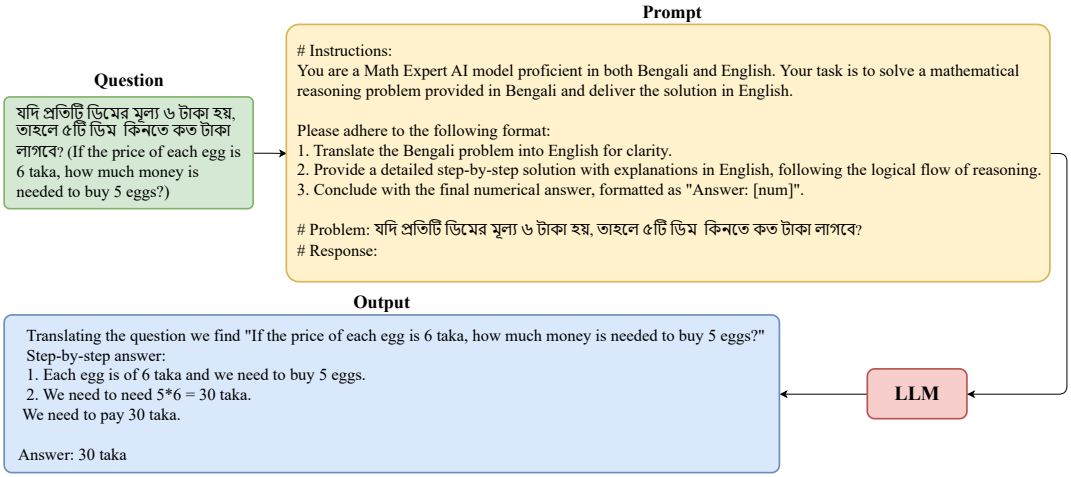

Figure 6: System overview with an example: A numerical reasoning task in Bengali is paired with the XCoT prompt template to create a language-independent prompt, which is then input to the LLM for generating a step-by-step solution in the desired format.

inspired by Huang et al. (2023), where the model assumes specialized roles as an expert in Bengali, English, and mathematics. The process begins with translating the question from Bengali to English for better comprehension, followed by generating a step-by-step solution using the *CoT* reasoning method (Qin et al., 2023). The final answer is presented in the format `"Answer: [num]"`.

Cross-Lingual Prompting (XLP) allows LLMs to solve mathematical reasoning tasks across languages by focusing on the logic behind the task (Zhou et al., 2022). The *XLP* directly translates the problem while preserving its mathematical reasoning (Qin et al., 2023), with the model returning the answer in a specified format, such as `"Answer: [num]"`.

Bengali Native Prompting (BNaP) allows LLMs to directly engage with numerical reasoning tasks in Bengali, preserving the language's integrity. By presenting both instructions and problems in Bengali, *BNaP* tests the model's ability to understand and process the language while accurately performing mathematical operations. To maintain clarity and enable systematic evaluation, the model is instructed to generate responses in a standardized format, such as "উত্তর: [সংখ্যা]". See Tables 3, 4 and 5 in Appendix B for detailed prompt specifications.

5.3 Evaluation

For a robust evaluation, each model underwent two independent runs, each employing a distinct random sample of 1,000 from the *BenNumEval* dataset. This dataset was sampled to ensure a balanced representation across task categories: 150 each from CA, CQ, FiB, QNLI, and AWP, and 250 from DS. This rigorous sampling and testing procedure was designed to enhance reliability and mitigate bias, with reported results (see Table 2) representing the average scores across the two runs. Detailed results of independent runs are given in Table 7 in Appendix B.

To measure model performance, we report accuracy scores on the *BenNumEval* benchmark using the *Exact-Match* accuracy metric, following a similar approach mentioned in Dua et al. (2019) and Qin et al. (2023).

Human Baseline. To establish a robust human performance baseline, we randomly selected a subset of 120 test samples, ensuring balanced representation across all six task categories. Each sample was independently solved by three native Bengali-speaking annotators with strong mathematical background. The final human accuracy was computed by averaging the scores of all annotators. This baseline serves as a crucial upper bound for model evaluation, highlighting the performance gap between LLMs and human expertise in numerical reasoning tasks in Bengali.

6 Results and Discussions

The *BenNumEval* dataset is utilized to assess the performance of various LLMs (see Table 2), revealing several key insights into their capabilities and behavioral patterns.

Prompting	Models	CA	DS	CQ	FiB	QNLI	AWP	Avg.
BNaP	**Mathstral-7B**	6.67	2.60	14.67	6.67	11.33	3.00	7.49
	Llama-3.3-70B	57.67	24.00	50.33	40.33	17.67	41.33	38.56
	DeepSeek-R1-Distill-Llama-70B	9.67	1.00	8.67	2.67	5.67	1.00	4.78
	Gpt-4o	84.00	58.40	83.00	76.67	56.33	80.33	73.13
	Gemini-2.0-flash	88.00	74.20	81.67	80.00	54.00	86.00	77.31
XLP	**Mathstral-7B**	35.33	17.20	44.33	31.33	35.33	30.33	32.32
	Llama-3.3-70B	82.00	61.60	84.00	76.67	53.00	82.00	73.21
	DeepSeek-R1-Distill-Llama-70B	65.33	27.80	35.67	49.67	26.00	47.00	41.91
	Gpt-4o	87.33	56.00	86.00	73.67	61.67	76.67	73.56
	Gemini-2.0-flash	90.00	68.80	86.67	86.00	57.33	89.00	79.63
XCoT	**Mathstral-7B**	36.00	18.20	47.67	28.00	33.67	34.00	32.93
	Llama-3.3-70B	82.33	56.60	80.67	73.67	49.00	78.67	70.16
	DeepSeek-R1-Distill-Llama-70B	57.67	18.00	48.00	35.00	21.67	32.33	35.45
	Gpt-4o	86.33	51.60	85.33	65.67	59.33	73.67	70.32
	Gemini-2.0-flash	83.33	62.00	88.67	75.33	55.00	75.00	73.02
	Human Baseline (Avg., N=3)	98.33	96.67	100	96.67	98.33	98.33	98.05

Table 2: Comparison of various language models across different prompting techniques on Bengali numerical reasoning tasks. The table reports average accuracy (%) scores from two evaluation runs for six task categories: CA (Commonsense + Arithmetic), DS (Domain-Specific), CQ (Commonsense + Quantitative), FiB (Fill-in-the-Blanks), QNLI (Quantitative Natural Language Inference), and AWP (Arithmetic Word Problems). Results are shown for three prompting techniques: *BNaP (Bengali Native Prompting), XLP (Cross-Lingual Prompting),* and *XCoT (Cross-Lingual Chain-of-Thought Prompting)*. A human baseline (N=3) is included for reference.

LLM Performance Across Prompting Strategies. Advanced language models such as GPT-4o and Gemini-2.0-flash consistently exhibit robust performance across various prompting methods. For instance, Gemini-2.0-flash attains average accuracies of 77.31, 79.63, and 73.02 under *BNaP*, *XLP*, and *XCoT*, respectively, with GPT-4o closely matching its performance across all tasks and prompts. A detailed error analysis based on prompting strategies is further provided in Appendix A.

In contrast, smaller or distilled models such as Mathstral-7B and DeepSeek-R1-Distill-Llama-70B show notably lower performance, especially under the *BNaP* setting. Mathstral-7B achieves a modest 7.49% accuracy, while DeepSeek-R1-Distill-Llama-70B performs even worse at 4.78%. However, both models benefit significantly from the *XLP* prompting strategy, with Mathstral-7B's accuracy rising to 32.32% and DeepSeek-R1-Distill-Llama-70B improving to 41.91%. This suggests that *XLP* prompts offer better linguistic structure and contextual grounding, potentially aligning more closely with the distribution of the models' pretraining data.

Another noteworthy observation is that, despite its larger size, DeepSeek-R1-Distill-Llama-70B underperforms compared to Llama 3.3-70B on Bengali text, particularly under the *BNaP* setting. In certain cases, it even trails behind the smaller Mathstral-7B. Qualitative analysis reveals that DeepSeek-R1-Distill-Llama-70B struggles to follow Bengali instructions and to generate coherent Bengali text. This discrepancy is likely due to the distillation process used to create DeepSeek-R1-Distill-Llama-70B, which prioritized transferring reasoning capabilities from its parent model, DeepSeek-R1. While this may improve performance on reasoning-intensive tasks in high-resource languages, it may have inadvertently compromised the model's generalization ability in multilingual settings, particularly for low-resource languages such as Bengali (Soltan et al., 2021). Distillation often involves trade-offs, and in this case, multilingual performance may have been deprioritized (Payoungkhamdee et al., 2024; Diddee et al., 2022).

Challenges in Different Tasks. Domain-specific tasks (DS) remain the hardest, with even top models like GPT-4o scoring only 51.20–58.40 and Gemini-2.0-flash peaking at 74.20 under *BNaP*. Quantitative NLI (QNLI) is similarly challenging, with no model surpassing 61.67, highlighting struggles with structured numerical reasoning. In contrast, models perform well on Arithmetic Word Problems (AWP) and Fill-in-

the-Blanks (FiB), where Gemini-2.0-flash reaches 89.00 in AWP. Notably, under *XLP*, it achieves 90.00 in Commonsense + Arithmetic (CA), showing strong integration of commonsense and numerical reasoning.

Cross-Lingual Prompting Impact. Switching from *BNaP* to *XLP* significantly boosts performance, especially for lower and mid-tier models, highlighting the advantages of multilingual cues in mitigating limitations of single-language prompts. However, the added complexity of chain-of-thought reasoning *(XCoT)* does not always improve results, suggesting that its integration in multilingual settings requires further optimization.

Gap to Human-Level Performance. Despite the advances seen with top-tier models, there remains a significant gap between LLM performance and human accuracy. Humans achieve near-perfect scores across all tasks, underscoring that even the best-performing models have room to grow, particularly in nuanced areas like domain-specific reasoning and quantitative inference.

7 Conclusion

This paper introduced *BenNumEval*, a novel benchmark dataset comprising six diverse tasks designed to assess the numerical reasoning capabilities of large language models (LLMs) in Bengali—a resource-constrained language with limited computational linguistic resources. Our extensive evaluation highlights a significant performance gap between state-of-the-art LLMs and human proficiency in numerical reasoning tasks, particularly in Bengali. These findings underscore the pressing need for more robust and linguistically inclusive AI models capable of handling mathematical reasoning in low-resource languages. By advancing research in Bengali Language Processing, *BenNumEval* serves as a critical stepping stone toward developing more equitable and effective AI systems for underrepresented languages.

Limitations

While *BenNumEval* marks a vital first step in evaluating Bengali numerical reasoning in LLMs, it has several limitations. First, although the dataset is diverse, its size can be expanded to include a broader range of problem types and real-world contexts for more thorough evaluation. Second, the exclusive focus on Bengali restricts cross-lingual insights; extending the framework to other low-resource languages will address this gap. Third, conducting a comparative analysis with existing numerical reasoning datasets would provide valuable context and, by establishing a more robust human performance baseline, yield deeper insights into how LLMs perform on *BenNumEval* relative to humans. Lastly, although we adopt various zero-shot prompting strategies, future work could benefit from exploring few-shot and more optimized chain-of-thought (*CoT*) prompting techniques, as well as incorporating advanced reasoning models. Evaluating larger and more capable LLMs also holds promise for enhancing the benchmark's overall depth and robustness.

Acknowledgements

We sincerely appreciate the generous support from the CUET NLP Lab, whose funding played a crucial role in facilitating this research. We also extend our gratitude to Mukul Chandra (Math Teacher, Saidpur High School, Saidpur), Mst. Sadia, Aksa Karim (Bangladesh Army University of Science and Technology (BAUST), Saidpur), and Fihima Tabassum (Chittagong College, Chittagong) for their valuable assistance in manual correction and human evaluation of our dataset.

Ethical Considerations

The *BenNumEval* dataset was meticulously curated by a team of native Bengali-speaking experts and trained student annotators aged 22-26, all with at least one year of research experience. To ensure data quality and consistency, comprehensive annotation guidelines were provided, covering numerical reasoning principles, data sources, collection procedures, task formatting, and rigorous cross-checking and manual correction protocols. All annotators received fair compensation as per institutional standards.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Harshita Diddee, Sandipan Dandapat, Monojit Choudhury, Tanuja Ganu, and Kalika Bali. 2022. Too brittle to touch: Comparing the stability of quantization and distillation towards developing low-resource mt models. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 870–885.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–533.

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

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

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597.

Nate Kushman, Yoav Artzi, Luke Zettlemoyer, and Regina Barzilay. 2014. Learning to automatically solve algebra word problems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 271–281.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2021. A diverse corpus for evaluating and developing english math word problem solvers. *arXiv preprint arXiv:2106.15772*.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. 2022a. LILA: A unified benchmark for mathematical reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5807–5832, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022b. Numglue: A suite of fundamental yet challenging mathematical reasoning tasks. *ACL*.

Sanchita Mondal, Debnarayan Khatua, Sourav Mandal, and Arif Ahmed Sekh. 2023. Bmwp: The first bengali math word problems dataset for operation prediction and solving.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*.

Patomporn Payoungkhamdee, Peerat Limkonchotiwat, Jinheon Baek, Potsawee Manakul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2024. An empirical study of multilingual reasoning distillation for question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7739–7751.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *arXiv preprint arXiv:2310.14799*.

Subhro Roy and Dan Roth. 2016. Solving general arithmetic word problems. *arXiv preprint arXiv:1608.01413*.

Eli Schwartz, Leshem Choshen, Joseph Shtok, Sivan Doveh, Leonid Karlinsky, and Assaf Arbelle. 2024. Numerologic: Number encoding for enhanced llms' numerical reasoning. *arXiv preprint arXiv:2404.00459*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Y Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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

Saleh Soltan, Haidar Khan, and Wael Hamza. 2021. Limitations of knowledge distillation for zero-shot transfer learning. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 22–31.

Shyam Upadhyay, Ming-Wei Chang, Kai-Wei Chang, and Wen-tau Yih. 2016. Learning from explicit and implicit supervision jointly for algebra word problems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 297–306.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

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

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, Yudong Wang, Zijian Wu, Shuaibin Li, Fengzhe Zhou, Hongwei Liu, Songyang Zhang, Wenwei Zhang, Hang Yan, Xipeng Qiu, Jiayu Wang, Kai Chen, and Dahua Lin. 2024. Internlm-math: Open math large language models toward verifiable reasoning.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? *arXiv preprint arXiv:2010.05345*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Meng Zhou, Xin Li, Yue Jiang, and Lidong Bing. 2022. Enhancing cross-lingual prompting with dual prompt augmentation. *arXiv preprint arXiv:2202.07255*.

A Error Analysis

To gain deeper insights into the limitations of various prompting strategies and model behaviors, we performed a comprehensive error analysis. Specifically, we categorized errors into two types: *wrong format* (wf%)—indicating deviations from the expected output structure—and *wrong calculation* (wc%)—capturing arithmetic or logical inaccuracies. The wf% reflects the proportion of total predictions that do not conform to the expected format, while wc% measures the proportion of calculation errors within the subset of predictions that are correctly formatted. The following equations were used:

$$WP = T - EM \tag{1}$$

$$WF\% = \frac{WF}{T} \times 100 \tag{2}$$

$$CF = T - WF \tag{3}$$

$$WC\% = \frac{WP - WF}{CF} \times 100 \tag{4}$$

Here, T denotes total samples, EM is exact matches, WP is wrong predictions, WF is wrong format, CF is correct format, $WF\%$ is the percentage of wrong format predictions, and $WC\%$ is the percentage of content errors among correctly formatted predictions.

To ensure the robustness of our findings, we performed two independent runs for each experimental setting. The corresponding error percentages for each model and prompting strategy are reported in Table 8 and Table 9.

***BNaP* Prompting.** As shown in Table 8, models like DeepSeek-R1-Distill-Llama-70B exhibited a notably high wf% (e.g., 98.8% in DS), indicating poor compliance with output constraints under naive prompting. However, its wc% remained relatively low, e.g., 0.0% in AWP, due to the low number of predictions eligible for reasoning evaluation. This highlights how wf% inflates total error by disqualifying outputs early. In contrast, GPT-4o and Gemini-2.0-Flash demonstrated exceptional formatting discipline (wf% consistently below 6%), yet their wc% reached over 40% in QNLI. This reflects a typical failure pattern for high-capacity models: they understand and follow instructions but occasionally miscompute results.

***XLP* Prompting.** This prompting strategy helped improve format adherence significantly for all models. Mathstral-7B showed a reduction in wf% from 66.4% (*BNaP* Task 2) to 34.8% (*XLP* Task 2), while Llama-3.3-70B consistently kept wf% below 12%. However, despite improved format alignment, wc% remained substantial across several models—e.g., Llama-3.3-70B reached 39.85% wc% in QNLI—revealing that *XLP* alone does not resolve deeper reasoning limitations.

***XCoT* Prompting.** The *XCoT* strategy, emphasizing step-by-step reasoning, maintained low wc% for most models, especially Gemini-2.0-Flash, which achieved wf% below 16% across all tasks. However, verbose rationales often confused rigid format parsers, increasing wf% in models like DeepSeek-R1-Distill-Llama-70B (wf% > 30% in all tasks) and occasionally in GPT-4o. Llama-3.3-70B again struck a favorable balance between low wf% (average 10%) and moderate wc% (22%).

This analysis highlights that formatting adherence and reasoning accuracy do not always align. A high wf%—as observed in DeepSeek-R1-Distill-Llama-70B—suggests a notable weakness in following structured instructions, regardless of the model's reasoning capabilities. Meanwhile, high wc% suggests that even well-structured outputs may hide subtle logical or numerical fallacies. Prompting strategies like *XCoT* reduce reasoning errors but may increase formatting violations unless models are fine-tuned for *CoT* outputs. These findings underscore the value of disaggregated error analysis in diagnosing model-specific and prompt-specific weaknesses.

B Experimental Details – (Prompt Templates, Examples, and Run-Specific Results)

To ensure reproducibility and provide further insight into our experimental setup, this appendix contains supplementary information. Specifically, we include the prompt templates utilized for each prompting technique: *XCoT* (Table 3), *XLP* (Table 4), and *BNaP* (Table 5). Table 6 showcases additional examples of arithmetic problems from the *BenNumEval* dataset, expanding upon the examples used in the main paper. For a comprehensive view of our experimental outcomes, Table 7 presents the detailed results for each of the two evaluation runs.

Prompt template for task 1 (CA), task 2 (DS), task 4 (FiB), and task 6 (AWP)	*#Instructions*: You are a Math Expert AI model proficient in Bengali and English. Your task is to solve a mathematical reasoning problem provided in Bengali and deliver the solution in English. Please adhere to the following format: 1. Translate the Bengali problem into English for clarity. 2. Provide a detailed step-by-step solution with explanations in English, following the logical flow of reasoning. 3. Conclude with the final numerical answer, formatted as "Answer: [num]". *# Problem:* <question> *# Response:* <LLM Response>
Template for task 3 (CQ)	*#Instructions*: You are a Math Expert AI model proficient in Bengali and English. Your task is to solve a mathematical reasoning problem provided in Bengali and choose the correct option. Please adhere to the following format: 1. Translate the Bengali question and options into English for clarity. 2. Provide a detailed step-by-step solution with explanations in English, following the logical flow of reasoning. 3. Conclude by selecting the correct option, formatted as "Answer: [Option]". The possible options are "Option 1" or "Option 2". *# Question:* <question> *# Option 1:* <option1> *# Option 2:* <option2> *# Response:* <LLM Response>
Template for task 5 (QNLI)	*#Instructions*: You are a Math Expert AI model proficient in Bengali and English. Your task is to solve a Quantitative Natural Language Inference (QNLI) problem presented in Bengali. You need to determine the relationship between the premise and the hypothesis. Please adhere to the following format: 1. Translate the premise and hypothesis from Bengali to English for clarity. 2. Provide a step-by-step explanation of your reasoning process in English. 3. Conclude by selecting the correct option, formatted as "Answer: [Option]". The possible options are "Entailment", "Neutral", or "Contradiction". *# Premise:* <premise> *# Hypothesis:* <hypothesis> *# Response:* <LLM Response>

Table 3: XCoT prompt templates for different tasks

Prompt template for task 1 (CA), task 2 (DS), task 4 (FiB), and task 6 (AWP)	*#Instructions*: You are a Math Expert AI model proficient in Bengali and English. Your task is to solve a mathematical reasoning problem provided in Bengali and deliver the solution in English. Please adhere to the following format: 1. Translate the Bengali problem into English for clarity. 2. Conclude with the final numerical answer, formatted as "Answer: [num]". *# Problem:* <question> *# Response:* <LLM Response>
Template for task 3 (CQ)	*#Instructions*: You are a Math Expert AI model proficient in Bengali and English. Your task is to solve a mathematical reasoning problem provided in Bengali and choose the correct option. Please adhere to the following format: 1. Translate the Bengali question and options into English for clarity. 2. Conclude by selecting the correct option, formatted as "Answer: [Option]". The possible options are "Option 1" or "Option 2". *# Question:* <question> *# Option 1:* <option1> *# Option 2:* <option2> *# Response:* <LLM Response>
Template for task 5 (QNLI)	*#Instructions*: You are a Math Expert AI model proficient in Bengali and English. Your task is to solve a Quantitative Natural Language Inference (QNLI) problem presented in Bengali. You need to determine the relationship between the premise and the hypothesis. Please adhere to the following format: 1. Translate the premise and hypothesis from Bengali to English for clarity. 2. Conclude by selecting the correct option, formatted as "Answer: [Option]". The possible options are "Entailment", "Neutral", or "Contradiction". *# Premise:* <premise> *# Hypothesis:* <hypothesis> *# Response:* <LLM Response>

Table 4: Prompt templates for different tasks in XLP setting

Prompt template for task 1 (CA), task 2 (DS), task 4 (FiB), and task 6 (AWP)	# নির্দেশাবলী: আপনি একজন গণিত বিশেষজ্ঞ এআই মডেল, যিনি বাংলা ভাষায় সম্পূর্ণভাবে দক্ষ। আপনার কাজ হলো প্রদত্ত গাণিতিক সমস্যার বিশদভাবে সমাধান করা এবং উত্তরটি বাংলায় প্রদান করা। *(You are an AI model specializing in mathematics, fully proficient in the Bengali language. Your task is to solve given mathematical problems in detail and provide the answer in Bengali.)* আপনাকে অবশ্যই চূড়ান্ত সাংখ্যিক উত্তরটি নিম্নলিখিত ফরম্যাটে উপস্থাপন করতে হবে: *(You must present the final numerical answer in the following format:)* **উত্তর: [সংখ্যা]**। *(**Answer: [number]**)* # সমস্যা *(Problem)*: <question> # সমাধান *(Solution)*: <LLM Response>
Template for task 3 (CQ)	# নির্দেশাবলী: আপনি একজন গণিত বিশেষজ্ঞ এআই মডেল, যিনি বাংলা ভাষায় সম্পূর্ণভাবে দক্ষ। আপনার কাজ হলো প্রদত্ত গাণিতিক সমস্যার বিশদভাবে সমাধান করা এবং উত্তরটি বাংলায় প্রদান করা। *(You are an AI model specializing in mathematics, fully proficient in the Bengali language. Your task is to solve the given mathematical problems in detail and provide the answer in Bengali.)* আপনার উত্তর অবশ্যই *নির্দিষ্ট বিন্যাসে* প্রদান করতে হবে: *(You must provide your answer in a *specific format*.)* **উত্তর: [সঠিক সম্ভাব্য উত্তর]** *(**Answer: [Correct Possible Answer]**)* যেখানে **[সঠিক সম্ভাব্য উত্তর]** হবে **উত্তর ১** অথবা **উত্তর ২**। *(Where **[Correct Possible Answer]** will be either **Answer 1** or **Answer 2**.)* — # সমস্যা *(Problem:* <question> # সম্ভাব্য উত্তরসমূহ *(Possible answers):* # সম্ভাব্য উত্তর ১ *(Possible Answer 1):* <option1> # সম্ভাব্য উত্তর ২ *(Possible Answer 2):* <option2> — # সমাধান *(Solution):* <LLM Response>
Template for task 5 (QNLI)	# নির্দেশাবলী: আপনি একজন গণিত বিশেষজ্ঞ AI মডেল, যিনি বাংলা ভাষায় দক্ষ। আপনার কাজ হলো একটি গাণিতিক ভাষাগত অনুমান সমস্যার সমাধান করা। আপনাকে প্রদত্ত পূর্বধারণা ও অনুমান এর মধ্যে সম্পর্ক নির্ধারণ করতে হবে। *(You are an AI model specializing in mathematics and proficient in the Bengali language. Your task is to solve a mathematical linguistic inference problem. You must determine the relationship between the given premise and hypothesis.)* অনুগ্রহ করে নিম্নলিখিত বিন্যাস অনুসরণ করুন: *(Please follow the below format:)* প্রদত্ত পূর্বধারণা ও অনুমান এর মধ্যে সম্পর্ক বিষয়ে চূড়ান্ত সিদ্ধান্ত নিন এবং সঠিক উত্তরটি **উত্তর: [সম্ভাব্য সঠিক উত্তর]** এই ভাবে প্রদান করুন। উত্তরের সম্ভাব্য বিকল্পগুলি হল: **সমর্থন**, **নিরপেক্ষ**, অথবা **বিরোধ**। *(Make the final decision regarding the relationship between the given premise and hypothesis, and provide the correct answer in the form **Answer: [Correct Possible Answer]**. The possible alternatives for the answer are: **Entailment**, **Neutral**, or **Contradiction**.)* # পূর্বধারণা *(Premise):* <premise> # অনুমান *(Hypothesis):* <hypothesis> # সমাধান: <LLM Response>

Table 5: BNaP prompt templates with English translations

Tasks	Examples
Commonsense + Arithmetic (CA)	**Question:** ফেব্রুয়ারি মাসের ৩০ তারিখ জিসানের জন্মদিন হওয়ার সম্ভাবনা কত? *(What is the probability of Jisan's birthday being on February 30?)* **Answer:** ০ *(0)* **Question:** একটি কারখানায় ৩ দিনে ৩৩৯টি মোটরসাইকেল তৈরি হয়। ৪ সপ্তাহে ওই কারখানায় কতটি মোটরসাইকেল তৈরি হবে। *(In a factory, 339 motorcycles are produced in 3 days. How many motorcycles will be produced in 4 weeks in that factory?)* **Answer:** ৩১৬৪ টি *(3164 pcs)* **Question:** একটি খামারে ৩৮টি প্রাণীর পা এবং মোট ১২টি প্রাণী রয়েছে। কেউ মুরগি আবার কেউ ভেড়া। মুরগির সংখ্যা নির্ণয় কর। *(In a farm, there are 38 animal legs and a total of 12 animals. Some are chickens, and some are sheep. Determine the number of chickens.)* **Answer:** ৫ *(5)*
Domain-Specific (DS)	**Question:** বিপাশার ক্লাসে ১৩(অক্টালে) জন শিক্ষার্থী আছে। প্রত্যেকের কাছে ৫৬(হেক্সাডেসিম্যালে) টি করে কলম আছে। ডেসিম্যাল সংখ্যা পদ্ধতি অনুযায়ী তাদের সবার মোট কয়টি কলম আছে? *(There are 13 (in octal) students in Bipasha's class. Each of them has 56 (in hexadecimal) pens. According to the decimal number system, how many pens do they all have in total?)* **Answer:** ৯৪৬ টি *(946 pcs)* **Question:** সালফিউরিক এসিডের ২০০ মিলি এর ০.৫ মোলার দ্রবণ তৈরী করা হলো। দ্রবণে ১০ গ্রাম সোডিয়াম হাইড্রোক্সাইড যোগ করলে কি পরিমাণ সোডিয়াম হাইড্রোক্সাইড দ্রবণে থেকে যাবে? *(A 0.5 molar solution of sulfuric acid is prepared with 200 mL. If 10 grams of sodium hydroxide are added to the solution, how much sodium hydroxide will remain in the solution?)* **Answer:** ২ গ্রাম *(2 grams)* **Question:** ৪ সে.মি. ব্যাসার্ধের একটি গোলক আকৃতির বল একটি সিলিন্ডার আকৃতির বাক্সে ঠিকভাবে এঁটে যায়। সিলিন্ডারটির অনধিকৃত অংশের আয়তন কত ঘন সে.মি.? *(A spherical ball with a radius of 4 cm fits perfectly inside a cylindrical box. What is the volume of the unused part of the cylinder in cubic centimeters?)* **Answer:** ১৩৪.০৪ ঘন সে.মি. *(134.04 cubic cm)*
Commonsense + Quantitative (CQ)	**Question:** একটি বোলিং বলের ভর ১৯ পাউন্ড এবং একটি বেসবলের ভর ৬ পাউন্ড। কোনটির শক্তিশালী মাধ্যাকর্ষণ আছে? (A) বেসবল (B) বোলিং বল *(The mass of a bowling ball is 19 pounds, and the mass of a baseball is 6 pounds. Which one has stronger gravitational force? (A) Baseball (B) Bowling ball)* **Answer:** বোলিং বল *(Bowling ball)* **Question:** একটি খরগোশ একটি ক্ষেতে ৩১ কিমি/ঘন্টা এবং একটি ক্যাকটাস ক্ষেতে ৪৭ কিমি/ঘন্টা বেগে দৌড়াতে পারে। এর মানে কোথায় খরগোশটি বেশী বাধা পায়? (A) ক্যাকটাস ক্ষেত্র (B) ক্ষেতে *(A rabbit can run at 31 km/h in a field, and a cactus field at 47 km/h. What does this mean? Where does the rabbit face more resistance? (A) Cactus field (B) Field)* **Answer:** ক্ষেতে *(Field)* **Question:** একটি বইয়ে ৪০০ পৃষ্ঠা এবং আরেকটি বইয়ে ৩৫০ পৃষ্ঠা আছে। কোন বইটি বেশি পুরু? (A) দ্বিতীয়টি (B) প্রথমটি *(One book has 400 pages and another book has 350 pages. Which book is thicker? (A) The second one (B) The first one)* **Answer:** প্রথমটি *(The first one)*
Fill-in-the-blanks (FiB)	**Question:** পিতা পুত্রের বর্তমান বয়সের পার্থক্য ১৫ বছর । ৩০ বছর পর তাদের বয়সের পার্থক্য ___ বছর হবে। *(The difference in the current ages of the father and son is 15 years. After 30 years, the difference in their ages will be ___ years.)* **Answer:** ১৫ বছর *(15 years)* **Question:** ১৫টি লজেন্সের মধ্যে ৩টি লজেন্স নিলাম। মোট লজেন্সের শতকরা ___ ভাগ লজেন্স নিলাম দেখি। *(I took 3 out of 15 candies. I took ___ percentage of the total candies.)* **Answer:** ২০ *(20)* **Question:** এমন একটি ক্ষুদ্রতম স্বাভাবিক সংখ্যা আছে যেটি দিয়ে ১০৮ কে গুণ করলে পূর্ণবর্গ সংখ্যা পাব। সংখ্যাটি হলো ___। *(There is a smallest natural number such that when multiplied by 108, it gives a perfect square. The number is ___.)* **Answer:** ৩ *(3)*
Quantitative NLI (QNLI0)	**Premise:** ৫ বছর পর অরুণের বয়স হবে ২৫ বছর।*(After 5 years, Arun's age will be 25 years.)* **Hypothesis:** ৭ বছর পর অরুণের বয়স হবে ২৫ বছর। *(After 7 years, Arun's age will be 25 years.)* **Answer:** Contradiction **Premise:** শ্রীধর একসঙ্গে ১৯৮০০ আয় করেন। *(Shreedhar earns 19,800 at once.)* **Hypothesis:** শ্রীধর একসাথে ৭৯৮০০ এর কম আয় করেছে। *(Shreedhar has earned less than 79,800 at once.)* **Answer:** Entailment **Premise:** স্যান্ডি মলির থেকে ৬৪ বছরেরও কম বয়সী। *(Sandy is less than 64 years old compared to Moli.)* **Hypothesis:** স্যান্ডি মলির চেয়ে ১৪ বছরের ছোট। *(Sandy is 14 years younger than Moli.)* **Answer:** Neutral
Arithmetic Word Problems (AWP)	**Question:** ৪টি মুরগি এবং ৩টি হাঁসের দাম একত্রে ৬৩৯ টাকা। ১টি হাঁসের দাম ৮৫ টাকা হলে ১টি মুরগির দাম কত ? *(The total cost of 4 chickens and 3 ducks is 639 Taka. If the cost of one duck is 85 Taka, what is the cost of one chicken?)* **Answer:** ৯৬ টাকা *(96 taka)* **Question:** ভাজ্য ৮৯০৩, ভাজক ৮৭ এবং ভাগশেষ ২৯। ভাগফল কত? *The dividend is 8903, the divisor is 87, and the remainder is 29. What is the quotient?)* **Answer:** ১০২ *(102)* **Question:** গিতার কাছে ১১/৬ লিটার ও মামুনের কাছে ১৩/৬ লিটার জুস আছে। মামুনের কত লিটার বেশি জুস আছে? *(Gita has 11/6 liters of juice and Mamun has 13/6 liters of juice. How many liters more juice does Mamun have?)* **Answer:** ১/৩ *(1/3)*

Table 6: Examples of numerical reasoning problems from the *BenNumEval* dataset

Run 1								
Prompting	**Models**	**CA**	**DS**	**CQ**	**FiB**	**QNLI**	**AWP**	**Avg.**
BNaP	**Mathstral-7B**	8.0	2.40	14.67	6.0	12.67	3.33	7.84
	Llama-3.3-70B	54.67	22.40	49.33	38.67	20.67	42.67	38.07
	DeepSeek-R1-Distill-Llama-70B	8.0	0.80	9.33	2.0	4.0	2.0	4.36
	Gpt-4o	84.0	60.0	83.33	74.0	58.0	82.67	73.67
	Gemini-2.0-flash	88.0	76.0	84.67	78.0	54.67	88.67	78.33
XLP	**Mathstral-7B**	30.0	18.80	42.67	33.33	36.67	32.67	32.36
	Llama-3.3-70B	82.67	62.40	84.0	77.33	53.33	82.67	73.73
	DeepSeek-R1-Distill-Llama-70B	64.0	27.60	38.0	47.33	28.0	53.33	43.04
	Gpt-4o	86.0	55.60	85.33	73.33	62.67	79.33	73.71
	Gemini-2.0-flash	90.67	67.20	88.67	85.33	56.0	88.67	79.42
XCoT	**Mathstral-7B**	36.67	20.0	47.33	28.67	37.33	38.67	34.78
	Llama-3.3-70B	83.33	59.20	79.33	72.0	47.33	80.67	70.31
	DeepSeek-R1-Distill-Llama-70B	60.67	18.80	41.33	36.67	24.0	34.0	35.91
	Gpt-4o	86.0	51.20	84.67	66.67	58.0	74.67	70.20
	Gemini-2.0-flash	84.67	62.40	88.67	78.0	55.33	73.33	73.33
Run 2								
Prompting	**Models**	**CA**	**DS**	**CQ**	**FiB**	**QNLI**	**AWP**	**Avg.**
BNaP	**Mathstral-7B**	5.33	2.80	14.67	7.33	10.0	2.67	7.13
	Llama-3.3-70B	60.67	25.60	51.33	42.0	14.67	40.0	39.04
	DeepSeek-R1-Distill-Llama-70B	11.33	1.20	8.0	3.33	7.33	0.0	5.20
	Gpt-4o	84.0	56.80	82.67	79.33	54.67	78.0	72.58
	Gemini-2.0-flash	88.0	72.40	78.67	82.0	53.33	83.33	76.29
XLP	**Mathstral-7B**	40.67	15.60	46.0	29.33	34.0	28.0	32.27
	Llama-3.3-70B	81.33	60.80	84.0	76.0	52.67	81.33	72.69
	DeepSeek-R1-Distill-Llama-70B	66.67	28.0	33.33	52.0	24.0	40.67	40.78
	Gpt-4o	88.67	56.40	86.67	74.0	60.67	74.00	73.40
	Gemini-2.0-flash	89.33	70.40	84.67	86.67	58.67	89.33	79.84
XCoT	**Mathstral-7B**	35.33	16.40	48.0	27.33	30.0	29.33	31.07
	Llama-3.3-70B	81.33	54.0	82.0	75.33	50.67	76.67	70.0
	DeepSeek-R1-Distill-Llama-70B	54.67	17.20	54.67	33.33	19.33	30.67	34.98
	Gpt-4o	86.67	52.0	86.0	64.67	60.67	72.67	70.44
	Gemini-2.0-flash	82.0	61.60	88.67	72.67	54.67	76.67	72.71

Table 7: Detailed accuracy scores (%) for two evaluation runs, showing performance of various language models (Mathstral-7B, Llama-3.3-70B, DeepSeek-R1-Distill-Llama-70B, Gpt-4o, Gemini-2.0-flash) on six reasoning tasks (CA, DS, CQ, FiB, QNLI, AWP) using *BNaP, XLP,* and *XCoT* prompting.

Error Analysis Breakdown by Prompting Strategy (Run 1)									
Prompting	**Models**	**Error Type**	**CA**	**DS**	**CQ**	**FiB**	**QNLI**	**AWP**	**Avg.**
BNaP	**Mathstral-7B**	wf%	56.00	66.40	75.33	63.33	62.00	66.67	64.95
		wc%	81.82	92.86	40.54	83.64	66.67	90.00	75.92
	Llama-3.3-70B	wf%	32.67	68.40	40.67	50.00	73.33	52.00	52.84
		wc%	18.81	29.11	16.85	22.67	22.50	11.11	20.17
	DeepSeek-R1-Distill-Llama-70B	wf%	91.33	98.80	88.00	97.33	91.33	98.00	94.13
		wc%	7.69	33.33	22.22	25.00	53.85	0.00	23.68
	GPT-4o	wf%	2.67	10.40	2.67	5.33	1.33	5.33	4.62
		wc%	13.70	33.04	14.38	21.83	41.22	12.68	22.81
	Gemini-2.0-Flash	wf%	3.33	2.00	0.00	1.33	0.00	0.00	1.11
		wc%	8.97	22.45	15.33	20.95	45.33	11.33	20.73
XLP	**Mathstral-7B**	wf%	32.00	34.80	26.67	30.00	17.33	34.67	29.25
		wc%	55.88	71.17	41.82	52.38	55.65	50.00	54.48
	Llama-3.3-70B	wf%	2.00	9.60	3.33	4.00	11.33	2.67	5.49
		wc%	15.65	30.97	13.10	19.44	39.85	15.07	22.35
	DeepSeek-R1-Distill-Llama-70B	wf%	24.67	56.00	52.00	38.67	56.67	34.00	43.67
		wc%	15.04	37.27	20.83	22.83	35.38	19.19	25.09
	GPT-4o	wf%	3.33	11.20	0.67	4.67	1.33	5.33	4.42
		wc%	11.03	37.39	14.09	23.08	36.49	16.20	23.05
	Gemini-2.0-Flash	wf%	2.67	1.20	0.00	0.67	1.33	0.67	1.09
		wc%	6.85	31.98	11.33	14.09	43.24	10.74	19.71
XCoT	**Mathstral-7B**	wf%	28.67	37.60	19.33	38.67	17.33	30.00	28.60
		wc%	48.60	67.95	41.32	53.26	54.84	44.76	51.79
	Llama-3.3-70B	wf%	4.00	8.80	10.67	6.67	24.00	6.00	10.02
		wc%	13.19	35.09	11.19	22.86	37.72	14.18	22.37
	DeepSeek-R1-Distill-Llama-70B	wf%	32.00	71.20	52.00	54.00	65.33	56.67	55.20
		wc%	10.78	34.72	13.89	20.29	30.77	21.54	21.99
	GPT-4o	wf%	4.00	22.80	2.67	16.67	6.67	10.67	10.58
		wc%	10.42	33.68	13.01	20.00	37.86	16.42	21.90
	Gemini-2.0-Flash	wf%	9.33	13.60	0.67	12.67	11.33	16.67	10.71
		wc%	6.62	27.78	10.74	10.69	37.59	12.00	17.57

Table 8: Breakdown of error rates (Run 1) across BNaP, XLP, and XCoT prompting strategies. wf% denotes Wrong Format errors, and wc% denotes Wrong Calculation errors. Lower values indicate better performance. "Avg." represents the mean error across all tasks.

Error Analysis Breakdown by Prompting Strategy (Run 2)									
Prompting	**Models**	**Error Type**	**Task 1**	**Task 2**	**Task 3**	**Task 4**	**Task 5**	**Task 6**	**Avg.**
BNaP	**Mathstral-7B**	wf%	51.33	60.80	73.33	56.00	59.33	68.00	61.46
		wc%	89.04	92.86	45.00	83.33	75.41	91.67	79.55
	Llama-3.3-70B	wf%	30.00	67.20	38.67	48.67	78.67	57.33	53.42
		wc%	13.33	21.95	16.30	18.18	31.25	6.25	17.88
	DeepSeek-R1-Distill-Llama-70B	wf%	88.67	98.80	89.33	96.67	90.67	99.33	94.78
		wc%	0.00	0.00	25.00	0.00	21.43	100.00	24.41
	GPT-4o	wf%	3.33	11.60	2.67	5.33	0.00	8.67	5.27
		wc%	13.10	35.75	15.07	16.20	45.33	14.60	23.34
	Gemini-2.0-Flash	wf%	4.00	1.20	0.00	2.00	0.00	3.33	1.76
		wc%	8.33	26.72	21.33	16.33	46.67	13.79	22.20
XLP	**Mathstral-7B**	wf%	36.67	32.80	25.33	31.33	22.00	37.33	30.91
		wc%	35.79	76.79	38.39	57.28	56.41	55.32	53.33
	Llama-3.3-70B	wf%	5.33	10.00	2.67	4.00	6.67	5.33	5.67
		wc%	14.08	32.44	13.70	20.83	43.57	14.08	23.12
	DeepSeek-R1-Distill-Llama-70B	wf%	23.33	58.00	59.33	36.00	58.00	50.67	47.55
		wc%	13.04	33.33	18.03	18.75	42.86	17.57	23.93
	GPT-4o	wf%	4.00	12.00	0.00	10.67	0.00	8.00	5.78
		wc%	7.64	35.91	13.33	17.16	39.33	19.57	22.16
	Gemini-2.0-Flash	wf%	4.67	1.60	0.67	2.00	0.67	3.33	2.16
		wc%	6.29	28.46	14.77	11.56	40.94	7.59	18.27
XCoT	**Mathstral-7B**	wf%	30.67	33.20	26.00	32.67	16.67	36.00	29.20
		wc%	49.04	75.45	35.14	59.41	64.00	54.17	56.20
	Llama-3.3-70B	wf%	5.33	10.40	8.00	6.67	21.33	8.00	9.96
		wc%	14.08	39.73	10.87	19.29	35.59	16.67	22.71
	DeepSeek-R1-Distill-Llama-70B	wf%	33.33	75.20	34.67	58.67	65.33	66.00	55.53
		wc%	18.00	30.65	16.33	19.35	44.23	9.80	23.06
	GPT-4o	wf%	6.67	21.60	2.67	19.33	6.67	13.33	11.71
		wc%	7.14	33.67	11.64	19.83	35.00	16.15	20.57
	Gemini-2.0-Flash	wf%	14.00	17.60	0.67	19.33	6.00	16.00	12.27
		wc%	4.65	25.24	10.74	9.92	41.84	8.73	16.85

Table 9: Comparative error analysis (from Run 2) evaluating BNaP, XLP, and XCoT prompt strategies. Metrics include wrong format (wf%) and wrong calculation (wc%) errors. Lower values reflect better adherence to format and reasoning correctness. Avg. indicates the average error across all evaluated tasks.