A Multilingual Evaluation of LLM-Based High-School Mathematics Assistance Across Indian Languages

Yash Kumar and Subhajit Roy
Indian Institute of Technology Kanpur, India
{yashk, subhajit}@iitk.ac.in

Abstract

Large Language Models (LLMs) are increasingly used by students as informal assistants for problem solving and concept clarification. Despite growing adoption, the multilingual reliability of these systems for educational mathematics remains underexplored, particularly across Indian regional languages. In this work, we present a multilingual evaluation of contemporary LLMs on high-school mathematics problems across six Indian languages: English, Hindi, Bengali, Urdu, Gujarati, and Malayalam. To guide the design of lightweight prompting conditions, we conduct a small exploratory prompt elicitation exercise with Grade 11–12 students and observe that students predominantly use concise direct-solution requests and lightweight expertise-oriented phrasing. Guided by these observations, we evaluate GPT-4o-mini, LLaMA 3-8B, and DeepSeek-R1 using simplified student-inspired prompting strategies on a parallel corpus of numerically perturbed textbook mathematics problems. Our results reveal consistent multilingual performance disparities, with accuracy strongest in English and lower across several regional languages. We further analyze the effects of prompting strategy and mathematical topic, and provide qualitative error analysis identifying recurring arithmetic, reasoning, and regional-language comprehension failures. These findings highlight the importance of evaluating multilingual educational reliability as LLM-based learning assistance becomes increasingly accessible across diverse linguistic settings.

1 Introduction

Large Language Models (LLMs) are increasingly being used as informal educational assistants for problem solving and concept clarification. Recent advances in conversational AI have made such systems easily accessible through freely available web interfaces and lightweight deployable models. While prior work has extensively studied the mathematical and reasoning capabilities of LLMs in English, comparatively limited attention has been given to their multilingual reliability across Indian regional languages in educational settings. This gap is particularly important in multilingual countries such as India, where a substantial number of students pursue school education in regional languages.

In this work, we present a multilingual evaluation of contemporary LLMs on high-school mathematics assistance across six Indian languages: English, Hindi, Bengali, Urdu, Gujarati, and Malayalam. We focus on high-school mathematics because it requires multi-step reasoning rather than simple factual recall, making it a useful setting for studying multilingual mathematical reasoning reliability. We additionally focus on *off-the-shelf* LLMs that are already widely accessible to students and educational institutions, namely GPT-4o mini, LLaMA 3-8B, and DeepSeek-R1.

To better ground our prompting strategies in plausible educational usage patterns, we conduct a small exploratory prompt elicitation exercise involving Grade 11–12 students. The collected responses commonly reflected concise direct-solution requests and lightweight expertise-oriented phrasing. Guided by these observations, we design simplified student-inspired prompting strategies rather than relying on heavily engineered prompting pipelines. Our goal is not to optimize model performance through sophisticated prompting techniques, but rather to study multilingual reliability under lightweight prompting conditions resembling typical educational interactions.

The scope of our study is defined by the following research question (**RQ**):

> *How reliably do contemporary LLMs provide high-school mathematics assistance across Indian languages under*

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 170–184
November 6-10, 2026 ©2026 Association for Computational Linguistics

lightweight student-inspired prompting conditions?

To conduct this study, we construct a parallel corpus of textbook mathematics problems aligned across six Indian languages. We extract problems from publicly available school textbooks, manually validate cross-lingual alignment, and generate numerically perturbed variants while preserving the underlying mathematical structure.

Our primary evaluation uses a single response per query to better approximate realistic educational usage. We analyze multilingual reliability across languages, topics, and prompting strategies using exact-answer accuracy and qualitative error analysis, while supplementary repeated-query experiments examine the gap between realistic and best-case performance.

Our results reveal substantial variation across languages, topics, and prompting strategies. Performance is consistently strongest in English and weaker in several regional languages, with recurring failures including arithmetic inconsistencies and regional-language comprehension errors.

Our work focuses on multilingual educational reliability under lightweight student-inspired prompting rather than advanced reasoning pipelines or optimized prompting techniques. Prior studies (Frieder et al., 2024; Anand et al., 2024; Imani et al., 2023; Li, 2024; Srivastava and Kim, 2024) have explored mathematical reasoning using sophisticated prompting and fine-tuned models. In contrast, we evaluate off-the-shelf LLMs in realistic multilingual educational settings rather than broader multilingual capabilities beyond mathematics assistance.

The primary contributions of this work are:

- A multilingual benchmark of high-school mathematics problems aligned across six Indian languages using textbook-derived and numerically perturbed questions.
- A multilingual evaluation of contemporary LLMs using lightweight student-inspired prompting strategies.
- An analysis of the effects of prompting strategy, mathematical topic, and language on multilingual mathematical reasoning reliability.
- A qualitative error analysis identifying recurring multilingual reasoning and interpretation failures.

Supplementary Material. We submit the complete datasets, prompting templates, experimental outputs, and supplementary analyses used in our study. The link to the github repository with all the scripts and the resources are mentioned here: Github repository link

2 Dataset

2.1 Dataset Description

To conduct the study, we created a dataset comprising 50 problems per mathematical topic (250 problems in total across five topics) from NCERT and SCERT Class 11 mathematics textbooks, yielding 1,500 problem instances across five topics and six languages. NCERT textbooks provide aligned mathematics content across English, Hindi, and Urdu, making them a natural source for constructing a multilingual parallel corpus. To further expand language coverage, we additionally collected corresponding problems from Bengali, Gujarati, and Malayalam SCERT textbooks. We sourced textbooks from the official NCERT and respective SCERT websites (nce, 2017; Mathematicians and experts, 2020; ben, 2018; Mathematicians and experts, 2017), and manually aligned problems across languages using textbook context, translation support, and manual verification.

Based on chapters containing sufficiently large numbers of parallel problems, we selected five mathematical topics (referred to as *problem categories*) along with representative textbook subtopics (referred to as *problem subcategories*):

- (Chapter 6) **Permutations and Combinations**: Fundamental Principle of Counting [10], Permutations [22], Combinations [18].
- (Chapter 9) **Straight Lines**: Distance of a point from a line [19], Slope of a line [14], Forms of Equation of a line [17].
- (Chapter 8) **Sequences and Series**: A&G Mean [10], Arithmetic Progression [15], Geometric Progression [15], Sequences [5], Series [5].
- (Chapter 10) **Limits and Derivatives**: Limits [20], Derivatives [25], Limits of Trigonometric Functions [5].
- (Chapter 12) **Conic Sections**: Circle [13], Parabola [13], Ellipse [13], Hyperbola [11].

Table 1 summarises the dataset structure, showing the number of problems per topic, the total per-language count, and the total instance count.

To reduce the effect of potential memorization from textbook exposure during pretraining, we created perturbed variants of the original problems by modifying the numerical values while preserving the underlying mathematical structure. Figure 1 shows an example of an original and perturbed

Topic	Problems	Instances (×6 langs)
P & C	50	300
Straight Lines	50	300
Sequences & Series	50	300
Limits & Derivatives	50	300
Conic Sections	50	300
Total	**250**	**1,500**

Table 1: Dataset statistics: 50 problems per topic across 6 languages yield 1,500 total instances.

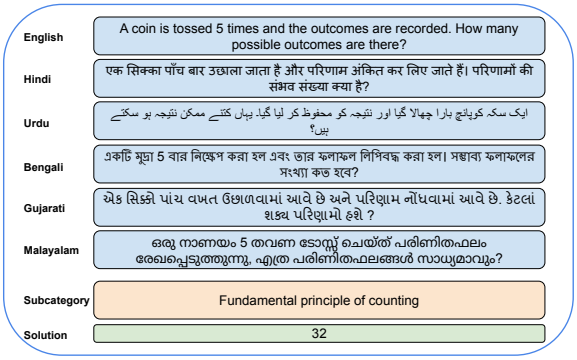


Figure 1: Sample problem from our dataset

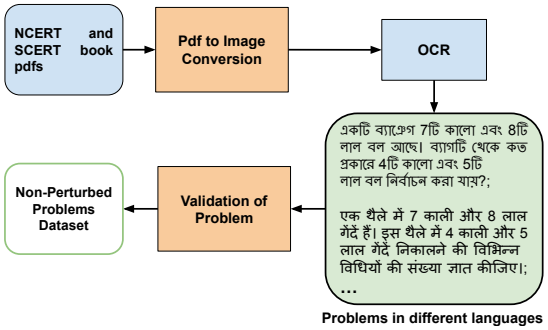


Figure 2: Dataset Creation Phase 1: Collect Problems

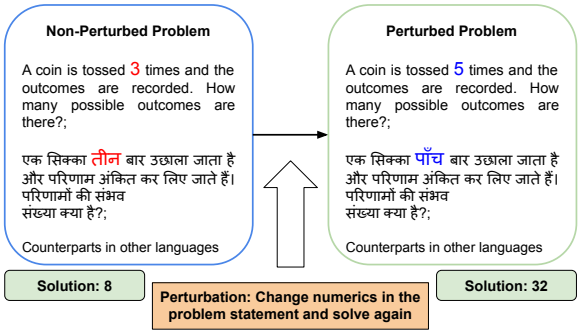


Figure 3: Dataset Creation Phase 2: Perturbations

problem across multiple languages. In this example, the associated subcategory is “Fundamental Principle of Counting”, and the manually derived solution to the perturbed problem is 32.

2.2 Dataset Creation Methodology

We use Optical Character Recognition (OCR) on the PDF files of textbooks to extract text in Unicode. This involved converting the textbook files of the book to an image format (JPEG) and exporting these images to Google Drive for use by the inbuilt OCR from Google. For the SCERT books, we had to search for problems that also appeared in the NCERT books; further, we used Google Translate (Jha and Patil, 2023), (Mondal et al., 2019) to validate that the problems were indeed the same using combination of Google Translate and our understanding of those particular languages. This procedure is shown in Figure 2.

We used the above procedure to extract 50 problems along with their numerical solutions from five different mathematical categories: Combinatorics, Straight Lines, Sequences and Series, Limits and Derivatives, and Conic Sections. We categorized these problems based on the section they appear in. For instance, in the Straight Lines category, we identified three subcategories in the NCERT/SCERT textbook: "Distance of a Point from a Line", "Slope of a Line", and "Forms of Equation of a Line".

These problems were perturbed by changing the numerical values (see Figure 3). The reason for introducing these perturbations is that the LLM may have encountered these problems during training, and hence, may employ *recollection* rather than *reasoning*. We first validated the perturbation procedure on 25 P&C problems, observing performance drops of 4–20 pp across prompting strategies, confirming reduced memorization effects. We then applied the identical perturbation methodology to all remaining problems across all five topics and six languages.

After introducing these perturbations, the perturbed problems were solved manually to find the ground-truth solutions. All perturbed variants were manually verified to confirm mathematical validity.

3 Prompting Strategies

3.1 Prompt Elicitation Exercise

To better ground our prompting strategies in realistic educational usage, we conducted a small prompt elicitation exercise with 53 Grade 11–12 students from schools in Uttar Pradesh, India. Students were shown a representative mathematics problem and asked how they would prompt an LLM-based assistant such as ChatGPT.

Most responses consisted of concise direct-solution requests, lightweight expertise-oriented prompts, or prompts with minimal contextual framing. Only a small number of students explicitly referred to mathematical subtopics or structured prompting techniques.

Students were additionally asked to select their preferred prompting format from lightweight prompting styles corresponding to the strategies explored in this work. Table 2 summarizes the observed preferences.

Prompt Style Preference	Count	%
Zero-Shot with topic-level expertise framing	21	40%
Plain direct-solution prompt	14	26%
Zero-Shot with subtopic information	9	17%
Multiple / other preferences	8	15%
Bare problem statement only	1	2%

Table 2: Observed prompt-style preferences from the exploratory student prompt elicitation exercise.

The responses suggest that students generally prefer concise prompting styles with lightweight contextual framing rather than heavily engineered prompting techniques. These observations informed our selection of simplified student-inspired prompting strategies for the experiments in this work.

Limitations of the elicitation sample. The 53 participants are Grade 11–12 students from Uttar Pradesh, reflecting a predominantly Hindi/English-medium interaction context. This sample cannot speak for Bengali, Gujarati, or Malayalam medium students, precisely the sub-populations for whom we observe the largest multilingual performance gaps. We acknowledge that regional-language students may prompt in-script, code-switch differently, or exhibit different tool-use competence.

Additional details regarding the survey instrument and anonymized responses are provided in the supplementary material.

Now, we discuss our prompting strategies. Recall that we refer to the chapter and section in the chapter from which the problem is taken as the *category*, and *subcategory* of the problem, respectively.

3.2 [Strategy 1] Zero-Shot Prompting

In this case, we use Zero-Shot prompts that requests the LLM to solve the problem, followed by the problem. We do mention that the LLM should use its expertise in the respective category (Combinatorics, Limits and Derivatives, etc.). The prompt remains the same for all the languages, except for the problem. Figure 4 shows a sample prompt. The instruction wrapper (e.g., "You are expert in solving problems in [Category]...") is written in English, with only the mathematical problem itself in the target language.

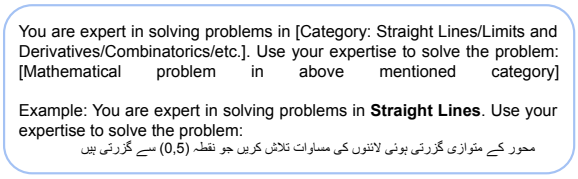


Figure 4: Zero Shot Prompt

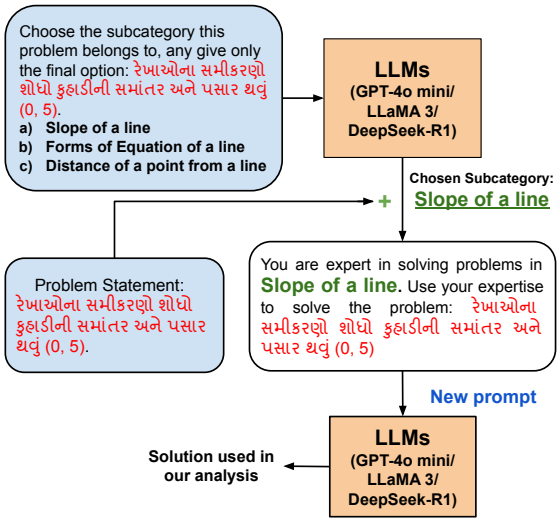


Figure 5: Self-Ask Prompt

3.3 [Strategy 2] Self-Ask Prompting

The Self-Ask prompting was proposed in (Press et al., 2023). In this strategy, we initiate a dialog with the LLM by first asking it to choose a suitable subcategory from respective subcategories of the problem's category (see Section 2). Next, we create another prompt using the subcategory chosen by the LLM and prompt it to solve that problem. We show an illustration in Figure 5: the LLM chooses "Slope of a line" as the subcategory, which is used to create a new prompt. We prepend the complete history of the conversation in each prompt to provide the complete context.

Note on Self-Ask motivation. Self-Ask was included as a lightweight conversational decomposition strategy motivated by expert-framing patterns observed in the elicitation responses, rather than as a direct instantiation of a surveyed student

You are expert in solving problems in [Subcategory of the problem to which it belongs to]. Use your expertise to solve the problem: [Mathematical problem in a chosen language from this subcategory]

Example: You are expert in solving problems in **Forms of Equation of a line**. Use your expertise to solve the problem:
محور کے متوازی گزرتی ہوئی لائنوں کی مساوات تلاش کریں جو نقطہ (0,5) سے گزرتی ہیں

Figure 6: Zero Shot Prompt using Subcategory of the Problem

Table 3: LLMs in our experiments

Model	Hosting	License	Purpose	Params.
GPT-4o Mini	*remote*	*proprietary*	*general*	*undisclosed*†
LLaMA 3-8B	*local*	*open-source*	*general*	*8B*
DeepSeek-R1	*local*	*open-source*	*reasoning*	*7B*

†OpenAI has not publicly disclosed the parameter count for GPT-4o mini.

preference.

3.4 [Strategy 3] Zero-Shot Prompting with Topic Information

The prompts created here are similar to Strategy 1 (Zero-Shot prompting), but instead of providing the category of a problem, we use the subcategory of that problem. We show an illustration in Figure 6: we provide the information that the problem belongs to the subcategory to which this problem belongs (limits and derivatives, etc.).

3.5 Structured Output for Evaluation

For automated evaluation, some prompt variants additionally requested answers in a lightweight structured format to simplify multilingual answer extraction and consistency checking. The underlying mathematical queries remained unchanged across prompting variants. Simplified prompt examples are shown in the main paper, while the exact prompts are provided in the supplementary material.

4 Experiments

We employ the perturbed dataset (see Section 2), which contains 50 parallel mathematical problems across five topics and six languages (1,500 instances total). We use three LLMs: the general-purpose remotely-hosted proprietary GPT-4o mini (Achiam et al., 2023), the general-purpose locally-hosted open-source LLaMA 3-8B (lla, 2024), and the open-source locally-hosted reasoning model DeepSeek-R1 (7B) (Guo et al., 2025) (see Table 3). We conduct all experiments on an NVIDIA RTX A4000 GPU (16 GB card), using the prompting strategies discussed in Section 3.

Evaluation Metric. Since LLM responses are stochastic, each prompt is queried over three independent trials, with each trial commencing from an empty context. Our **primary metric is single-trial accuracy** (**Trial-1**): the percentage of problems for which the model's first-trial response matches the ground-truth numerical answer. This metric best approximates a realistic educational interaction, in which a student submits one query and acts on the response received. As a supplementary measure, we additionally report **any-of-3 accuracy**: the percentage of problems answered correctly in at least one of the three trials. This upper-bound metric quantifies the additional coverage gained through repeated querying and allows us to assess the gap between best-case and realistic performance. As we show in Section 4.2, this gap is substantially larger for weaker models, motivating our use of single-trial accuracy as the primary measure.

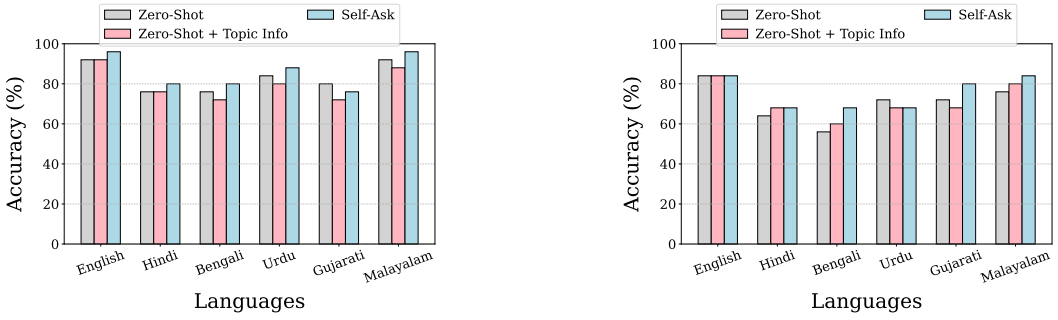
Given the relatively small benchmark size, our analysis focuses primarily on descriptive trends rather than formal statistical significance testing.

4.1 Justification of Perturbed Dataset

We introduce perturbations in our dataset instead of using the original problems, as an LLM may have encountered these problems during training and employ *recollection* rather than *reasoning*. To validate this, we compared model performance on 25 original (unmodified) versus perturbed problems across languages in the *Permutations and Combinations* category. Figure 7 shows the results for GPT-4o mini: performance declines by 4–20 percentage points on the perturbed problems across all prompting strategies. This is consistent with reduced memorization effects and encourages generalization beyond textbook recall. The identical perturbation methodology was subsequently applied to all 1,500 problem instances across all five topics and six languages.

4.2 Results

Table 4 summarizes overall Trial-1 accuracy across models and prompting strategies, while Figure 9 summarizes topic-wise performance trends. Detailed language-wise and topic-wise Trial-1 breakdown plots are provided in Appendix A, while Appendix B presents the corresponding any-of-3 analyses. Unless otherwise noted, all discussion refers to single-trial (Trial-1) accuracy; any-of-3 results are discussed in Section 4.2.6 as a supplementary measure.



(a) **Original** problem set results

(b) **Perturbed** problem set results

Figure 7: Performance of GPT-4o mini on: Original and Perturbed Problem Set

Model	Prompting Strategy				Language Avg.		
	Zero-Shot	Topic Info	Self-Ask	Overall	English	Hindi	Malayalam
GPT-4o mini	62.2	64.7	65.7	**64.2**	66.8	59.6	49.5
LLaMA 3-8B	17.8	20.7	24.5	**21.0**	31.7	21.0	18.0
DeepSeek-R1	47.9	47.1	43.7	**46.2**	55.6	47.8	31.8

Table 4: Overall Trial-1 accuracy (%) across prompting strategies and representative languages. All models perform best on English and degrade substantially on regional languages such as Malayalam. Bootstrap 95% CIs for the English–Malayalam gap are non-overlapping across all models (e.g., GPT-4o mini: English [41.9, 47.9] vs. Malayalam [27.0, 32.7] under equivalence-aware scoring), confirming this finding is robust. Intermediate contrasts (e.g., Hindi vs. Bengali) overlap and should be treated as descriptive trends.

4.2.1 Overall Model Performance

GPT-4o mini consistently achieves the strongest multilingual mathematical performance across all prompting strategies, obtaining an overall Trial-1 accuracy of **64.2%**. DeepSeek-R1 follows with **46.2%**, while LLaMA 3-8B achieves **21.0%**. Despite comparable parameter scales (7B–8B), the models exhibit substantial differences in multilingual mathematical reliability, suggesting that instruction tuning, multilingual representation quality, and reasoning optimization contribute more strongly than parameter count alone.

4.2.2 Effect of Prompting Strategy

The effect of prompting strategy is strongly model-dependent. For GPT-4o mini, Self-Ask prompting improves Trial-1 accuracy by approximately **+3.5 pp** over Zero-Shot prompting, while LLaMA 3-8B benefits more substantially with a gain of approximately **+6.7 pp**. In contrast, DeepSeek-R1 performs strongest under standard Zero-Shot prompting and slightly degrades under Self-Ask prompting by approximately $-$**4.2 pp**. This suggests that conversational prompting may benefit general-purpose instruction-tuned models more than reasoning-oriented models optimized for direct single-turn inference.

Across all models, Zero-Shot prompting with explicit topic information performs comparably to or slightly below Self-Ask prompting, indicating that lightweight conversational decomposition may provide more useful reasoning guidance than simply specifying the mathematical subtopic upfront.

4.2.3 Effect of Problem Topic

As shown in Figure 9, topic difficulty is reasonably consistent across models. Averaged across all models, languages, and prompting strategies, Trial-1 accuracy ranks as follows: Limits and Derivatives (**60.3%**) $>$ Sequences and Series (**42.8%**) $>$ Conic Sections (**41.3%**) $>$ Straight Lines (**30.8%**) $>$ Permutations and Combinations (**28.4%**). Permutations and Combinations emerges as the most challenging topic overall, while Limits and Derivatives consistently achieves the strongest performance across all models. Given the small number of problems per topic, these rankings should be interpreted as descriptive trends; bootstrap CIs show the Limits vs. P&C gap is distinguishable (Limits: 61.3% [58.1, 64.4] vs. P&C: 28.0% [26.1, 30.0]), while intermediate topic contrasts overlap.

GPT-4o mini and DeepSeek-R1 achieve particularly strong performance on Limits and Derivatives, exceeding **80%** and **65%** average Trial-1 accuracy respectively. LLaMA 3-8B exhibits the largest variation across topics, performing compar-

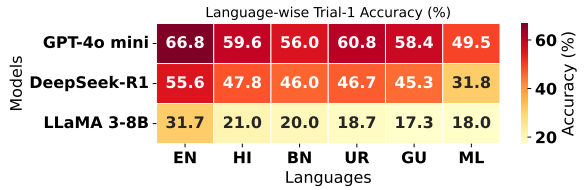


Figure 8: Heatmap of overall Trial-1 accuracy (%) across models and languages. English consistently achieves the strongest performance, while Malayalam exhibits the largest multilingual degradation. GPT-4o mini shows the strongest overall multilingual robustness.

atively better on procedural symbolic manipulation tasks while struggling on combinatorial and geometric reasoning problems. Overall, these results suggest that current LLMs handle procedural algebraic transformations more reliably than problems requiring deeper combinatorial reasoning and arithmetic consistency.

4.2.4 Effect of Language

All three models perform best on English problem statements, consistent with the dominance of English in LLM pretraining corpora (Figure 8). Performance across the six languages follows a rough hierarchy: English $>$ Hindi/Urdu $>$ Bengali/Gujarati $>$ Malayalam. Malayalam is consistently the weakest language across all models, exhibiting the largest degradation relative to English. Hindi and Urdu generally show stronger multilingual performance, while Bengali and Gujarati occupy an intermediate position. GPT-4o mini demonstrates the strongest multilingual robustness overall, whereas LLaMA 3-8B exhibits substantial degradation across regional languages. The language gap is present across all prompting strategies, although Self-Ask prompting partially reduces this gap for GPT-4o mini and LLaMA 3-8B in several non-English settings.

4.2.5 Equivalence-Aware Accuracy

Exact-match evaluation can penalise mathematically correct answers that differ only in symbolic form (e.g., -8 vs. $x + 8 = 0$). To quantify this effect, we applied a three-layer deterministic equivalence checker to all Trial-1 responses: (1) numeric extraction with comma normalisation (e.g., 60,480 = 60480; 3/5 = 0.6); (2) symbolic equation solving (e.g., $x + 8 = 0$ vs. -8); and (3) SymPy algebraic simplification. Table 5 reports exact-match and equivalence-aware accuracy with bootstrap 95% confidence intervals (2,000 resamples).

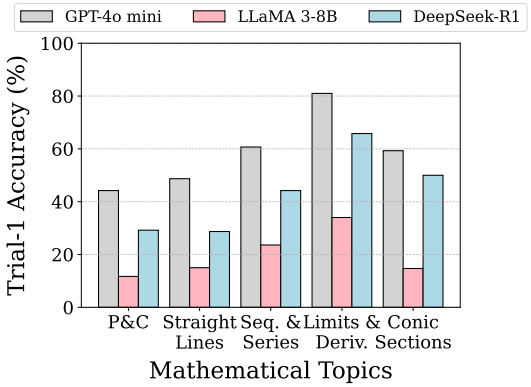


Figure 9: Average Trial-1 accuracy (%) across mathematical topics. Limits and Derivatives achieves the strongest overall performance, while Permutations and Combinations remains the most challenging topic.

Model	Exact-match	Equiv.-aware	Gain
GPT-4o mini	40.2% [37.0, 43.3]	50.9% [47.7, 54.2]	+10.7 pp
DeepSeek-R1	21.9% [20.6, 23.3]	39.0% [37.5, 40.6]	+17.1 pp
LLaMA 3-8B	15.9% [14.3, 17.7]	18.3% [16.5, 20.1]	+2.3 pp

Table 5: Equivalence-aware vs. exact-match Trial-1 accuracy with bootstrap 95% CIs. The English–Malayalam hierarchy is fully preserved under robust scoring; the largest gains appear for Conic Sections (+19.3 pp), consistent with the equivalent-form mismatch failures identified in Section 4.3.

The language hierarchy of Section 4.2 is fully preserved under equivalence-aware scoring: English 44.9% [41.9, 47.9] $>$ Hindi 35.2% [32.3, 38.3] $>$ Bengali 34.0% [31.1, 36.8] $>$ Urdu 32.4% [29.7, 35.2] $>$ Gujarati 32.5% [29.7, 35.3] $>$ Malayalam 29.8% [27.0, 32.7]. The English–Malayalam gap remains non-overlapping, confirming the core multilingual disparity finding is robust to metric choice. Format failures (unparseable outputs) are slightly more frequent in regional languages (avg. 11.8%) than in English (9.6%), with LLaMA 3-8B and DeepSeek-R1 showing the largest regional overhead (up to 15.2% for Malayalam).

4.2.6 Any-of-3 Inflation

To quantify the gap between realistic and best-case performance, we compare Trial-1 accuracy to any-of-3 accuracy (the percentage of problems solved in at least one of three trials). Table 6 compares Trial-1 and any-of-3 accuracy across models. The inflation from single-trial to any-of-3 is strongly model-dependent: GPT-4o mini gains **+6.4 pp** under Zero-Shot and **+5.0 pp** under Self-Ask, while LLaMA 3-8B gains **+10.5 pp** and **+12.5 pp** respec-

Model	Trial-1	Any-of-3	Inflation
GPT-4o mini	64.2%	73.5%	+9.3 pp
DeepSeek-R1	46.2%	61.4%	+15.2 pp
LLaMA 3-8B	21.0%	34.3%	+13.3 pp

Table 6: Comparison between realistic single-query performance (Trial-1) and repeated-query upper-bound performance (Any-of-3). Weaker models exhibit substantially larger inflation under repeated querying.

tively, and DeepSeek-R1 gains **+11.2 pp** and **+15.4 pp**. Weaker models benefit disproportionately from repeated querying, meaning that any-of-3 accuracy substantially overstates their reliability in a realistic single-query educational setting. This motivates our use of Trial-1 as the primary metric.

4.3 Error Analysis

To better understand multilingual model failures beyond aggregate accuracy scores, we manually examined incorrect Trial-1 responses across models, languages, and mathematical topics. Our analysis revealed several recurring categories of failure.

Arithmetic and reasoning failures. A common failure mode involved models identifying an appropriate solution strategy but making arithmetic mistakes during intermediate computations. For example, in a Bengali Permutations and Combinations problem with ground-truth answer 49,086,048, LLaMA 3 generated the answer 1858 despite partially correct reasoning steps. Similarly, in several multi-step algebraic problems, models produced expressions that were structurally close to the correct answer but contained incorrect coefficients or constants. These observations suggest that maintaining symbolic and arithmetic consistency across multiple reasoning steps remains challenging, particularly for smaller models.

Equivalent-form mismatches. For several Conic Sections problems, models generated mathematically equivalent but syntactically different answers. For example, for a problem whose expected answer was represented as $x + 8 = 0$, GPT-4o-mini generated only the equivalent directrix value -8. In another example, GPT-4o-mini produced the expression $y = \frac{1}{3}x^2 + \frac{2}{3}x + 4$ instead of the ground-truth form $3y = x^2 + 4x + 12$. Such cases highlight the difficulty of evaluating symbolic mathematical reasoning using strict answer matching alone and partially explain the comparatively low accuracy observed for Conic Sections. The equivalence-aware scorer in Section 4.2.5 recovers these cases automatically.

Format compliance and truncation failures. We additionally observed cases where models failed to generate parseable final answers. DeepSeek-R1 occasionally echoed portions of the prompt template itself (e.g., returning phrases such as "Numeric answer" instead of a numerical value), while LLaMA 3 sometimes produced truncated or incomplete outputs without a clearly identifiable final answer. These failures were substantially more common for smaller models and under regional-language prompting conditions, consistent with the higher format-failure rates observed in Section 4.2.5.

Regional-language comprehension failures. Some multilingual failures appeared to stem from language understanding rather than mathematical reasoning alone. In several Gujarati and Malayalam examples, models appeared to misinterpret the problem statement or mathematical terminology, occasionally solving a different mathematical formulation than the one posed. We also observed occasional language interference phenomena where reasoning traces partially switched into English or transliterated Hindi despite the original query being written entirely in a regional script. These observations suggest that multilingual mathematical reasoning failures may arise from upstream comprehension limitations in addition to symbolic reasoning errors. Note that OCR was used only for dataset construction and was not part of the model inference pipeline; Google Translate was used only to validate NCERT/SCERT problem alignment, not to generate or translate problems.

4.4 Summary

GPT-4o mini is the strongest model overall, followed by DeepSeek-R1, while LLaMA 3-8B exhibits substantially weaker multilingual mathematical reasoning performance. All models perform best on English problem statements, with Malayalam showing the largest multilingual degradation. The core English–Malayalam disparity is robust to metric choice, holding under both exact-match and equivalence-aware evaluation with non-overlapping bootstrap confidence intervals.

Prompting effectiveness is strongly model-dependent: Self-Ask prompting improves performance for GPT-4o mini and LLaMA 3-8B, whereas DeepSeek-R1 performs strongest under simpler

Lang.	Topic	Model	Error Type	Expected	Model Output	Observation
Bengali	P&C	LLaMA 3	Arithmetic failure	49,086,048	"1858"	Correct reasoning path but incorrect arithmetic computation.
English	Conics	GPT-4o-mini	Eqv.-form mismatch	$x + 8 = 0$	"-8"	Correct directrix value generated instead of symbolic equation form.
English	Conics	GPT-4o-mini	Reasoning breakdown	$3y = x^2 + 4x + 12$	"$y = \frac{1}{3}x^2 + \frac{2}{3}x + 4$"	Incorrect coefficient introduced during algebraic transformation.
Gujarati	Conics	DeepSeek-R1	Comprehension failure	Parabola equation	Solved as ellipse problem	Gujarati mathematical description misinterpreted.
Malayalam	Straight Lines	LLaMA 3	Language interference	Numerical answer	Mixed English reasoning with incomplete Malayalam output	Code-mixed and incomplete final response.
English	Multiple Topics	DeepSeek-R1	Format failure	Numerical answer	"Numeric answer"	Prompt template echoed instead of numerical solution.

Table 7: Representative multilingual failure cases observed during manual analysis of Trial-1 responses.

Zero-Shot prompting. Across mathematical domains, Permutations and Combinations emerges as the most challenging topic, while Limits and Derivatives achieves the strongest performance.

Finally, the substantial gap between Trial-1 and any-of-3 evaluation for weaker models shows that repeated querying can considerably inflate apparent reliability relative to realistic single-query educational usage.

5 Related Work

Prior work has extensively explored the mathematical reasoning capabilities of LLMs, particularly in English. Several studies have evaluated LLMs on mathematical word problems and benchmark datasets (Frieder et al., 2024; Srivastava and Kim, 2024; Imani et al., 2023). Mathify (Anand et al., 2024) introduced the MathQuest dataset based on English NCERT mathematics problems for evaluating and fine-tuning open-source models. In contrast, our work focuses on multilingual educational mathematics across multiple Indian languages using lightweight student-inspired prompting strategies.

Recent multilingual reasoning studies have explored mathematical reasoning across languages such as Chinese, Vietnamese, Arabic, Spanish, Russian, and Japanese (Li, 2024; Wu et al., 2024; Le et al., 2024). The closest work to ours is (Kumar and Roy, 2025), which evaluates LLMs on Hindi combinatorics problems. Our study substantially expands this setting across multiple mathematical domains and Indian regional languages.

Several studies have additionally explored the broader use of LLMs in education, including reading comprehension generation, tutoring systems, and database education (Xiao et al., 2023; Zhai, 2022; Prakash et al., 2024). Prior work has also examined risks associated with LLM usage in educational settings, including AI-generated content detection in examinations (Orenstrakh et al., 2024).

Multilingual NLP for Indian languages. Several efforts have focused on improving NLP support for Indian languages through multilingual pretraining and benchmark creation. IndicNLP-Suite (Kakwani et al., 2020), MuRIL (Khanuja et al., 2021), IndicTrans2 (Gala et al., 2023), and Airavata (Gala et al., 2024) improved multilingual resources, translation, and conversational capabilities for Indian languages. However, comparatively little work has investigated multilingual mathematical reasoning reliability in educational settings across Indian regional languages, which our work begins to address. Future work should evaluate Indic-focused models such as Airavata (Gala et al., 2024) and MuRIL (Khanuja et al., 2021) directly on this benchmark.

6 Conclusions

Our experiments reveal consistent multilingual reliability disparities across Indian languages, substantial variation across mathematical topics, and strong model dependence on prompting strategy. We additionally observe that repeated-query evaluation can substantially inflate apparent model reliability, particularly for weaker models. Qualitative analysis further suggests that multilingual degradation reflects not only arithmetic errors, but also regional-language comprehension challenges in under-resourced settings. These findings highlight the importance of careful multilingual evaluation when deploying LLM-based educational assistance systems across linguistically diverse environments.

Ethical Considerations

The prompt elicitation exercise was conducted with the knowledge and consent of participants. We collected only minimal academic metadata and excluded personally identifying information from the released data. All survey responses were anonymized prior to release.

Our work evaluates the multilingual reliability of existing LLMs in educational mathematics settings and does not advocate replacing qualified instruction with automated systems. We note that multilingual reliability disparities may disproportionately affect students studying in regional languages.

Limitations

Our study has several limitations. First, the benchmark contains only 50 parallel problems per mathematical topic across six Indian languages making a total of 1500 problems, which limits statistical robustness and topic coverage. Expanding the dataset to additional mathematical domains and educational sources would improve generalizability.

Second, we evaluate only lightweight prompting strategies representative of typical student interactions and do not explore model fine-tuning or advanced reasoning pipelines. Future work may investigate whether smaller instruction-tuned or fine-tuned multilingual models including Indic-focused systems such as Airavata (Gala et al., 2024) and MuRIL (Khanuja et al., 2021) can provide stronger and more reliable educational assistance or not.

Third, the exact-match evaluation metric can conflate reasoning failures with format or extraction failures. We address this partially through the equivalence-aware scorer described in Section 4.2.5, though residual measurement uncertainty remains for edge cases not covered by symbolic simplification.

Fourth, the instruction wrapper for all prompting strategies is written in English, with only the mathematical problem in the target language. Adapting instruction language per condition is a valuable ablation for future work. Since this design is consistent across all conditions and models, cross-lingual comparisons within our study remains internally valid.

Fifth, the prompt elicitation sample consists of 53 Grade 11–12 students from Uttar Pradesh, reflecting a predominantly Hindi/English-medium interaction context. This sample cannot represent Bengali, Gujarati, or Malayalam-medium students, whose prompting behaviour, in-script usage, and tool competence may differ substantially from the observed patterns.

Acknowledgments

We would like to thank the anonymous reviewers and the meta-reviewer for their insightful comments and suggestions. This research work was partially supported by the Research-I Foundation of the Department of CSE at IIT Kanpur.

References

2017. National council of educational research and training. https://ncert.nic.in/textbook.php.

2018. Bengali scert books. https://www.scerttripura.org/ebooks.php.

2024. Meta llama team. introducing meta llama 3: The most capable openly available llm to date. (accessed on this url). https://ai.meta.com/blog/meta-llama-3/.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report.

Avinash Anand, Mohit Gupta, Kritarth Prasad, Navya Singla, Sanjana Sanjeev, Jatin Kumar, Adarsh Raj Shivam, and Rajiv Ratn Shah. 2024. Mathify: Evaluating large language models on mathematical problem solving tasks. *arXiv preprint arXiv:2404.13099*.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. 2024. Mathematical capabilities of chatgpt. *Advances in neural information processing systems*.

Jay Gala, Pranjal A. Chitale, A. K. Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth M. Kumar, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. 2024. Airavata: Introducing hindi instruction-tuned llm. *arXiv preprint arXiv:2401.15006*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. In *ACL (industry)*. Association for Computational Linguistics.

Abhinav Jha and Hemprasad Yashwant Patil. 2023. A review of machine transliteration, translation, evaluation metrics and datasets in indian languages. *Multimedia Tools and Applications*, 82(15):23509–23540.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 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*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Yash Kumar and Subhajit Roy. 2025. Studying the capabilities of large language models in solving combinatorics problems posed in hindi. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 90–99.

Nguyen-Khang Le, Dieu-Hien Nguyen, Dinh-Truong Do, Chau Nguyen, and Minh Le Nguyen. 2024. Vietnamese elementary math reasoning using large language model with refined translation and dense-retrieved chain-of-thought. In *JSAI International Symposium on Artificial Intelligence*.

Alkhouli Tamer Bonadiman Daniele Pappas Nikolaos Mansour Saab Li, Bryan. 2024. Eliciting better multilingual structured reasoning from LLMs through code. In *Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Mathematicians and Gujarati Language experts. 2020. Gujarati scert books. https://gujarat-education.gov.in/TextBook/Textbooks/new-syllabus2020.htm.

Mathematicians and Malayalam Language experts. 2017. Malayalam scert books. https://samagra.kite.kerala.gov.in.

Himel Mondal, Shaikat Mondal, and Sarika Mondal. 2019. Feasibility of using "google translate" in adaptation of survey questionnaire from english to bengali: a pilot study. *Indian Journal of Social Psychiatry*, 35(2):119–124.

Michael Sheinman Orenstrakh, Oscar Karnalim, Carlos Anibal Suarez, and Michael Liut. 2024. Detecting llm-generated text in computing education: Comparative study for chatgpt cases. In *IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE.

Kishore Prakash, Shashwat Rao, Rayan Hamza, Jack Lukich, Vatsal Chaudhari, and Arnab Nandi. 2024. Integrating llms into database systems education. *Data Systems Education: Bridging education practice with education research*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. Association for Computational Linguistics.

Nilesh Srivastava and Seongchan Kim. 2024. Enhancing mathematical reasoning in math word problems: A numerical masking approach for encoder-decoder models. *Available at SSRN 4790013*.

Yanan Wu, Jie Liu, Xingyuan Bu, Jiaheng Liu, Zhanhui Zhou, Yuanxing Zhang, Chenchen Zhang, Zhiqi Bai, Haibin Chen, Tiezheng Ge, and 1 others. 2024. Conceptmath: A bilingual concept-wise benchmark for measuring mathematical reasoning of large language models. *arXiv preprint arXiv:2402.14660*.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by llms: A showcase of chatgpt in education applications. In *BEA 2023*.

Xiaoming Zhai. 2022. Chatgpt user experience: Implications for education. *Available at SSRN 4312418*.

A Appendix

A.1 Detailed Topic-wise Multilingual Results

Figures 10–12 provide detailed language-wise Trial-1 accuracy breakdowns across prompting strategies for all evaluated mathematical topics and models. Each figure contains three subplots corresponding to GPT-4o mini, LLaMA 3-8B, and DeepSeek-R1 respectively.

A.2 Any-of-3 Accuracy Across Topics and Languages

Figures 13–14 show the corresponding any-of-3 accuracy (%) results, where a problem is considered solved if the model produces the correct answer in at least one of three independent trials. Compared to Trial-1 accuracy, weaker models exhibit substantially larger gains under repeated querying.

B Supplementary Material

All resources for this study are publicly available in our GitHub repository: Github repository link

Datasets. The `Datasets/` folder contains five CSV files, one per mathematical topic (`PnC.csv`, `StraightLines.csv`, `SeqnSeries.csv`, `LimitsnDerivatives.csv`, `conic_sections.csv`). Each file contains 50 parallel problems across six languages (English, Hindi, Bengali, Urdu, Gujarati, Malayalam), with columns for problem text in each language, the perturbed ground-truth solution, and the original unperturbed solution.

Inference Scripts. The `Supplementary_Materials/` folder contains one inference script per model: `GPT-4o/test_openai.py`, `LLaMA 3/llama_fast.py`, and `DeepSeek-R1/deepseek_r1.py`. Each script accepts a topic and prompting strategy as arguments and queries the respective model over three independent trials from an empty context, saving per-record results as CSV files. Results are organized by model, topic, and prompting strategy (`One_Shot`, `One_Shot_Subcat`, `CoT` corresponding to Zero-Shot, Zero-Shot + Topic Info, and Self-Ask respectively).

Equivalence-Aware Scoring and Bootstrap CIs. `robust_equivalence_scoring.py` implements the three-layer deterministic equivalence checker described in Section 4.2.5: numeric extraction with comma normalisation, symbolic equation solving, and SymPy algebraic simplification. Running it over the results produces per-record flags and summary CSVs aggregated by model, language, and topic, available in the `output/` folder. `bootstrap_ci.py` reads these outputs and computes bootstrap 95% confidence intervals (2,000 resamples), reproducing the CI tables reported in Section 4.2.5.

Survey Results. The `Supplementary_Materials/Survey_Results/` folder contains the anonymized responses from the prompt elicitation exercise conducted with 53 Grade 11–12 students. The CSV includes student responses and prompt-style preferences. Email addresses and all personally identifying fields were removed prior to release, consistent with the anonymization procedure described in the Ethical Considerations section.

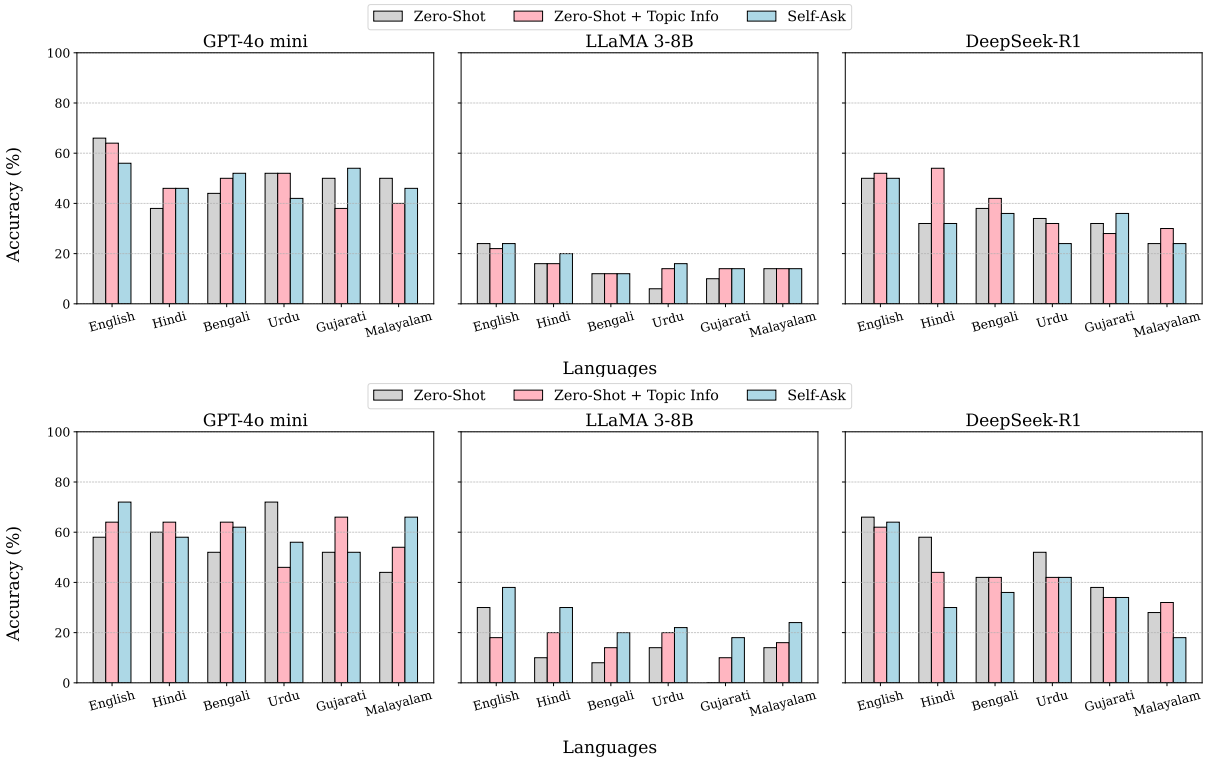


Figure 10: Trial-1 accuracy (%) on Permutations and Combinations (top) and Straight Lines (bottom) problems across languages and prompting strategies.

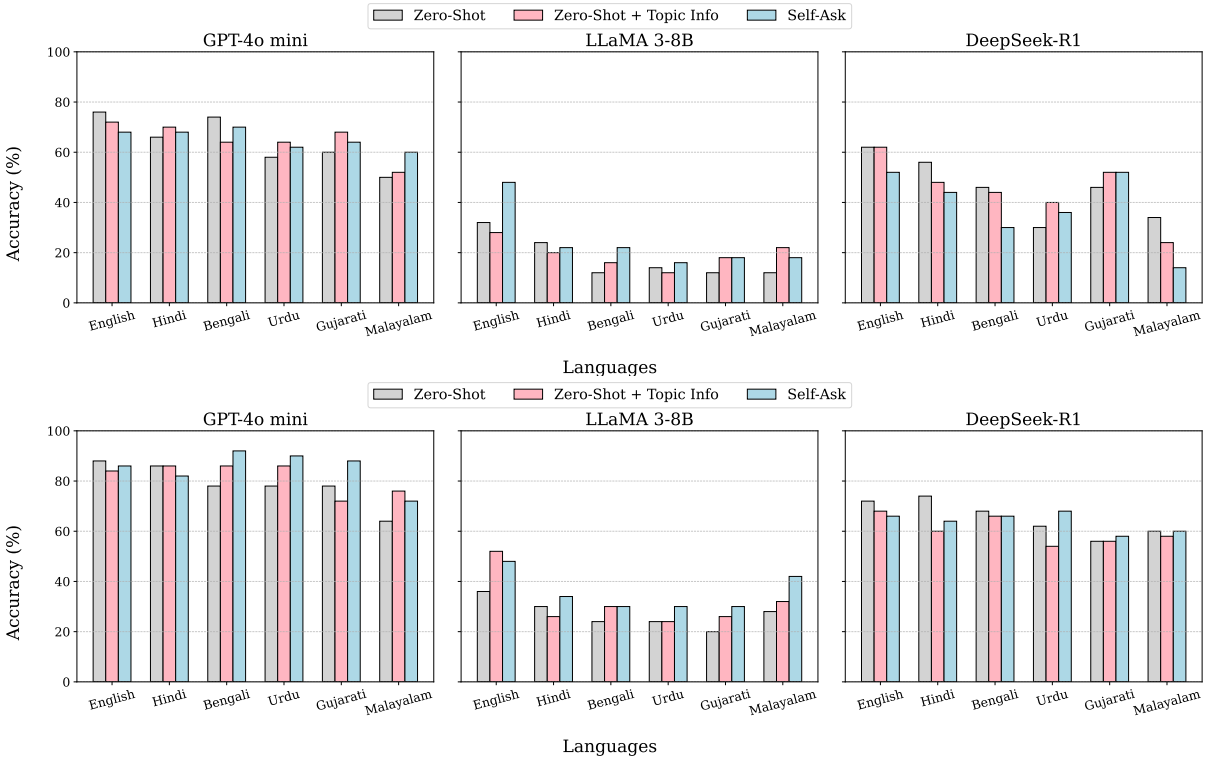


Figure 11: Trial-1 accuracy (%) on Sequences and Series (top) and Limits and Derivatives (bottom) problems across languages and prompting strategies.

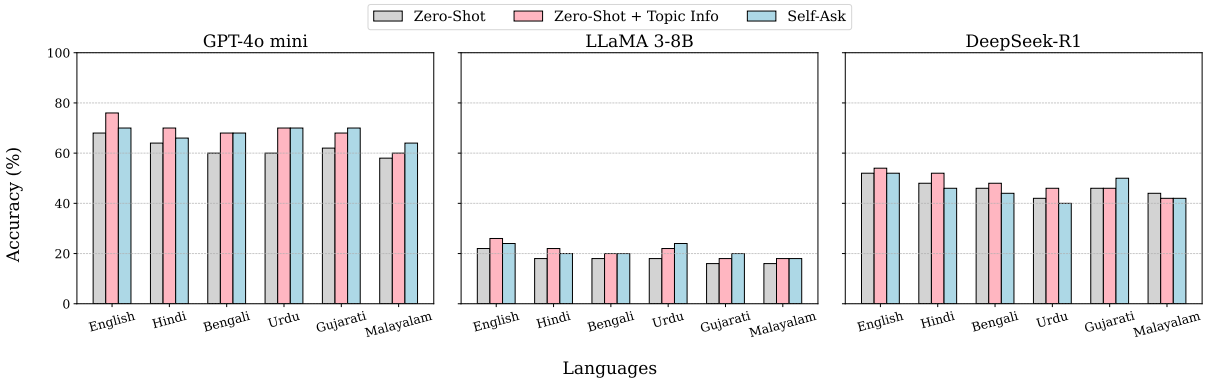


Figure 12: Trial-1 accuracy (%) on Conic Sections problems across languages and prompting strategies.

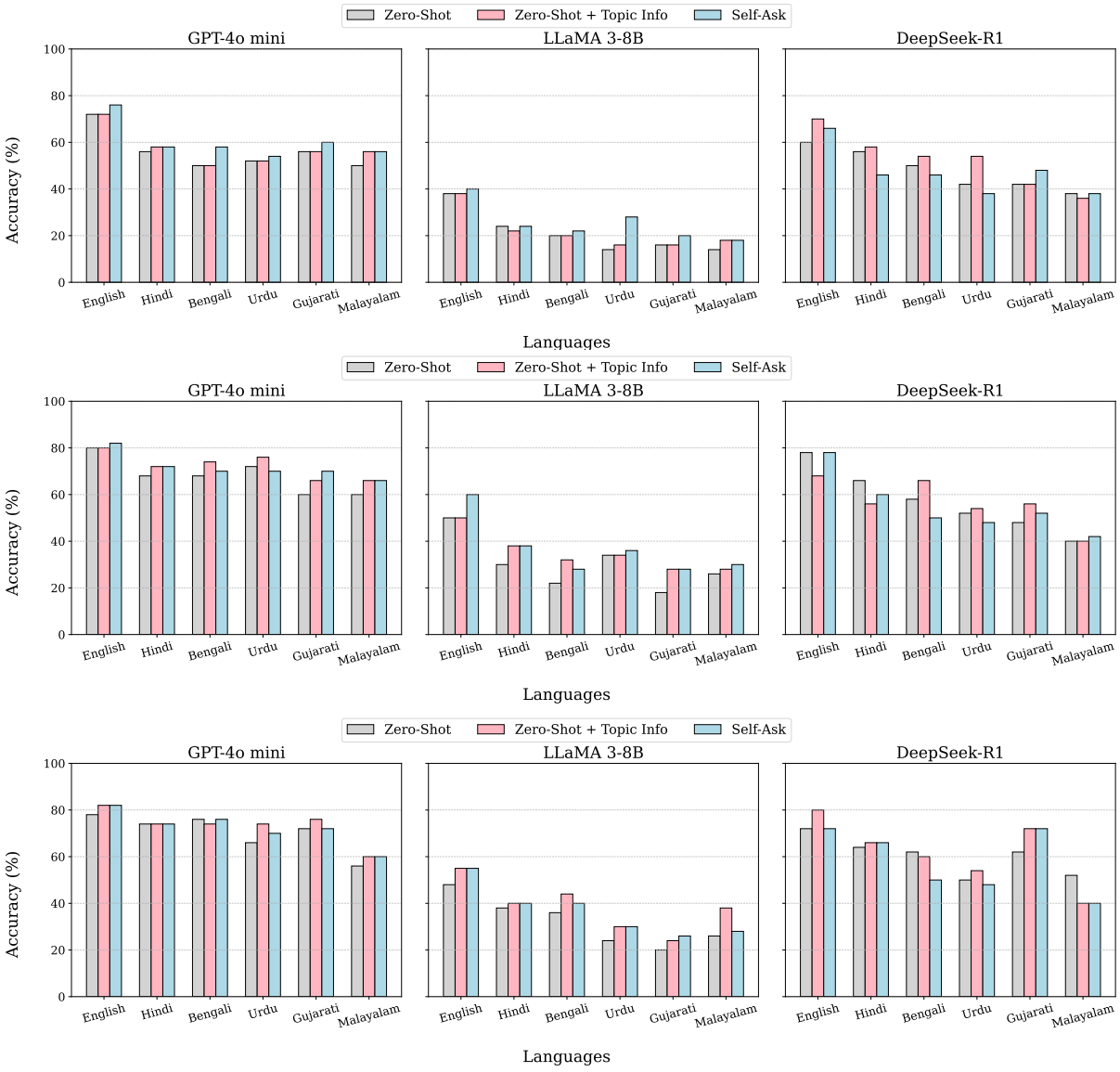


Figure 13: Any-of-3 accuracy (%) on Permutations and Combinations (top) and Straight Lines (middle) and Sequences and Series (bottom) problems across languages and prompting strategies.

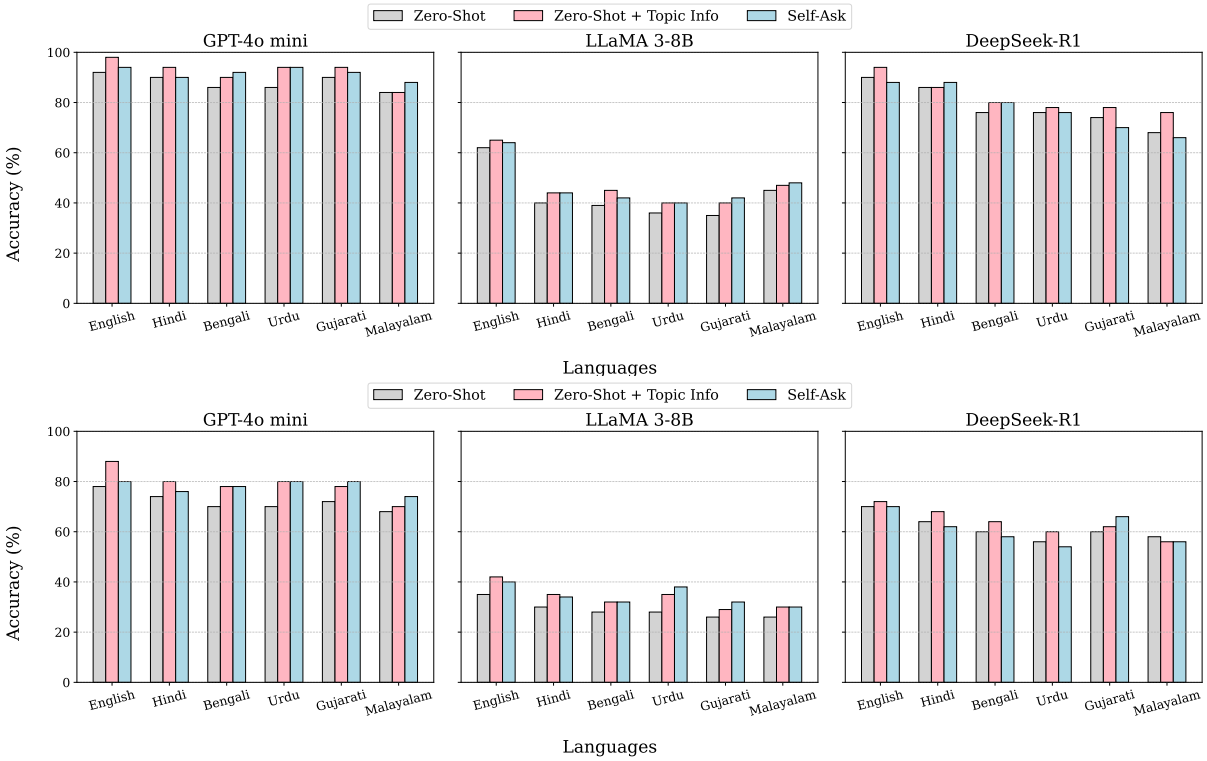


Figure 14: Any-of-3 accuracy (%) on Limits and Derivatives (top) and Conic Sections (bottom) problems across languages and prompting strategies.