

Simulating Student-Tutor Interactions with LLMs for Multilingual Math Feedback

Junior Cedric Tonga^{1*} KV Aditya Srivatsa^{2†} Kaushal Kumar Maurya^{3†}
Fajri Koto¹ Ekaterina Kochmar^{4†}

¹Mohamed bin Zayed University of Artificial Intelligence

²ETH Zürich ³IIT Delhi Abu Dhabi ⁴University of Victoria

Abstract

Large language models (LLMs) have demonstrated the ability to generate formative feedback and instructional hints in English, making them increasingly relevant for AI-assisted education. However, their ability to provide effective instructional support across different languages, especially for mathematically grounded reasoning tasks, remains largely unexamined. In this work, we present the *first* large-scale simulation of multilingual student-tutor interactions using LLMs. A stronger model plays the role of the tutor, generating feedback in the form of hints, while a weaker model simulates the student. We explore 352 experimental settings across 11 typologically diverse languages, four state-of-the-art LLMs, and multiple prompting strategies to assess whether language-specific feedback leads to measurable learning gains.¹ Our study examines how student input language, tutor feedback language, model choice, and language resource level jointly influence performance. Results show that English hints are generally preferred, even by multilingual student LLMs, with few exceptions that lead to improved student learning outcomes. These findings offer practical insights for developing multilingual, LLM-based educational tools that are both effective and inclusive.

1 Introduction

Large language models (LLMs) have demonstrated strong reasoning abilities in solving mathematical problems, particularly when prompted in English (Kojima et al., 2022; Guo et al., 2025; Bandyopadhyay et al., 2025). A common benchmark for evaluating such capabilities is GSM8K (Cobbe et al., 2021), which consists of grade-school-level math word problems. Its multilingual counterpart,

MGSM8K (Shi et al., 2022), extends this evaluation to a typologically diverse set of languages. Both English-centric and multilingual LLMs perform substantially worse on MGSM8K than on its English-only GSM8K benchmark (Shi et al., 2022; Ko et al., 2025), highlighting a gap in LLMs’ cross-linguistic reasoning abilities. This discrepancy raises questions about the use of LLMs as instructional agents beyond English. Recent work has explored their role as proxy tutors, generating formative feedback and pedagogical hints to support weaker student models or human learners (Wang et al., 2024b; Meyer et al., 2024). One widely studied form of support is hinting: a concise prompt aimed at guiding problem-solving without directly providing the answer. While such interventions have been shown to improve learning outcomes in English (Kochmar et al., 2022), their impact in multilingual settings remains largely unexamined. Moreover, these explorations are conducted in isolation, where either the tutor generates a hint or the student merely consumes it.

In this work, we take a step further by *simulating student-tutor interactions* entirely using LLMs: a stronger model generates hints as a tutor, while a weaker model attempts to solve the problem as a student. This simulation setup allows us to explore the effects of language, hint quality, and prompting strategy in a scalable and reproducible way. Moreover, such LLM-to-LLM simulations can serve as a valuable proxy for real-world educational scenarios, offering insights into how multilingual feedback might impact learning before deploying these systems with actual students. To the best of our knowledge, this is **the first work to simulate multilingual student-tutor interactions between LLMs across a broad range of languages and settings**. Given that effective feedback depends on both linguistic and reasoning proficiency, this raises a central question: *Does multilingual feedback from LLM tutors lead to measurable learning*

*Correspondence to: junior.tonga@mbzuai.ac.ae

†Work done while at MBZUAI.

¹Code on GitHub

gains in student models?

Our key contributions are as follows:

1. We simulate student-tutor interactions entirely using LLMs, modeling multilingual feedback across 11 languages, multiple prompting strategies, and four LLMs, yielding a large-scale experimental space of 352 settings.
2. We investigate how language-specific feedback influences student performance, examining the interplay between student input language, tutor feedback language, model selection, and whether the language is high- or low-resource, within the domain of mathematically grounded reasoning tasks.
3. We offer practical recommendations for designing LLM-based systems that support effective multilingual feedback and hint generation, highlighting considerations for both research and real-world educational deployment.

2 Related Work

LLMs in Math Reasoning LLMs have demonstrated strong performance in mathematical reasoning tasks, particularly in English, with benchmarks like GSM8K (Cobbe et al., 2021) driving much of this progress. Specialized models such as Minerva (Lewkowycz et al., 2022), trained in scientific texts, achieve state-of-the-art results without relying on external tools. However, these advances are still largely focused on English. Recent benchmarks like MGSM8K (Shi et al., 2022) reveal that multilingual LLMs underperform significantly, especially in low-resource languages, due to limited language coverage and weaker alignment between linguistic and mathematical representations.

Automated Hint Generation Before neural language models, automatic hint generation was typically formulated as a Markov Decision Process (MDP), selecting optimal hints based on student states (Stamper et al., 2008). Subsequent work improved scalability by structuring large hint sets into solution paths, and Paaßen et al. (2017) advanced this by modeling hint policies in continuous edit-distance spaces for better generalization. With LLMs, research has shifted toward direct hint generation: GPT-4 has been paired with a GPT-3.5 “student-validator” to filter poor hints (Phung

et al., 2024), and smaller models like LLaMA-3-8B were shown to approach GPT-4o performance when prompts target specific error types (Tonga et al., 2025). Studies show ChatGPT-generated hints can yield learning gains comparable to human hints in mathematics (Pardos and Bhandari, 2024), though quality varies by domain and task complexity, and McNichols et al. (2024) note LLMs still struggle to generalize beyond training data to unseen student errors.

However, much of prior research has focused exclusively on English. This narrow scope limits the applicability of LLM-based tutoring in multilingual learning environments, where students often benefit more from feedback in their native language. Recent work such as MathOctopus (Chen et al., 2024) shows that multilingual tuning can significantly improve math reasoning across languages. However, the generation of effective instructional hints in languages beyond English, especially for low-resource contexts, remains an open challenge.

Simulating student behavior with LLMs. Although persona simulation has long been studied, it has recently gained momentum in LLM-based multi-agent systems for modeling agents with distinct personalities and goal-driven behaviors (Tran et al., 2025). In education, early approaches like SimStudent (Matsuda et al., 2013) aimed for faithful student simulation. Modern LLMs allow for richer, controllable simulations: Benedetto et al. (2024) adjust skill levels when generating exam answers, MathVC creates virtual classrooms with LLM “students” exhibiting diverse misconceptions (Yue et al., 2024), and error-focused datasets like LEMMA help models learn to self-repair (Pan et al., 2025). Recent work, such as Macina et al. (2023), also explores multi-turn dialogues where LLMs act as students in conversations with human tutors. These developments provide ethically appropriate alternatives to real students. *Our work builds on these efforts by treating LLMs as simulated students, aligned with established guidelines and best practices.*

3 Methodology

This section describes our modeling framework, system architecture, and prompting strategies to simulate student-tutor interactions with LLMs. The overall design is illustrated in Figure 1 and consists of the following key components:

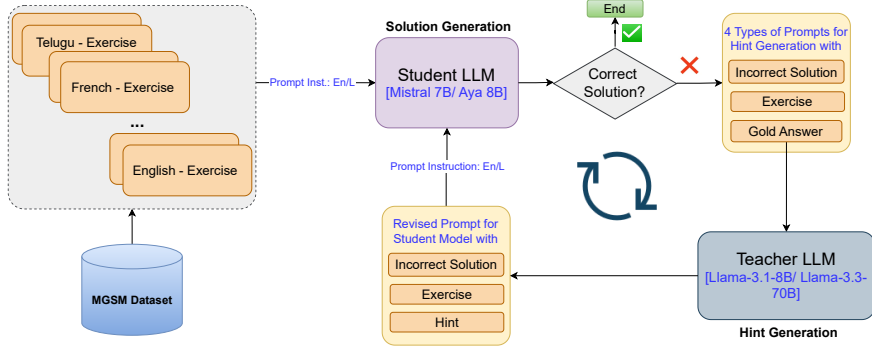


Figure 1: Overview of the student-tutor interaction flow.

Solution Generation The first stage of our pipeline involves generating a candidate student solution to a given multilingual exercise. Let $x \in \mathcal{X}$ represent an exercise in a given language L , and let LLM_S denote the student language model. The model is prompted with x to produce a candidate solution $\hat{y}_S = LLM_S(x)$. To mimic real-world learner behavior, we consider multilingual scenarios where the instructional prompt is given either in English or in the native language of x (i.e., L). This design choice allows us to evaluate the language sensitivity of LLM_S and its downstream performance. We experiment with two student models: the instruction-tuned Mistral-7B and the multilingual Aya-8B, chosen for their balance between model capacity and efficiency.

Hint Generation For exercises where the generated solution $\hat{y}_S$ is found to be incorrect (i.e., $\hat{y}_S \neq y^*$, where y^* is the reference or gold solution), we employ a tutor model LLM_T to generate pedagogically helpful hints. The motivation here is to simulate intelligent tutoring interventions that guide students toward the correct solution path without directly revealing the answer. The tutor model takes as input a triplet $\langle x, \hat{y}_S, y^* \rangle$ and produces a hint $h = LLM_T(x, \hat{y}_S, y^*)$ under one of four controlled prompting strategies (described in the next paragraph on prompting). GPT-4o (Hurst et al., 2024) was used to validate the correctness of both the initial and revised solutions by comparing them to the gold solution (see prompts in Appendix E.3). We employ large LLaMA-3.3-70B LLM as the main tutor but also experiment with the smaller LLaMA-3.1-8B LLM.

Prompting We consider two prompting setups for the LLM_S : (1) **Multilingual prompting**, where the student prompt is written in the same language L as the exercise. To operationalize this, we translate the base prompts (provided in Figure 4 and Figure 5) into the 11 languages of the MGSM

benchmark (detailed discussion in Section 4) using Google Translate API;² (2) **English-only prompting**, where the student prompt remains in English, regardless of the exercise’s language L . This serves as a control condition to isolate the impact of prompt language on downstream performance.

For hint generation, we explore four strategies by varying the **input prompt language** to the tutor model LLM_T and the **output hint language**:

1. **English-to-English (EN→EN)**: The tutor model (LLM_T) is prompted in English and instructed to generate a hint in English, regardless of the language of the original exercise. This setup uses the prompt format illustrated in Figure 6 of the Appendix.
2. **English-to-English with Translation (EN→EN→L)**: The hint generated in the EN→EN setup is machine-translated with Google Translate into the exercise’s native language L . This configuration controls for content while varying the delivery language, enabling analysis of whether presenting hints in the student’s native language improves comprehension and learning outcomes.
3. **Native-to-Native (L→L)**: The tutor model is prompted in the native language L of the exercise, using the translated version of the hint prompt provided in Appendix (see Figure 6), and is instructed to generate the hint in the same language L . This aims to explore the impact of native language interaction and hints on the tutor model.
4. **English-to-Native (EN→L)**: The tutor model is prompted in English, following the format shown in Figure 6 of the Appendix, but is explicitly instructed to generate the hint in

²Translation performed using Google Translate API from <https://github.com/nidhaloff/deep-translator>.

the target language L . This explores how instruction in English and hinting in the native language affects the tutoring outcomes.

Student-Tutor Interaction Flow The full pipeline of student-tutor interaction is summarized in Algorithm 1. It outlines the interaction between the student and tutor models, including the hint-guided revision loop. Due to computational constraints, we primarily experimented with a single hint iteration ($N = 1$). However, we also include a small-scale analysis for $N > 1$ (see Section ??), which shows that the key observations made with $N = 1$ largely hold across higher values of N as well.

Algorithm 1 Student-Tutor Interaction Flow

Require: Exercise x in language L , reference solution y^* , maximum hint attempts N

- 1: Choose student prompting mode (P_S): *Multilingual* or *English-only*
- 2: Generate solution $\hat{y}_S \leftarrow LLM_S(P_S(x))$
- 3: **if** $\hat{y}_S = y^*$ **then**
- 4: **return** Correct solution
- 5: **end if**

▷ *Hint-Guided Revision Loop*

- 6: **for** $i = 1$ to N **do**
- 7: Choose hint generation prompt strategy P_T
- 8: Generate hint $h \leftarrow LLM_T(P_T(x, \hat{y}_S, y^*))$
- 9: Provide hint h to LLM_S and generate revised solution $\hat{y}_S \leftarrow LLM_S(P_S(x, \hat{y}_S, h))$
- 10: **if** $\hat{y}_S = y^*$ **then**
- 11: **return** Correct solution
- 12: **end if**
- 13: **end for**
- 14: **return** Final student solution $\hat{y}_S$ after N attempts

4 Experimental Setup

Dataset and Languages We use the Multilingual Grade School Math (MGSM) dataset (Shi et al., 2022), a multilingual extension of GSM8K (Cobbe et al., 2021), consisting of 250 grade-school-level math problems translated from English into 11 typologically and geographically diverse languages (English (en), Bengali (bn), Chinese (zh), French (fr), German (de), Japanese (ja), Russian (ru), Spanish (es), Swahili (sw), Telugu (te), and Thai (th)). These examples are representative of the full GSM8K dataset (see Appendix B). Following prior work (Shi et al., 2022), we group the languages into High-Resource Languages (HRLs: en, zh, fr, de, ja, ru, es) and Low-Resource Languages (LRLs: bn, th, te, sw).

Models We use four LLMs: two tutor models, LLaMA-3.3-70B and LLaMA-3.1-8B, selected for their multilingual strength, and two student models,

Mistral-7B (monolingual) and Aya-8B (multilingual). These cover a range of model sizes and capabilities. Appendix Section C provides solution accuracy scores.

Evaluation Metric We use student gain G as the main evaluation metric, defined as the relative improvement in a student model’s accuracy after receiving hints. Let A_{before} and A_{after} denote accuracy before and after hinting (over N iterations). The Student Gain G is:

$$G = \frac{A_{\text{after}} - A_{\text{before}}}{A_{\text{before}}} \times 100 \quad \bar{G}_{\text{category}} = \frac{1}{L} \sum_{i=1}^L G_i$$

This percentage-based metric accounts for the model’s initial performance A_{before} and allows for meaningful comparisons (Törnqvist et al., 1985). We average the gain G across all languages within each category (i.e, HRLs and LRLs) to obtain $\bar{G}_{\text{category}}$, with L being the number of languages in the category, and G_i is the gain for language i .

Experimental Space Our setup spans 11 languages, 2 student models, 2 prompt types, 2 tutor models, and 4 hint strategies—yielding **352 unique configurations** (see Table 10 in Appendix). See Appendix Section D for implementation details.

5 Results

In this section, we discuss the impact of different student models and prompts, tutor models and prompts, and variations of student gains across HRLs and LRLs. We present key observations relevant to our core research question on multilingual feedback in mathematical reasoning tasks.

Does the size of the tutor model impact student gains? Based on Appendix Figure A, we observe that LLaMA-3.3-70B consistently yields greater gains across both student models and hint generation prompts, outperforming LLaMA-3.1-8B for HRLs and LRLs. Specifically, LLaMA-3.3-70B achieves higher median gains – for instance, 8.6% in the *multilingual* prompt setup and 10% in the *English-only* with Mistral – compared to 7.7% and 8% for LLaMA-3.1-8B. This effect is especially pronounced for LRLs, with median improvements reaching up to 31% (Aya-8B, *multilingual*) and 38% (Mistral-7B, *English-only* setup). While LLaMA-3.1-8B also shows strong improvements for LRLs, its performance exhibits higher variability than LLaMA-3.3-70B, as evidenced by the presence of outliers and larger interquartile ranges in

		En→En	En→En→L	En→L	L→L
MULTILINGUAL SETUP: Student input prompt in native language					
HRLs	Mistral-7B	11.30	10.90	9.50	11.00
	Aya-8B	4.00	3.60	3.60	3.90
LRLs	Mistral-7B	12.10	6.60	5.60	14.90
	Aya-8B	27.60	16.90	17.40	20.50
Avg. HRLs		7.65	7.25	6.55	7.45
Avg. LRLs		19.85	11.75	11.50	17.70
Avg. Overall		13.75	9.50	9.03	12.57
Δ HRLs	Mistral-7B–Aya-8B	7.30	7.30	5.90	7.10
Δ LRLs	Mistral-7B–Aya-8B	-15.50	-10.30	-11.80	-5.60
ENGLISH-ONLY SETUP: Student input prompt in English					
HRLs	Mistral-7B	8.70	11.00	12.10	10.60
	Aya-8B	5.90	5.50	4.70	4.70
LRLs	Mistral-7B	46.00	7.00	43.10	8.80
	Aya-8B	35.00	15.10	13.40	15.70
Avg. HRLs		7.30	8.25	8.40	7.65
Avg. LRLs		40.50	11.05	28.25	12.25
Avg. Overall		23.90	9.64	18.33	9.95
Δ HRLs	Mistral-7B–Aya-8B	2.80	5.50	7.40	5.90
Δ LRLs	Mistral-7B–Aya-8B	11.00	-8.10	29.70	-6.90

Table 1: Mean student gains (%) across two experimental setups using LLaMA-3.3-70B as the tutor model. The **best** average values and **second-best** values are highlighted. Δ rows indicate performance differences between Mistral-7B and Aya-8B within each language resource category (HRLs and LRLs).

the boxplots, indicating a less equitable distribution for LRLs. *These results suggest that larger tutor models are more effective at generating helpful hints, and that model size plays a key role in mitigating the challenges of low-resource settings.* This is expected as larger models are more capable overall. To simplify the interpretation of our findings, we focus the remainder of our discussion on the large tutor model only.

Does the multilingual student model prompting lead to higher student gains compared to English-only prompting? The Avg. labeled rows from Table 1 (distilled from Figure 3) show that *English-only* prompting consistently outperforms *multilingual* prompting across all hint types, except for the L→L setting. This indicates that delivering instructions in English is generally beneficial regardless of hint language. However, when instructions are provided in a native language, the corresponding hints should also be in the native language, especially for LRLs, where this effect is more pronounced. For HRLs, the average gain is comparable across all hint prompts for both setups. Finally, these observations are largely invariant to the type of student LLM, whether monolingual or multilingual. *Overall, multilingual instruction does not necessarily lead to higher student gains unless paired carefully with the hint language.*

Which student model leads to greater gains – multilingual or English-centric? The rows labeled with Δ in Table 1 provide clarity on this question. A negative Δ indicates better performance by the multilingual Aya-8B model, while a positive value favors the monolingual Mistral-7B. For LRLs, Aya-8B tends to be more effective, likely due to stronger language representation. In contrast, Mistral-7B generally performs better on HRLs. Interestingly, the difference in student gains is more pronounced on the average in the *multilingual* setting compared to the *English-only* setting, as indicated by higher absolute values of Δ scores. *Overall, as expected, multilingual LLMs tend to perform better for LRLs, while monolingual LLMs are better suited for HRLs.*

Which hint generation prompt strategy performs best with LLaMA-3.3-70B? Figure 3 and Table 1 present the comparative performance of various hint generation strategies under both *multilingual* and *English-only* settings. The EN→EN strategy yields the highest average improvement across both setups, outperforming other hint prompting strategies. This indicates that models are highly responsive to English prompts. This trend generally holds across both HRLs and LRLs, as well as across different student LLM types. An exception is observed in the *English-only* setting for HRLs when using Mistral-7B as the student LLM, where EN→EN is less effective, possibly due to limited representation of HRLs in Mistral-7B. *Overall, tailoring hint strategies to model- and language-specific characteristics is essential for maximizing student gains, though EN→EN remains the most preferred strategy.*

Does the language of the hint (English vs. the native) influence student gains? Using the larger tutor model, LLaMA-3.3-70B, we find that hint language substantially impacts student model gains as shown in Table 1. In the *multilingual* prompt setup, English hints (generated using the EN→EN tutor prompt) consistently yield the highest median improvements across both student models. Further, for LRLs, Aya-8B achieves its best and most consistent improvement of 31.12%, while Mistral-7B reaches its highest median improvement of 7.6%. In the *English-only* setup, the trend holds: Mistral-7B and Aya-8B obtain their best results for LRLs with English hints, reaching 38.19% and 16.24%, respectively. For HRLs, Aya-8B performs slightly better with translated hints using the

EN→EN→L prompt, whereas Mistral-7B benefits more from hints provided directly in the target language (L→L). In contrast, when using the smaller tutor model, LLaMA-3.1-8B, no consistent patterns emerge regarding hint language effectiveness. *Overall, English-language hints tend to be more effective when generated by larger tutor models. However, native-language hints can be competitive or even superior in specific cases.*

Student gains across languages. The Avg. row in Table 1 highlights that student models achieve higher gains on LRLs than HRLs in both *multilingual* and *English-only* setups. Multilingual LLMs perform better on LRLs, while monolingual LLMs are more effective for HRLs, consistent with earlier findings. Additionally, among the four hint prompting strategies, the EN→EN prompt yields the highest overall gains across all languages. Further, Table 12 (Appendix) shows that Mistral-7B struggles with Telugu and Bengali in the *multilingual* setup but achieves better gains in these languages in the *English-only* setup (EN→EN, EN→L). In contrast, Aya-8b demonstrates more consistent and higher gains across LRLs (Telugu, Swahili, Bengali, and Thai) in both setups (see Table 13, Appendix).

Final Takeaways Based on the results and discussion, we summarize our key findings regarding multilingual student–tutor interactions:

1. *Student Prompt:* English-only prompts generally perform well; however, when either the hint generation prompt or the hint is in the native language, multilingual prompting may be preferable.
2. *Student Model:* Monolingual models perform better for HRLs, while multilingual models are more effective for LRLs.
3. *Hint Generation Prompt:* EN→EN remains the most preferred strategy, with a few exceptions.
4. *Tutor Model:* Larger models such as LLaMA-3.3-70B are generally more effective and should be preferred.
5. *Hint Language:* English hints are generally preferred; however, for HRLs and monolingual student models, native-language hints can be more effective.

Models	En→En		En→En→L		En→L		L→L	
	L-3.1	L-3.3	L-3.1	L-3.3	L-3.1	L-3.3	L-3.1	L-3.3
HRLs								
Mistral-7B	0.51	0.80	0.51	0.80	0.51	0.74	1.43	0.63
Aya-8B	0.46	0.51	0.46	0.51	0.17	0.29	0.57	0.06
LRLs								
Mistral-7B	0.70	1.90	0.70	1.90	0.50	1.60	4.70	2.20
Aya-8B	1.10	1.80	1.10	1.80	1.20	1.30	8.30	1.60

Table 2: Answer leakage proportions (%) across LLaMa-3.1-8B (as L-3.1) and LLaMa-3.3-70B (as L-3.3).

6 Further Analyses

Gold Answer Leakage. Despite explicit instructions to the tutor model to avoid revealing the final answer while generating hints, gold answer leakage may still occur due to LLM hallucinations, potentially compromising our findings. To assess this risk, we conduct a *Gold Answer Leakage* test—i.e., *checking whether the final answer appears verbatim within the generated hint*. This is a challenging task, as it requires precise extraction of the gold answer from free-form hints; we adopt a regex-based approach for detection. During hint generation, we flagged any hints that included the gold answer as a *stand-alone number* (i.e., not embedded in a longer number or decimal). Detection used the regex,³ which matches the exact integer or decimal token when it is delimited by non-digit boundaries and not attached to a decimal point. Table 2 reports the proportion of such hints among the total samples for HRLs (250×7 = 1750) and LRLs (250×4 = 1000), across the four hint generation strategies and student models. The highest observed answer leakage was approximately 8% for LLaMA-3.3-8B and around 2% for LLaMA-3.3-70B. Since our primary tutor model is LLaMA-3.3-70B, this level of leakage is unlikely to significantly impact our findings, especially considering that the regex-based extractors tend to produce some false positives. Notably, the leakage rate is higher for low-resource languages, suggesting greater difficulty in handling those languages.

We further investigated whether the hints that helped students revise their initial answers tended to contain the gold answer. Focusing on the two best-performing strategies in the *multilingual* setup—EN→EN and L→L—we computed the leakage ratio, defined as the number of helpful hints that included the gold answer divided by the total number of helpful hints. These results are presented in Figure 9. The findings suggest that helpful hints rarely reveal the gold answer, with leakage ratios close to zero for most languages. No-

³Regex pattern = `r'(?!\d)'` + `r'\b'` + `re.escape(answer_str)` + `r'\b'` + `r'(!\d)'`

	En→En		En→En→L		En→L		L→L	
Models	Aya-8B	Mistral-7B	Aya-8B	Mistral-7B	Aya-8B	Mistral-7B	Aya-8B	Mistral-7B
HRLs	L-3.1	99.99	99.88	99.88	99.31	97.62	99.87	99.92
	L-3.3	100.00	99.77	99.94	99.94	99.71	99.94	99.94
LRLs	L-3.1	99.64	98.90	97.92	96.90	98.40	97.85	98.15
	L-3.3	100.00	99.80	98.95	99.17	99.40	99.37	99.47

Table 3: Mean language identification accuracy for hints with LLaMA-3.3-70B (as L.3.3) and LLaMA-3.1-8B (as L.3.1). The lowest number is bold.

		En→En	En→En→L	En→L	L→L
Initial solution					
HRLs	Mistral-7b	99.8	99.8	99.8	99.8
	Aya	100.0	100.0	100.0	100.0
LRLs	Mistral-7b	96.8	96.8	96.7	96.8
	Aya	98.6	98.6	98.6	98.6
Revised solution					
HRLs	Mistral-7b	93.1	93.7	94.9	94.3
	Aya	92.0	92.9	93.4	93.0
LRLs	Mistral-7b	90.6	92.1	92.3	91.8
	Aya	99.4	99.1	98.7	98.9

Table 4: Mean language identification accuracy (in %) of student models’ initial and revised solutions. Lowest value is bold for both initial and revised solutions.

table exceptions include Thai and Swahili LRLs, where over 2% of helpful hints contained the gold answer.

Language Consistency of Hints and Student Outputs. *Are the initial solution, generated hints, and revised solution in the intended language?* To verify this, we conducted sanity checks using a FastText-based language identification (LID) method (Bojanowski et al., 2017) with the pre-trained lid.176 model. Table 5 reports the mean LID accuracy of hints for HRLs and LRLs across hint generation strategies and tutor models. With a minimum LID accuracy of $\sim 97\%$, most hints are in the intended language with minor code-mixing. Table 6 presents LID accuracy for initial and revised student solutions. Initial solutions are mostly in the expected language, with a minimum LID accuracy of $\sim 97\%$. However, revised solutions show a drop in accuracy to $\sim 91\%$, indicating increased code-mixing or language switching. Manual inspection reveals that both tutor models struggle to preserve language consistency in revised solutions, particularly for LRLs—most notably Mistral-7B, likely due to its English-centric bias. This degradation may propagate from minor code-mixing in the hints themselves (as previously observed). Interestingly, the EN→L strategy yields the highest LID accuracy in revised outputs, suggesting it is more effective in maintaining language fidelity.

Translation Quality. To evaluate the quality of translated hints in the En→En→L setting—where English hints are translated into target languages using Google Translate—we perform back-

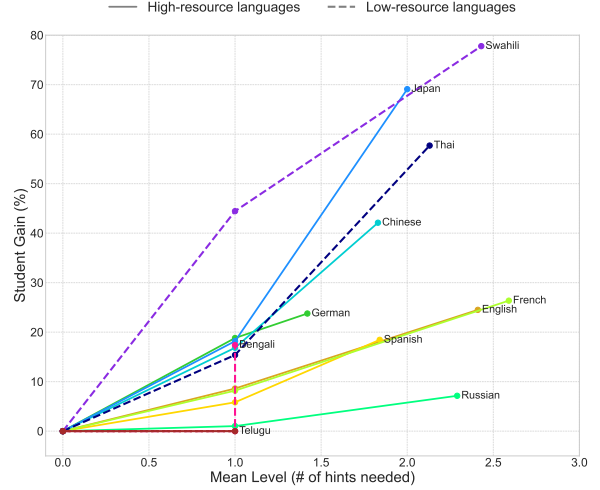


Figure 2: Student Gain scores as a function of the mean number of hints required per language.

	En→En		En→En→L		En→L		L→L	
Models	Aya-8B	Mistral-7B	Aya-8B	Mistral-7B	Aya-8B	Mistral-7B	Aya-8B	Mistral-7B
HRLs	L-3.1	99.99	99.88	99.88	99.31	97.62	99.87	99.92
	L-3.3	100.00	99.77	99.94	99.94	99.71	99.94	99.94
LRLs	L-3.1	99.64	98.90	97.92	96.90	98.40	97.85	98.15
	L-3.3	100.00	99.80	98.95	99.17	99.40	99.37	99.47

Table 5: Mean language identification accuracy for hints with LLaMA-3.3-70B (as L.3.3) and LLaMA-3.1-8B (as L.3.1). The lowest number is bold.

translation to English and compute BLEU scores (Papineni et al., 2002) over all samples for each language. As shown in Table 11, translation quality is generally high across languages, with the exception of Telugu, which shows only moderate quality. This further validates the findings with the En→En→L prompt.

Impact of Multiple Hints on Student Gains (N>1). To evaluate whether providing multiple hints enhances student model performance, we consider single setup with L→L strategy—reflecting realistic multilingual tutoring scenarios—using LLaMA-3.3-70B as the tutor and Mistral-7B as the student. We extend the interaction up to N=5 iterations (i.e., up to five hints), terminating early if the student model produces the correct answer. Figure 2 shows the relative improvement based on the number of hints required. For most languages, performance improves after the first hint, with Swahili showing the largest gain (over 70%). In contrast, Telugu shows no improvement, suggesting that Mistral struggles significantly with this language—likely due to its English-centric training. Interestingly, the average number of hints needed across most languages is around two, indicating that a second hint often contributes meaningfully to student performance. Among high-resource lan-

		En→En	En→En→L	En→L	L→L
Initial solution					
HRLs	Mistral-7b	99.8	99.8	99.8	99.8
	Aya	100.0	100.0	100.0	100.0
LRLs	Mistral-7b	96.8	96.8	96.7	96.8
	Aya	98.6	98.6	98.6	98.6
Revised solution					
HRLs	Mistral-7b	93.1	93.7	94.9	94.3
	Aya	92.0	92.9	93.4	93.0
LRLs	Mistral-7b	90.6	92.1	92.3	91.8
	Aya	99.4	99.1	98.7	98.9

Table 6: Mean language identification accuracy (in %) of student models’ initial and revised solutions. Lowest value is bold for both initial and revised solutions.

guages, German requires the lowest number of hints, with the student typically improving after just one.

7 Conclusion

In this work, we present the first large-scale study involving 352 unique experiments on multilingual student–tutor interactions powered by LLMs, aimed at understanding the effect of language-specific hints across 11 typologically diverse languages, 4 models, and multiple prompting strategies. Our findings reveal that English-centric feedback can enhance student performance, but the most effective configurations vary depending on both the language and the LLM used. Additionally, we observed that even a few iterations of feedback can significantly improve problem solvability. This study offers key insights for designing equitable educational technologies and lays the groundwork for future research in multilingual feedback generation and evaluation. In future work, we aim to extend this research to subject areas beyond mathematics.

Limitations

Although our work advances inclusive, multilingual LLM-based tutoring, several limitations remain and point to fruitful avenues for future research.

LLMs as Tutor–Student Simulators. As in prior studies (Macina et al., 2023; Tran et al., 2025; Wang et al., 2024a), we use LLMs to simulate both the tutor and the student, enabling large-scale, controlled experiments across languages and instructional settings. This strategy yields consistency and broad coverage, but it cannot capture the full diversity of real learners’ misconceptions, language proficiency, or problem-solving styles. Likewise, LLM tutors lack the nuanced pedagogical instincts of experienced tutors. Introducing human-in-the-

loop evaluations, interactions with real students, and richer student models will increase validity and speed progress toward truly adaptive AI tutors.

Evaluating Hint Quality Beyond Student Gains.

We treat a hint as “good” if it leads the student to the correct final answer – that is, if it results in student gains. While practical and easy to measure, correctness alone misses key pedagogical dimensions such as conceptual scaffolding, clarity, and alignment with learning objectives. Expert reviews and rubric-based assessments by mathematics educators could supply these missing perspectives and help refine what counts as a high-quality hint.

Coarse Gain Measurement. Step-level verification of math reasoning with current LLMs is far harder than judging overall solution correctness (Daheim et al., 2024). Consequently, we evaluate feedback with a binary metric – does the student’s solution become fully correct or not? This ignores cases where feedback fixes the current error but a later, independent errors may still be present. Developing reliable, fine-grained metrics for partial progress is an important direction for future work.

Lack of Phase-wise Evaluation. Our pipeline follows the two-phase paradigm of first verifying the student’s work and then generating a hint (Macina et al., 2023). Ideally, each phase should be evaluated separately. Yet automated assessment of both verification quality and hint usefulness is still unreliable; adding further sub-steps may boost overall performance but it compounds the evaluation challenge.

Limited Data and Language Coverage. Translating math-word problems is non-trivial: real-world contexts must be preserved, and many cultural references lack direct equivalents (Shi et al., 2022). We therefore rely on the manually curated MGSM dataset, which contains only 250 problems per language across 11 languages. While sufficient for our experiments, this scale limits analyses such as comparing gains across typologically related languages or training models on parallel corpora to induce language-agnostic pedagogy. Expanding high-quality, multilingual datasets, especially for low-resource languages, remains a pressing need.

Acknowledgments

We thank the supercomputing center at MBZUAI for providing resources for this research.

References

- Dibyanayan Bandyopadhyay, Soham Bhattacharjee, and Asif Ekbal. 2025. Thinking machines: A survey of llm based reasoning strategies. *arXiv preprint arXiv:2503.10814*.
- Luca Benedetto, Giovanni Aradelli, Antonia Donvito, Alberto Lucchetti, Andrea Cappelli, and Paula Buttery. 2024. Using LLMs to simulate students’ responses to exam questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. [Enriching word vectors with subword information](#). *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. [Breaking language barriers in multilingual mathematical reasoning: Insights and observations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.
- Karl Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *ArXiv*, abs/2110.14168.
- Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. [Stepwise verification and remediation of student reasoning errors with large language model tutors](#). *Preprint*, arXiv:2407.09136.
- 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*.
- OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mkadry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alexander Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alexandre Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braundstein, Andrew Cann, Andrew Codis-poti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, An drey Mishchenko, Angela Baek, Angela Jiang, An toine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, B. Ghorbani, Ben Leimberger, Ben Rossen, Benjamin Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Chris Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mély, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Phong Duc Nguyen, Duncan Findlay, Edele Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Hai-Biao Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Pondé de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian D. Kivlichan, Ian O’Connell, Ian Osband, Ian Silber, Ian Sohl, İbrahim Cihangir Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub W. Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Ryan Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quiñero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Joshua Gross, Josh Kaplan, Josh Snyder, Josh Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Hariman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Ouyang Long, Louis Feuvrier, Lu Zhang, Lukasz Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Madeline Simens, Madeleine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljube, Ma teusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Ali Yatbaz, Mengxue Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Castro,

- Mikhail Pavlov, Miles Brundage, Miles Wang, Mina Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Na talie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nikolas A. Tezak, Niko Felix, Nithanth Kudige, Nitish Shirish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Phil Tillet, Prafulla Dhariwal, Qim ing Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Raphael Gontijo Lopes, Raul Puri, Reah Miyara, Reimar H. Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Ramilevich Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal A. Patwardhan, Thomas Cunninghamman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiye Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. [Gpt-4o system card](#). *ArXiv*, abs/2410.21276.
- Feihu Jin, Yifan Liu, and Ying Tan. 2024. [Zero-shot chain-of-thought reasoning guided by evolutionary algorithms in large language models](#). *ArXiv*, abs/2402.05376.
- Hyunwoo Ko, Guijin Son, and Dasol Choi. 2025. Understand, solve and translate: Bridging the multilingual mathematical reasoning gap. *arXiv preprint arXiv:2501.02448*.
- Ekaterina Kochmar, Dung Do Vu, Robert Belfer, Varun Gupta, Iulian Vlad Serban, and Joelle Pineau. 2022. Automated data-driven generation of personalized pedagogical interventions in intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 32(2):323–349.
- 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.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.
- Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. *arXiv preprint arXiv:2305.14536*.
- Noburo Matsuda, Michael Yudelson, and Kenneth R. Koedinger. 2013. Simstudent: An agent that learns by tutoring. In *Proceedings of the 16th International Conference on Artificial Intelligence in Education*.
- Hunter McNichols, Jaewook Lee, Stephen Fancsali, Steve Ritter, and Andrew Lan. 2024. Can large language models replicate its feedback on open-ended math questions? *arXiv preprint arXiv:2405.06414*.
- Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students’ text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199.
- Benjamin Paaßen, Barbara Hammer, Thomas W. Price, Tiffany Barnes, Sebastian Gross, and Niels Pinkwart. 2017. The continuous hint factory: Providing hints in vast and sparsely populated edit-distance spaces. *Journal of Educational Data Mining*, 9(1):1–35.
- Zhuoshi Pan, Yu Li, Honglin Lin, Qizhi Pei, Zinan Tang, Wei Wu, Chenlin Ming, H. Vicky Zhao, Conghui He, and Lijun Wu. 2025. Lemma: Learning from errors for mathematical advancement in llms. *arXiv preprint arXiv:2503.17439*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Zachary A Pardos and Shreya Bhandari. 2024. Chatgpt-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *Plos one*, 19(5):e0304013.
- Huy Phung, Michael Khosravani, Tuan Nguyen, Zameer Ul Hassan, and Zhihao Wu. 2024. Leveraging gpt-4 tutor model for hint generation and gpt-3.5 student model for hint validation. In *Proceedings of the*

2024 Learning Analytics & Knowledge Conference, Kyoto, Japan.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4780–4797.

Kv Aditya Srivatsa and Ekaterina Kochmar. 2024. **What makes math word problems challenging for LLMs?** In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1138–1148, Mexico City, Mexico. Association for Computational Linguistics.

John Stamper, Tiffany Barnes, Marvin Croy, and Lisa F. Lehman. 2008. The hint factory: Automatic generation of contextualized help for existing computer-aided instruction. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, pages 373–382.

Junior Cedric Tonga, Benjamin Clement, and Pierre-Yves Oudeyer. 2025. **Automatic generation of question hints for mathematics problems using large language models in educational technology.** In *Proceedings of Large Foundation Models for Educational Assessment*, volume 264 of *Proceedings of Machine Learning Research*, pages 61–102. PMLR.

Leo Törnqvist, Pentti Vartia, and Yrjö O Vartia. 1985. How should relative changes be measured? *The American Statistician*, 39(1):43–46.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O’Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demiszky. 2024a. **Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes.** In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199, Mexico City, Mexico. Association for Computational Linguistics.

Rose E Wang, Ana T Ribeiro, Carly D Robinson, Susanna Loeb, and Dora Demiszky. 2024b. Tutor copilot: A human-ai approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*.

Murong Yue, Wenhan Lyu, Wijdan Mifdal, Jennifer Suh, Yixuan Zhang, and Ziyu Yao. 2024. Mathvc: An llm-simulated multi-character virtual classroom for mathematics education. In *Proceedings of the 11th ACM Conference on Learning at Scale*.

Feature	Mean Feature Value		T-Statistic	p-value
	GSM8K	MGSM		
Gx_mean_numerical_word_rank	28638.85	28661.13	-0.413	0.680
Gx_word_arg_count	0.64	0.68	-0.407	0.684
Gx_world_knowledge	1.07	1.06	0.140	0.889
Qx_constituency_tree_depth	10.97	10.96	0.044	0.965
Qx_flesch_kincaid_grade	4.19	4.18	0.071	0.943
Qx_flesch_reading_ease	88.93	89.10	-0.211	0.833
Qx_mean_numerical_word_rank	22739.56	22592.54	0.664	0.507
Qx_mean_word_rank	10664.46	10596.28	0.466	0.641
Qx_multi_np_count	0.44	0.37	0.884	0.377
Qx_np_count	18.52	18.62	-0.208	0.835
Qx_prp_count	1.82	1.90	-0.688	0.492
Qx_sentence_length	3.46	3.44	0.134	0.893
Qx_token_length	67.43	67.49	-0.034	0.973
Qx_unique_np_count	3.50	3.41	0.862	0.389
Qx_word_arg_count	1.11	1.17	-0.599	0.549
Qx_word_length	46.91	46.90	0.008	0.994

Table 7: Pairwise T-Test results between feature level mean values for GSM8K and MGSM.

A Detailed Results

Figure 3 provides a comprehensive summary of the results obtained from our extensive experimental setup.

B Dataset Representation

The MGSM (Shi et al., 2022) dataset is built on top of the first 250 math problems from the GSM8K dataset (Cobbe et al., 2021). In order to make sure that this subset is representative of the larger set from GSM8K, we perform a feature level comparison between the two sets. For this, we borrow the feature set from Srivatsa and Kochmar (2024) spanning the phrasing of the math problem in English, the count and nature of math operations and arguments involved in the gold solution, and the count of variables which require world knowledge. After generating the feature values for all English questions from MGSM and GSM8K, we compare their mean values – see Table 7. The low t -statistics and high p -values for corresponding pairwise t -tests indicate that there is not a significant difference between the two sets along any of the features.

C Comparative Analysis of Zero-Shot Performance Across Models

To select our student and tutor models, we initially evaluated five models – Mistral-7B, Aya-8b, LLaMA-3.1-8B, LLaMA-3.3-70B, and Deepseek-R1-LLaMA-distill-70B⁴ – using zero-shot prompting across both setups. We opted for standard zero-shot prompting over zero-shot Chain-of-Thought (CoT) prompting (Jin et al.,

⁴Via Together.ai: <https://www.together.ai/models/deepseek-r1-distilled-llama-70b-free>

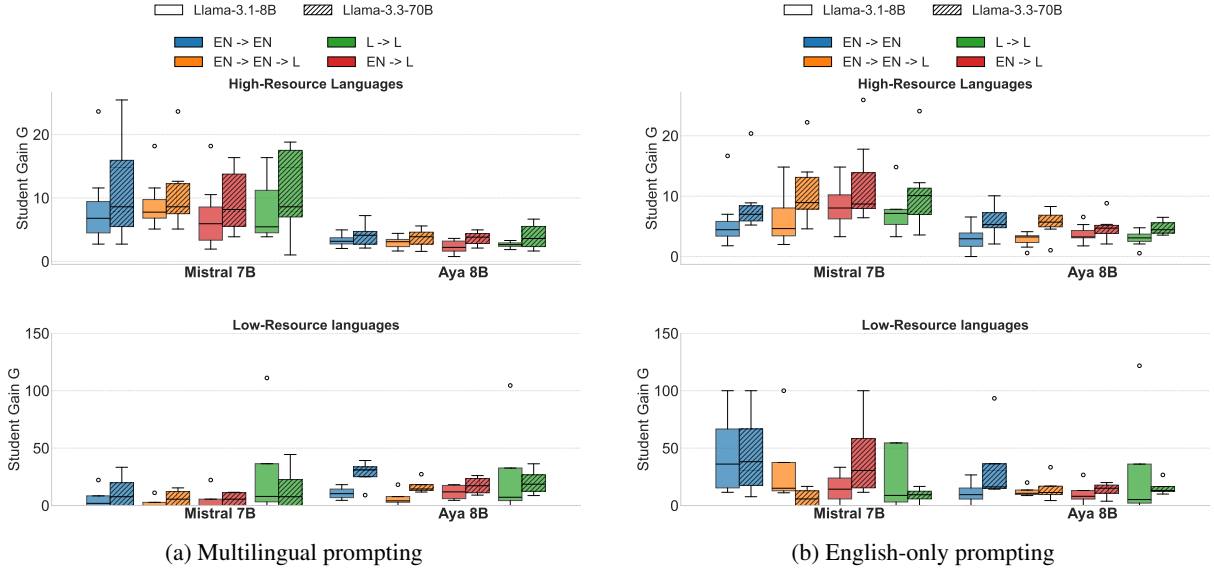


Figure 3: Relative improvement (Student gain G) in multilingual student-tutor interaction.

2024), as CoT led to correct answers, whereas we sought student models that produce a balanced mix of correct and incorrect responses. Based on these criteria, we selected Aya-8b and Mistral-7B as student models, as shown in Table 8, which reports their balanced zero-shot accuracy—the baseline for subsequent comparisons. Table 8 also shows that LLaMA-3.3-70B achieves the highest accuracy across both setups and outperforms Deepseek-R1-LLaMA-distill-70B, particularly with better output structure in low-resource languages. We therefore selected LLaMA-3.3-70B and LLaMA-3.1-8B as tutor models.

D Implementation Details

We set the temperature to 0 for the student models to ensure deterministic outputs, while the tutor models use a temperature of 1 to encourage diverse hint generation. GPT-4o (Hurst et al., 2024) was used with a temperature of 0 to evaluate the correctness of both the initial and revised answers by comparing them to the gold answer, as prior work has shown that temperatures above 0.2 can lead to unreliable results (Tonga et al., 2025). For the tutor models, we employed LLaMA-3.3-70B and LLaMA-3.1-8B. For the student models, we selected a monolingual model, Mistral-7B, and a multilingual model, Aya-8b. All models and their corresponding reproduction links are presented in Table 9.

E Prompts

In this section, we present the prompts used in our experiments.

E.1 Student prompts

Figure 4 shows the base prompt for generating a candidate solution, while Figure 5 displays the prompt for revising an initial answer via the student model. These prompts are used as-is in the English-only prompting setup. For the multilingual prompting setup, they are translated into the 10 target languages of the MGSM dataset.

System role
You are a high school student who must solve math exercises.
User role
Your goal is to answer the question asked in the exercise:
Exercise and question: {exercise}
Ensure that the response is in the specified language: {lang}
Required response format: use a JSON format with the following structure:
{{"raisonnement": "Explain your reasoning and provide your final answer to the exercise here..."}}
Respect the output format.

Figure 4: Prompt for generating a candidate solution.

E.2 Hint generation prompt

Figure 6 shows the base prompt used to generate a hint via the tutor model. For strategies where the tutor is prompted in English, the prompt is used as-is, with the variable *hint_lang* in the prompt replaced by either “English” (EN) or the exercise language (L), depending on the desired hint language. For strategy, where the tutor is prompted in the language of the exercise L, the prompt is translated into that language, and *hint_lang* is again set

English only prompting: when student input prompt is in English language					
Languages	Llama 3.1-8B	Aya-8b	Mistral-7B	Llama-3.3-70B	Deepseek-R1-Llama-distill-70B
English	85.2	77.2	60.8	93.2	85.2
Spanish	76.8	72.0	46.0	86.0	80.8
French	74.8	68.0	44.8	86.4	78.4
German	74.8	68.0	40.0	86.8	82.4
Russian	72.4	76.8	43.6	90.8	80.8
Chinese	72.4	67.2	36.0	88.0	84.0
Japanese	54.8	61.6	21.6	84.8	83.2
Thai	64.8	21.2	20.8	87.6	88.4
Swahili	60.8	9.2	3.6	84.8	79.2
Bengali	60.0	28.0	9.6	85.6	79.2
Telugu	53.2	6.0	0.4	81.6	58.0
AVG	68.2%	50.5%	29.7%	86.9%	80.0%
Multilingual prompting: when student input prompt is in native language					
Languages	Llama 3.1-8B	Aya-8b	Mistral-7B	Llama-3.3-70B	Deepseek-R1-Llama-distill-70B
English	85.2	77.6	60.4	93.6	82.8
Spanish	76.8	78.4	41.2	89.6	82.0
French	74.8	73.2	44.0	80.0	77.6
German	74.8	73.6	40.4	88.8	80.0
Russian	72.4	76.0	39.2	93.2	82.8
Chinese	72.4	72.0	38.0	88.0	86.8
Japanese	54.8	64.4	22.0	84.8	83.2
Thai	64.8	18.4	10.4	86.8	83.2
Swahili	60.8	8.8	3.6	87.6	82.0
Bengali	60.0	20.4	9.2	85.6	81.6
Telugu	53.2	4.4	0.4	84.8	75.2
AVG	68.2	51.5	28.1	87.5	81.6

Table 8: Zero-shot accuracy (%) of language models across languages. These zero-shot scores serve as the baseline for each model. The table shows results for both English-only prompting (top) and multilingual prompting (bottom) setups. Llama-3.3-70B consistently achieves the highest accuracy across most languages, demonstrating superior cross-lingual capabilities in zero-shot settings.

Model	Reproduction Link
Llama-3.3-70B	https://www.together.ai/models/llama-3-3-70b-free
Mistral-7B	https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
Llama-3.1-8B	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
Aya-8b	https://huggingface.co/CohereLabs/aya-expansive-8b
GPT-4o	gpt-4o-2024-08-06

Table 9: Models used in our experiments with their corresponding links for reproducibility.

based on whether the hint should be in English or the exercise language.

E.3 Prompts for Evaluating Student Outputs

Figure 7 shows the prompt used to evaluate the initial candidate solution against the gold solution,

System role
You are a high school student who must solve math exercises.
User role
Your previous answer to the math exercise below was incorrect. A teacher has provided you with a hint to help you understand your mistake and correct it. Your goal is to think about how the hint applies to your initial answer. Answer the question posed in the hint; then, use this answer to guide your thinking, correct your mistake in your initial response, and find the correct solution to the exercise.
Exercise and question: (exercise). Teacher's hint: (hint)
Do not include the answer to the question posed in the hint in the output. Just provide your revised answer to the question asked in the exercise.
Ensure that the response is in the specified language: (lang) Required response format: use a JSON format with the following structure: {{"response": "Provide your revised answer here..."}}
Make sure the generated output does not contain escape characters such as line breaks (\n) or slashes (/).
Provide a clean and readable output. I insist on this. Do not make formatting errors.
Respect the output format.

Figure 5: Prompt for revising the initial candidate solution

while Figure 8 presents the prompt for evaluating the revised solution via GPT-4o, after a hint is provided. We adopt the evaluation prompts from Tonga et al. (2025), but omit their error type categorization for the initial candidate solution, as it is beyond the scope of our work. However, we retain their approach of classifying hints as good (if the

System role
You are an expert in teaching mathematics.
User role
Your goal is to provide {num_indices} clear and relevant hints to help the student correct their errors and improve their answers in math exercises. This hint should be in the form of a question. Moreover, this hint should not include the correct answer to the exercise or fragments of the correct answer. Exercise and question: {exercise} The correct answer to the exercise: {answer} The student's reasoning, including their final answer to the exercise: {gpt_reasoning} Language of the {num_indices} hints: {hint_lang} Consider the following aspects to generate the hint: - Reasoning - Method - Concept application - Calculations - Problem interpretation JSON output format: {{"indice": "{indice1", "indice2": ..., "indice{num_indices}"}}}. Do not number the hints. Please generate exactly {num_indices} hint(s), no more. Follow this condition strictly. The {num_indices} hints should be in the specified language: {hint_lang} Ensure that the generated output does not contain escape characters such as new lines (\n) or slashes (/). Provide a clean and readable output. I insist on this. Do not make formatting errors. Follow the output format strictly. I emphasize that the hint should not be numbered and must be in the form of a question.

Figure 6: Prompt used by the tutor model to generate hints.

revision is correct) or bad (if incorrect) to track hint effectiveness.

System role
You are an expert in math teaching.
User role
Your task is to verify whether a student's answer to a math exercise is correct or not by comparing it with the provided correct answer. Exercise and question: {exercise} The correct answer to the exercise: {answer} The student's reasoning containing their final answer to the exercise: {reasoning} Extract the student's final answer from their reasoning, based on the question asked in the exercise. In order to compare it with the correct answer provided for the exercise, categorize the student's error. Here are some error categories and examples. You can add other error categories. If the reasoning contains multiple errors, it is important to list all the errors present. Specify each error distinctly, even if they belong to different categories or are combined. 1) Comprehension error: The student does not clearly understand the problem or the given instructions. 2) Partial answer: The student provides part of the expected answer but fails to complete it correctly. 3) Term grouping error: The student incorrectly combines or groups terms in a mathematical expression. 4) Simplification error: The student incorrectly simplifies a mathematical expression. 5) Calculation error: The student performs mathematical operations incorrectly. 6) Incorrect substitution error: The student substitutes an incorrect value in an expression or equation. 7) Interpretation error: The student misinterprets the instructions or data of a problem. 8) Algebraic error: The student makes a mistake in algebraic manipulations, such as distribution, factoring, or equation solving. Example: When simplifying the expression $3x + 2x + 5$, the student combines the terms $3x$ and $2x$ to obtain $5x+2$ instead of $5x$. Example: When simplifying $6x/2$, the student divides the numerator and denominator by x instead of 2, resulting in an incorrect simplification of $6/2x$. Example: When multiplying 7 by 8, the student gets 54 instead of 56. Example: In the equation $2x + 3y = 10$, the student substitutes $x = 4$ instead of $y = 2$, leading to an incorrect solution. Example: In a probability problem, the student confuses the probability of event A with that of the complementary event of A. Example: When solving $2(x + 3) = 10$, the student incorrectly divides 10 by $x + 3$ instead of 2, leading to an incorrect answer. 1) If the student's final answer does not match the correct answer, categorize the type of error by placing it in "type_derreur" field and do not put anything in the "bonne_reponse" field. 2) If the student's final answer matches the correct answer, place the student's answer in the "bonne_reponse" field. Provide the output in a JSON format with the following structure: {{"type_derreur": "", "bonne_reponse": ""}} Respect the output format, the evaluation criteria, and your role. Do not add anything else.

Figure 7: Prompt employed by GPT-4o to assess candidate answer correctness.

System role
You are an expert in math teaching.
User role
Your task is to verify whether a student's revised answer to a math exercise is correct or not by comparing it with the provided correct answer. The correct answer to the exercise: {answer} The student's revised answer: {revised_response} The hint: {hint} 1) If the student's revised answer does not match the correct answer, place the hint in the "mauvais_hint" field in the output. 2) If the student's revised answer matches the correct answer, place the hint in the "bon_hint" field in the output. 3) If the given hint contains the correct answer to the exercise, place the hint in the "mauvais_hint" field in the output. Provide the output in a JSON format with the following structure: {{"bon_hint": "", "mauvais_hint": ""}} Make sure the generated output does not contain escape characters such as line breaks (\n) or slashes (/). Provide a clean and readable output. I insist on this. Do not make formatting errors. Respect the output format, the evaluation criteria, and your role. Do not add anything else.

Figure 8: Prompt employed by GPT-4o to assess the correctness of the revised solution and categorize hints.

Axis	Values	Count
Languages	en, bn, de, es, fr, ja, ru, sw, te, th, zh	11
Student Prompts	English-only, Multilingual	2
Student Models	Mistral-7B, Aya-8B	2
Hint Strategies	EN→EN, EN→EN→L, L→L, EN→L	4
tutor Models	LLaMA-3.1-8B, LLaMA-3.3-70B	2
Total Configs	$11 \times 2 \times 2 \times 4 \times 2$	352

Table 10: Overview of the experimental space.

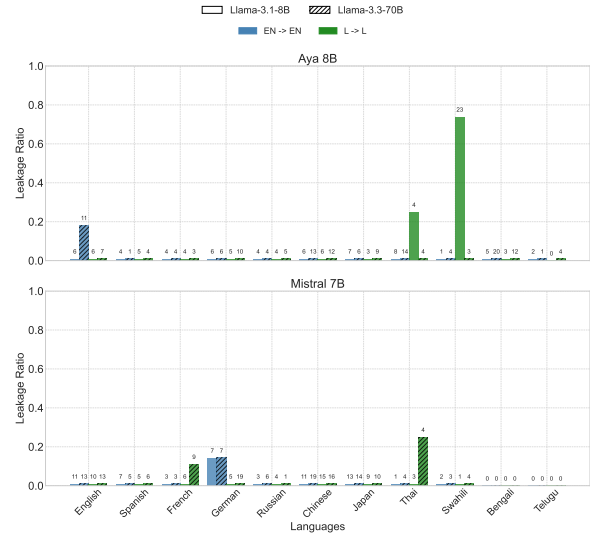


Figure 9: Leakage ratio per language; numbers above bars indicate total helpful hints.

F Analysis: Student gains across languages for the two different student models with LLaMA-3.3-70B

Table 13 and Table 12 show the student gains of Aya-8b and Mistral-7B, respectively, after giving a hint, when we used LLaMA-3.3-70B as the tutor model.

Language	BLEU	Language	BLEU
Spanish	62.7	Chinese	40.1
French	54.7	Japanese	45.6
German	46.7	Thai	40.0
Russian	49.3	Swahili	45.5
Bengali	43.0	Telugu	39.5

Table 11: BLEU scores by language

		English-only prompting						Multilingual prompting			
		EN→EN	EN→EN→L	EN→L	L→L			EN→EN	EN→EN→L	EN→L	L→L
HRLs	English	7.9	7.9	7.9	5.9	English		8.6	8.6	6.0	8.6
	Spanish	5.2	7.8	8.7	10.4	Spanish		4.9	7.8	3.9	5.8
	French	5.4	8.9	8.0	3.6	French		2.7	7.3	8.2	8.2
	German	7.0	14.0	10.0	8.0	German		11.9	11.9	13.9	18.8
	Russian	6.4	4.6	6.4	10.1	Russian		6.1	5.1	5.1	1.0
	Chinese	8.9	12.2	17.8	12.2	Chinese		20.0	12.6	13.7	16.8
	Japanese	20.4	22.2	25.9	24.1	Japanese		25.5	23.6	16.4	18.2
LRLs	Thai	7.7	11.5	11.5	7.7	Thai		15.4	15.4	11.5	15.4
	Swahili	55.6	0.0	44.4	11.1	Swahili		33.3	11.1	11.1	44.4
	Bengali	20.8	16.7	16.7	16.7	Bengali		0.0	0.0	0.0	0.0
	Telugu	100.0	0.0	100.0	0.0	Telugu		0.0	0.0	0.0	0.0

Table 12: Student gains G (%) of Mistral-7B after receiving a hint from LLaMA-3.3-70B across English-only and multilingual prompting setups.

		English-only prompting						Multilingual prompting			
		EN→EN	EN→EN→L	EN→L	L→L			EN→EN	EN→EN→L	EN→L	L→L
HRLs	English	5.7	5.7	3.1	4.1	HRLs	English	4.1	4.1	3.1	3.6
	Spanish	5.0	7.2	5.0	4.4		Spanish	5.1	5.1	2.6	2.0
	French	8.8	6.5	8.8	5.9		French	2.2	2.7	3.8	1.6
	German	5.3	5.3	4.7	5.3		German	3.3	2.7	4.9	5.4
	Russian	2.1	1.0	2.1	3.6		Russian	2.1	1.6	2.1	2.6
	Chinese	10.1	8.3	5.3	3.6		Chinese	7.2	3.9	3.9	6.7
	Japanese	4.5	4.5	4.5	6.5		Japanese	4.3	5.6	5.0	5.6
LRLs	Thai	15.1	11.3	3.8	13.2	LRLs	Thai	30.4	15.2	26.1	8.7
	Swahili	17.4	4.3	13.0	13.0		Swahili	31.8	13.6	22.7	13.6
	Bengali	14.3	11.4	17.1	10.0		Bengali	39.2	11.8	11.8	23.5
	Telugu	93.3	33.3	20.0	26.7		Telugu	9.1	27.3	9.1	36.4

(a) English-only prompting

(b) Multilingual prompting

Table 13: Student gain G (%) of Aya-8b after receiving a hint from LLaMA-3.3-70B in the two setups.