Are Large Language Models for Education Reliable for All Languages?

Vansh Gupta[*†] **Sankalan Pal Chowdhury**[*†]
Vilém Zouhar[†] **Donya Rooein**[‡] **Mrinmaya Sachan**[†]

{guptav,spalchowd,vzouhar,msachan}@ethz.ch donya.rooein@unibocconi.it
[†]ETH Zurich [‡]Bocconi University

Abstract

Large language models (LLMs) are increasingly being adopted in educational settings. These applications expand beyond English, though current LLMs remain primarily English-centric. In this work, we ascertain if their use in education settings in non-English languages is warranted. We evaluated the performance of popular LLMs on four educational tasks: identifying student misconceptions, providing targeted feedback, interactive tutoring, and grading translations in eight languages (Mandarin, Hindi, Arabic, German, Farsi, Telugu, Ukrainian, Czech) in addition to English. We find that the performance on these tasks somewhat corresponds to the amount of language represented in training data, with lower-resource languages having poorer task performance. However, at least some models are able to more or less maintain their levels of performance across all languages. Thus, we recommend that practitioners first verify that the LLM works well in the target language for their educational task before deployment.

1 Introduction

Education is a multilingual, multicultural endeavour. AI-based technologies have recently shown the potential to improve students' learning experiences, and educational systems worldwide are increasingly adopting these tools (Gligorea et al., 2023). From personalized instruction and targeted feedback to appropriate content generation and interactive tutoring, these tools offer solutions to key educational challenges (Leon, 2024; Rooein et al., 2024; Mosher et al., 2024). Large language models such as GPT, Gemini, and Llama (OpenAI, 2023; Team, 2024; Roumeliotis et al., 2023) have become particularly influential, with early evidence suggesting their ability to support teachers or scaffold student learning (Kasneci et al., 2023; Alqahtani et al., 2023).

Although most of these LLMs are trained on multilingual corpora (OpenAI, 2019; Nvidia, 2022; Peng et al., 2023; Gu and Dao, 2023), they are still overwhelmingly English-centric (Argoub, 2022; Ruder et al., 2022, Table 1). Inadequate adaptation to local languages in an educational setting risks diminishing their utility and exacerbating existing inequalities by privileging dominant languages and cultures. The question of multilingualism arises in every domain where LLMs are applied (Lai et al., 2023; Ahuja et al., 2023, 2024). However, it is especially important in the field of education, which has seen wide use of LLMs despite the high stakes (Alhafni et al., 2024; Raheja et al., 2023; Naismith et al., 2023). Without rigorous evaluation tailored to educational tasks across languages, deploying LLMs in classrooms may introduce new forms of harm, including misinformation, misalignment with curricula, or culturally inappropriate content (Almasoud et al., 2025).

In this work, we present an empirical investigation of the capabilities of frontier LLMs on educational tasks across several languages. We identify four education-related tasks (identifying student misconceptions, providing targeted feedback, interactive tutoring, and translation grading) with well-defined language-agnostic metrics. We then evaluate several frontier LLMs (Claude, Gemini, GPT4o, Llama, and Mistral) on these tasks in eight languages (Mandarin, Hindi, Arabic, German, Farsi, Telugu, Ukrainian, and Czech) in addition to English.

Our results show that though performance in English still dominates, other languages are not too far behind, at least for GPT4o and Gemini-2.0-flash, which emerge as the best models. We also find that using prompts in the language of the task

[*]Equal Contribution

[0]We release the collected dataset and code at github.com/eth-lre/multilingual-educational-llm-bias. The dataset comprises 313,500 automatically evaluated model outputs across seven languages, four tasks, and six models.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 612–631
July 31 - August 1, 2025

is rarely helpful compared to English prompts.

2 Methods

We select our set of tasks based on 3 desiderata:

- **Relevant to Education:** We focus on tasks that LLMs would encounter specifically in the role of tutors, teachers, or teaching assistants. We do not cover tasks like question answering or solving math questions, which, while possibly being relevant to education, are more general tasks that are primarily studied in other contexts.
- **Have a Language Component:** We avoid tasks whose formulation uses purely notation, for example, solving a math equation. If the equation is provided in mathematical notation, the task would remain unchanged between different languages, making the question of multilingual performance moot.
- **Language Invariant Evaluation:** Finally, we need evaluation metrics that remain comparable across languages to compare performance across different languages efficiently. This means we cannot rely on language-dependent metrics like BLEU or COMET (Papineni et al., 2002; Rei et al., 2020).

Based on these, we selected following four tasks:

Task 1: Misconception identification. An important aspect of teaching is fixing student misconceptions, which first requires identifying the student misconception (Liu et al., 2023). We build this task on the EEDI Math Questions Dataset, which contains thousands of multiple-choice questions with four answer choices. For many of the wrong choices, we have expert-annotated misconceptions that could lead to a student picking the said choice. We leverage these to build our task. The LLM is given a multiple-choice question, the student's (incorrect) answer, and four possible misconceptions. The candidate misconceptions include the true misconception identified by experts and three distractors chosen at random from the other misconceptions present in the dataset. The LLM must pick the correct misconception from these four options (see Example 1 for an example). We evaluate the LLM performance by reporting accuracy in predicting the student misconception. Since the model must pick one of four options, a random baseline has an accuracy of 25%.

Task 2: Feedback selection. A key step towards fixing students' misconceptions is generating feedback to alleviate them. The EEDI dataset discussed above also includes feedback for all the choices we use for this part. The LLM is again given a multiple-choice question, the student's answer, and this time, a set of four possible feedbacks, out of which the LLM must select the feedback corresponding to the student's answer. Note that while there are 4 possible feedbacks, one corresponds to the correct answer. This one is easily identifiable as it reinforces the student's answer, while the feedbacks corresponding to wrong answers all try to make the student realize their mistake. As an example, see Option C in both parts of 2, which are the only options in their respective questions that do not start with a negative tone. Therefore, if the selected answer is also the correct answer to the problem, the LLM might be able to pick the correct feedback using some shallow semantics, which we want to avoid. Therefore, we ensure that the selected answer is always incorrect. The random baseline has an accuracy of 25%, or 33% if choosing among responses to the wrong answer.

Task 3: Tutoring. For more complex misconceptions, a single-turn feedback often does not suffice, and fixing the misconception requires a multi-turn conversation between the student and the teacher, also known as tutoring. (Bloom, 1984; Cohen et al., 1982) This involves a teacher LLM trying to help the student identify and fix an error in their solution. We evaluate the tutoring ability of the LLM by having it tutor a weaker LLM, which acts as the student. Both the teacher and the student are given the question, but only the teacher LLM can access the correct answer. The student LLM is instructed to stick to the wrong solution unless it sees strong justification to shift. The teacher and the student take turns to send messages, with the teacher's goal being to get the student model to the correct answer, without revealing the answer themselves. The teacher LLM is considered to get a *success* if the student LLM states the answer. If the teacher reveals the answer before the student has gotten to it, it is counted as *telling*. An *adjusted success* occurs when there is a success but no telling. The task is finally evaluated by Tutoring score (Pal Chowdhury et al., 2024), which is the harmonic mean between success rate and adjusted success rate.

This task differs from the other tasks on this list

	Language family	Script	Wikipedia	CommonCrawl	Speakers
English	Germanic	Latin	6973K	42.8%	1500M
Mandarin	Sino-tibetan	Hanzi[1]	1480K	5.8%	1184M
Hindi	Indo-Iranian	Brahmic	165K	0.20%	609M
Arabic	Afro-Asiatic	Abjad	1259K	0.68%	411M
German	Germanic	Latin	3021K	5.5%	411M
Farsi	Indo-Iranian	Abjad	1034K	0.74%	134M
Telugu	Dravidian	Brahmic	111K	0.02%	96M
Ukrainian	Slavic	Cyrillic	1371K	0.62%	39M
Czech	Slavic	Latin	566M	0.10%	12M

Table 1: Language information, number of speakers (Ethnologue 2025), and global representations of tested languages in NLP (Wikipedia Articles and proportion in CommonCrawl in March 2025).

in at least two significant ways. First, it is a multi-turn conversation task, so there is no scope for guessing the answer. Secondly, the final evaluation depends on the performance of the student LLM, so the multilingual capabilities of the student LLM also restrict the applicability of this task. These factors make this task both slower to run and more complex for the LLMs.

Task 4: Translation grading. A common field of education that has seen an increase in the use of LLMs is Language learning (Klimova et al., 2024; Zhu et al., 2024). A representative task from this field is to assign a grade to a translation provided by a student. While we lack proper datasets across languages with translations and their appropriate grades, we can approximate this task by the fact that *the machine translation of a sentence should receive a higher grade than the exact translation with one word replaced by a random word.* We use English sentences from Duolingo's English→Spanish SLAM dataset (Settles, 2018), which are machine translated to other languages. We chose this dataset because it is meant to be used for translation, so it should contain fewer hard-to-translate sentences. We filter out simple sentences that do not end with a full stop or have fewer than five words. For each translated sentence, we then create a corresponding *perturbed translation* by replacing one of the words in the sentence with a different word selected at random from the other sentences in the dataset, disrupting both the fluency and adequacy of the translation. The LLM judges both the original and perturbed versions on a scale from 1 (completely incorrect) to 5 (perfect), with the expectation that it should assign a strictly lower score to the perturbed version. A model assigning all scores at random would therefore score around 40%.

[1]Alternately referred to as Kanji, Hanja or Hantu

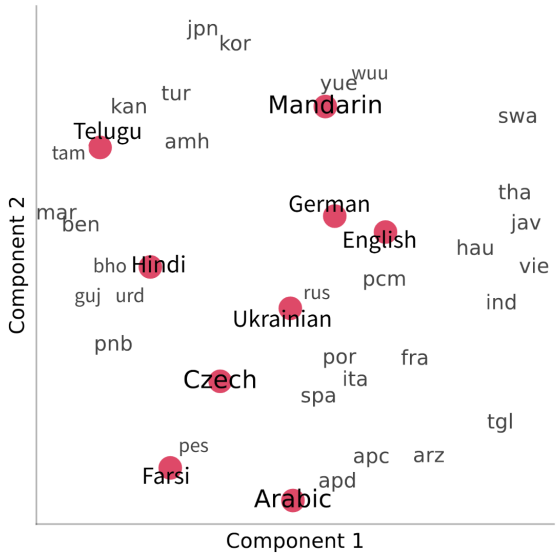

Figure 1: Multidimensional Scaling projection of languages based on syntax features from URIEL/lang2vec. Languages used in our experiments are highlighted and shown with full names, others are in ISO 639/set 2.

Language selection. We choose eight languages for experiments: Mandarin, Hindi, Arabic, German, Farsi, Telugu, Ukranian, Czech in addition to English for comparison. This language selection reflects diverse linguistic properties, varying levels of representation in training data, and different language families (Foundation, 2024), see Table 1. Hindi (Indo-Aryan) and Telugu (Dravidian) represent major languages from the Indian subcontinent that use the Brahmic script and are under-represented in both CommonCrawl and Wikipedia. German and Mandarin, on the other hand, are examples of languages well represented in both CommonCrawl and Wikipedia Farsi and Arabic offer insights into LLM performance on a right-to-left Abjad script, whereas Ukrainian and Czech allow us to study generalisation in medium resource morphologically rich languages, using the Cyrillic and

Language	Questions	Misconception	Feedback	Translation
Mandarin	0.593	0.607	0.666	0.781
Hindi	0.455	0.546	0.593	0.831
Arabic	0.596	0.659	0.605	0.793
German	0.623	0.642	0.697	0.792
Farsi	0.574	0.644	0.708	0.840
Telugu	0.518	0.569	0.578	0.639
Ukrainian	0.607	0.642	0.682	0.821
Czech	0.611	0.626	0.663	0.805

Table 2: Average $\text{COMET}_{23}^{\text{DA,XL}}$ scores for different languages for different components of the tasks. The **Questions** are used for both Misconception and Feedback tasks. The Tutoring task is not translated.

Latin scripts, respectively.

To assess the typological diversity of our selected languages, we used the URIEL typological database (Littell et al., 2017) with `lang2vec`, which provides dense vector representations of languages based on a range of typological, phylogenetic, and geographical features. As recommended by the package, we extracted `syntax` features with k-NN predictions for the missing values for a set of 40 languages, constructed as the union of our core experimental languages and the most widely spoken languages worldwide according to Ethnologue (Eberhard et al., 2025). We projected each language feature vector into two dimensions using Multidimensional Scaling, producing a 2D language similarity plot. This allows us to visualise (see Figure 1) the relative syntactic diversity of our selected languages and confirm that they span a broad typological space. The visualisation demonstrates that our language selection (highlighted) is well distributed across the typological landscape.

Translation. We obtain our tasks in all the abovementioned languages by machine translation. Following the GPT4 Technical Report (OpenAI, 2023, Figure 5), we use Azure Translate to translate all our examples to the target languages. However, this introduces an additional noise source for tasks performed in languages other than English. In fact, after reviewing some of the translations manually, it does look like the translations, though decent, are not as easy to follow as their English counterparts. This finding is further corroborated by $\text{COMET}_{23}^{\text{DA,XL}}$ (Rei et al., 2023) scores of the translations (see Table 7). This means that any differences we observe between English and other-language performance cannot be conclusively attributed to the LLM being tested. However, we can still compare the performance of different LLMs across the same language, as the same translation was used for all LLMs. Further, if at least one LLM performs well in a task on a given language, we can be reasonably certain that the translation for that task-language pair was also good enough.

Models and prompts. We evaluate six state-of-the-art LLMs praised for their multilingual capabilities: GPT-4o (OpenAI, 2023), Gemini 2.0 Flash (Team, 2024), Claude 3.7 Sonnet (Anthropic, 2024), Llama 3.1 405B (Grattafiori et al., 2024), Mistral Large 2407 (AI, 2024; Jiang et al., 2023), and Command-A (Cohere et al., 2025). We leave all sampling parameters to their defaults. For prompts, we use a simple chain of thought prompting method, where the model is first asked to explain why it would pick a certain answer, and then asked to choose it in a separate prompt. Based on literature (Mondshine et al., 2024; Huang et al., 2023), it is unclear whether or not it is beneficial to translate the prompt itself to the target language or keep it in English, so we try both options.[2,3]

For each task, we use 1000 examples for reporting our results, sampled at random from the dataset, except for 200 examples in the tutoring task, which is multi-turn.

3 Results

In this section, we describe the results of five popular large language models on the four tasks described in Section 2. The main results are shown in Tables 3 to 6.

English is easiest for LLMs. The gap between English and other languages is large in general. On

[2]A weaker model roleplays the student model used in the tutoring task to be consistent with the original work. We only use the original prompts because it does not work well with non-English prompts.

[3]We machine-translate the prompts and manually verify (with L1/L2 language knowledge) the translation adequacy.

Input	Options
Question: Which number is the greatest? **Student Answer:** 5.0001 **Right Answer**:5.2	**A:** Believes the mean is total frequency divided by something, **B (correct):** Thinks the more digits a number has the greater it is, regardless of place value, **C:** Believes parallel lines have gradients that multiply to give -1, **D:** When multiplying by a multiple of 10, gives an answer 10 times bigger than it should be
Question: What is the lowest common multiple of 8 and 4? **Student answer:** 4 **Right Answer:** 8	**A:** Subtracts instead of adds when answering worded problems, **B (correct):** Confuses factors and multiples, **C:** Rounds up instead of down, **D:** Adds instead of multiplying when expanding bracket

Example 1: Two examples of the misconception identification task (English).

Input	Options
Question: 6 pencils cost £1.50. How much do 3 pencils cost? **Student answer:** 25p	**A:** I think you have made an arithmetic error when halving £1.50. Use short division to divide by two, **B:** I think you have used the incorrect notation for money. Consider how the monetary values in the question are written, **C (correct answer):** If 6 pencils cost £1.50, then 3 pencils cost half of £1.50, which is £0.75 or 75p., **D (student answer):** I think you have found the cost for one pencil. The question asks for the cost of 3 pencils.
Question: A film starts at 8.50pm. The film lasts 2 hours and 52 minutes. What time does the film finish? **Student answer:** 11.02pm	**A (student answer):** This isn't quite right. Remember that there are 60 minutes in an hour, not 100 :), **B:** I think you've confused your method a little. Noticing that 2 hours and 52 minutes is just 8 minutes less than 3 hours is super, just make sure you add and subtract in the correct directions though :), **C:** Almost there! Take care to notice how many hours and minutes you're adding here. Is your answer 2 hours and 52 minutes later than 8.50pm?, **D (correct answer):** Adding 2 hours to 8.50pm gives 10.50pm. Adding 10 minutes on takes us to 11.00pm, and adding the remaining 42 minutes gives 11.42pm.

Example 2: Two examples of the feedback selection task (English).

Math Problem	Student's (Incorrect) Solution	Correct Solution
Sam sells bread. He has a target of selling 120 crates of bread in a week. One week he was closed on Monday and Friday. Over the weekend he sold 20 crates. On Tuesday he sold 15 crates, on Wednesday 12 crates, and Thursday 18 crates. By how many crates was Sam off from his target for the week?	Sam had 5 days to sell bread because he was closed on Monday and Friday. He sold a total of 20 + 15 + 12 + 18 = 65 crates of bread from Tuesday to Thursday. Adding the 20 crates he sold over the weekend, Sam sold a total of 65 + 20 = 85 crates of bread in a week. Sam was off from his target by 120 - 85 = 35 crates of bread.	During the whole week Sam sold 15 + 12 + 18 + 20 = 65 crates. Sam was off his target by 120 - 65 = 55 crates.
Sophia is thinking of taking a road trip in her car, and would like to know how far she can drive on a single tank of gas. She has traveled 100 miles since last filling her tank, and she needed to put in 4 gallons of gas to fill it up again. The owner's manual for her car says that her tank holds 12 gallons of gas. How many miles can Sophia drive on a single tank of gas?	Sophia used 4 out of the 12 gallons of gas in her tank, so there are 12-4 = 8 gallons of gas left in the tank. If Sophia can drive 100 miles on 4 gallons of gas, then she can drive 100/4 = 25 miles per gallon. Therefore, with 8 gallons of gas left in the tank, Sophia can drive 25 x 8 = 200 miles on a single tank of gas.	To find miles per gallon, divide 100 miles / 4 gallons = 25 miles per gallon. To find how far Olivia can go on a single tank, multiply 25 miles per gallon × 12 gallons = 300 miles.

Example 3: Two examples of the tutoring task.

English Source	Original Translation	Perturbed Translation	Language
It is a kind of tomato.	它是一种番茄	这位工程师有一个家庭	Mandarin
	वह एक तरह का टमाटर है।	भाई एक तरह का टमाटर है।	Hindi
	.هذا نوع من الطماطم	.هذا نوع من التفاح	Arabic
	Es ist eine Art Tomate	Katze ist eine Art Tomate	German
	.این نوعی گوجه‌فرنگی است	.این رنگ گوجه‌فرنگی است	Farsi
	ఇది ఒక రకమైన టమోటా.	ఇది కనుగొంటాము రకమైన టమోటా.	Telugu
	Він впливають різновидом томатів.	Він є різновидом томатів.	Ukrainian
	Je to druh rajčete.	matka to druh rajčete.	Czech

Example 4: A single example of the translation grading task for non-English languages.

Language	English prompt GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	Translated prompt GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
English	97.6%	96.2%	95.1%	94.0%	95.0%	95.3%	97.6%	96.2%	95.1%	94.0%	95.0%	95.3%
Mandarin	· 95.8%	95.2%	· 92.5%	· 92.1%	· 92.8%	· 92.9%	96.5%	95.0%	· 91.7%	93.8%	· 92.9%	94.1%
Hindi	· 94.5%	· 93.2%	· 91.9%	· 89.8%	· 91.8%	· 93.2%	· 95.5%	· 93.6%	· 89.6%	· 90.4%	· 90.6%	· 91.4%
Arabic	· 95.9%	· 93.0%	· 92.0%	⋆ 86.0%	· 92.6%	· 93.4%	· 95.9%	· 93.0%	· 92.8%	· 90.9%	· 92.0%	94.0%
German	· 96.0%	96.2%	94.6%	⋆ 84.6%	95.1%	95.2%	· 95.9%	96.6%	94.0%	★74.0%	94.9%	95.2%
Farsi	· 94.8%	· 93.3%	· 93.0%	· 87.5%	· 92.7%	· 93.1%	· 95.1%	· 94.4%	★68.0%	· 88.3%	★66.9%	· 93.6%
Telugu	· 95.2%	· 92.2%	· 89.9%	· 86.9%	· 89.7%	⋆ 85.5%	· 94.2%	· 90.8%	★68.6%	⋆ 83.6%	★35.5%	★77.9%
Ukranian	· 95.7%	94.9%	· 92.9%	93.3%	94.4%	94.9%	· 95.6%	· 94.3%	★56.6%	· 90.4%	94.2%	93.9%
Czech	96.9%	95.1%	94.5%	92.3%	94.5%	94.1%	96.6%	95.8%	★70.2%	⋆ 81.6%	★41.0%	94.5%

Table 3: Results (accuracy) for the **misconception identification** task. We mark results significantly lower (at least 10%=★, at least 5%=⋆, otherwise ·) than English with a one-sided 95% confidence t-test.

Language	English prompt GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	Translated prompt GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
English	53.4%	38.2%	17.0%	51.1%	48.5%	39.7%	53.4%	38.2%	17.0%	51.1%	48.5%	39.7%
Mandarin	· 49.6%	⋆ 29.7%	· 12.3%	· 43.0%	· 40.1%	· 31.8%	⋆ 41.1%	★19.2%	⋆ 5.8%	★30.3%	★30.3%	⋆ 27.8%
Hindi	· 48.7%	35.6%	· 13.0%	· 43.6%	· 40.5%	· 31.6%	★32.1%	★13.4%	⋆ 6.2%	· 44.3%	★18.6%	★18.8%
Arabic	· 49.6%	⋆ 28.7%	· 13.9%	· 45.3%	⋆ 38.8%	· 33.3%	· 48.8%	★10.7%	16.3%	48.1%	★27.8%	⋆ 28.9%
German	52.5%	· 32.1%	15.0%	· 46.4%	· 42.4%	· 32.8%	50.6%	· 30.8%	15.6%	· 44.4%	⋆ 39.4%	37.6%
Farsi	50.2%	⋆ 27.9%	· 11.3%	· 44.9%	· 41.3%	⋆ 30.9%	· 45.9%	· 31.6%	16.3%	· 44.0%	★33.5%	· 35.5%
Telugu	· 45.2%	⋆ 27.6%	· 10.4%	· 43.4%	★34.0%	⋆ 26.3%	★13.9%	★12.7%	⋆ 6.1%	⋆ 37.7%	★15.5%	★9.5%
Ukranian	50.3%	· 33.2%	· 13.0%	· 44.8%	· 41.3%	· 32.2%	★35.9%	★19.6%	⋆ 8.1%	52.8%	★31.0%	⋆ 27.2%
Czech	49.9%	37.8%	· 14.1%	· 46.5%	· 41.6%	⋆ 30.7%	⋆ 42.7%	⋆ 26.1%	19.2%	· 46.6%	⋆ 35.5%	· 35.6%

Table 4: Results (accuracy) for the **feedback selection** task. We mark results significantly lower (at least 10%=★, at least 5%=⋆, otherwise ·) than English with a one-sided 95% confidence t-test.

	Harmonic mean GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	Success/1-Telling GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
English	94.7%	97.0%	22.1%	93.0%	82.0%	95.5%	96.0/2.5%	97.5/1.0%	96.5/84.0%	93.5/1.0%	82.0/0.0%	96.0/1.0%
Mandarin	89.8%	89.0%	26.4%	79.7%	79.7%	88.2%	94.0/8.0%	90.5/3.0%	90.0/74.5%	80.5/1.5%	80.0/0.5%	93.0/9.0%
Hindi	90.5%	92.7%	24.2%	⋆ 72.2%	73.5%	· 88.4%	95.0/8.5%	93.0/0.5%	89.5/75.5%	77.5/10.0%	73.5/0.0%	91.0/5.0%
Arabic	91.4%	89.7%	24.3%	· 84.2%	75.2%	87.4%	94.5/5.9%	90.0/0.5%	91.0/77.0%	86.0/3.5%	75.5/0.5%	93.0/10.5%
German	90.7%	91.2%	23.4%	84.2%	77.2%	· 86.3%	92.5/3.5%	92.0/1.5%	88.0/74.5%	85.0/1.5%	77.5/0.5%	90.5/8.0%
Farsi	· 85.6%	⋆ 81.3%	28.7%	77.2%	· 65.8%	· 77.8%	89.0/6.5%	87.5/11.5%	91.5/74.5%	78.0/1.5%	69.5/7.0%	91.0/23.0%
Telugu	★50.1%	★39.5%	27.7%	· 58.9%	★2.9%	★40.7%	77.5/40.5%	77.5/51.0%	85.5/69.0%	61.0/4.0%	59.0/57.5%	63.5/33.5%
Ukranian	91.2%	91.5%	23.5%	· 81.2%	71.5%	90.9%	93.0/3.5%	92.0/1.0%	91.5/78.0%	84.0/5.5%	71.5/0.0%	93.5/5.0%
Czech	★43.8%	★44.1%	17.2%	70.2%	★2.9%	★21.5%	65.5/32.5%	73.5/42.0%	90.0/80.5%	71.5/2.5%	52.5/51.0%	77.0/64.5%

Table 5: Results (harmonic mean, success, and telling) for the **tutoring** task. We mark results significantly lower (at least 10%=★, at least 5%=⋆, otherwise ·) than English with a one-sided 95% confidence t-test when occurring in both success and telling. Telling is flipped such that higher is better.

Language	English prompt GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	Translated prompt GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
Mandarin	100.0%	99.3%	98.9%	99.9%	99.5%	99.9%	99.9%	99.4%	24.8%	99.6%	99.4%	99.9%
Hindi	91.5%	74.1%	92.1%	77.6%	82.4%	77.9%	93.8%	88.5%	56.5%	86.5%	87.6%	81.3%
Arabic	98.6%	97.9%	99.2%	98.8%	97.5%	99.0%	98.8%	98.3%	67.2%	98.6%	97.8%	97.9%
German	98.2%	97.9%	97.9%	98.2%	98.2%	98.2%	98.5%	98.3%	29.9%	98.0%	98.3%	97.8%
Farsi	95.3%	93.5%	96.0%	96.4%	92.3%	96.6%	96.8%	96.0%	67.0%	96.4%	94.1%	96.2%
Telugu	77.2%	33.7%	81.0%	51.9%	48.7%	25.2%	82.8%	46.8%	40.7%	82.1%	67.1%	15.6%
Ukranian	98.0%	97.3%	96.9%	96.5%	97.3%	98.3%	98.1%	97.9%	85.3%	97.7%	98.4%	98.2%
Czech	98.7%	98.3%	98.9%	98.3%	97.5%	98.8%	99.3%	98.8%	80.8%	98.7%	99.5%	99.2%

Table 6: Results (accuracy) for the **translation grading** task.

average[4] across all tasks (excluding translation) and models, English has 70.9%, in contrast to 63.1% (Hindi), 55.3% (Czech), 67.8% (Ukrainian), 49.7% (Telugu), 66.2% (Farsi), 66.8% (German), 64.6% (Mandarin) and 67.4% (Arabic). This in itself does not make it clear if the loss is due to the LLMs being weak or the translation quality being poor. The poor performance on Telugu is largely driven by Command-A and Mistral. The former is unsurprising as Telugu is the only language in our list that is not officially supported by it (Cohere et al., 2025). On the other hand, Mistral lists only 12 supported languages of which we test only Hindi, Arabic, German and Chinese. Telugu also has the lowest representation in CommonCrawl and Wikipedia,

[4]Averaging here is done to give a general idea, but we must note that the scores are not equivalent. We use Accuracy for tasks 1 and 2 but Tutoring Score for Task 3

Language	Questions	Misconception	Feedback	Translation
Mandarin	0.593	0.607	0.666	0.781
Hindi	0.455	0.546	0.593	0.831
Arabic	0.596	0.659	0.605	0.793
German	0.623	0.642	0.697	0.792
Farsi	0.574	0.644	0.708	0.840
Telugu	0.518	0.569	0.578	0.639
Ukrainian	0.607	0.642	0.682	0.821
Czech	0.611	0.626	0.663	0.805

Table 7: Average $\text{COMET}_{23}^{\text{DA,XL}}$ scores for different languages for different components of the tasks. The **Questions** are used for both Misconception and Feedback tasks. The Tutoring task is not translated.

so the result is expected. Manual analysis of the low tutoring performance for Czech reveals that the interactions switch between various language formality styles, to the point that it becomes distracting. Additionally, the language used in Czech classrooms is particular and likely not represented on the internet.

Model performance and consistency. Mistral is the most inconsistent across non-English languages (average deviation[5]=0.186). For example, it completely fails the tutoring task for both Czech and Telugu, despite performing reasonably on other languages in the same task. Command-A is not much better (average deviation[5]=0.161). On the other hand, Gemini is the most consistent (average deviation[5]=0.078) and also has the second-best performance (average score 75.0%). GPT4o, is the best performing model (average 78.6%) while Claude performs the worst (average 49.3%) mostly due to Feedback and Tutoring tasks.

Task difficulty. The worst performance is observed in the Feedback task despite the similarity to the Misconception identification task. While Claude is still the standout worst performer with a worse-than-random performance, all models struggle. Further analysis in Table 11 shows that all models tended to default the feedback corresponding to the correct answer, with the models' chain of thoughts being "regardless of the student's mistake, this is the feedback that gives the student the most information about the correct answer." Most models perform well in the Translation evaluation task, with the accuracy being even higher than human annotators, who were presented with attention checks with similar perturbations (Kocmi et al., 2024; Zouhar et al., 2025). They also do well in the Misconceptions task, with most percentage scores (at least in the English prompt setting) being in the 90s. The tutoring task seems to have the most inconsistent performance across models and languages. In general, all models struggle in Czech and Telugu, while Claude struggles in all languages. Avoiding telling seems to be the more challenging part of the problem for all the models, although success rates are not very consistent either.

English and translated prompts. Excluding for the Tutoring task (which did not use native prompts), using English prompts yields better performance than using translated prompts (averages 72.7% and 67.2%). The exceptions to these are GPT, Llama, Gemini, and Mistral in the translation task though in most cases, the difference is not very large. Note that some of the poor performance could be attributed to the prompts being translated and checked for correctness rather than being written in the target language directly, which could introduce some translationese. Regardless, we believe it is best to keep prompts in English. As a further note for English-speaking developers designing multilingual applications, keeping prompts in English ensures that the chains-of-thought remain English, making it easier to run sanity checks.

4 Related Work

LLMs, trained on vast multilingual texts, have dominated tasks such as text generation, translation, and dialogue (Brown et al., 2020), making them promising tools in Intelligent Tutoring Systems (ITS; Corbett et al., 1997; Pal Chowdhury et al., 2024). Prior work explores their use in educational contexts, such as dynamic student interactions (Schmucker et al., 2023), simulating expert and novice behavior (Liu et al., 2023), and math word problem reasoning (Opedal et al., 2023).

Beyond mathematical context, LLMs have also been explored for other forms of learning. Cui

[5]We calculate the standard deviation across the six languages for each task and then calculate the mean.

and Sachan (2023) investigate LLMs in adaptive and personalized exercise generation for language learners, while (Wang et al., 2023) examines how conversational tutoring strategies can aid student understanding. Additionally, LLMs have been used to assess grammatical correctness and translation accuracy (Kocmi and Federmann, 2023; Omelianchuk et al., 2024; Freitag et al., 2024), facilitate automated essay scoring (Pack et al., 2024), and provide corrective feedback in second language writing (Han et al., 2024). While LLMs excel in English, their abilities in other languages often vary, reflecting an over-representation of high-resource languages in pre-training corpora. For example, Koto et al. (2023) introduces IndoMMLU, which reveals significant performance disparities between Indonesian and English contexts. Similarly, Holtermann et al. (2024) examines LLMs across 137 languages and attributes discrepancies in performance to tokenisation strategies. Li et al. (2024); Armengol-Estapé et al. (2022) further find a strong correlation between pre-training data proportions and performance, reaffirming the gap between high- and low-resource languages. For Catalan, Armengol-Estapé et al. (2022) find that while GPT-3 performed well in generative tasks, its comprehension capabilities were limited by the language's moderate representation.

Recent research has increasingly explored the application of LLMs in multilingual educational contexts, though challenges persist in balancing performance across languages. Systematic reviews of AI-based language learning tools highlight the prevalence of NLP and machine learning techniques for error correction, feedback provision, and assessment in non-English contexts, though they note persistent gaps in dialogic competence and teacher preparedness (Alhusaiyan, 2025). Studies evaluating LLMs' cross-lingual capabilities reveal performance disparities, with models demonstrating stronger skill tagging accuracy for English-centric curricula compared to underrepresented languages like Irish or Marathi (Kwak and Pardos, 2024). Bibliometric analyses indicate growing research interest in AI for foreign language education, particularly in vocabulary acquisition and writing support, though most studies still focus on high-resource European and Asian languages (Doğan and Talan). These works collectively underscore both the transformative potential and current limitations of LLMs in achieving equitable multilingual educational support.

To address multilingual education more directly, projects like Kaleidoscope (Salazar et al., 2025) and Aya (Üstün et al., 2024) by Cohere For AI aim to support culturally diverse languages, while SEA-HELM (Susanto et al., 2025) and ECLeKTic (Goldman et al., 2025) emphasise culturally grounded evaluations in Southeast Asian and cross-lingual contexts, respectively. These efforts highlight the need for multilingual benchmarks that move beyond English-centric evaluations.

Prior pedagogical studies tend to assess single LLMs in monolingual settings. We fill this gap by benchmarking LLMs in multiple tasks. Specifically, we conduct zero-shot experiments across multiple models and languages to better analyze their real-world applicability.

5 Conclusion

We analyse the performance of six well-known state-of-the-art LLMs across six languages other than English on four educational tasks. We find that while performance in English continues to be better than in other languages, the drop to other models is not always large. In particular, we find that GPT4o and Gemini 2.0 perform consistently well across all languages, with a few exceptions. We also note that English prompts work as well, if not better, than prompts written in the target language, when solving multilingual tasks. This opens up opportunities for porting applications developed for English into different languages. However, we note that certain models perform poorly in some tasks and languages, so **we recommend** first verifying that a model works well in a particular language on a specific educational task before deployment. However, to answer the question posed by the title, we believe that *atleast some* language models **are** reliable across languages.

Limitations

The shown experiments could naturally be better extended to more languages. The selected languages reflect a balance between author familiarity, which is necessary for meaningful qualitative analysis, and linguistic diversity, as evidenced by their spread in URIEL feature space. Similarly, we only covered six LLMs. In both cases, the cost of experiments (see Table 8) becomes prohibitively expensive, which motivated the data release in this paper to enable further research.

Additionally, translation quality remains a con-

Model	API	Total	Miconception	Feedback	Tutoring	Translation
Mistral	Mistral API	$530	$170	$170	$120	$70
Claude	Anthropic	$600	$190	$190	$135	$85
Command	Cohere	$520	$165	$165	$120	$70
Llama	Together.ai	$600	$190	$190	$135	$80
GPT4o	Open AI	$80	$25	$25	$18	$12
Gemini	Google Genai	$30	$10	$10	$6	$4

Table 8: Approximate costs for the experiments. Does not include taxes or currency conversion charges. The total is about $2360 with approximately an additional $500 spent on preliminary experiments.

cern, as previously discussed. A more thorough evaluation would involve human translations for every task, similar to the MMLU multilingual benchmark (Xuan et al., 2025), but doing so for all our tasks would be resource-intensive.

Finally, the set of tasks is not a complete representation of problems in the education space, primarily because most of the more complex tasks lack well-defined language-agnostic metrics.

Acknowledgements

Sankalan Pal Chowdhury is partially funded by the ETH-EPFL JDPLS Program. Donya Rooein is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR).

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. 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 2598–2637. Association for Computational Linguistics.

Mistral AI. 2024. Large enough: Introducing mistral large 2. Accessed: 2024-09-08.

Bashar Alhafni, Sowmya Vajjala, Stefano Bannò, Kaushal Kumar Maurya, and Ekaterina Kochmar. 2024. LLMs in education: Novel perspectives, challenges, and opportunities. *Preprint*, arXiv:2409.11917.

Eman Alhusaiyan. 2025. A systematic review of current trends in artificial intelligence in foreign language learning. *Saudi Journal of Language Studies*, 5(1):1–16.

Abdullah M. Almasoud, Muhammad Rafay Naeem, Muhammad Imran Taj, Ibrahim Ghaznavi, and Junaid Qadir. 2025. Toward inclusive educational AI: Auditing frontier LLMs through a multiplexity lens. *ArXiv*, abs/2501.03259.

Tariq Alqahtani, H. Badreldin, Mohammed A. Alrashed, Abdulrahman I. Alshaya, S. Alghamdi, Khalid Bin saleh, Shuroug A. Alowais, Omar A. Alshaya, I. Rahman, Majed S Al Yami, and Abdulkareem M. Albekairy. 2023. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in social & administrative pharmacy : RSAP*.

Anthropic. 2024. Introducing claude 3.5 sonnet. Accessed: 2024-09-08.

Sabrina Argoub. 2022. The NLP divide: English is not the only natural language - polis.

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2022. On the multilingual capabilities of very large-scale English language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3056–3068. European Language Resources Association.

Benjamin S. Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Peter A. Cohen, James A. Kulik, and Chen-Lin C. Kulik. 1982. Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, 19(2):237–248.

Team Cohere, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, and 1 others. 2025. Command a: An enterprise-ready large language model. *arXiv preprint arXiv:2504.00698*.

Albert T. Corbett, Kenneth R. Koedinger, and John R. Anderson. 1997. Chapter 37 - intelligent tutoring systems. In Marting G. Helander, Thomas K. Landauer, and Prasad V. Prabhu, editors, *Handbook of Human-Computer Interaction (Second Edition)*, second edition edition, pages 849–874. North-Holland, Amsterdam.

Peng Cui and Mrinmaya Sachan. 2023. Adaptive and personalized exercise generation for online language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10184–10198. Association for Computational Linguistics.

Yunus Doğan and Tarık Talan. Artificial intelligence in foreign language learning: A bibliometric analysis. *Journal of Pedagogical Research*, 9(2):206–230.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. *Ethnologue: Languages of the World*, 28 edition. SIL International, Dallas, Texas. Online version: `http://www.ethnologue.com`.

Wikimedia Foundation. 2024. List of wikipedias by language group. Accessed: 2024-09-08.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81. Association for Computational Linguistics.

Ilie Gligorea, Marius Cioca, Romana Oancea, A. Gorski, Hortensia Gorski, and Paul Tudorache. 2023. Adaptive learning using artificial intelligence in e-learning: A literature review. *Education Sciences*.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, Laura Rimell, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2025. ECLeKTic: A novel challenge set for evaluation of cross-lingual knowledge transfer. *Preprint*, arXiv:2502.21228.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Albert Gu and Tri Dao. 2023. Mamba: Linear-Time sequence modeling with selective state spaces. *Preprint*, arXiv:2312.00752.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. 2024. LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student-LLM interaction. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 284–293. Association for Computational Linguistics.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with multiq. *Preprint*, arXiv:2403.03814.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Enkelejda Kasneci, Kathrin Seßler, S. Küchemann, M. Bannert, Daryna Dementieva, F. Fischer, Urs Gasser, G. Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, J. Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, T. Seidel, and 4 others. 2023. ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*.

Blanka Klimova, Marcel Pikhart, and Liqaa Habeb Al-Obaydi. 2024. Exploring the potential of ChatGPT for foreign language education at the university level. *Frontiers in Psychology*, 15:1269319.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach

for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453. Association for Computational Linguistics.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in indonesia: A comprehensive test on IndoMMLU. *Preprint*, arXiv:2310.04928.

Yerin Kwak and Zachary A. Pardos. 2024. Bridging large language model disparities: Skill tagging of multilingual educational content. *British Journal of Educational Technology*, 55(5):2039–2057.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189. Association for Computational Linguistics.

Maikel Leon. 2024. Leveraging generative AI for on-demand tutoring as a new paradigm in education. *International Journal on Cybernetics & Informatics*.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2024. Quantifying multilingual performance of large language models across languages. *Preprint*, arXiv:2404.11553.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14. Association for Computational Linguistics.

Naiming Liu, Shashank Sonkar, Zichao Wang, Simon Woodhead, and Richard G. Baraniuk. 2023. Novice learner and expert tutor: Evaluating math reasoning abilities of large language models with misconceptions. *Preprint*, arXiv:2310.02439.

Itai Mondshine, Tzuf Paz-Argaman, Asaf Achi Mordechai, and Reut Tsarfaty. 2024. HeSum: a novel dataset for abstractive text summarization in Hebrew. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 26–36. Association for Computational Linguistics.

Maggie A. Mosher, Lisa Dieker, and Rebecca Hines. 2024. The past, present, and future use of artificial intelligence in teacher education. *Journal of Special Education Preparation*.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403. Association for Computational Linguistics.

Nvidia. 2022. Transformer model.

Kostiantyn Omelianchuk, Andrii Liubonko, Oleksandr Skurzhanskyi, Artem Chernodub, Oleksandr Korniienko, and Igor Samokhin. 2024. Pillars of grammatical error correction: Comprehensive inspection of contemporary approaches in the era of large language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 17–33. Association for Computational Linguistics.

Andreas Opedal, Niklas Stoehr, Abulhair Saparov, and Mrinmaya Sachan. 2023. World models for math story problems. *Preprint*, arXiv:2306.04347.

OpenAI. 2019. Language models are unsupervised multitask learners.

OpenAI. 2023. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Austin Pack, Alex Barrett, and Juan Escalante. 2024. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, page 5–15, New York, NY, USA. Association for Computing Machinery.

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. Association for Computational Linguistics.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Jiaju Lin, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, and 15 others. 2023. RWKV: Reinventing RNNs for the transformer era. *Preprint*, arXiv:2305.13048.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text editing by task-specific instruction tuning. *Preprint*, arXiv:2305.09857.

Ricardo Rei, Nuno M. Guerreiro, Jos
textasciitilde A© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics.

Donya Rooein, Paul Röttger, Anastassia Shaitarova, and Dirk Hovy. 2024. Beyond flesch-kincaid: Prompt-based metrics improve difficulty classification of educational texts. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 54–67. Association for Computational Linguistics.

Konstantinos I. Roumeliotis, Nikolaos D. Tselikas, and Dimitrios K. Nasiopoulos. 2023. Llama 2: Early adopters' utilization of meta's new open-source pre-trained model. *Preprints*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2022. Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2340–2354. Association for Computational Linguistics.

Israfel Salazar, Manuel Fernández Burda, Shayekh Bin Islam, Arshia Soltani Moakhar, Shivalika Singh, Fabian Farestam, Angelika Romanou, Danylo Boiko, Dipika Khullar, Mike Zhang, Dominik Krzemiński, Jekaterina Novikova, Luísa Shimabucoro, Joseph Marvin Imperial, Rishabh Maheshwary, Sharad Duwal, Alfonso Amayuelas, Swati Rajwal, Jebish Purbey, and 25 others. 2025. Kaleidoscope: In-language exams for massively multilingual vision evaluation. *Preprint*, arXiv:2504.07072.

Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2023. Ruffle&riley: Towards the automated induction of conversational tutoring systems. *Preprint*, arXiv:2310.01420.

Burr Settles. 2018. Data for the 2018 duolingo shared task on second language acquisition modeling (SLAM).

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xian Bin Yong, Weiqi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. SEA-HELM: Southeast asian holistic evaluation of language models. *Preprint*, arXiv:2502.14301.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939. Association for Computational Linguistics.

Lingzhi Wang, Mrinmaya Sachan, Xingshan Zeng, and Kam-Fai Wong. 2023. Strategize before teaching: A conversational tutoring system with pedagogy self-distillation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2268–2274. Association for Computational Linguistics.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, Nan Liu, Qingyu Chen, Douglas Teodoro, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. *Preprint*, arXiv:2503.10497.

Tiffany Zhu, Kexun Zhang, and William Yang Wang. 2024. Embracing AI in education: Understanding the surge in large language model use by secondary students. *Preprint*, arXiv:2411.18708.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. AI-assisted human evaluation of machine translation. *Preprint*, arXiv:2406.12419.

Language	English prompt GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	Translated prompt GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
English	0.0%	0.7%	0.1%	2.1%	0.0%	0.9%	0.0%	0.7%	0.1%	2.1%	0.0%	0.9%
Mandarin	0.5%	1.6%	0.0%	3.2%	0.0%	0.2%	0.5%	1.6%	0.0%	3.2%	0.0%	0.2%
Hindi	0.0%	1.6%	0.4%	2.3%	0.0%	0.4%	0.0%	0.9%	0.2%	2.5%	0.0%	0.1%
Arabic	0.1%	1.7%	0.2%	2.1%	0.0%	0.2%	0.0%	1.1%	0.3%	2.2%	0.0%	0.1%
German	0.5%	1.6%	0.3%	2.3%	0.0%	0.2%	0.5%	1.6%	0.3%	2.3%	0.0%	0.2%
Farsi	0.0%	1.8%	0.2%	2.0%	0.0%	0.3%	0.0%	1.6%	0.2%	2.9%	0.0%	0.1%
Telugu	0.0%	0.1%	0.0%	2.2%	0.0%	0.4%	0.0%	0.3%	0.1%	1.7%	0.0%	0.0%
Ukranian	0.1%	1.6%	0.1%	2.2%	0.0%	0.3%	0.0%	1.8%	0.4%	1.6%	0.0%	0.1%
Czech	0.1%	1.6%	0.1%	1.9%	0.0%	0.7%	0.0%	1.4%	0.0%	0.7%	0.0%	0.5%

Table 9: Response error rate for the **misconception identification** task.

Language	English prompt GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	Translated prompt GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
English	0.0%	0.3%	0.0%	1.3%	0.0%	0.0%	0.0%	0.3%	0.0%	1.3%	0.0%	0.0%
Mandarin	0.0%	0.1%	0.0%	1.5%	0.0%	0.2%	0.0%	0.0%	0.1%	1.6%	0.0%	0.1%
Hindi	0.0%	0.0%	0.0%	1.1%	0.0%	0.1%	0.0%	0.0%	0.0%	1.0%	0.0%	0.2%
Arabic	0.0%	0.0%	0.0%	1.5%	0.0%	0.1%	0.0%	0.0%	0.0%	2.1%	0.0%	0.2%
German	0.0%	0.0%	0.0%	1.1%	0.0%	0.1%	0.0%	0.0%	0.0%	0.8%	0.0%	0.2%
Farsi	0.0%	0.0%	0.0%	1.2%	0.0%	0.0%	0.0%	0.0%	0.0%	1.1%	0.0%	0.1%
Telugu	0.0%	0.0%	0.0%	1.7%	0.0%	0.1%	0.0%	0.2%	0.0%	1.8%	0.0%	0.1%
Ukranian	0.0%	0.0%	0.0%	0.9%	0.0%	0.0%	0.0%	0.0%	0.0%	1.3%	0.0%	0.0%
Czech	0.0%	0.0%	0.0%	1.1%	0.0%	0.0%	0.0%	0.0%	0.0%	3.0%	0.0%	0.0%

Table 10: Response error rate for the **feedback selection** task.

Language	English prompt GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	Translated prompt GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
English	23.7%	45.8%	75.0%	27.8%	23.0%	35.1%	23.7%	45.8%	75.0%	27.8%	23.0%	35.1%
Mandarin	26.9%	54.5%	81.5%	36.9%	33.8%	47.7%	32.6%	71.9%	89.7%	24.3%	49.3%	54.5%
Hindi	30.3%	42.3%	79.6%	34.9%	29.9%	47.9%	55.5%	78.9%	87.7%	33.2%	68.9%	71.3%
Arabic	28.0%	54.1%	79.5%	35.3%	32.9%	44.0%	21.9%	81.7%	73.6%	22.4%	36.4%	49.5%
German	25.1%	48.8%	79.4%	32.7%	30.4%	45.4%	22.3%	54.5%	77.0%	29.6%	32.9%	36.0%
Farsi	28.5%	52.5%	82.5%	31.6%	32.6%	45.5%	21.6%	52.1%	75.1%	29.3%	30.3%	28.9%
Telugu	29.2%	55.6%	81.5%	35.2%	33.1%	52.4%	78.3%	73.8%	89.5%	37.9%	70.9%	78.8%
Ukranian	27.3%	49.4%	80.3%	32.7%	33.5%	45.2%	47.3%	69.7%	87.1%	20.7%	46.7%	49.7%
Czech	27.9%	39.5%	80.2%	30.2%	31.8%	49.0%	34.9%	59.5%	67.8%	23.0%	33.1%	38.0%

Table 11: Rate of defaulting to the correct answer for the **feedback selection** task.

Language	English prompt GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A	Translated prompt GPT4o	LLama	Claude	Gemini	Mistral	Cmd-A
Mandarin	0.0%	0.1%	0.5%	0.0%	0.0%	0.0%	0.0%	0.1%	4.2%	0.0%	0.0%	0.0%
Hindi	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%
Arabic	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3.7%	0.0%	0.0%	0.0%
German	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	3.5%	0.0%	0.0%	0.0%
Farsi	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Telugu	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%	0.0%	0.0%	0.0%
Ukranian	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.5%	0.0%	0.0%	0.0%
Czech	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.8%	0.0%	0.0%	0.0%

Table 12: Response error rate for the **translation grading** task.

A Experiment Prompts

A.1 Task: Misconception Identification

We used a sequence of 3 prompts:

```
System prompt:
You are an expert math tutor who knows about all grade-school level math misconceptions. Your task is
 to select the accurate type of misconceptions your student has based on the (incorrect) answer he/
she gives to a multiple-choice math question. You will be given 4 misconceptions types. Your selected
 misconception type should correspond to the given question and answer. Explain your reasoning
```

```
User message 1:
Question: {QUESTION}
Selected Answer: {SELECTED_ANSWER}
Misconceptions:
A. {Misconception 1}
B. {Misconception 2}
C. {Misconception 3}
D. {Misconception 4}
```

The position of the Misconception corresponding to the selected answer rotates from question to question. The subsequent assistant message is stored as the chain-of-thought. Thereafter, we sent the second user message.

```
User message 2:
Now based on your above explanation, output the option corresponding to the correct misconception.
Only say 'A', 'B', 'C', or 'D' without any other text. Do not say anything else.
```

The response to this part is the final answer. We regenerate until an answer of 'A', 'B', 'C', or 'D' is received, up to 20 times. If no answer is received, a response of 'E' is saved.

This method is used for all models except Gemini. In case of Gemini, we use the `generate_content` method, which is recommended for non-chat tasks and allows for a single user message. In this case, after obtaining the chain-of-thought, we make a new query with the same system prompt but with the following user message:

```
Gemini message:
You have previously given the following answer and explanation:
{COT}
Now based on your above explanation, output the option corresponding to the correct misconception.
Only say 'A', 'B', 'C', or 'D' without any other text. Do not say anything else.
```

Note that the last part is identical to User Message 2
When using translated prompts, the System Prompt and, User Message 2 and Gemini Message are translated to the target language.

A.2 Task: Feedback Selection

```
System prompt:
You are an expert math tutor who specialises in providing precise and helpful feedback for grade-
school level math questions. Your task is to select the correct explanation for a student's given
answer to a multiple-choice math question.

You will be provided with:
- A math question
- A specific answer chosen by the student (which can be correct or incorrect).
- Four possible explanations (labelled A, B, C, and D).
Your selected explanation should accurately correspond to the given answer. Provide your reasoning
for selecting the explanation.
```

```
User message 1:
Question: {QUESTION}
Selected Answer: {SELECTED_ANSWER}
Feedbacks:
A. {Feedback 1}
B. {Feedback 2}
C. {Feedback 3}
D. {Feedback 4}
```

The position of the Feedback corresponding to the selected answer rotates from question to question. If it is placed at positions A, B, or C, the feedback corresponding to the correct answer is at position D. Otherwise, it is at C. The subsequent assistant message is stored as the chain-of-thought. Thereafter, we sent the second user message.

```
User message 2:
Now based on your above explanation, output the option corresponding to the correct explanation. Only
 say 'A', 'B', 'C', or 'D' without any other text. Do not say anything else.
```

The response to this part is the final answer. We regenerate until an answer of 'A', 'B', 'C', or 'D' is received, up to 20 times. If no answer is received, a response of 'E' is saved.

This method is used for all models except Gemini. In case of Gemini, we use the `generate_content` method, which is recommended for non-chat tasks and allows for a single user message. In this case, after obtaining the chain-of-thought, we make a new query with the same system prompt but with the following user message:

```
Gemini message:
You have previously given the following answer and explanation:
{COT}
Now based on your above explanation, output the option corresponding to the correct explanation. Only
 say 'A', 'B', 'C', or 'D' without any other text. Do not say anything else.
```

Note that the last part is identical to User Message 2. When using translated prompts, the System Prompt, User Message 2, and Gemini Message are translated to the target language. We manually made sure that the formatting was maintained after the translation.

A.3 Task: Tutoring

Student system prompt:

```
Student Persona: {STUDENT_PERSONA}

Math problem: {MATH_PROBLEM}

Student solution: {STUDENT_SOLUTION}

Context: You need to role-play the student, {STUDENT_NAME}, while the user roleplays the tutor. {
STUDENT_NAME} thinks their answer is correct. Only when the teacher provides several good reasoning
questions, {STUDENT_NAME} understands the problem and corrects the solution. {STUDENT_NAME} can use
calculator and thus makes no calculation errors. Send <EOM> tag at end of the student message.
```

Teacher system prompt:

```
A tutor and a student work together to solve the following math word problem.
Math problem: {MATH_PROBLEM}
The correct solution is as follows:
{GROUND_TRUTH}

You need to role-play the tutor while the user roleplays the student, {STUDENT_NAME}. The tutor is a
soft-spoken empathetic man who dislikes giving out direct answers to students, and instead likes to
answer questions with other questions that would help the student understand the concepts, so that
she can solve the problem themselves.

{STUDENT_NAME} has come up with a solution, but it is incorrect. Please start the conversation, one
line at a time, aiming to figure out what is {STUDENT_NAME}'s solution and what is wrong with it.
Then try to get her to fix it.
```

The dialogue history was formatted as user-assistant message pairs for teacher and student roles. We manually set the initial messages to initiate conversations in the target language.

A.4 Task: Translation Grading

```
System prompt:
You are a language translation evaluator. Your task is to assess the quality of a translation from
English to {LANGUAGE}. You will be provided with two sentences:
1. An original English sentence.
2. A translated sentence in {LANGUAGE}.
Your goal is to rate the translation on a scale from 1 to 5 based on the following criteria:
1: The translation is incorrect, incomprehensible, or completely unrelated to the original English
sentence.
2: The translation has significant errors and distorts the meaning of the original English sentence.
3: The translation is understandable but contains notable errors or awkward phrasing.
4: The translation is mostly accurate with minor errors or slightly awkward phrasing.
5: The translation is fluent, natural, and accurately conveys the meaning of the original English
sentence without errors.
Explain your decision
```

```
User message 1:
English: {ENGLISH_SENTENCE}
{LANGUAGE}: {TRANSLATED_SENTENCE}
```

The subsequent assistant message is stored as the chain-of-thought. Thereafter, we sent the second user message.

```
User message 2:
Now based on your above explanation, output the final score from 1 to 5. Only say '1', '2', '3', '4',
 or '5' without any other text. Do not say anything else.
```

The response to this part is the final answer. We regenerate until an answer of '1', '2', '3', '4', or '5' is received, up to 20 times. If no answer is received, a response of '0' is saved.

This method is used for all models except Gemini. In case of Gemini, we use the `generate_content` method, which is recommended for non-chat tasks and allows for a single user message. In this case, after obtaining the chain-of-thought, we make a new query with the same system prompt but with the following user message:

```
Gemini message:
You have previously given the following answer and explanation:
{COT}
Now based on your above explanation, output the final score from 1 to 5. Only say '1', '2', '3', '4',
 or '5' without any other text. Do not say anything else.
```

Note that the last part is identical to User Message 2

This sequence is repeated twice for each sentence, once with the original translation and once with the perturbed translation. The scores are then compared. When using English prompts, the LANGUAGE fields are set to their English exonyms, i.e., Mandarin, Hindi, Arabic, German, Farsi, Telugu, Ukrainian, and Czech. When using translated prompts, the System Prompt, User Message 2, and Gemini Message are translated to the target language. We manually made sure that the formatting was maintained after the translation. We also use the language endonyms, namely 中文, हिन्दी, العربية, Deutsch, فارسی, తెలుగు, Українська, and Čeština·

B Translation Quality

As we mentioned in Limitations, an LLM performing poorly in a given language does not necessarily mean that the LLM itself is bad. It could also mean that information was lost during translation. This is particularly problematic because the machine translation systems likely suffer from the same resource limitations that plague the LLMs in the first place. As such, we manually investigated a small subset of translated questions for the languages they we are fluent in, namely Persian, Arabic, Czech, and Hindi. For each language, we analysed 10 questions each for the Feedback and Misconception tasks, and 20 questions for the Translation Grading task.

In the case of Persian, the only recurring error was with mathematical notation, particularly that the minus sign gets placed to the right of the numbers instead of the left, where it should be. This, however, seems to be a rendering issue, which is a result of the fact that the minus sign ('−', U+2212) is often replaced by the similar-looking hyphen ('-', U+002D), confusing the rendering program into believing that it is rendering text. This should not be an issue since LLMs take raw Unicode encodings as input. Beyond this, there were some minor tense errors, but the meanings were clear.

The issue with sign placement was also observed in Arabic. In addition, there seem to be some translation errors. For example, the word 'travel' used here in the context of the movement of a graph was translated to 'liyusaafir', which is more like 'taking a trip'. We found no errors in the sentences for the translation task. In Czech, the primary source of errors was improper context-dependent terminology. For example, when translating the word 'co-interior (angles)', it missed the 'co' prefix and translated only the 'interior' part. While this is fine in regular speech, in Mathematical terminology, this can be confusing. Despite making the translation harder to follow, the core meaning of the question is preserved.

In Hindi we found several cases where the Hindi sentence was difficult to follow for the Hindi speaking author due to misinterpretation of polysemes by the translator e.g. the word 'round', which was being used in the sense of 'approximate' was translated to the sense of 'circle' and 'property' which was being used in the sense of 'quality', was translated as 'possessions'. Also, the phrase 'Not Quite' was translated to something like 'Not Enough', perhaps due to the word 'quite' not having a Hindi equivalent. However, given the context, using the word for 'Almost' would have been more tonally accurate. However, quite a few translations were hard for the annotator to follow, but backtranslating them yielded reasonably good results, meaning there was no information loss.

The translation exercises showed few errors, perhaps due to the sentences being easy to translate by design. There were one or two mistranslations, but otherwise it worked well. One minor issue was that word boundary detection, which was performed in Python using the regex `'\b\w+\b'`, sometimes identified individual characters in Hindi rather than whole words. However, the resulting sentence still had errors, just not the type of errors that we expected.

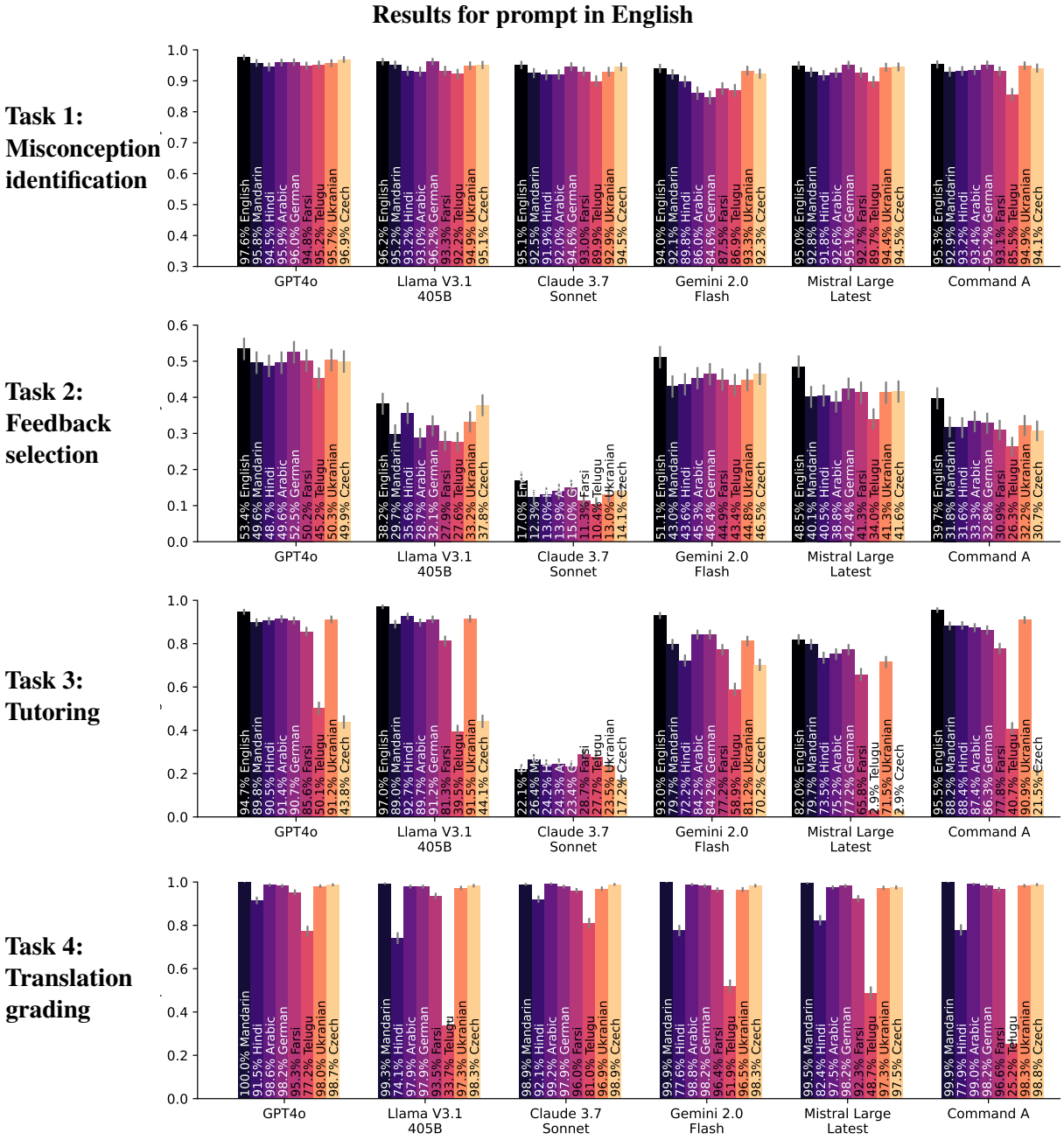

Figure 2: Evaluation results of the four tasks across five lare language models. The error bars show a 95% confidence interval (t-test). MathDial Graphs show *tutoring score after five turns*, most models flatline after 5 utterance pairs. The English language column is absent because translation evaluation uses English as the source. All scores range from 0.0 to 1.0, with higher being better, though they are not comparable with each other. Note the truncated y-axes for better detail. Visualizes Tables 3 to 6.

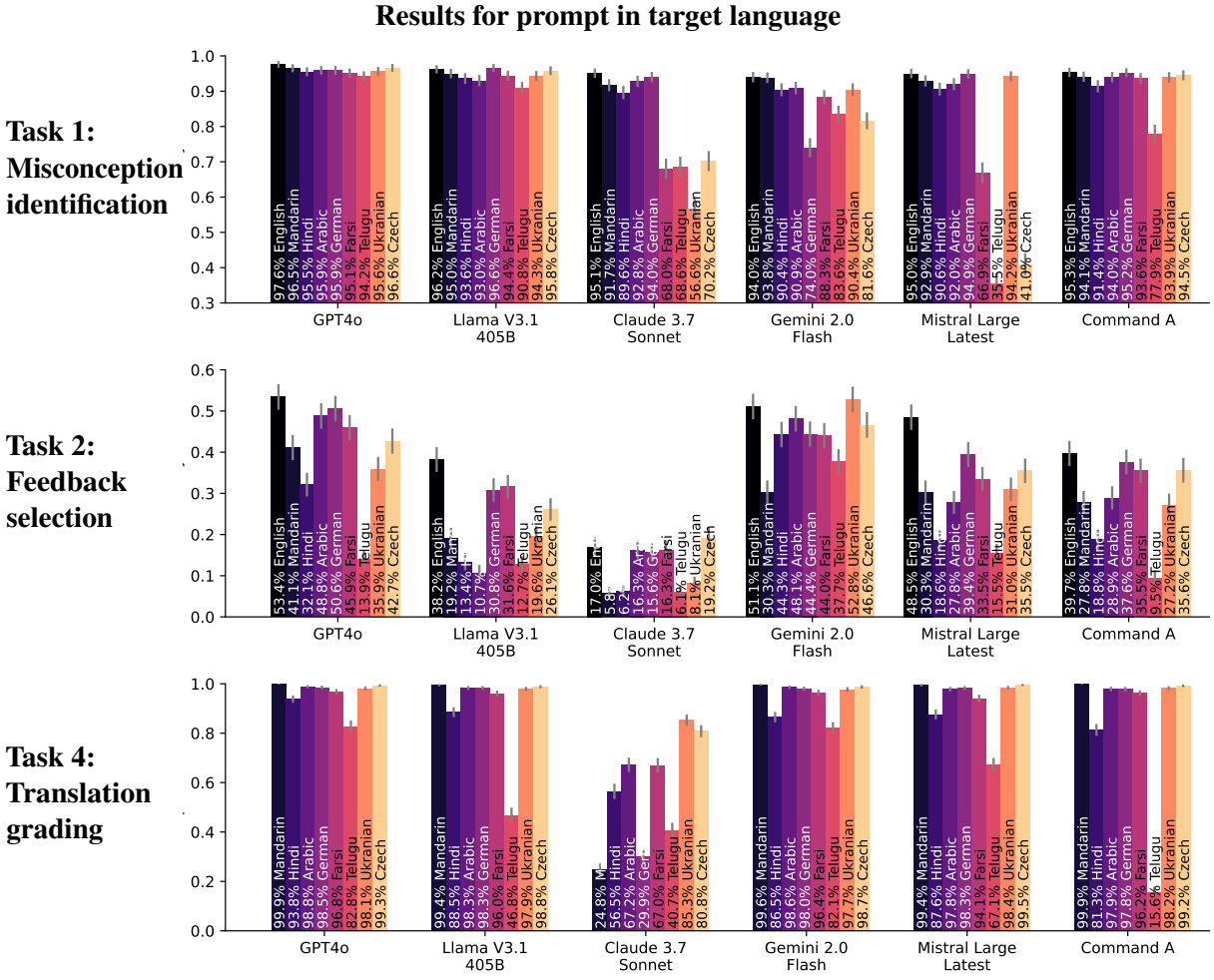

Figure 3: Evaluation results of the four tasks across five large language models. The error bars show 95% confidence interval (t-test). MathDial Graphs show *tutoring score after five turns*, most models flatline after 5 utterance pairs. The English language column is absent because translation evaluation uses English as the source. All scores range from 0.0 to 1.0, with higher being better, though they are not comparable with each other. Note the truncated y-axes for better detail. Visualizes Tables 3 to 6.