

mmPISA-bench: Do LLMs Reason Equally Well Across 43 Languages?

Yerzhan Sapenov
Independent Scholar
ysapenov@gmail.com

Jaromir Savelka
School of Computer Science,
Carnegie Mellon University,
Pittsburgh, PA, USA
jsavelka@cs.cmu.edu

Abstract

We introduce *mmPISA-bench*, a compact high-quality multilingual reasoning benchmark derived from the OECD Programme for International Student Assessment (PISA). The benchmark consists of 25 multiple-choice questions that require reasoning in order to be answered correctly. Each question is provided in official human translations to 43 languages and complemented with machine-translated counterparts (i.e., 2,150 data points in total). We evaluate two mainstream proprietary LLMs across languages, reasoning effort levels, and translation types in terms of their ability to answer the questions correctly. Our results show that modern LLMs can reason effectively across all evaluated languages, achieve accuracy comparable to human test-takers, with some performance variations across covered languages. We further find that machine-translated questions do not degrade accuracy relative to official human translations which suggests that high-quality machine translation (synthetic data) might often be adequate for large-scale multilingual reasoning evaluations where official translations are not available. Finally, we analyze token usage and related inference cost and find that LLMs usage in some languages is simultaneously more expensive and less accurate.

1 Introduction

Large language models (LLMs) have demonstrated strong reasoning capabilities, yet their reasoning ability in many languages remains comparatively under-explored. Despite substantial investments in multilingual modeling, today’s LLM ecosystem is still largely dominated by English (Wu et al., 2025). Persistent challenges include limited data for many languages, uneven performance across language communities, and tokenizer-induced disparities that can affect both effectiveness and cost (Qin et al., 2025). As a

result, reasoning in diverse languages remains at an early stage of evaluation and understanding (Ghosh et al., 2025).

To study reasoning in many languages in a controlled setting, we leverage the OECD Programme for International Student Assessment (PISA), which provides a rigorous framework for assessing student competencies and collects contextual data to explain performance differences (OECD, 2023a). PISA is a worldwide study that measures the proficiency of 15-year-old students in reading, mathematics, and science. Crucially, PISA items undergo extensive translation and validation procedures to ensure cross-country comparability. This includes localization workflows (adaptation, translation, and validation), explicit translatability assessment, and reconciliation practices designed to reduce language-specific artifacts and improve equivalence across versions (OECD, 2024). These properties make PISA questions a high-quality source for evaluating whether LLM performance on questions requiring reasoning is stable across languages, rather than being confounded by low-quality or inconsistent translations.

We investigate the following research questions:

- (RQ1) How stable are frontier LLMs in answering questions requiring reasoning across many different languages?
- (RQ2) How does model performance on machine-translated questions compare to performance on official human translations?
- (RQ3) Do reasoning length vary systematically across languages, and how does this relate to accuracy?

We release a dataset of 2,150 questions requiring reasoning drawn from PISA (25 multiple-choice questions in 43 languages with official human and matched machine translations). Further, we provide an analysis of stability and reasoning effort across human and machine translations in the 43

covered languages for selected frontier LLMs.

2 Related Work

PISA-based evaluation of LLMs. PISA questions have only been used sparingly for LLM evaluation, primarily because most items are not publicly available and access is restricted to a subset of released questions. Takami (2023) evaluated ChatGPT using PISA multiple-choice questions, but limited the comparison to English and Japanese. Başaran et al. (2025) focused exclusively on English reading items to assess reading proficiency. The most extensive prior use of PISA is PISA-Bench (Haller et al., 2025), which adapts PISA questions for evaluating vision language models; however, their benchmark relies on English source items that are machine-translated into five additional languages. In contrast, our work focuses on text-only reasoning and leverages official human translations across 43 languages.

Massively multilingual benchmarks. A broad range of benchmarks has been proposed to evaluate multilingual capabilities of LLMs. Global-MMLU (Singh et al., 2025) extends MMLU to 42 languages with a culturally sensitive subset, though a portion of the data is machine-translated. M3Exam (Zhang et al., 2023) evaluates models across multiple modalities and difficulty levels in nine languages. MMLU-ProX (Xuan et al., 2025) emphasizes reasoning complexity using translated items in 29 languages, while BUFFET (Asai et al., 2024) unifies 15 tasks across 54 languages using machine-translated instructions. GlotEval (Luo et al., 2025) provides a framework for integrating and comparing results from 27 specialized multilingual benchmarks.

Several benchmarks target specific skills or modalities. Belebele (Bandarkar et al., 2024) focuses on reading comprehension in 122 language variants based on FLORES-200 (Costa-jussà et al., 2024). Shi et al. (2023) introduce a multilingual grade-school mathematics benchmark in 10 languages, demonstrating chain-of-thought reasoning beyond English. mSTEB (Beyene et al., 2025) combines text and speech evaluation across many languages using FLORES-200 and FLEURS (Conneau et al., 2022). In addition, AI Language Proficiency Monitor (Pomerence et al., 2025) aggregates results across multiple multilingual benchmarks to track model progress over time. Prior work has also examined trade-offs between ac-

curacy and the language used for reasoning (Qi et al., 2025), as well as approaches to improve non-English reasoning efficiency (Huang et al., 2024) or to disentangle language processing from reasoning (Zhao et al., 2025). Finally, identifying and covering low-resource languages remains an active area of research, with recent progress reaching over a thousand languages (Kargaran et al., 2023).

Compared to these benchmarks, our dataset emphasizes high-quality, fully human-translated questions with explicit difficulty levels, enabling direct comparison to the performance of 15-year-old students across 43 languages.

Reasoning length and cost across languages.

Recent studies indicate that multilingual differences in tokenization create systematic disparities in token counts across languages (Petrov et al., 2023). For models that externalize their reasoning, these disparities manifest not only as higher cost but also as differences in *reasoning length*. It is the amount of generated intermediate text used to justify an answer. Such variation matters because longer reasoning traces increase inference cost and may reflect different internal strategies or difficulty in a given language. While prior work has emphasized the cost implications of token inflation (Ahia et al., 2023), we center our analysis on reasoning length itself and show that some languages elicit longer, more expensive reasoning while still exhibiting reduced accuracy.

3 Dataset

The collected dataset¹ comprises 25 multiple-choice reasoning questions represented in 43 languages, derived from publicly available materials from the OECD Programme for International Student Assessment (PISA). Specifically, the collection includes 11 mathematics items from PISA 2022 and 14 reading comprehension items from PISA 2018 (OECD, 2026).

All publicly accessible PISA items were manually reviewed across available assessment years by the authors. The questions were hand-selected according to the following criteria:

1. availability of a broad set of official language translations;
2. exclusive reliance on textual information, excluding items requiring images or interactive components;

¹<https://github.com/ysapenov/mmPISA-bench>

3. multiple-choice format, excluding items that require evaluation of free-form responses.

These constraints resulted in the inclusion of reading items from PISA 2018 and mathematics items from PISA 2022. Only languages for which complete translations existed for both reading and mathematics questions were retained. Question texts were obtained by structured scraping of interactive, language-specific subpages in [OECD, 2026](#), followed by manual verification against the original English sources to ensure textual fidelity.

PISA items are annotated with eight difficulty levels, comprising six major levels, with level 1 further subdivided into three sublevels (1a, 1b, and 1c). The two assessed competencies are defined by PISA as follows. Mathematics is defined as students’ capacity to reason mathematically and to formulate, employ, and interpret mathematics to solve problems in a variety of real-world contexts, encompassing concepts, procedures, facts, and tools to describe, explain, and predict phenomena. Reading is defined as students’ capacity to understand, use, evaluate, reflect on, and engage with written texts in order to achieve goals, develop knowledge and potential, and participate in society ([OECD, 2023b](#)).

This massively multilingual dataset enables systematic evaluation of LLM performance on questions requiring reasoning across 43 languages. It also supports analysis of machine translation effects, as models can be evaluated on both official human translations and machine-translated variants of the same questions. For mathematics items, the availability of rationales further enables the construction of auxiliary or derived reasoning tasks. As shown in section 5, model performance on machine-translated questions is not lower than on human-translated versions, suggesting that large-scale machine translation could be used to extend the dataset to hundreds of additional languages and to probe the breadth of multilingual reasoning capabilities.

Because each item is decomposed into context, question, and answer options with line-level consistency across languages, the dataset also enables controlled experiments involving mixed-language inputs at the component or line level. In the present experiments, models were not explicitly informed of the input language, leaving language identification implicit; providing such information may represent a potential avenue for improving perfor-

mance. Finally, the dataset supports extensions that increase task difficulty, such as introducing adversarial or incorrect answer options to study robustness across languages ([Góral et al., 2025](#)).

4 Experimental Design

Across all experiments, we issued a total of 107,500 API calls to the evaluated LLMs. Unless explicitly stated otherwise, temperature and all other model parameters were kept at their defaults.

The total number of evaluations is given by:

$$\begin{aligned} 107,500 \text{ data points} = & \\ & 25 \text{ questions} \times 43 \text{ languages} \\ & \times 2 \text{ models} \times 2 \text{ translation types} \\ & \times 5 \text{ reasoning effort levels} \times 5 \text{ repetitions.} \end{aligned} \tag{1}$$

Among the 25 questions, 11 assess mathematical reasoning and 14 assess reading comprehension. With respect to difficulty, 10 questions are at levels 1–2, 6 at levels 3–4, and 9 at levels 5–6, following the official PISA difficulty annotations ([OECD, 2023b](#)).

We evaluated two proprietary frontier LLMs, OpenAI’s GPT and Anthropic Claude, under multiple reasoning effort settings. For GPT, we used the GPT-5.1-2025-11-13 model and evaluated five effort configurations: *none*, *none with double prompt*, *low*, *medium*, and *high*. The double-prompt, no-reasoning configuration was included to test the effect of prompt repetition, following the methodology of [Leviathan et al. \(2025\)](#). The most recent Opus-4-5-20251101 model does not support disabling reasoning effort. Haiku-4-5-20251001 was used to approximate the no-reasoning setting, while higher effort levels were evaluated using the Opus model.

Two translation conditions were considered. The first uses the human translations provided by PISA. The second uses machine-translated versions produced with Google Translate. Because English and French both serve as source languages in the original PISA materials, machine-translated English and French items were obtained by translating each language into the other.

Each question–language–model–configuration combination was evaluated independently five times to assess answer stability and estimate accuracy under stochastic generation. The system prompts used for both models are provided in Appendix B. These prompts were restricted to enforcing a uniform multiple-choice answer format.

Language	Claude		GPT	
	none ^a	high	none	high
Albanian	88.0	94.4	80.0	98.4
Arabic	86.4	98.4	81.6	94.4
Azerbaijani	84.0	95.2	76.0	96.0
Basque	88.8	96.8	80.8	96.0
Bokmål	85.6	96.0	76.0	99.2
Bosnian	87.2	99.2	75.2	96.8
Bulgarian	85.6	96.8	84.0	90.4
Catalan	86.4	99.2	79.2	96.0
Chinese	94.4	96.8	73.6	96.8
Croatian	88.8	96.8	78.4	96.0
Czech	84.8	96.8	79.2	94.4
Danish	89.6	98.4	81.6	93.6
Dutch	90.4	96.0	84.8	95.2
English	90.4	100.0	76.8	96.8
Estonian	89.6	96.0	75.2	93.6
Finnish	88.0	100.0	83.2	98.4
French	91.2	99.2	79.2	96.8
Galician	88.8	98.4	74.4	93.6
Georgian	84.8	100.0	80.0	99.2
German	85.6	96.8	81.6	100.0
Greek	84.0	88.0	79.2	90.4
Hebrew	88.0	100.0	73.6	99.2
Hungarian	80.8	98.4	84.8	98.4
Icelandic	88.8	92.0	73.6	92.0
Indonesian	85.6	95.2	76.0	96.0
Italian	88.8	92.0	74.4	92.8
Japanese	88.0	99.2	78.4	96.8
Kazakh	80.0	93.6	74.4	92.8
Korean	88.0	92.0	73.6	93.6
Latvian	86.4	96.0	77.6	96.8
Lithuanian	91.2	95.2	80.8	97.6
Malay	90.4	96.0	80.0	96.0
Nynorsk	85.6	92.0	80.8	98.4
Polish	87.2	94.4	78.4	96.0
Portuguese	89.6	100.0	82.4	96.8
Russian	85.6	95.2	84.8	94.4
Serbian	87.2	99.2	77.6	96.0
Slovak	88.0	96.0	74.4	95.2
Slovenian	89.6	95.2	80.0	96.0
Spanish	87.2	98.4	82.4	90.4
Swedish	86.4	100.0	81.6	94.4
Thai	85.6	94.4	84.0	96.8
Turkish	92.0	100.0	80.8	96.8
Total	87.5	96.6	78.9	95.7

Table 1: Comparative Language Performance Accuracy, %: Claude vs GPT.

^a Haiku model was used for none reasoning case

Model	Effort	Original			Machine		
		Acc	In	Out	Acc	In	Out
Claude	high	96.6	4219	2258	97.4	4101	2230
	medium	96.2	4767	796	96.6	4650	778
	low	94.6	4767	515	95.5	4650	513
	double	86.6	8304	2708	87.0	8068	2708
	none	87.5	4219	2822	87.5	4101	2805
	Total		92.3	26277	9098	92.8	25569
GPT	high	95.7	3087	3709	96.1	2974	3368
	medium	94.7	3087	1434	94.9	2970	1325
	low	93.7	3087	582	94.3	2970	562
	double	84.2	6089	59	85.5	5854	59
	none	78.9	3087	59	79.8	2970	59
	Total		89.5	18439	5843	90.1	17738

Table 2: Model accuracy in %, thousands of input and output tokens across model reasoning effort levels.

Aside from the system prompt, no additional instructions or contextual information were supplied to the models. All evaluations were conducted in a zero-shot setting. Cost is computed as the sum of input tokens multiplied by the input token price and output tokens multiplied by the output token price.

5 Results

Subsections 5.1 and 5.3 report results on the original, human-translated questions only. Subsection 5.2 compares performance on original questions against their machine-translated counterparts.

5.1 RQ1: Consistency across languages

Table 1 reports accuracy by language for both models under no-reasoning and high-reasoning settings. Both models’ performance varies somewhat across languages, and these differences persist across reasoning levels. While performance differences across languages clearly exist the relatively bounded extent of the variations indicates reasonably capable multilingual reasoning behavior across the 43 studied languages.

5.2 RQ2: Human vs machine translation

Table 2 compares accuracy and token usage between original human translations and machine-translated questions across reasoning effort levels. Overall, accuracy on machine-translated questions is not lower than on the original versions for either model, and in several configurations it is marginally higher. This suggests that machine translation does not introduce performance degradation for the evaluated reasoning tasks.

Table 3 further breaks down results by PISA difficulty level. As expected, accuracy decreases

Model	Difficulty	Original			Machine		
		Acc	In	Out	Acc	In	Out
Claude	level 5-6	80.0	869	415	80.5	843	411
	level 3-4	99.1	952	302	99.6	931	305
	level 1-2	99.3	1092	292	99.7	1061	288
GPT	level 5-6	75.6	610	475	76.7	586	429
	level 3-4	93.9	683	112	94.8	659	117
	level 1-2	99.2	756	49	99.4	727	44

Table 3: Model accuracy in %, average token usage per question across difficulty levels.

Model	Category	Original			Machine		
		Acc	In	Out	Acc	In	Out
Claude	Math	85.1	557	355	84.9	542	357
	Reading	98.0	1308	325	98.9	1273	320
GPT	Math	82.4	400	372	83.4	381	344
	Reading	95.0	911	96	95.4	879	86

Table 4: Model accuracy in %, average token usage per question across categories.

with increasing difficulty, while average token usage per question increases. This trend is consistent across both original and machine-translated questions, indicating that difficulty effects dominate translation effects.

Table 4 shows results by question category. Both models perform better on reading comprehension than on mathematics, with similar patterns observed for original and machine-translated inputs. Token usage differs substantially between categories, particularly the GPT model.

5.3 RQ3: Reasoning length

Table 5 summarizes the overall accuracy–cost trade-off. Claude achieves higher average accuracy, but at more than double the cost of GPT.

Figure 1 illustrates the relationship between input and output token usage across languages. Claude exhibits a strong positive correlation between input and output tokens, whereas this relationship is notably weaker for GPT. The correlation between input and output tokens is extremely high for Claude (0.950) but much weaker for GPT (0.334). This indicates that Claude’s reasoning length scales closely with the length of the input across languages, while GPT exhibits more

Model	Accuracy	Cost (\$)
Claude	92.3%	198.1
GPT	89.5%	81.5

Table 5: Comparison of Model Accuracy and Cost.

decoupled behavior in which longer prompts do not consistently result in longer generated reasoning.

Figure 2 plots accuracy against cost by language. For both models, higher cost is generally associated with lower accuracy, yielding negative correlations (Claude: -0.484 ; GPT: -0.339). Importantly, cost reflects two distinct sources: tokenization-driven input inflation and variation in reasoning length (output tokens). For Claude, languages with higher input token counts also tend to elicit longer reasoning traces (Figure 1), amplifying cost and contributing to lower accuracy. In contrast, GPT exhibits weaker coupling between input length and output length, suggesting that cost–accuracy degradation cannot be attributed to tokenization alone but also reflects language-dependent differences in generated reasoning.

5.4 Additional results

Figure 3 compares accuracy between mathematics and reading questions. The correlation between category-specific accuracies for Claude is weak (0.210), and for GPT it is negative (-0.324), suggesting that performance on one category does not reliably predict performance on the other.

Although LLM outputs are stochastic, qualitative inspection reveals that models can exhibit distinct reasoning behaviors across languages for the same question. Tables 6, 8, and 7 show that the model’s reasoning is not necessarily carried out in prompt language or English. In some cases, the input text is implicitly translated from the prompt language into another language used for reasoning. Notably, for Kazakh prompts, Claude performs parts of its reasoning in Russian rather than English and may switch between languages within a single response. These behaviors were identifiable because one of the authors is fluent in Kazakh and Russian. Similar cross-language reasoning may occur in other languages but remain difficult to detect without native-language expertise.

6 Discussion

This section interprets the empirical results in light of the three research questions, focusing on multilingual reasoning robustness, translation effects, and cost–accuracy trade-offs.

6.1 RQ1: Consistency across languages

The results indicate that leading LLMs can reason across all evaluated languages at a level compara-

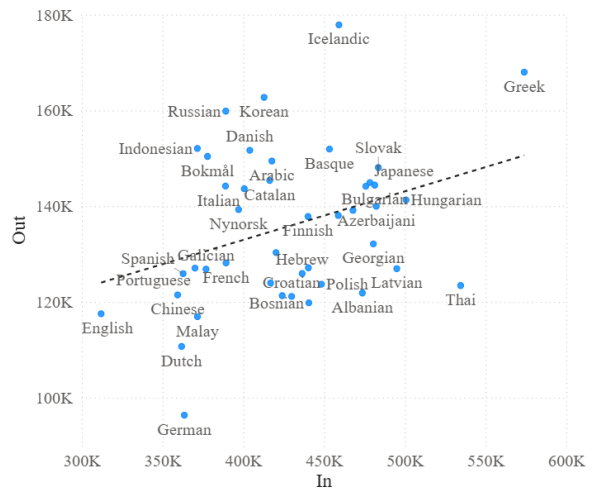
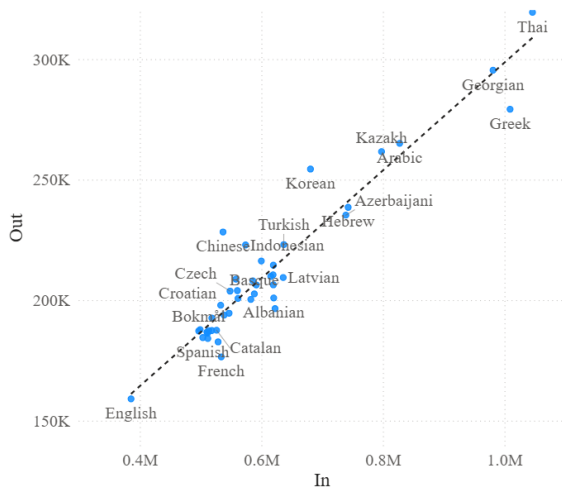


Figure 1: Claude (left) and GPT (right) input and output tokens.

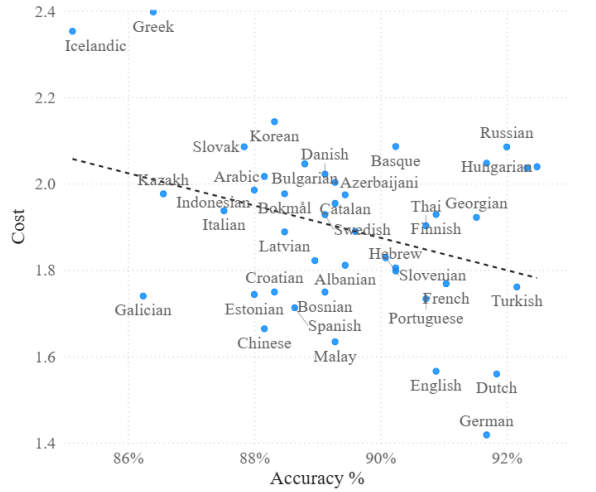
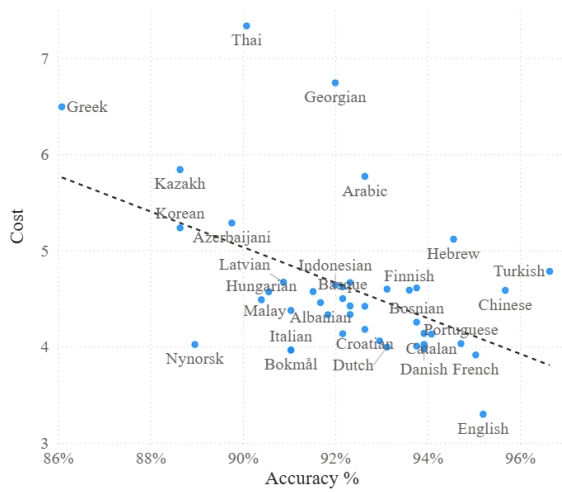


Figure 2: Claude (left) and GPT (right) accuracy (%) versus cost (\$).

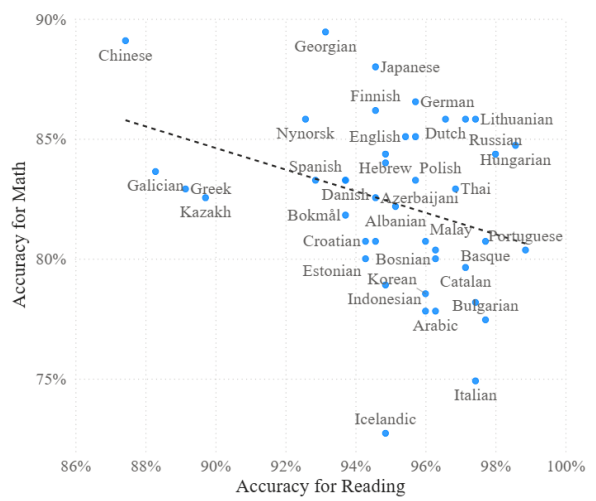
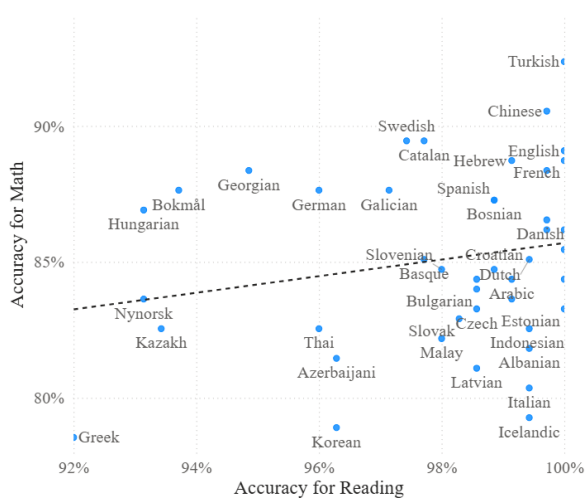


Figure 3: Claude (left) and GPT (right) accuracy (%) for Math and Reading categories.

Response A (Russian)	Response B (Kazakh)
Мне нужно определить, какие комментарии в этом форуме относятся к проблеме Индира_88 (вопрос о том, можно ли давать аспирин курице).	Тауықтарға аспирин беру туралы сұрақ бойынша әрбір хабарламаның проблемасына қатысты немесе қатыстас емес екенін анықтау керек.
1. НұргүлВ79 - "Не знаю, безопасен ли аспирин для птиц...". Это прямое выступление по поводу проблемы Индиры. Иә (Да)	1. НұргүлВ79 хабарламасы: "...сұрақты анықтамайды, бірақ қауіпсіздік мәселесін қозғайды" - Иә
2. Мөлдiр - "Я дал аспирин своей курице когда-то...". Это прямой опыт, связанный с проблемой Индиры. Иә (Да)	2. Мөлдiр хабарламасы: "...БҰЛ ПРОБЛЕМАСЫНА ҚАТЫСТЫ (тәжірибе пен дозалау туралы)" - Иә
3. Айдың ұсыныстар - "Не забудьте посмотреть мои отличные низкие цены...". Это спам/реклама, не относится к проблеме Индиры. Жок (Нет)	3. Айдың ұсыныстар хабарламасы: "ҚАТЫСТАС ЕМЕС (жөнсіз сату ойтарымы)" - Жок
4. Бағылан - "Вы можете объяснить мне, как узнать, что курица больна?". Это новый вопрос, косвенно связанный... Жок (Нет)	4. Бағылан хабарламасы: "ПРОБЛЕМАСЫНА ҚАТЫСТАС ЕМЕС (өзгеше сұрақ)" - Жок
5. Фархат - "Привет Индира, я ветеринар...". Это прямой профессиональный ответ на вопрос Индиры. Иә (Да)	5. Фархат хабарламасы: "БҰЛ ПРОБЛЕМАСЫНА ТІКЕЛЕЙ ҚАТЫСТЫ (түсінік дәрігерден)" - Иә
Получается: ['Иә', 'Иә', 'Жок', 'Жок', 'Иә']	Сонымен: ['Иә', 'Иә', 'Жок', 'Жок', 'Иә']

Table 6: Comparison of Claude reasoning for the same question #23 in Kazakh language across independent runs (same prompt, identical settings).

ble to that expected of 15-year-old students. Compared to earlier evaluations, cross-lingual performance gaps appear reduced, suggesting improved multilingual robustness in recent models (Petrov et al., 2023). As shown in Table 1, under high reasoning effort GPT achieves 100% accuracy only for German, whereas Claude reaches perfect accuracy in seven languages.

Claude’s accuracy spans 88%–100% under high effort and 80%–94.4% without explicit reasoning, corresponding to ranges of 12.0%–14.4%, respectively. GPT exhibits narrower ranges: 90.4%–100% under high effort and 73.6%–84.8% without reasoning, corresponding to ranges of 9.6%–11.2%. Thus, although Claude achieves higher average accuracy, GPT displays less variability across languages.

6.2 RQ2: Human vs machine translation

Prompt repetition improves accuracy for GPT in the no-reasoning setting, as shown in Table 2. In contrast, Claude Haiku does not benefit from double prompting, consistent with prior findings (Leviathan et al., 2025). These results suggest that prompt repetition is model-dependent and primarily beneficial for architectures that do not expose internal reasoning by default.

Increasing question difficulty leads to lower accuracy and higher output token usage, as shown in Table 3. Notably, higher-difficulty questions

are often shorter in terms of input tokens, yet they elicit longer outputs, indicating more elaborate reasoning processes that more closely resemble human problem-solving behavior.

Table 4 further shows category-specific differences. GPT uses substantially more output tokens for mathematics questions than for reading, whereas Claude’s output token usage is comparatively similar across categories. This suggests different internal strategies for handling numerical reasoning versus textual comprehension.

6.3 RQ3: Reasoning length

Token usage varies substantially across languages, corroborating earlier observations disparities based on tokenization premium (Petrov et al., 2023). For Claude, the highest tokenization premium is observed for Thai in Figure 1, with factors of 2.71 for input tokens and 2.01 for output tokens relative to English. Crucially, we also observe systematic variation in reasoning length across languages, operationalized as output tokens under a fixed reasoning-effort setting. For some languages, models produce substantially longer rationales even when answering the same items. These values are lower than previously reported maxima for the `cl100k_base` tokenizer (Petrov et al., 2023), indicating partial mitigation of extreme token inflation.

For GPT, the largest input token premium oc-

curs for Greek (1.84), while the largest output token premium is observed for Icelandic (1.51) in Figure 1. In the Claude case, English consistently yields the lowest token usage and exhibits a strong correlation between input and output tokens, suggesting that languages that are “longer to read” also tend to elicit longer reasoning traces. In contrast, GPT shows weaker coupling between input and output length, indicating that cross-linguistic differences in generated reasoning verbosity are not fully explained by tokenization alone.

Figure 2 illustrates the relationship between cost and accuracy. For both models, languages that are less costly to process tend to yield higher accuracy. For example, Claude incurs $2.22\times$ higher cost on Thai than on English while achieving 5.1 percentage points lower accuracy. Similarly, GPT spends $1.69\times$ more on Greek than on German, with a corresponding 5.3 percentage point accuracy decrease. These findings indicate that some languages are simultaneously more expensive and less accurate, reinforcing the importance of cost-aware multilingual evaluation. Taken together, these results suggest that multilingual evaluation should report not only accuracy but also reasoning length, since some languages systematically induce longer and sometimes less effective reasoning.

6.4 Additional results

Figure 3 compares performance on mathematics and reading questions. The weak correlation between category-specific accuracies suggests that strong performance in one category does not necessarily transfer to the other. Notably, GPT achieves one of its lowest reading accuracies overall, yet performs strongly on reading questions in Chinese, highlighting language–category interactions that merit further investigation.

The qualitative examples in Tables 6, 7 and 8 show that multilingual differences extend beyond accuracy and reasoning length to the linguistic behavior of model reasoning. In particular, Claude exhibits variation in reasoning structure across languages, including explicit translation during reasoning and cross-language reasoning. Notably, in some cases the model switches to a language different from both the input language and English (e.g., Russian when prompted in Kazakh), suggesting that intermediate reasoning may occur in a latent pivot language. These behaviors were identifiable only because one of the authors is fluent in Kazakh and Russian, highlighting a broader eval-

uation challenge: such phenomena may remain invisible without native-language expertise. This suggests that multilingual LLM evaluation would benefit from qualitative inspection in addition to aggregate accuracy metrics.

7 Conclusion

Our results show that leading proprietary LLMs are capable of reasoning across all 43 evaluated languages. While overall reasoning accuracy remains high, both performance and inference cost vary substantially across languages, reflecting differences in multilingual robustness and tokenization efficiency. These findings underscore the importance of evaluating LLM reasoning beyond English and of jointly considering accuracy and cost when assessing multilingual capabilities.

8 Future Work

Several directions follow naturally from this study. First, the dataset can be extended to a substantially larger number of languages by leveraging machine translation systems, such as Google Translate, which support more than 250 languages. Targeted human translations could be obtained for selected ultra low-resource languages to assess reasoning capabilities under extreme data scarcity. Second, future work may compare the observed trends with those of open-source LLMs, enabling analysis of how architectural choices and training regimes affect multilingual reasoning. Finally, a longitudinal analysis tracking multilingual reasoning performance of LLMs over the past several years would provide insight into the pace and nature of progress in this area.

Acknowledgments

We thank the Organisation for Economic Cooperation and Development (OECD) for conducting the PISA assessments worldwide and for providing open access to the test questions in numerous languages. The second author acknowledges generous support of Carnegie Mellon-Accenture Center of Excellence in AI-Enabled Workforce Training (ACE-AI). The content of this paper does not necessarily reflect the position or the policy of the funding organization and no official endorsement should be inferred.

References

- Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. [Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.
- Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. [BUFFET: Benchmarking Large Language Models for Few-shot Cross-lingual Transfer](#). 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 1771–1800.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. [The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.
- Mehmet Başaran, Ömer Faruk Vural, and Cennet Tandırcı. 2025. [Assessing AI in Educational Evaluation: A Comprehensive Analysis of ChatGPT’s Performance on PISA Reading Skills](#). *Technology, Knowledge and Learning*.
- Luel Hagos Beyene, Vivek Verma, Min Ma, Jesujoba O. Alabi, Fabian David Schmidt, Joyce Nakatumba-Nabende, and David Ifeoluwa Adelani. 2025. [mSTEB: Massively Multilingual Evaluation of LLMs on Speech and Text Tasks](#). In *Proceedings of the 2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.
- Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. [FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech](#). In *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.
- Marta Ruiz Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Mailard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630(8018):841–846.
- Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. [A Survey of Multilingual Reasoning in Language Models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Gracjan Góral, Emilia Wiśnios, Piotr Sankowski, and Paweł Budzianowski. 2025. [Wait, that’s not an option: LLMs Robustness with Incorrect Multiple-Choice Options](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1495–1515.
- Patrick Haller, Fabio Barth, Jonas Golde, Georg Rehm, and Alan Akbik. 2025. [PISA-Bench: The PISA Index as a Multilingual and Multimodal Metric for the Evaluation of Vision-Language Models](#). *arXiv preprint arXiv:2510.24792*.
- Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024. [MindMerger: Efficiently Boosting LLM Reasoning in non-English Languages](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. [GlotLID: Language Identification for Low-Resource Languages](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2025. [Prompt Repetition Improves Non-Reasoning LLMs](#). *arXiv preprint arXiv:2512.14982*.
- Hengyu Luo, Zihao Li, Joseph Attieh, Sawal Devkota, Ona de Gibert, Xu Huang, Shaoxiong Ji, Peiqin Lin, Bhavani Sai Praneeth Varma Mantina, Ananda Sreenidhi, Raúl Vázquez, Mengjie Wang, Samea Yusufi, Fei Yuan, and Jörg Tiedemann. 2025. [GlotEval: A Test Suite for Massively Multilingual Evaluation of Large Language Models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 602–614.
- OECD. 2023a. [PISA 2022 Assessment and Analytical Framework](#). PISA. OECD Publishing, Paris.
- OECD. 2023b. [PISA 2022 Results \(Volume I\): The State of Learning and Equity in Education](#). PISA. OECD Publishing, Paris.
- OECD. 2024. [PISA 2022 Technical Report](#). PISA. OECD Publishing, Paris.
- OECD. 2026. [PISA Test Examples: Mathematics and Reading](#). <https://www.oecd.org/en/about/programmes/pisa/pisa-test.html>. Accessed: December 5, 2025.
- Aleksandr Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. [Language Model Tokenizers Introduce Unfairness Between Languages](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023)*, pages 36963–36990. Curran Associates Inc.

- David Pomeranke, Jonas Nothnagel, and Simon Ostermann. 2025. [The AI Language Proficiency Monitor – Tracking the Progress of LLMs on Multilingual Benchmarks](#). *arXiv preprint arXiv:2507.08538*.
- Jirui Qi, Shan Chen, Zidi Xiong, Raquel Fernández, Danielle S. Bitterman, and Arianna Bisazza. 2025. [When Models Reason in Your Language: Controlling Thinking Language Comes at the Cost of Accuracy](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20279–20296.
- Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. [A survey of multilingual large language models](#). *Patterns*, 6(1):101118.
- 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. 2023. [Language Models are Multilingual Chain-of-Thought Reasoners](#). In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.
- Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Raymond Ng, Shayne Longpre, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. [Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.
- Kyosuke Takami. 2023. [Exploring ChatGPT Performance on PISA Multiple Choice Sample Questions Comparing English and Japanese Expression](#). In *Proceedings of the Workshop on The Applications of Generative Artificial Intelligence (GAI) in Education, 31st International Conference on Computers in Education (ICCE)*, Matsue, Japan.
- Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. [The Bitter Lesson Learned from 2,000+ Multilingual Benchmarks](#). *arXiv preprint arXiv:2504.15521*.
- Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. [MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation](#). *arXiv preprint arXiv:2503.10497*.
- Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. [M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models](#). In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.
- Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, Tat-Seng Chua, and Ting Liu. 2025. [When Less Language is More: Language-Reasoning Disentanglement Makes LLMs Better Multilingual Reasoners](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Limitations

The evaluation in this study is limited to two proprietary large language models. While these models are representative of current frontier systems, their training data, architectures, and inference mechanisms are not publicly documented. Consequently, the observed patterns in multilingual reasoning, reasoning length, and cost may not generalize to other proprietary models or to open-source models developed under different training regimes.

In addition, the benchmark draws exclusively on publicly released PISA questions from specific assessment years (PISA 2018 for reading and PISA 2022 for mathematics). Although these items were selected to maximize language coverage and comparability, they represent only a subset of PISA competencies and formats. In particular, the exclusion of constructed-response items, visual prompts, and interactive tasks limits the scope of reasoning behaviors that can be evaluated. Extending the benchmark to additional PISA cycles or complementary assessment frameworks would help capture a wider spectrum of multilingual reasoning skills.

B System Prompts

The system prompt for GPT models: "Reply format: <LETTER>". Interesting that the same prompt did not work for Claude. It looks like GPT can easily have hidden reasoning tokens, while Claude displays all the tokens. So, the system prompt for Claude models: "Reply text inside the <reasoning> tags. Output only the letter answer outside the tags."

C Additional Results

Response A	Response B
Looking at the data for Suður-Kórea (South Korea): 64.42, 64.08, 63.69	Skoðum gögnin fyrir Suður-Kóreu: 64,42 (2005), 64,08 (2010), 63,69 (2015).
Comparing to Panama: 64.33, 63.21, 62.11	Samanborið við Panama: 64,33 (2005), 63,21 (2010), 62,11 (2015).
In 2005: Suður-Kórea (64.42) > Panama (64.33)	Árið 2005: Suður-Kórea (64,42) > Panama (64,33) - já, Suður-Kórea hæst
In 2010: Suður-Kórea (64.08) > Panama (63.21)	Árið 2010: Suður-Kórea (64,08) > Panama (63,21) - já, Suður-Kórea hæst
In 2015: Suður-Kórea (63.69) > Panama (62.11)	Árið 2015: Suður-Kórea (63,69) > Panama (62,11) - já, Suður-Kórea hæst
South Korea has the highest values for all three years.	Suður-Kórea er hæst í öllum árum.

Table 7: Comparison of Claude reasoning for the same question #11 in Icelandic language across independent runs (same prompt, identical settings).

Response A	Response B
The text clearly states: "Тоғыз ай бойы зерттеген моайды, яғни аралдың мүсіндерін және құздарын бүгін кешке аралап, олармен қоштасамын."	Мәтінде профессор өзінің жұмысы туралы қандай ақпарат берген бөлімді іздейміз.
This translates to: "For nine months I have been studying moai, that is, the statues and caves of the island, and today in the evening I will walk around and say goodbye to them."	Негізгі сөйлем: "Тоғыз ай бойы зерттеген моайды, яғни аралдың мүсіндерін және құздарын бүгін кешке аралап, олармен қоштасамын."
The blog was published on "23 мамыр" (May 23rd), and the professor states that he has been working there for "тоғыз ай" (nine months) and is now finishing his work.	Және сонымен қатар: "Аймақтағы жұмысымды аяқтадым, енді үйге қайтуым керек."
The question asks when the professor started his work in the region according to the blog data.	Профессор мәтінінде анық айтады: - Бүгін (23 мамыр) солай жарияланды - Тоғыз ай бойы зерттегенін - Бүгін аралдан кетеленің соңғы күні болғанын
If the professor has been working for nine months and is finishing now (on May 23rd), then he would have started approximately nine months before May 23rd.	Осы ақпаратқа сүйене отырып, егер өзі "тоғыз ай бойы" зерттеген болса, ол тоғыз ай бұрын басталған деген сөз.
The answer is clearly stated in the text: the professor worked there for nine months ("Тоғыз ай бойы").	Басқа опцияларды тексерейік: - A: 1990 жылдар - бұл қимасы қаралған мүсіндерді қалай жылжыттығы ашылған уақыт, профессордың жұмысының уақыты емес - C: Бір жыл бұрын - тоғыз ай нақты айтылған - D: Мамыр айының басында - мәтін 23 мамырда жарияланды бірақ жұмыс тоғыз ай бұрын басталды
Therefore, the answer is B. Тоғыз ай бұрын (Nine months ago).	Түзік жауап - B: Тоғыз ай бұрын

Table 8: Comparison of Claude reasoning for the same question #14 in Kazakh language across independent runs (same prompt, identical settings).