The Reasoning Lingua Franca: A Double-Edged Sword for Multilingual AI

Alan Saji[1‡], **Raj Dabre**[1,2,3] **Anoop Kunchukuttan**[1,4], **Ratish Puduppully**[5‡]

[1]Nilekani Centre at AI4Bharat, [2]Indian Institute of Technology Madras, India,
[3]Google, [4]Microsoft, India, [5]IT University of Copenhagen

Abstract

Large Reasoning Models (LRMs) achieve strong performance on mathematical, scientific, and other question-answering tasks, but their multilingual reasoning abilities remain underexplored. When presented with non-English questions, LRMs often default to reasoning in English, raising concerns about interpretability and the handling of linguistic and cultural nuances. We systematically compare an LRM's reasoning in English versus the language of the question. Our evaluation spans two tasks: MGSM and GPQA Diamond. Beyond measuring answer accuracy, we also analyze cognitive attributes in the reasoning traces. We find that English reasoning traces exhibit a substantially higher presence of these cognitive behaviors, and that reasoning in English generally yields higher final-answer accuracy, with the performance gap increasing as tasks become more complex. However, this English-centric strategy is susceptible to a key failure mode - getting **"Lost in Translation,"** where translation steps lead to errors that would have been avoided by question's language reasoning. Code is available at `https://github.com/AI4Bharat/multilingual_reasoning_analysis`.

1 Introduction

Large reasoning models (LRMs) (Guo et al., 2025; Yang et al., 2025; Jaech et al., 2024) have become popular over the past year thanks to their superior ability to analyze and answer questions across a range of tasks. Unlike traditional Large Language Models (LLMs) (Touvron et al., 2023; Team et al., 2024), LRMs have a two-phase response generation. First, they generate a *reasoning sequence*, which resembles a human's step-by-step reasoning process. This phase allows an LRM to do in-depth analysis, explore potential solutions and verify intermediate steps. After completing this reasoning stage, they generate an *answering sequence* that succinctly presents the final answer derived from the reasoning sequence.

LRMs tend to generate their reasoning sequences predominantly in English (Yong et al., 2025). For multilingual tasks, ideally the reasoning should be generated in the question's language: this improves interpretability, lets users follow the model's chain of thought directly, and preserves cultural and linguistic nuance (Aggarwal et al., 2025). Motivated by these considerations, we ask: *how does reasoning in English compare to reasoning in the question's language across multilingual tasks?*

In this study, we evaluate the above question on two benchmark datasets, MGSM (Shi et al., 2022) and GPQA Diamond, which vary in difficulty (Rein et al., 2024) and together provide a comprehensive assessment of LRMs' reasoning capabilities. We analyze results from three perspectives: (1) which reasoning language yields correct final answers more frequently; (2) which reasoning sequences exhibit richer cognitive behaviors; and (3) in which conditions reasoning in the question's language outperforms reasoning in English.

2 Related Work

Recent studies have contrasted reasoning in English with reasoning in non-English languages (Wang et al., 2025; Yong et al., 2025; Tam et al., 2025; Qi et al., 2025). Our study extends this line of inquiry by quantifying how much reasoning in the question's language diverges from reasoning in English across tasks of varying difficulty. In addition, we examine the distribution of cognitive behaviors within the reasoning traces for both settings. Finally, we highlight scenarios where reasoning in the question's language can actually surpass English reasoning, providing insights that complement prior work. See Appendix A for more related work.

[‡] **Correspondence:** Alan Saji (alansaji2001@gmail.com), Ratish Puduppully (rapu@itu.dk)

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 329–344
March 24-29, 2026 ©2026 Association for Computational Linguistics

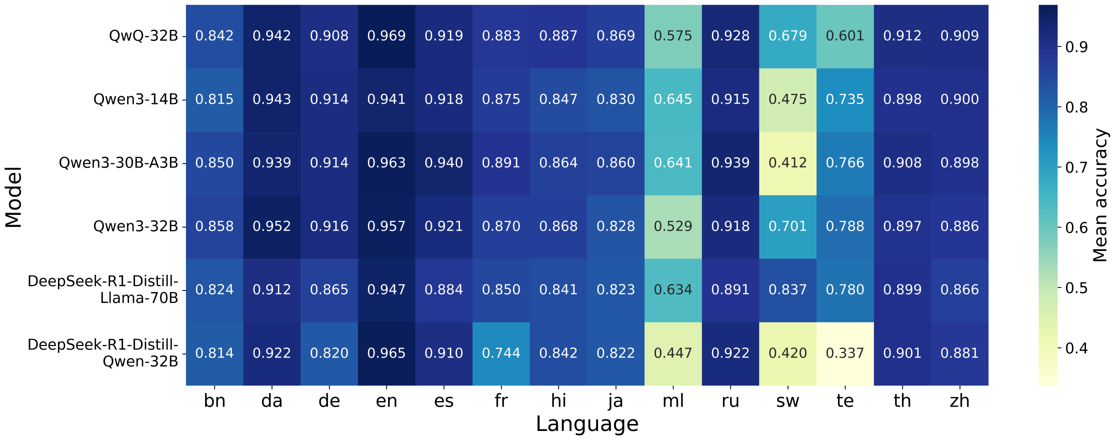


Figure 1: **MGSM - Reasoning in the Question's Language.** Final answer accuracy of LRMs on MGSM when reasoning in question's language. Y-axis: models; X-axis: languages; each cell reports the avg@4 accuracy. Standard deviations are below 0.05 (c.f. Appendix C.2).

3 Methodology

We compare how an LRM reasons in English versus reasoning in the language of the question for multilingual tasks, focusing on differences in final answer accuracy across language. Using a system prompt, the model is explicitly instructed to present its final answer within a `\boxed{}` format, which is then used for answer extraction. To encourage models to reason in the question's language, we employ a system prompt (c.f. Appendix B.6) and prepend prefix tokens to the reasoning trace (c.f. Appendix B.5), an approach shown to be effective in prior work (Yong et al., 2025).

4 Experimental Settings

The study employs MGSM (Shi et al., 2022) and GPQA-Diamond (Rein et al., 2024) benchmark datasets, the latter translated into Danish and five Indic languages(c.f. Appendix B.1). We utilize open-weight reasoning models like Qwen3 32B (Yang et al., 2025) and DeepSeek-R1-Distill-Llama-70B (Guo et al., 2025) (c.f. Appendix B.2).

For LRMs, sampling-based decoding is generally favored over greedy decoding. We adopt the conventional hyperparameter setting of $T = 0.6$, $p = 0.95$, and $k = 20$ (Wang et al., 2025) for LRMs. Each model is evaluated four times using these fixed hyperparameters, and we report both the mean and standard deviation across the four runs. We performed API-based inference using the Deepinfra API (Deepinfra, 2025). Additional details about experimental settings, API cost and system prompts are present in Appendix B. Languages involved in this study are described in Appendix B.3.

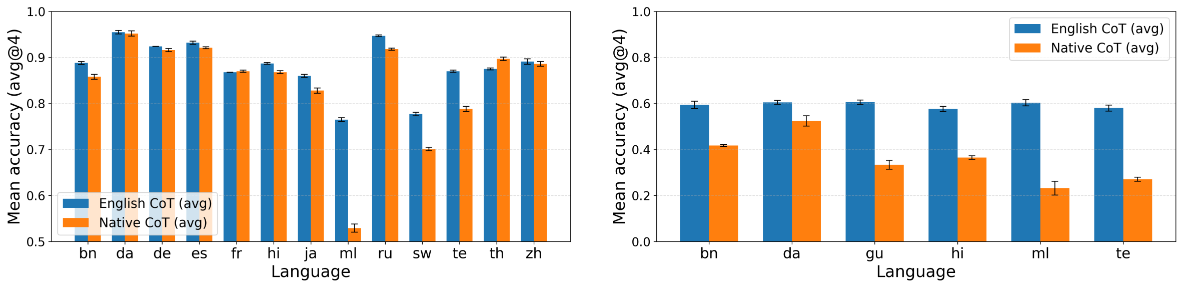
4.1 Grouping Languages by Resource Availability

Among the languages analyzed in this study, English, Chinese, French, German, Japanese, Russian, Danish and Spanish are generally considered high-resource languages, as they have abundant training data, digital content, and linguistic resources available for large language models. In contrast, Hindi, Bengali, Malayalam, Gujarati, Telugu, Swahili, and Thai are typically regarded as low-resource languages, with comparatively limited high-quality text corpora and fewer NLP tools available. Within this group, languages such as Hindi and Bengali can be viewed as relatively better resourced (mid-resource) than others like Malayalam, Gujarati, Swahili, and Telugu, which remain low resource.

5 Results

Figure 1 presents the final answer accuracy of LRMs on the MGSM task when reasoning in the language of the question. We find that performance is highest for English, remains strong for high-resource languages, and declines progressively for lower-resource languages.

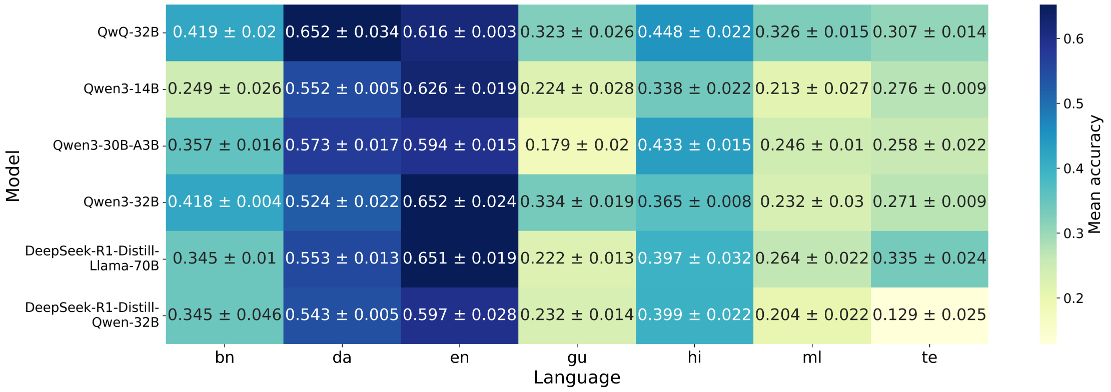
Figure 2a compares reasoning in English with reasoning in the question's language for the MGSM task. As we move from high-resource to low-resource languages, the accuracy gap between English and native-language reasoning widens, indicating that reasoning in English generally yields higher final-answer accuracy.



(a) MGSM

(b) GPQA Diamond

Figure 2: **Reasoning in English vs Question's Language** is compared via final answer accuracy for **Qwen3 32B** for MGSM and GPQA diamond task. Error bars denote standard deviation.



	bn	da	en	gu	hi	ml	te
QwQ-32B	0.419 ± 0.02	0.652 ± 0.034	0.616 ± 0.003	0.323 ± 0.026	0.448 ± 0.022	0.326 ± 0.015	0.307 ± 0.014
Qwen3-14B	0.249 ± 0.026	0.552 ± 0.005	0.626 ± 0.019	0.224 ± 0.028	0.338 ± 0.022	0.213 ± 0.027	0.276 ± 0.009
Qwen3-30B-A3B	0.357 ± 0.016	0.573 ± 0.017	0.594 ± 0.015	0.179 ± 0.02	0.433 ± 0.015	0.246 ± 0.01	0.258 ± 0.022
Qwen3-32B	0.418 ± 0.004	0.524 ± 0.022	0.652 ± 0.024	0.334 ± 0.019	0.365 ± 0.008	0.232 ± 0.03	0.271 ± 0.009
DeepSeek-R1-Distill-Llama-70B	0.345 ± 0.01	0.553 ± 0.013	0.651 ± 0.019	0.222 ± 0.013	0.397 ± 0.032	0.264 ± 0.022	0.335 ± 0.024
DeepSeek-R1-Distill-Qwen-32B	0.345 ± 0.046	0.543 ± 0.005	0.597 ± 0.028	0.232 ± 0.014	0.399 ± 0.022	0.204 ± 0.022	0.129 ± 0.025

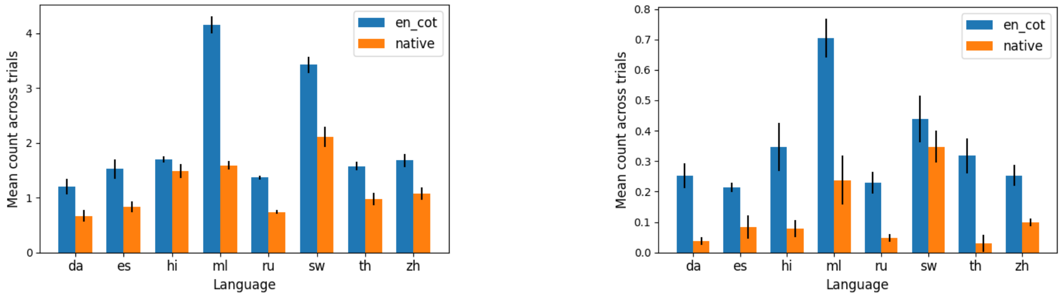
Figure 3: **GPQA diamond - Reasoning in the Question's Language.** Final answer accuracy of LRMs on GPQA diamond when reasoning in question's language. Y-axis: models; X-axis: languages; each cell reports the avg@4 accuracy and standard deviation.

We next examine whether this pattern holds for the GPQA diamond task which requires expert-level domain knowledge. In Figure 3 we observe that the contrast between performance in English and non-English languages is much higher than for the MGSM task during reasoning in the question's language. In addition, in Figure 2b, we could see the contrast between reasoning in English and question's language for multilingual gpqa questions is much larger than the same for MGSM (c.f. Figure 2). Notably, Danish exhibits a smaller final-answer accuracy gap between reasoning in the question's original language and reasoning in English (c.f. Figure 2b). Danish performs relatively well on GPQA-Diamond compared with other non-English languages (c.f. Figure 3). This likely reflects Danish's relatively high resource status among the languages analyzed (c.f. Section. 4.1). These results highlight that leveraging English is particularly important for tapping an LLM's domain knowledge and reasoning ability on expert-level questions. Additional results are shown in appendix C.

6 Analyzing cognitive behaviors in the reasoning trace

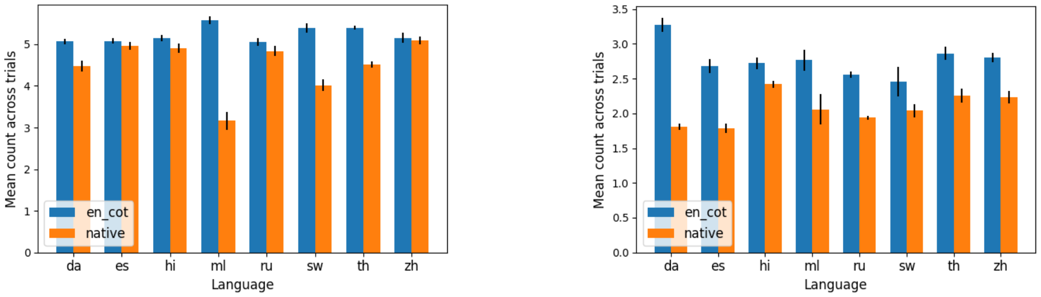
Evaluating multilingual reasoning by final answer accuracy alone overlooks key aspects of an LRM's response generation. While final answer accuracy typically captures the essential behavior of LLMs, LRMs introduce additional complexity through their reasoning traces.

To capture this, following prior studies (Gandhi et al., 2025), we analyze reasoning traces enclosed in <think></think> tokens for cognitive attributes such as *Sub-goal setting* (breaking a task into smaller, manageable steps), *Verification* (performing systematic checks to catch and correct errors), *Back-tracking* (abandoning unsuccessful strategies and selecting alternative approaches) , and *Backward chaining* (reasoning from the desired outcome back toward the necessary initial inputs). These behaviors reflect how expert practitioners tackle difficult problems; for example, a software engineer tests and reviews each change, rewrites code when a design fails, and splits large features into



(a) Cognitive behavior: backtracking

(b) Cognitive behavior: backward chaining



(c) Cognitive behavior: subgoal setting

(d) Cognitive behavior: verification

Figure 4: **Cognitive behaviors in the reasoning chain** are averaged and compared for each language across reasoning in English and in the question's language. Error bars denote standard deviation. Model: Qwen QwQ-32B

smaller, testable modules.

We use gpt-4o-mini as a judge to identify and count these attributes, employing evaluation prompts (c.f. Appendix D) adapted from prior studies (Gandhi et al., 2025). For the MGSM task, we compute the average frequency of each attribute for the Qwen QwQ-32B model across languages, and report means and standard deviations over four trials. Appendix D.1 details the human-LLM agreement study for this task.

Figure 4 illustrates the distribution of cognitive traits within the reasoning traces of an LRM. These traits appear more frequently when the model reasons in English than when it reasons in the question's language, and they coincide with higher final-answer accuracy. Among the attributes, sub-goal setting and verification appear more prominently in question's language reasoning trace, whereas backtracking and backward chaining are comparatively less frequent.

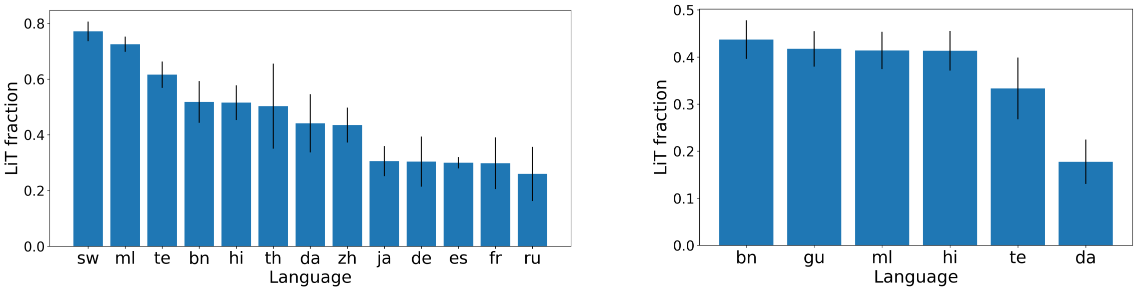
7 Making a case for reasoning in the Question's language

In analyzing reasoning in English versus the question's language, we generally observe stronger performance in English. Yet, a key question remains: *are there cases where native-language reasoning surpasses English?* We find such cases, typically when translation introduces errors into the English reasoning. We term this phenomenon "Lost in Translation," and provide a detailed explanation and example in Appendix E.

To quantify "Lost in Translation," we measure the fraction of incorrect answers from English reasoning caused by translation mistakes; errors that would not arise if reasoning occurred directly in the question's language. These cases are identified using gpt-4o-mini as evaluator. The evaluation setup and prompts are described in Appendix E.1. We average the proportion across four trials with the Qwen QwQ-32B model and report the standard deviation. This averaged proportion is referred to as the LiT (Lost in Translation) fraction. Appendix E.2 details the human-LLM agreement study for the Lost in Translation (LiT) task.

Figure 5a reports the LiT fraction for MGSM, decreasing from 0.77 in low-resource languages to 0.30 in high-resource languages, showing that translation errors are more common in low-resource settings. Similarly, Figure 5b shows LiT fractions for GPQA, ranging from 0.44 to 0.33.

A substantial portion of incorrect answers can be traced to errors introduced during translation, exposing a systemic weakness in reasoning in English. While our prior experiments have shown that reasoning in English allows an LLM to better



(a) MGSM

(b) GPQA

Figure 5: **Lost in Translation.** Visualization of fraction of incorrect answers occurring due to errors in translation when reasoning in English. Error bars represents standard deviation.

leverage its domain-specific knowledge and reasoning capabilities, this experiment reveals a critical vulnerability: **in multilingual tasks, English reasoning risks being "Lost in Translation".**

8 Conclusion

Overall, we find that models generally reason better in English than in the language of the question. This advantage is not limited to final-answer accuracy, but also in the presence of cognitive behaviors in the reasoning chain. However, relying on English introduces a critical weakness: translation of multilingual inputs into English sometimes produces errors that would likely be avoided if models reasoned directly in the question's language. These findings underscore the need to develop and benchmark native-language reasoning capabilities with the same rigor and resources devoted to English.

We address an important question: are LRMs that primarily reason in English by translating non-English inputs sufficient for robust multilingual reasoning? Our results suggest they are not. Achieving reliable native-language reasoning will require targeted efforts in dataset construction, training objectives, and evaluation that preserve language-specific reasoning patterns. This study contributes an empirical diagnosis of the problem and establishes baselines that future work can use to measure progress toward native-language reasoning.

9 Limitations

Our evaluation of multilingual reasoning is confined to mathematical and scientific tasks; whether the patterns we observe extend to open-ended questions or other task types remains an open question. Likewise, our experiments are limited to open-weight LRMs, so it remains to be seen if these findings generalize to closed/proprietary models such as o1 (Jaech et al., 2024) and Gemini 2.5 Pro (Comanici et al., 2025). Future work should evaluate these axes to determine the broader applicability of our results.

10 Ethics Statement

Through this work, our aim is to analyze the performance of LRMs when reasoning in English vs reasoning in questions language for multilingual tasks. We emphasize that it is not our intention to diminish the value or significance of non-English languages included in this study.

The code and datasets created in this work will be made available under permissible licenses. Generative AI systems were only used for assistance purely with the language of the paper, e.g., paraphrasing, spell-check, polishing the author's original content, and for writing boiler-plate code.

11 Acknowledgments

We thank AI4Bharat lab members Deepon Halder, Irina Agastin, and Atharva Gawane for their assistance with annotation on the cognitive behavior analysis and Lost in Translation task. We would also like to thank EkStep Foundation and Nilekani Philanthropies for their generous grant towards research at AI4Bharat.

References

Tushar Aggarwal, Kumar Tanmay, Ayush Agrawal, Kumar Ayush, Hamid Palangi, and Paul Pu Liang. 2025. Language models' factuality depends on the language of inquiry. *arXiv preprint arXiv:2502.17955*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Deepinfra. 2025. Deepinfra api. `https://deepinfra.com/docs/api-reference`. Accessed 2024 - 09-21.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. Blog post. Accessed: 2025-8-06.

Jirui Qi, Shan Chen, Zidi Xiong, Raquel Fernández, Danielle S Bitterman, and Arianna Bisazza. 2025. When models reason in your language: Controlling thinking trace language comes at the cost of accuracy. *arXiv preprint arXiv:2505.22888*.

QwenTeam. 2025. Qwen qwq-32b. `https://qwen.ai/blog?id=6aed6aa257238a0b6c77a6753f180350c2fecc4a&from=research.research-list`. Accessed: 2025-8-06.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

SarvamAI. 2025. Sarvam translate. `https://www.sarvam.ai/blogs/sarvam-translate`. Accessed: 2025-7-06.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Zhi Rui Tam, Cheng-Kuang Wu, Yu Ying Chiu, Chieh-Yen Lin, Yun-Nung Chen, and Hung-yi Lee. 2025. Language matters: How do multilingual input and reasoning paths affect large reasoning models? abs/2505.17407.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, and 1 others. 2025. Polymath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15366–15394. Association for Computational Linguistics.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zheng-Xin Yong, M Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H Bach, and Alham Fikri Aji. 2025. Crosslingual reasoning through test-time scaling. *arXiv preprint arXiv:2505.05408*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

A Related Work - additional details

Prior work has shown that English-centric LLMs carry out intermediate reasoning in a language-agnostic latent space that is nevertheless biased toward English (Wendler et al., 2024; Zhao et al., 2024). Wu et al. (2025) further demonstrates that this English bias extends across modalities.

LRMs amplify this tendency by favoring explicit reasoning in English. In this paper we present a careful, empirical comparison of reasoning performed in English versus reasoning in the question's original language, and we highlight a seldom-discussed shortcoming of English-centric reasoning; namely, that translation-related errors can undermine correctness even when English reasoning appears stronger overall.

B Experimental settings: Additional details

Our evaluation framework is based on lm-evaluation-harness library (Gao et al., 2024). We update the framework to make it suitable for evaluating reasoning models. This includes capturing the final answer in a `\boxed{}` format and extracting the answer from there.

We set the maximum output token limit to 10,000 for MGSM and 30,000 for GPQA diamond dataset. These limits were chosen based on pilot runs. GPQA-Diamond required a larger limit because the higher task difficulty here led to substantially longer reasoning traces; MGSM's responses were much shorter on average.

Because we used sampling-based decoding rather than greedy decoding, we ran the experiment four times and report the mean to ensure reproducibility.

B.1 Dataset

The MGSM dataset (Shi et al., 2022) involves 250 mathematical question-answer pairs at graduate-level for 12 languages. We extend this dataset to Hindi and Malayalam using Sarvam - Translate (SarvamAI, 2025) and to Danish using Comanici et al. (2025). The quality of these translations are manually verified to ensure high quality.

GPQA Diamond: The GPQA Diamond dataset (Rein et al., 2024) contains 198 challenging science questions that require advanced subject expertise. It is significantly harder than the MGSM dataset. In this study, we use an in-house translation of GPQA Diamond into five Indic languages: Hindi, Gujarati, Bengali, Malayalam, and Telugu. Additionally, we translate GPQA Diamond dataset into Danish using Comanici et al. (2025).

B.2 Models

In this study, we consider open weight reasoning models including Qwen QwQ-32B (QwenTeam, 2025), Qwen3-32B (Yang et al., 2025), Qwen3-14B (Yang et al., 2025), Qwen3-30-A3B (Yang et al., 2025), DeepSeek-R1-Distill-Llama-70B (Guo et al., 2025), DeepSeek-R1-Distill-Qwen-32B (Yang et al., 2025).

B.3 Languages Involved

MGSM involves a typologically diverse set of ten languages other than English (en), spanning eight language families. They are Bengali (bn), Chinese (zh), French (fr), German (de), Japanese (ja), Russian (ru), Spanish (es), Swahili (sw), Telugu (te), and Thai (th). We extended MGSM to Malayalam (ml), Hindi (hi) and Danish (da). GPQA diamond is originally in English and is translated to Danish (da) and 5 Indic languages: Hindi (hi), Malayalam (ml), Gujarati (Gu), Bengali (bn) and Telugu (te).

Most of the European and South Asian languages listed belong to the Indo-European family: French and Spanish are Romance languages, German is West Germanic, Danish is North Germanic, Russian is East Slavic, and Bengali, Hindi and Gujarati are Indo-Aryan languages. The remaining languages come from other families: Malayalam and Telugu are Dravidian, Chinese is Sino-Tibetan, Thai is Kra-Dai, Japanese is Japonic, and Swahili is Niger–Congo.

B.4 API Cost Overview

Qwen QwQ-32B (QwenTeam, 2025), Qwen3-32B (Yang et al., 2025), Qwen3-14B (Yang et al., 2025), Qwen3-30-A3B (Yang et al., 2025), DeepSeek-R1-Distill-Llama-70B (Guo et al., 2025), DeepSeek-R1-Distill-Qwen-32B (Yang et al., 2025). was accessed via deepinfra API (Deepinfra, 2025). gpt-4o-mini (OpenAI, 2024) was accessed via openai API. Table 1 provides detailed cost incurred for each model. As inferred the total cost incurred via API usage for this project was 295 dollars.

B.5 Prepending Prefix tokens

The LLM's reasoning is enclosed within <think></think> tokens. Immediately after the <think> token, we insert a prefix token (a

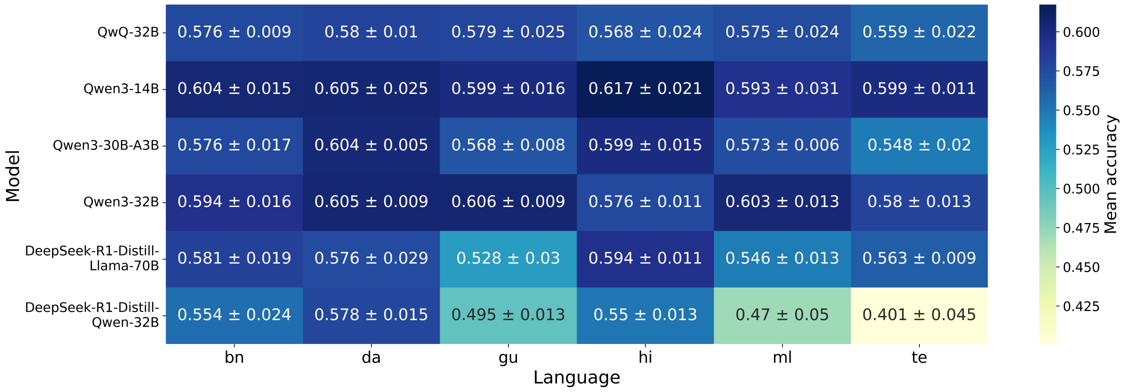


Figure 6: Here we compare performance of reasoning models on GPQA diamond task. Y axis represent LLMs, X axis represents languages and each entry is avg@4 accuracy score. Here the models are explicitly instructed to reason in English via prefix tokens.

Model	**Provider**	**Cost (USD)**
Qwen QwQ-32B	Deepinfra	37
Qwen3-32B	Deepinfra	29
Qwen3-14B	Deepinfra	26
Qwen3-30-A3B	Deepinfra	28
DeepSeek-R1-Distill-Llama-70B	Deepinfra	42
DeepSeek-R1-Distill-Qwen-32B	Deepinfra	29
gpt-4o-mini	OpenAI	104
Grand Total (USD)		**295**

Table 1: API usage and cost for models utilized in this study.

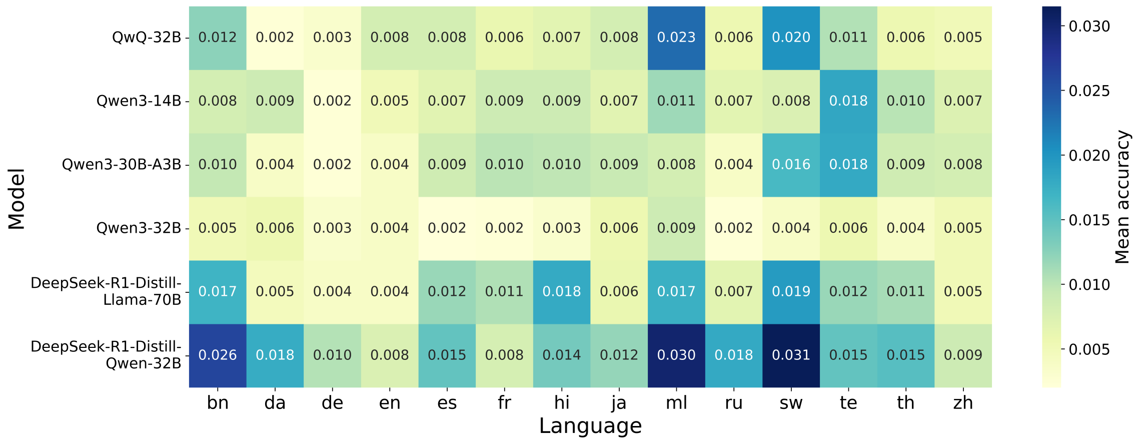


Figure 7: **Standard deviation** of accuracy of reasoning models on MGSM task. Y axis represent LLMs, X axis represents languages. Here the models are explicitly instructed to reason in question's language via prefix tokens.

translation of the phrase "Ok, let me think step by step") to force reasoning in a desired language. Figure 2 illustrates the addition of the prefix token.

B.6 System prompt

Table 3 showcases system prompts used for MGSM and GPQA Diamond tasks.

C Results: Additional details

In Figure 3 we observe a substantial contrast between performance in English and non-English languages when reasoning is conducted in the question's language. However, as we go to Figure 6 where the model reasons in English irrespective of the language of the questions, LLM's performance across languages seems more comparable.

C.1 English vs Question's language reasoning - Additional Models

Comparison of reasoning in English vs question's language for additional models can be found in Figures 8 - 12. These observations align with our findings for the Qwen QwQ-32B model; specifically, the performance gap between reasoning in English and in the question's language for multilingual questions is substantially larger for GPQA Diamond than for MGSM.

C.2 Standard deviation of accuracy for MGSM

In Figure 7 we present the standard deviation of accuracy of reasoning models on MGSM over 4 trials. The models reason in questions language for the multilingual task. Standard deviation remain below **0.05** in all instances.

D Analyzing Cognitive Behaviors: Additional details

Table 5 presents the prompt templates used to count the occurrence of cognitive behaviors in the reasoning trace when an LRM reasons in questions language. We replace "questions language" with "English" in the prompt when we analyze reasoning trace in English. Using gpt-4o-mini, we count the occurrences of these attributes by enclosing them within <count> and </count> tags, then compute the average across samples and subsequently across four trials.

D.1 Human - LLM agreement study

To validate the LLM-as-a-judge evaluator, we conducted a human annotation study for the cognitive behavior analysis task. We randomly sampled 30 evaluation prompts for each of the two low-resource languages - Hindi and Malayalam, yielding 60 examples in total. Each example was independently solved and labeled by native-speaker annotators with strong grade-school mathematics skills. All annotators were computer science undergraduate students and each annotated 30 questions in their respective native language. Annotators were instructed to count and report the occurrence of each of the four cognitive behaviors in the reasoning trace using the same instructions given to the LLM evaluator. We compared the LLM judgments to the human labels and computed percent agreement and Cohen's Kappa which are detailed in Table 4. The results indicate that the LLM evaluator delivers reasonably reliable assessments of cognitive behaviors, even for low-resource languages like Malayalam.

<table>
<tr><th>Scenario</th><th>Token Sequence</th></tr>
<tr><td>Without Language Forcing</td><td><|im_start|>assistant\n<think>\n</td></tr>
<tr><td>With Language Forcing</td><td><|im_start|>assistant\n<think>\nD'accord, laissez-moi essayer de résoudre ce problème étape par étape.</td></tr>
</table>

Table 2: Illustration of language forcing to control the language of thought. By inserting a specific language prefix (e.g., French) after the `<think>` token, the model is prompted to generate its internal reasoning in that language.

Task	**Reasoning Language**	**System Prompt**
MGSM	question's language	"You are a helpful assistant. You should reason and analyze the question in the language of the question and wrap your thought process in `<think>...</think>` tags. Then provide the answer in the language of the question, and keep the final answer in \boxed{}."
	English	"You are a helpful assistant. You should reason and analyze the question in English and wrap your thought process in `<think>...</think>` tags. Then provide the answer in the language of the question, and keep the final answer in \boxed{}."
GPQA Diamond	question's language	"You are a helpful assistant. You should reason and analyze the question in the language of the question and wrap your thought process in `<think>...</think>` tags. Outside the `<think>...</think>` tags summarize the answer in the language of the question. Finally, select the correct option out of the given options A,B,C or D and keep this option in \boxed{}."
	English	"You are a helpful assistant. You should reason and analyze the question in English and wrap your thought process in `<think>...</think>` tags. Outside the `<think>...</think>` tags summarize the answer in the language of the question. Finally, select the correct option out of the given options A,B,C or D and keep this option in \boxed{}."

Table 3: System Prompt Variations for MGSM and GPQA Benchmarks

E Lost in Translation

In our experiments we investigated scenarios where reasoning in question's language outperformed reasoning in English. Across languages and across models, we observed that reasoning in question's language outperformed reasoning in English in scenarios where reasoning in English was incorrect due to errors introduced in translation. We refer to this as "Lost in Translation".

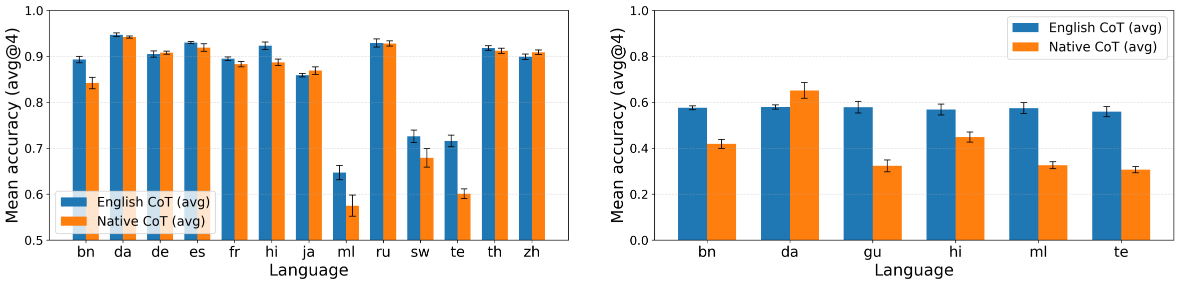
Figure 13 illustrates an example of Lost in Translation. Here, the English reasoning fails to arrive at the correct final answer due to a mistranslation. Specifically, the English reasoning interprets a key detail as “he sent two letters in total” instead of the correct meaning, “he sent two letters to each of them.” In contrast, the reasoning in the original question's language does not encounter this issue, as no translation is involved, and therefore arrives at the correct answer.

E.1 Lost in Translation - Quantitative Analysis

We use gpt-4o-mini as an evaluator LLM, employing the prompt template shown in Table 7 to determine whether an incorrect answer from reasoning in English occurred due to a mistranslation or not. From gpt-4o-mini's output we extract a "Yes" or "No" enclosed in <classification></classification> tags, compute the fraction of incorrect answers attributable to mistranslation for each run, and report the mean value over four trials.

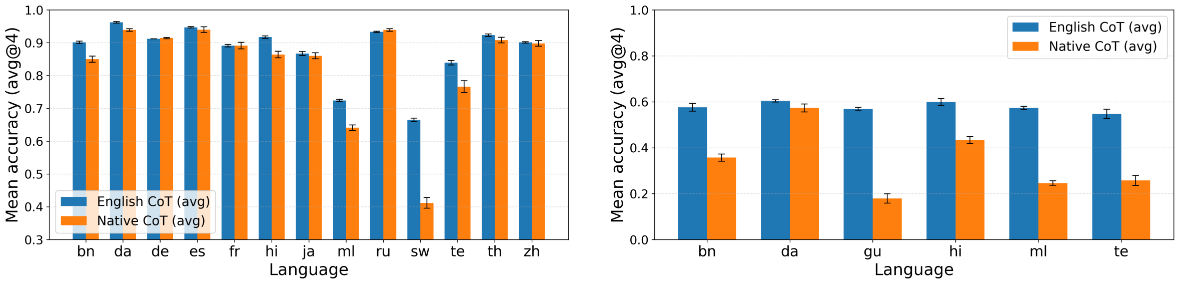
E.2 Lost in Translation: Human-LLM agreement study

To validate the LLM-as-a-judge evaluator, we conducted a human annotation study for "Lost in Translation" (LiT) task. We randomly sampled 50 LiT evaluation prompts from MGSM for each of three low-resource languages - Hindi, Bengali and Malayalam, yielding 150 examples in total. Each example was independently solved and labeled by native-speaker annotators with strong grade-school mathematics skills. All annotators were computer science undergraduate students and each annotated 50 questions in their respective native language. Annotators were instructed to classify a reasoning trace as “Lost in Translation” or not based on the same instructions provided to the LLM judge. We compared the LLM judgments to the human labels and computed percent agreement, precision, recall, and Cohen’s Kappa which are detailed in Table 6. These results indicate that the LLM evaluator produces reasonably reliable evaluations for the LiT task even for low resource languages like Malayalam.



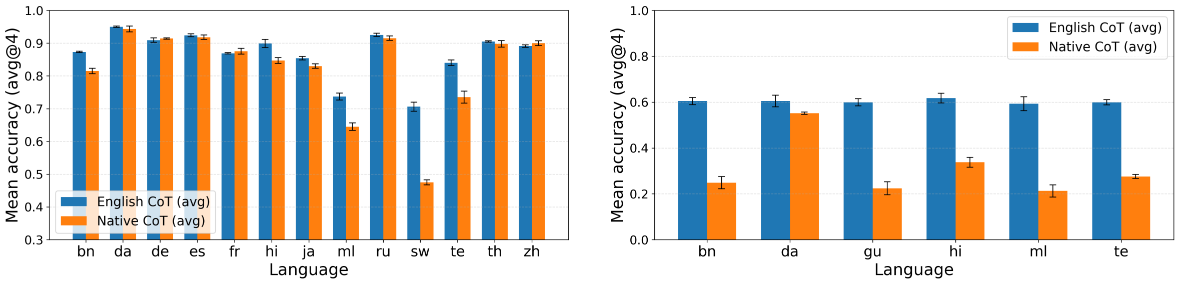
(a) MGSM (b) GPQA Diamond

Figure 8: **Reasoning in English vs Question's Language** is compared via final answer accuracy for Qwen QwQ 32B for MGSM and GPQA diamond task. Error bars denote standard deviation.



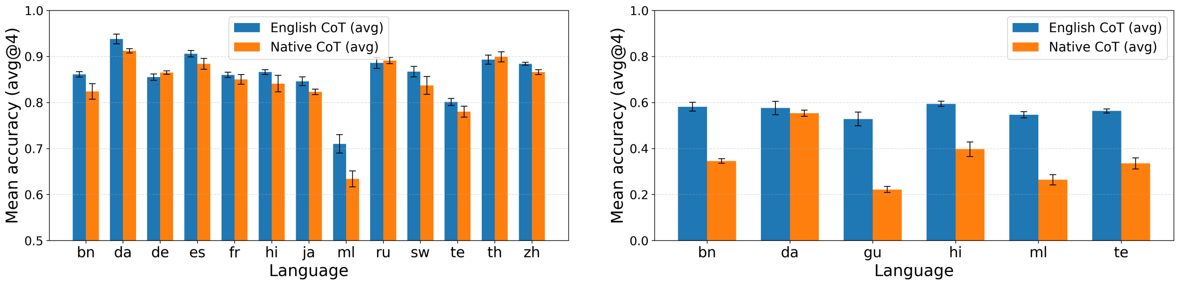
(a) MGSM (b) GPQA Diamond

Figure 9: **Reasoning in English vs Question's Language** is compared via final answer accuracy for **Qwen3 30B-A3B** for MGSM and GPQA diamond task. Error bars denote standard deviation.



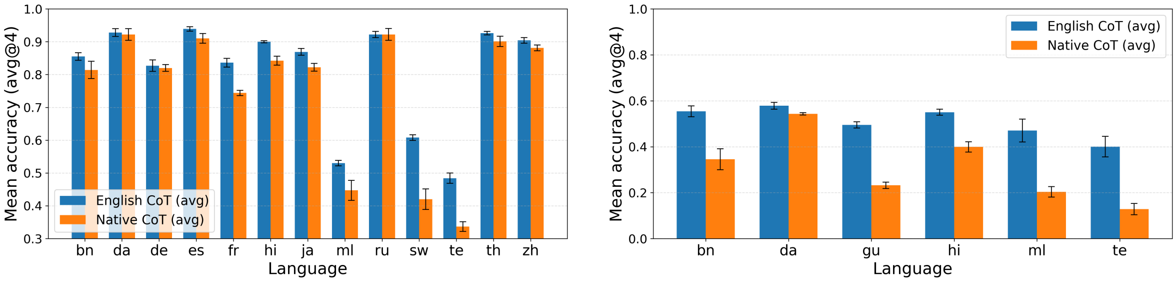
(a) MGSM (b) GPQA Diamond

Figure 10: **Reasoning in English vs Question's Language** is compared via final answer accuracy for **Qwen3 14B** for MGSM and GPQA diamond task. Error bars denote standard deviation.



(a) MGSM (b) GPQA Diamond

Figure 11: **Reasoning in English vs Question's Language** is compared via final answer accuracy for **DeepSeek-R1-Distill-Llama-70B** for MGSM and GPQA diamond task. Error bars denote standard deviation.



(a) MGSM

(b) GPQA Diamond

Figure 12: **Reasoning in English vs Question's Language** is compared via final answer accuracy for **DeepSeek-R1-Distill-Qwen-32B** for MGSM and GPQA diamond task. Error bars denote standard deviation.

Cognitive Behavior	**Hindi (30 samples)**		**Malayalam (30 samples)**	
	% Agreement	**Kappa**	**% Agreement**	**Kappa**
Subgoal	0.73	0.68	0.67	0.62
Verification	0.80	0.69	0.70	0.58
Backtracking	0.80	0.58	0.733	0.54
Backward Chaining	0.83	0.52	0.87	0.52

Table 4: Human-LLM agreement for cognitive analysis task. Percent agreement and Cohen's Kappa are reported for Hindi and Malayalam.

Cognitive Attribute	**Prompt**
Verification	Here is a chain-of-reasoning in the question's language that a Language Model generated for the question: "*{question}*" The model's reasoning (between <think>…</think>) is: *{chain_of_reasoning}* Evaluate whether this chain-of-reasoning contains any explicit answer-verification steps. An answer-verification step is any instance where the model checks its intermediate numeric result and asks itself if the answer is correct or not and probably goes on to re-check it. If you find any of these instances, count them and put the number between <count>and </count>. Otherwise, output <count>0</count>.
Backtracking	Here is a chain-of-reasoning in the question's language '*{L}*' that a Language Model generated for the question: "*{question}*" The model's reasoning (between <think>…</think>) is: *{chain_of_reasoning}* Evaluate whether this reasoning contains any backtracking behavior—i.e., places where the model decides that its previous approach won't reach the correct answer and explicitly abandons that path, starting fresh on an alternative intermediate step. Count the number of such backtracking instances and put the result between <count>and </count>. If none, output <count>0</count>.
Subgoal setting	Here is a chain-of-reasoning in the question's language '*{L}*' that a Language Model generated for the question: "*{question}*" The model's reasoning (between <think>…</think>) is: *{chain_of_reasoning}* Evaluate whether this reasoning explicitly sets any subgoals (e.g., "First I will try to isolate x…", "Next I aim to simplify …", etc.) on the way toward *{target}*. Count how many distinct subgoals appear and put that number between <count>and </count>. If none, output <count>0</count>.
Backward chaining	Here is a chain-of-reasoning in the question's language '*{L}*' that a Language Model generated for the question: "*{question}*" The model's reasoning (between <think>…</think>) is: *{chain_of_reasoning}* Evaluate whether this reasoning uses backward-chaining—i.e., it starts from the final answer and works backward to earlier steps. Count how many distinct backward-chaining instances occur and put that number between <count>and </count>. If none, output <count>0</count>.

Table 5: Cognitive Attributes Analysis Prompt Templates

Language (Samples)	Percent agreement	Precision	Recall	Cohen’s kappa
Hindi (50)	0.78	0.83	0.80	0.55
Malayalam (50)	0.76	0.79	0.79	0.50
Bengali (50)	0.78	0.72	0.82	0.56
Total (150)	**0.77**	**0.78**	**0.80**	**0.54**

Table 6: Lost in Translation: Human–LLM agreement across 150 samples (50 per language).

Question: माइक 5 लोगों के साथ एक पेन पाल था। वह उनमें से 2 के साथ पेन पाल्स नहीं रहा। वे प्रत्येक सप्ताह में 2 पत्र भेजते हैं जो 5 पृष्ठ लंबे होते हैं। वह बदले में ऐसा ही करता है। वह 6 मिनट में एक पृष्ठ लिख सकता है। वह एक सप्ताह में कितने घंटे लिखने में बिताता है?

English translation: Mike was a pen pal with 5 people. He is no longer pen pals with 2 of them. They each send 2 letters a week that are 5 pages long. He does the same in return. He can write a page in 6 minutes. How many hours a week does he spend writing?

Reasoning in English

Okay, let me try to figure out this problem step by step. So, Mike has a pen pal arrangement with 5 people. But he's not pen pals with 2 of them. Hmm, wait, does that mean he's only writing to 3 people? Because 5 minus 2 equals 3. Yeah, that makes sense. So he's actively corresponding with 3 pen pals. Now, each week, he sends 2 letters, each 5 pages long. And he does the same in return. Wait, the problem says ""वह बदले में ऐसा ही करता है"" which translates to ""he does the same in return."" So maybe that means he receives letters too, and has to write responses? Or does it mean he sends the same amount? Hmm, this part is a bit ambiguous.

Final answer: 2 hours ✗

Reasoning in Hindi

"ठीक तो, मैंने प्रश्न पढ़ा। माइक के पास 5 लोगों के साथ पेन पाल है, लेकिन उनमें से 2 के साथ वह पेन पाल्स नहीं रहा। तो पहले ही स्पष्ट होता है कि वह वास्तव में 5 - 2 = 3 लोगों के साथ पत्र लिख रहा है। अब, प्रत्येक सप्ताह में वह प्रत्येक से 2 पत्र भेजता है। तो 3 लोगों के लिए, कुल पत्रों की संख्या 3 * 2 = 6 पत्र प्रति सप्ताह। पत्रों के प्रत्येक पत्र 5 पृष्ठों का है, तो प्रत्येक पत्र के लिए 5 पृष्ठ। इसलिए कुल पृष्ठों की संख्या 6 * 5 = 30 पृष्ठ प्रति सप्ताह। फिर, माइक प्रत्येक पृष्ठ 6 मिनट में लिखता है। तो 30 पृष्ठों के लिए कुल समय 30 * 6 = 180 मिनट। अब, मिनट को घंटों में बदलना है, तो 180 मिनट ÷ 60 = 3 घंटे।

Translation: “Okay, so I read the question. Mike has pen pals with 5 people, but with 2 of them, he is no longer pen pals. So it’s already clear that he’s actually writing to 5 - 2 = 3 people. Now, each week he sends 2 letters to each of them. So for 3 people, the total number of letters is 3 * 2 = 6 letters per week. Each letter is 5 pages long, so 5 pages per letter. Therefore, the total number of pages is 6 * 5 = 30 pages per week. Then, Mike writes each page in 6 minutes. So for 30 pages, the total time is 30 * 6 = 180 minutes. Now, to convert minutes into hours: 180 minutes ÷ 60 = 3 hours.

Final answer: 3 hours ✓

Analysis: When reasoning in English the model mistranslate "वे प्रत्येक सप्ताह में 2 पत्र भेजते हैं" as "Now, each week, he sends 2 letters" whereas the actual translation should be "Now, each week he sends 2 letters to each of them". We refer to this kind of mistake as "Lost in Translation".

Figure 13: **Lost in Translation.** Here, we compare the reasoning traces in English and Hindi generated by Qwen QwQ-32B for a question originally written in Hindi. We observe that the English reasoning contains an error caused by a mistranslation (highlighted in red), while the corresponding correct sentence in the Hindi reasoning is highlighted in green. For the reader’s clarity, English translations of the Hindi question and reasoning are also provided.

Task	**Prompt Template**

MGSM/ GPQA Diamond	Here is a chain-of-reasoning that a Language Model generated for the question:
	`"{question}"`
	The model's reasoning (between `<think>...</think>`) is:
	`{chain_of_reasoning}`
	The chain-of-reasoning here has led to incorrect answer. Here the model reasons in English, check if here some information available in the question has been omitted/ mis-translated while translating the question to English. If so output 'Yes' else 'No". Put the Output in between `<classification>` and `</classification>` tags. Do not generate anything else.

Table 7: Lost in Translation - Prompt templates