

EfficientXLang: Towards Improving Token Efficiency Through Cross-Lingual Reasoning

Sanchit Ahuja^{1*}, Praneetha Vaddamanu^{2*}, Barun Patra²

¹Northeastern University ²Microsoft Corporation

{ahuja.sanc@northeastern.edu, pvaddamanu, bapatra@microsoft.com}

Abstract

Despite recent advances in Reasoning Language Models (RLMs), most research focuses solely on English, even though many models are pretrained on multilingual data. In this work, we investigate: *Is English the most token-efficient language for reasoning?* We evaluate three open-source RLMs: DEEPSEEK R1, QWEN 2.5, and QWEN 3, across four math datasets and seven typologically diverse languages. We find that reasoning in non-English languages not only reduces token usage, but also preserves accuracy. These gains persist even after translating the reasoning traces into English, suggesting genuine shifts in reasoning behavior rather than surface-level linguistic effects. The extent of improvement, however, depends on the model’s multilingual strength. Our findings motivate a broader view of reasoning in language models, highlighting the potential of multilingual reasoning and the importance of strong multilingual foundations. The code for our work can be found [here](#).

1 Introduction

Recent advancements have shown that increasing inference-time compute can significantly enhance the downstream performance of large language models. In particular, RLMs benefit from extended test-time computation through long chains-of-thought, resulting in improved reasoning capabilities and task performance (DeepSeek-AI, 2025; OpenAI et al., 2024b). Despite this progress, the majority of prior work has focused on reasoning in English, implicitly assuming it as the default or optimal language for thought (Wendler et al., 2024).

A key limitation of current RLMs lies in the number of “thinking tokens” they consume during inference. Since test-time compute scales with the number of tokens generated, longer reasoning traces di-

rectly impact latency and cost. This inefficiency is often exacerbated by the phenomenon of overthinking (Cuadron et al., 2025b), where models produce unnecessarily verbose intermediate reasoning. To address this, several works have explored length-aware rewards in reinforcement learning setups to encourage brevity in reasoning traces (Aggarwal and Welleck, 2025; Sui et al., 2025b). However, these approaches typically require model retraining, which is computationally expensive and infeasible for many practitioners.

In this work, we take a different perspective on language and reasoning by asking a simple yet underexplored question: *Is English the most efficient language for reasoning?* We explore whether reasoning in other languages can produce shorter but equally effective reasoning traces, potentially reducing inference-time compute. This question is timely for two reasons: (1) many reasoning language models (RLMs) are multilingual by design (Team, 2025a); and (2) prior work suggests that languages vary in how clearly they express mathematical ideas, affecting reasoning performance (Han, 2001). Our key contributions are:

- We show that across a range of mathematical reasoning datasets, reasoning in alternative languages consistently results in fewer tokens, often matching the final English accuracy.
- We demonstrate that even when translated back to English, reasoning traces originally generated in other languages remain shorter than those natively generated in English. This suggests that the reduction in token count is not merely a function of tokenizer fertility or linguistic compactness, but reflects genuine differences in the characteristics of reasoning traces compared to their English counterparts.
- We provide quantitative evidence that English is not always the most effective language for reasoning, with traces in other languages consistently preferred for their conciseness, clarity, and struc-

Equal Contribution.

Work done while at Microsoft Corporation.

ture.

2 Experiments

To assess whether reasoning in other languages can enhance thinking token efficiency, we experiment with seven typologically diverse languages: Chinese, Russian, Spanish, Hindi, Arabic, Korean, and Turkish, along with English. We prompt the reasoning models in English, instructing them to carry out their reasoning in the target language while providing the final answer in English. This setup allows us to isolate the effect of language on the reasoning process itself. For inferencing, we used the SGLang framework (Zheng et al., 2024), which facilitates the extraction of intermediate reasoning traces.¹ We now detail our experimental setup.

Datasets: We evaluate our hypothesis in a variety of math datasets, which we categorize into *easy* and *difficult* subsets based on the complexity of the problem and the typical performance of existing models. The *difficult* datasets include AIME2024 (Veeraboina, 2023) and AIME2025 (Art of Problem Solving, 2025a,b), which feature competition-level problems that require advanced reasoning. The *easy* datasets include AMC23 (Art of Problem Solving, 2023a,b), and MATH500 (Lightman et al., 2023) comprising of simpler problems.

Models: We explore three open-weight reasoning models for our experiments: DEEPSEEK R1 (DeepSeek-AI, 2025), QwQ (based on the 32B variant of QWEN 2.5) (Yang et al., 2024; Team, 2025b), and QWEN3 (Qwen3-235B-A22B), the latest model focused on reasoning of the Qwen family (Team, 2025a).

Evaluation: Following the evaluation setup of DEEPSEEK R1, we set the maximum generation length to 32,768 tokens. For response generation, we use a sampling temperature of 0.6 and a top- p value of 0.95, generating $n = 16$ responses across all models and datasets.

Metrics: Different RLMS vary in their ability to follow prompt instructions for generating reasoning traces in the indicated target language. To better evaluate multilingual reasoning performance, we introduce two complimentary metrics: Target Language Consistency (TLC), and Target Language Pass @ k (TLP@ k) in conjunction with the commonly used Pass@ k metric. We use $k = 4$ in all our results.

¹The exact prompt we use can be found in Appendix A.

Target Language Consistency (TLC): measures how well models adhere to the target language constraint when prompted. Given a dataset of size m , with question-answer pairs (q_i, y_i) , and target language j , let $\mathcal{R}^{(k)}(i, j) = \{r_1, \dots, r_k\}$ be the set of generated reasoning traces under the prompt to reason in language j . TLC for a sample is the fraction of traces in the correct language, and the overall TLC is the average across all samples. We use a LID model (Joulin et al., 2016) to detect the language of a reasoning trace.

Target Language Pass @ k (TLP@ k): Analogous to the Pass@ k metric, we propose TLP@ k to measure the fraction of samples where at least one of the top- k reasoning traces both elicits the correct answer and is in the target language. Extending Chen et al. (2021), for measuring TLP@ k , we use a low variance estimator $\frac{1}{m} \sum_{i=1}^m \left[1 - \frac{\binom{n-c}{k}}{\binom{n}{k}} \right]$, where c is the number of reasoning traces that yield the correct answer and are in the correct target language among the n generated traces. The final TLP@ k is the average across all samples. Further details about the implementation of this estimator are provided in the Appendix A.2.

3 Analysis

This section explores a series of research questions aimed at validating our thesis.

RQ1: Is English the optimal language for reasoning for LRMs ? Table 1 shows per-language model performance relative to English (absolute scores in Table 8). Across models, high TLC correlates with smaller performance drops when reasoning in non-English languages—despite using significantly fewer tokens. In contrast, models with weaker multilingual foundations (e.g., QWEN 2.5) show larger performance gaps.

To analyze task and language effects in more detail, we examine DEEPSEEK R1 (due to its strong TLC) across datasets (Tables 2, 4, Fig. 1, 2). Detailed results for QWEN 2.5 and QWEN 3 are provided in Appendix A.3. We observe that

- For *easy* tasks (e.g., MATH500), language choice has little effect on accuracy, while non-English reasoning uses fewer tokens.
- As task difficulty \uparrow , we see a relative drop in performance of up to 12.52% on average in AIME25 compared to English.
- For harder tasks, high-resource languages (Zh, Ru) outperform low-resource ones (Hi).

Model	en			zh			ru			es		
	TLP@4	TLC	Tok.	$\Delta\%$ TLP@4	TLC	$\Delta\%$ Tok	$\Delta\%$ TLP@4	TLC	$\Delta\%$ Tok	$\Delta\%$ TLP@4	TLC	$\Delta\%$ Tok
DEEPSEEK R1	0.87	1.000	3816	4.6%	1.000	21.1%	4.3%	1.000	14.1%	-0.9%	0.985	29.9%
QWEN 2.5	0.87	1.000	6249	3.8%	0.796	21.4%	23.3%	0.340	38.2%	82.0%	0.056	88.1%
QWEN 3	0.89	0.998	6123	0.9%	0.987	27.7%	-0.3%	0.996	35.8%	1.7%	0.583	73.0%

(a) Performance relative to English (TLP@4 and Tok. as % change).

Model	ar			tr			hi			ko		
	$\Delta\%$ TLP@4	TLC	$\Delta\%$ Tok	$\Delta\%$ TLP@4	TLC	$\Delta\%$ Tok	$\Delta\%$ TLP@4	TLC	$\Delta\%$ Tok	$\Delta\%$ TLP@4	TLC	$\Delta\%$ Tok
DEEPSEEK R1	7.4%	0.901	28.2%	8.3%	1.000	8.3%	8.0%	0.971	7.7%	9.5%	0.997	43.9%
QWEN 2.5	57.2%	0.177	68.1%	29.3%	0.444	52.3%	58.3%	0.097	41.3%	81.4%	0.059	90.1%
QWEN 3	15.7%	0.666	47.4%	18.5%	0.721	59.6%	16.3%	0.887	23.7%	17.1%	0.574	47.4%

(b) Performance relative to English (TLP@4 and Tok. as % change).

Table 1: Performance relative to English: TLP@4 and mean token count are reported as percentage change from English; Target Language Consistency (TLC) is reported directly.

Dataset	en		ko		es		zh		hi		ru		tr		ar	
	TLP@4	Tokens	$\Delta\%$ TLP@4	$\Delta\%$ Tok	$\Delta\%$ TLP@4	$\Delta\%$ Tok	$\Delta\%$ TLP@4	$\Delta\%$ Tok	$\Delta\%$ TLP@4	$\Delta\%$ Tok	$\Delta\%$ TLP@4	$\Delta\%$ Tok	$\Delta\%$ TLP@4	$\Delta\%$ Tok	$\Delta\%$ TLP@4	$\Delta\%$ Tok
MATH500	0.77	1404.24	0.00	37.80	0.00	32.85	0.00	18.41	0.00	17.81	0.00	36.96	0.00	5.78	0.00	26.98
AMC23	1.00	3571.1	-3.00	42.09	0.00	25.20	-1.00	14.44	-2.00	-0.83	-3.00	27.22	-4.00	-12.29	-4.00	25.87
AIME2024	0.89	5637.27	-12.36	42.40	-4.49	21.95	-11.24	27.07	-12.36	-13.58	-8.99	0.57	-11.24	-11.08	-11.24	17.43
AIME2025	-0.81	7979.17	-24.69	47.78	-2.47	36.54	-4.94	20.43	-18.52	25.79	-4.94	13.29	-18.52	32.13	-13.58	37.24

Table 2: Both TLP@4 and mean token count are reported as percentage changes from English. Color intensity ranges from orange to green/yellow, where brighter shades (green/yellow) indicate greater deviation from English performance, while duller shades like orange represent minimal or no deviation.

- Multilingual reasoning consistently achieves significant token savings across all settings.

In conclusion, while English remains the top-performing language for reasoning in most RLMs, it isn't always optimal - especially in terms of token efficiency. Models with strong multilingual capabilities can reason effectively in other languages with minimal accuracy loss and significant token savings. For high-resource languages, especially, multilingual reasoning offers a competitive and efficient alternative to English.

RQ2: Do multilingual reasoning traces differ meaningfully from their English counterparts?

To answer this, we examine whether token efficiency gains arise from linguistic conciseness or deeper semantic differences in reasoning. We use GPT-4o (temperature = 0.6, top-p = 0.95) to translate non-English traces into English and compare their token counts to the original English versions. As shown in Figure 1, the translated traces remain significantly shorter, indicating that non-English prompts can elicit more concise reasoning paths. This suggests that multilingual prompting can uncover efficient strategies not naturally produced in English.

In conclusion, multilingual reasoning traces are meaningfully different and often more concise.

Their brevity persists after translation, suggesting more efficient reasoning strategies—not just linguistic differences.

RQ3: Are Multilingual Reasoning Traces Better Than English?

To further assess whether non-English reasoning traces are *better overall* than their English counterparts, we conduct an automated pairwise comparison using GPT-4o as the judge. For each language, we select cases where both English and non-English reasoning are correct. We then compare the English trace to the GPT-4o-translated non-English trace, leveraging GPT-4o's high translation quality (OpenAI et al., 2024a) to control for linguistic brevity. Evaluations use GPT-4o with greedy decoding and a prompt (Appendix A.1) focused on clarity, logic, and efficiency. To reduce position bias, we assess each pair in both orders, (A, B) and (B, A) and retain only consistent preferences. We leverage a one-sided binomial test to establish if reasoning in the native language is (statistically) significantly better than reasoning in English. Table 3 reports results for DEEPSEEK R1. Across benchmarks, most languages show significant gains when reasoning in the native language, with Arabic, Korean, and Spanish exceeding 90% win rates. QWEN 3 shows similar gains across all non-English languages (Ta-

Token Counts Across All Datasets (Problems Correct in Both Languages)

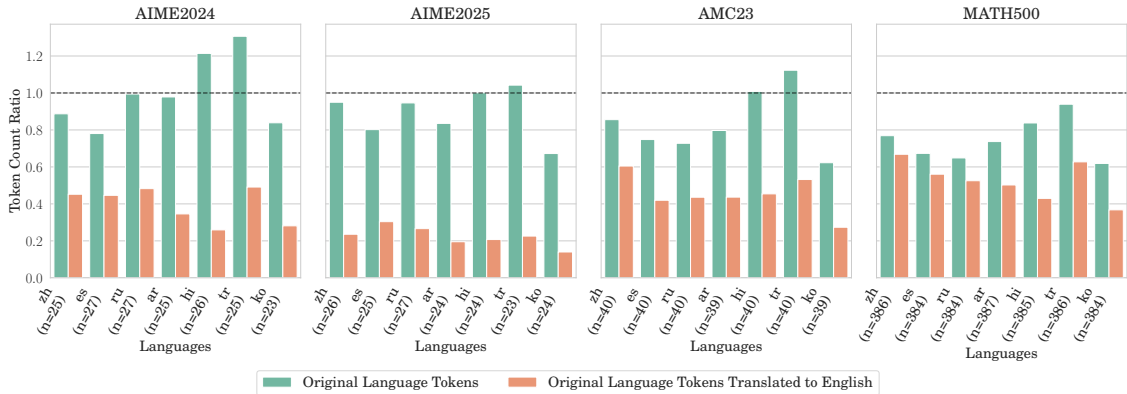


Figure 1: Token Count Ratio vs. number of problems with at least one correct answer in both English and the target language (min. 5 shared cases), computed relative to the average English token count per problem for DEEPSEEK R1.

	AIME2024			AIME2025			AMC23			MATH500		
Lang	N	Win%	Sig?	N	Win%	Sig?	N	Win%	Sig?	N	Win%	Sig?
ar	5	100.0	✓	16	93.8	✓	21	100.0	✓	250	92.8	✓
zh	5	100.0	✓	17	94.1	✓	12	58.3	✗	192	55.7	✗
hi	11	100.0	✓	17	88.2	✓	16	93.8	✓	263	95.4	✓
ko	12	100.0	✓	21	100.0	✓	27	96.3	✓	273	90.5	✓
ru	6	83.3	✗	17	82.4	✓	21	90.5	✓	243	93.8	✓
es	6	100.0	✓	12	91.7	✓	22	90.9	✓	241	97.1	✓
tr	8	75.0	✗	16	93.8	✓	20	85.0	✓	199	85.9	✓

Table 3: Binomial test results comparing non-English vs English reasoning responses from DEEPSEEK R1. ✓ indicates $p < 0.05$, ✗ indicates no significant difference. N denotes the number of samples wherein both the answers in English and the target language are correct.

ble 9, Appendix A.4). We also observe qualitative difference between English outputs and language translated outputs (Ref Appendix A.5 for a QWEN 3 example). We omit QWEN 2.5 due to its low Target Language Consistency, which makes such comparisons unreliable (Table 1).

These findings suggest that multilingual reasoning can enhance the quality of a model’s thought process. Since all responses are translated to English for evaluation, structural differences are controlled for, indicating that gains likely arise from improved clarity, structure, and conciseness.

4 Related Work

Efficient Reasoning: Recent interest in LRMs has spurred investigations into reasoning efficiency as to why some tasks demand more reasoning tokens and whether models can answer correctly without explicit reasoning steps (Cuadron et al., 2025a;

Hashemi et al., 2025; Sui et al., 2025a; Chen et al., 2025; Ma et al., 2025; Marjanović et al., 2025). This work explores reasoning efficiency from a multilingual perspective, analyzing whether certain reasoning paths across languages are both accurate and token-efficient.

Multilingual Reasoning: Prior studies showed that prompting models in the target language often underperforms compared to reasoning in English (Qin et al., 2023; Shi et al., 2022; Ko et al., 2025). However, RLMs like Deepseek R1 and Qwen have shown stronger multilingual reasoning without explicit tuning (DeepSeek-AI, 2025; Team, 2025a; Yang et al., 2024). We build on this by introducing tighter evaluation bounds and fine-grained trace-level analysis, uncovering language-specific reasoning paths that are both accurate and token-efficient.

Concurrently, several recent works have also begun

to explore multilingual reasoning in depth (Gao et al., 2025; Bajpai and Chakraborty, 2025).

5 Conclusion and Future Work

In this work, we explore two central questions: (1) Is English the optimal language for reasoning? and (2) Are multilingual reasoning traces meaningfully different from their English counterparts?

We find that multilingual prompting can reduce token usage by 20–40% without compromising accuracy, presenting a simple yet effective strategy for improving inference efficiency without retraining. These efficiency gains persist even after translating the outputs back to English, suggesting that the improvements stem from genuine differences in reasoning behavior rather than mere linguistic compactness.

Importantly, the effectiveness of this approach is tied to the model’s underlying multilingual strength. Our findings highlight the need to improve base multilingual alignment to better support efficient and high-quality reasoning across languages. Future work should explore training and preference optimization across more diverse languages to unlock these benefits more broadly.

Limitations

In this work, we focus exclusively on reasoning traces generated through English prompts. The behavior of models may vary when prompted in target languages or when guided using techniques like teacher-forcing with a few target-language tokens. Our analysis is limited to seven languages and relies solely on open-source reasoning models. Proprietary models such as OpenAI’s o1 (OpenAI et al., 2024b) may exhibit different behavior from those studied here.

References

Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*.

Art of Problem Solving. 2023a. 2023 amc 12a. https://artofproblemsolving.com/wiki/index.php/2023_AMC_12A. Accessed: 2025-05-04.

Art of Problem Solving. 2023b. 2023 amc 12b. Accessed: 2025-05-04.

Art of Problem Solving. 2025a. 2025 aime i. Accessed: 2025-05-04.

Art of Problem Solving. 2025b. 2025 aime ii. Accessed: 2025-05-04.

Prasoon Bajpai and Tanmoy Chakraborty. 2025. [Multilingual test-time scaling via initial thought transfer](#).

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Heben Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. [Evaluating large language models trained on code](#).

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuqi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. [Do not think that much for 2+3=? on the overthinking of o1-like llms](#).

Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica, Ana Klimovic, Graham Neubig, and Joseph E. Gonzalez. 2025a. [The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks](#).

Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, et al. 2025b. [The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks](#). *arXiv preprint arXiv:2502.08235*.

DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#).

Changjiang Gao, Xu Huang, Wenhao Zhu, Shujian Huang, Lei Li, and Fei Yuan. 2025. [Could thinking multilingually empower llm reasoning?](#)

Yi et.al Han. 2001. Chinese and english mathematics language: The relation between linguistic clarity and mathematics performance. *Mathematical Thinking and Learning*, 3(2-3):201–220.

Masoud Hashemi, Oluwanifemi Bamgbose, Sathwik Tejaswi Madhusudhan, Jishnu Sethumadhavan Nair, Aman Tiwari, and Vikas Yadav. 2025. [Dnr bench: Benchmarking over-reasoning in reasoning llms](#).

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Hyunwoo Ko, Guijin Son, and Dasol Choi. 2025. [Understand, solve and translate: Bridging the multilingual mathematical reasoning gap.](#)

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. [Let’s verify step by step.](#) *arXiv preprint arXiv:2305.20050*.

Wenjia Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025. [Reasoning models can be effective without thinking.](#)

Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Kroger, Xing Han Lü, Nicholas Meade, Dongchan Shin, Amirhossein Kazemnejad, Gaurav Kamath, Marius Mosbach, Karolina Stańczak, and Siva Reddy. 2025. [Deepseek-r1 thoughtology: Let’s think about llm reasoning.](#)

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, and Alex Kirillov et. al. 2024a. [Gpt-4o system card.](#)

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and Ally Bennett et.al. 2024b. [Openai o1 system card.](#)

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. [Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages.](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. [Language models are multilingual chain-of-thought reasoners.](#)

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, and Xia Hu. 2025a. [Stop overthinking: A survey on efficient reasoning for large language models.](#)

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, et al. 2025b. [Stop overthinking: A survey on efficient reasoning for large language models.](#) *arXiv preprint arXiv:2503.16419*.

Qwen Team. 2025a. [Qwen3.](#)

Qwen Team. 2025b. [Qwq-32b: Embracing the power of reinforcement learning.](#)

Hemish Veeraboina. 2023. [Aime problem set 1983-2024.](#)

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)*, pages 15366–15394.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. [Qwen2.5 technical report.](#) *arXiv preprint arXiv:2412.15115*.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. 2024. [Sglang: Efficient execution of structured language model programs.](#) In *Advances in Neural Information Processing Systems*, volume 37, pages 62557–62583. Curran Associates, Inc.

A Appendix

A.1 Prompts

Model Prompt

You **MUST** give your reasoning AND start in **{lang}** between `<think>` and `</think>` tags only. Keep the tags for the thinking part in your response. Do not generate anything outside these tags except for the final answer, which you **MUST** give in **ENGLISH**. Please reason step by step in **{lang}** and put your final answer within `boxed{{}}` in **ENGLISH**.

A.2 Unbiased Estimator for TLP@k

For a given language l and k , we want to estimate The naive estimator for the above on a dataset $\mathcal{D} = \{(x_0, y_0), \dots, (x_m, y_m)\}$ can be defined as

However, as shown in (Chen et al., 2021), this estimator has a high variance. Consequently, we extend the low variance estimator proposed in (Chen et al., 2021) for estimating $Pass@k(l)$ as

$$Pass@k(l, n) = \frac{1}{m} \sum_{i=1}^m \left[1 - \frac{\binom{n-c(x_i, y_i)}{k}}{\binom{n}{k}} \right] \text{ with}$$

Evaluation Prompt

Evaluation Criteria for a Better Response:

You will be comparing two responses that **both arrive at the correct final answer**. Your task is to evaluate the *quality of the reasoning and explanation* leading to that answer. A better response generally excels in the following areas:

1. Clarity & Logic:

- Presents reasoning in a clear, step-by-step, and logical manner.
- Is easy to follow and understand.
- Avoids ambiguity in its explanations.

2. Efficiency & Conciseness (Relative to Clarity):

- Avoids unnecessary detours, excessive repetition, or significant over-explanation of simple concepts *that detract from clarity or add significant bloat*.

3. Structure & Presentation:

- Organizes the solution in a well-structured way.
- May include helpful checks or clarifications that genuinely aid understanding without being redundant or overly verbose.

Important Note on Positional Bias: The order in which responses are presented below (A then B) is arbitrary. Please strive to evaluate each response on its own merits against the criteria before making a comparative judgment. Your final decision should not be influenced by this presentation order.

Your Task:

You are shown a math question and two responses, labeled “A” and “B”. Please follow these steps in your evaluation process:

1. Assess Individually:

- First, carefully read and assess the response labeled “A” based *solely* on the “Evaluation Criteria” above. Note its strengths and weaknesses.
- Next, carefully read and assess the response labeled “B” based *solely* on the “Evaluation Criteria” above. Note its strengths and weaknesses.

2. Compare and Decide:

- Now, compare your independent assessments of “A” and “B”.
- Decide **which response is better overall**.
- If they are of very similar quality, or if their different strengths and weaknesses balance out making a clear preference difficult, you may indicate they are “*Neither – Roughly Equal*”.

3. Explain Your Decision:

- Provide a **brief but specific explanation** for your final comparative choice. Justify why the chosen response is superior, or why they are roughly equal, referencing specific aspects of the criteria and your comparative judgment.
Example: “A was ultimately preferred because its clearer logical flow outweighed its slight verbosity compared to B.”
Example: “Neither – Roughly Equal because while A was more concise, B offered a more intuitive structural breakdown needed for this question.”

Question: {question}

Response A: {response_1_text}

Response B: {response_2_text}

Respond ONLY in this JSON format:

```
{
  "better_response_label": "A" or "B" or "Neither - Roughly Equal",
  "explanation": "<Your specific explanation for the final comparative decision,
  reflecting the comparison>"
}
```

$$Pass@k(l) = \mathbb{E}_x \left[\mathbb{1} \left(\exists r \in \mathcal{R}^{(k)}(x) \mid LLM(x, r) = y \wedge LID(r) = l \right) \right]$$

$$Pass\hat{@}k(l) = \frac{1}{m} \sum_{i=1}^m \mathbb{1} \left(\exists r \in \mathcal{R}^{(k)}(x_i) \mid LLM(x_i, r) = y_i \wedge LID(r) = l \right)$$

$c(x_i, y_i) = \sum_{r \in R^{(n)}(x_i)} \mathbb{1}(LLM(x_i, r) = y_i \wedge LID(r) = l)$. Thus, for a fixed k by increasing the number of reasoning traces n , we reduce the variance of the estimator. Calculating this estimator directly results in very large numbers and numerical instability. Instead, we use A.2 below, as done in (Chen et al., 2021).

```
def pass_at_k(n, c, k):
    """
    :param n: total number of samples
                generated
    :param c: number of correct samples
                as defined above
    :param k: k in TLP@k%
    """
    if n - c < k: return 1.0
    return 1.0 - np.prod(1.0 - k /
                        np.arange(n - c + 1, n + 1))
```

A.3 TLA Results

To provide a finer-grained view, we also define Target Language Accuracy (TLA) as the average accuracy over faithful traces:

$$TLA(j) = \frac{1}{m} \sum_{i=1}^m \frac{\sum_{r \in \mathcal{T}(i,j)} \mathbb{1}[LLM(r) = y_i]}{|\mathcal{T}(i, j)|} \quad (1)$$

Detailed TLA results for DEEPSEEK R1 in Table 4 and Figure 2, for QWEN 2.5 in Table 5 and Figure 4. For QWEN 3 in Table 6 and Figure 3.

A.4 Detailed Comparison Test Results

Check Table 9 for the results for QWEN 3.

A.5 Qualitative analysis done for a QWEN 3 output in Chinese

Check Figure 5 and Table 10 for a qualitative check on QWEN 3 output.

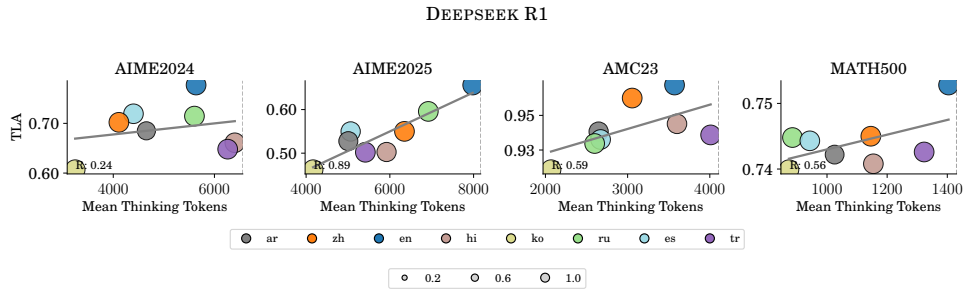
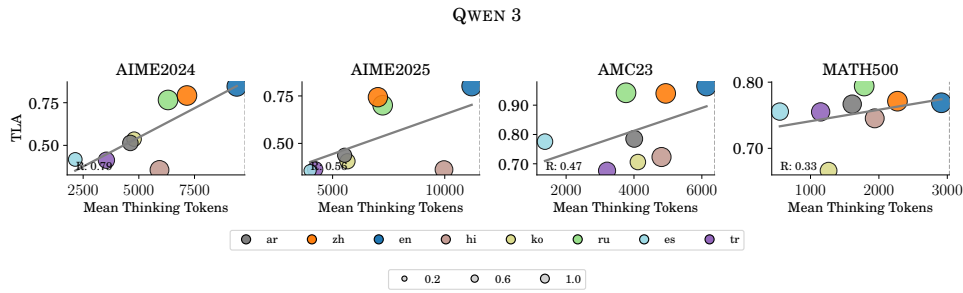
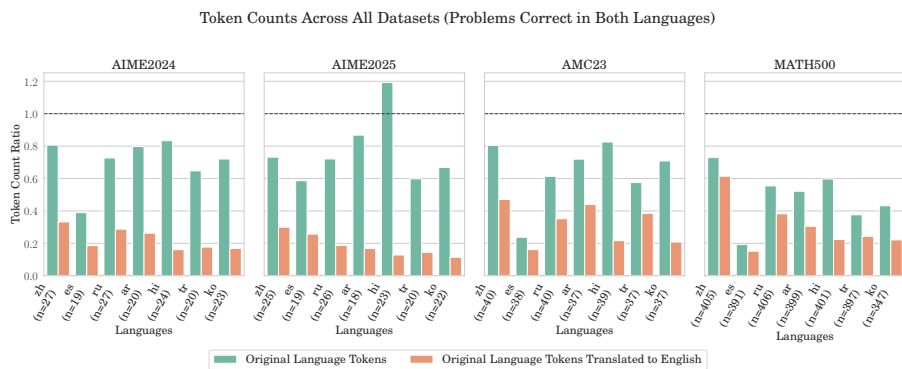


Figure 2: DEEPSEEK R1 analysis: TLA versus mean thinking tokens per language, with point size indicating how often the target language was used (out of 16 samples).

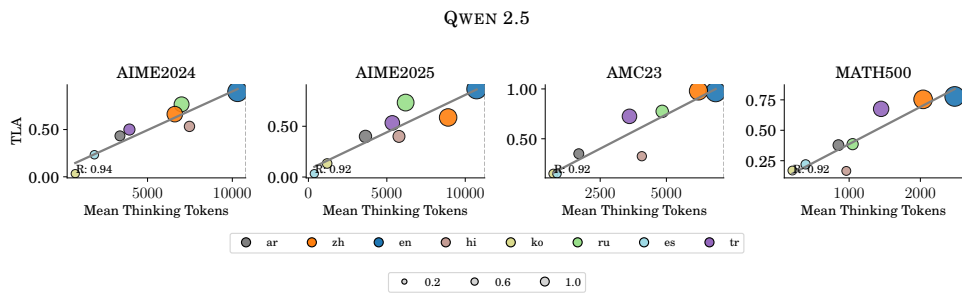


(a) Figure 3a shows TLA versus mean thinking tokens per language, with point size indicating how often the target language was used (out of 16 samples).

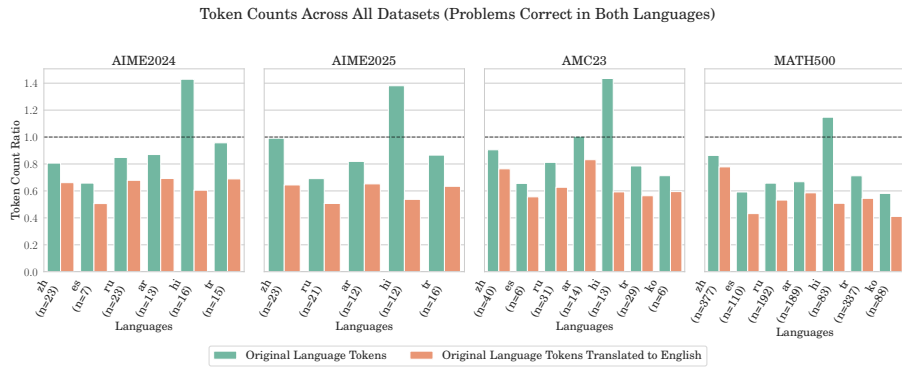


(b) Figure 3b plots the Token Count Ratio against the number of problems with at least one correct answer in both English and the target language (minimum 5 shared instances). The ratio is computed relative to the average English token count for each problem.

Figure 3: QWEN 3 analysis



(a) Figure 4a shows TLA versus mean thinking tokens per language, with point size indicating how often the target language was used (out of 16 samples).



(b) Figure 4b plots the Token Count Ratio against the number of problems with at least one correct answer in both English and the target language (minimum 5 shared instances). The ratio is computed relative to the average English token count for each problem.

Figure 4: QWEN 2.5 analysis

Example problem from MATH500 solved by QWEN 3 in Chinese

Question: Simplify $\sqrt{242}$.

Response in English: Okay, so I need to simplify the square root of 242. Hmm, let me think. First, I remember that simplifying a square root involves factoring the number into its prime factors and then looking for pairs of the same number because each pair can be taken out of the square root as a single number.

Alright, starting with 242. Let me try dividing by small primes first. The number 242 is even, right? Because it ends with a 2, so it should be divisible by 2. Let me check: 242 divided by 2 equals 121. So that gives me 2 times 121. Now, I need to factor 121. Hmm, 121... I think that's a square number. Wait, 11 times 11 is 121. Yes, that's right! So putting that all together, the prime factorization of 242 is $2 \times 11 \times 11$.

So, writing that out as a product of primes: 2×11^2 . Now, when simplifying square roots, every pair of primes can be taken out of the square root. Here, we have a pair of 11s and a single 2. That means the square root of 242 can be written as $\sqrt{11^2 \times 2}$.

Using the property of square roots that $\sqrt{a \times b} = \sqrt{a} \times \sqrt{b}$, this becomes $\sqrt{11^2} \times \sqrt{2}$. Since $\sqrt{11^2}$ is 11, the simplified form should be $11 \times \sqrt{2}$.

Wait a second, let me double-check. If I square $11\sqrt{2}$, I should get back the original number. Let's see: $(11\sqrt{2})^2 = 11^2 \times (\sqrt{2})^2 = 121 \times 2 = 242$. Yep, that checks out.

But hold on, did I factorize correctly? Let me confirm. $2 \times 121 = 242$. Then 121 is 11 squared. Correct. So there are no other factors. There's a single 2 and a pair of 11s. So taking the pair out, leaving the 2 inside.

Is there a possibility that 242 could be factored differently? Let's see. Suppose I didn't start with 2. What if I tried dividing by 11 first? Let's try: 242 divided by 11 is 22. Then 22 divided by 11 is 2. So that gives the same result: $11 \times 11 \times 2$. So regardless of the order, the prime factors are the same. Therefore, regardless of the method, the simplified square root should be $11\sqrt{2}$.

I think that's all. I don't see any mistakes in my reasoning. So the answer should be 11 times the square root of 2.

Translated Response for Chinese to English: First, I need to simplify $\sqrt{242}$. To simplify the square root, I need to find the square factors of the number under the root. First, I will factorize 242 into its prime factors. Dividing 242 by 2 gives 121, so $242 = 2 \times 121$. Next, 121 is a known square number because $11 \times 11 = 121$. Therefore, $\sqrt{242}$ can be broken down into $\sqrt{2 \times 121}$.

According to the multiplication property of square roots, $\sqrt{a \times b} = \sqrt{a} \times \sqrt{b}$, so $\sqrt{2 \times 121} = \sqrt{2} \times \sqrt{121}$. Since $\sqrt{121} = 11$, the entire expression can be simplified to $11\sqrt{2}$.

Therefore, the simplified form of $\sqrt{242}$ is $11\sqrt{2}$.

Figure 5: Example problem solved by QWEN 3

Dataset	en	zh	ru	es	ar	tr	hi	ko
MATH500	0.75	0.75	0.74	0.74	0.74	0.74	0.74	0.74
AMC23	0.97	0.96	0.93	0.93	0.94	0.94	0.94	0.91
AIME2024	0.78	0.70	0.71	0.72	0.69	0.65	0.66	0.61
AIME2025	0.66	0.55	0.60	0.55	0.53	0.50	0.50	0.46

Table 4: TLA scores as defined in Equation: A.3 for DEEPSEEK R1

Dataset	en	zh	ru	es	ar	tr	hi	ko
MATH500	0.77	0.77	0.79	0.76	0.77	0.76	0.75	0.67
AMC23	0.97	0.94	0.94	0.78	0.79	0.68	0.72	0.71
AIME2024	0.85	0.79	0.77	0.42	0.51	0.41	0.36	0.54
AIME2025	0.80	0.74	0.70	0.36	0.44	0.36	0.36	0.41

Table 5: TLA scores as defined in Equation: A.3 for QWEN 2.5

Dataset	en	zh	ru	es	ar	tr	hi	ko
MATH500	0.77	0.77	0.79	0.76	0.77	0.76	0.75	0.67
AMC23	0.97	0.94	0.94	0.78	0.79	0.68	0.72	0.71
AIME2024	0.85	0.79	0.77	0.42	0.51	0.41	0.36	0.54
AIME2025	0.80	0.74	0.70	0.36	0.44	0.36	0.36	0.41

Table 6: TLA scores as defined in Equation: A.3 for QWEN 3

Dataset	en		zh		ru		es		ar		tr		hi		ko	
	TLP@k	Tokens	$\Delta\%TLP$	$\Delta\%Tok$	$\Delta\%TLP$	$\Delta\%Tok$	$\Delta\%TLP$	$\Delta\%Tok$	$\Delta\%TLP$	$\Delta\%Tok$	$\Delta\%TLP$	$\Delta\%Tok$	$\Delta\%TLP$	$\Delta\%Tok$	$\Delta\%TLP$	$\Delta\%Tok$
MATH500	0.78	2482.11	-78.21	61.31	-14.10	41.62	-3.85	17.86	-71.79	84.31	-51.28	65.76	-51.28	57.73	-76.92	91.80
AMC23	1.00	6860.88	-67.00	40.64	-27.00	47.45	0.00	9.48	-85.00	87.35	-65.00	75.34	-22.00	29.43	-85.00	89.34
AIME2024	0.90	10310.63	-41.11	27.53	-44.44	61.88	-14.44	35.90	-74.44	81.92	-52.22	67.30	-14.44	32.04	-96.67	92.85
AIME2025	0.87	10725.53	-54.02	46.16	39.08	50.12	-11.49	16.95	-96.55	96.49	-54.02	66.08	-19.54	42.21	-85.06	88.86

(a) QWEN 2.5

Dataset	en		zh		ru		es		ar		tr		hi		ko	
	TLP@k	Tokens	$\Delta\%TLP$	$\Delta\%Tok$	$\Delta\%TLP$	$\Delta\%Tok$	$\Delta\%TLP$	$\Delta\%Tok$	$\Delta\%TLP$	$\Delta\%Tok$	$\Delta\%TLP$	$\Delta\%Tok$	$\Delta\%TLP$	$\Delta\%Tok$	$\Delta\%TLP$	$\Delta\%Tok$
MATH500	0.81	2913.68	0.00	22.05	0.00	38.74	-3.70	81.08	-1.23	44.75	-2.47	60.59	-1.23	33.46	-14.81	56.61
AMC23	1.00	6138.45	0.00	19.60	0.00	38.67	-5.00	77.58	-7.00	34.71	-7.00	47.76	-2.00	21.61	-7.00	32.97
AIME2024	0.93	9399.33	-3.23	23.83	-3.23	32.86	-32.26	77.18	-27.96	50.84	-27.96	62.29	-13.98	36.94	-17.20	49.07
AIME2025	0.93	11200.1	-10.75	37.32	-6.45	35.43	-32.26	64.32	-35.48	50.57	-27.96	62.32	-17.20	11.02	-21.51	49.44

(b) QWEN 3

Table 7: Performance and token efficiency of native-language reasoning using QWEN models. In each subtable, TLP@4 and mean token count are reported as percentage change from English

Model	en			zh			ru			es		
	TLP@4	TLC	Tok.	TLP@4	TLC	Tok.	TLP@4	TLC	Tok.	TLP@4	TLC	Tok.
DEEPSEEK R1	0.87	1.00	3816	0.83	1.00	3009	0.83	1.00	3276	0.88	0.99	2677
QWEN 2.5	0.87	1.00	6249	0.84	0.80	4912	0.67	0.34	3861	0.16	0.06	746
QWEN 3	0.89	1.00	6123	0.88	0.99	4427	0.89	1.00	3932	0.88	0.58	1656

(a) Performance across English, Chinese, Russian, and Spanish.

Model	ar			tr			hi			ko		
	TLP@4	TLC	Tok.	TLP@4	TLC	Tok.	TLP@4	TLC	Tok.	TLP@4	TLC	Tok.
DEEPSEEK R1	0.80	0.90	2738	0.80	1.00	3498	0.80	0.97	3521	0.79	1.00	2140
QWEN 2.5	0.37	0.18	1996	0.62	0.44	2978	0.36	0.10	3668	0.16	0.06	616
QWEN 3	0.75	0.67	3223	0.73	0.72	2476	0.75	0.89	4675	0.74	0.57	3221

(b) Performance across Arabic, Turkish, Hindi, and Korean.

Table 8: Performance metrics (TLP@4, TLC, Mean Token count) across different languages and models.

Lang	AIME2024			AIME2025			AMC23			MATH500		
	N	Win (%)	Sig?	N	Win (%)	Sig?	N	Win (%)	Sig?	N	Win (%)	Sig?
ar	12	100.0	✓	10	100.0	✓	21	100.0	✓	295	98.6	✓
zh	11	90.9	✓	13	100.0	✓	11	81.8	✓	94	73.4	✓
hi	20	100.0	✓	18	94.4	✓	34	94.1	✓	368	98.1	✓
ko	17	100.0	✓	19	94.7	✓	31	96.8	✓	298	99.3	✓
ru	12	91.7	✓	15	100.0	✓	21	95.2	✓	252	99.6	✓
es	16	93.8	✓	12	91.7	✓	36	100.0	✓	370	100.0	✓
tr	18	94.4	✓	15	93.3	✓	27	100.0	✓	345	99.1	✓

Table 9: Binomial test results comparing non-English vs English reasoning responses obtained from QWEN 3, using consistent GPT-4o judgments. ✓ indicates $p < 0.05$, and ✗ indicates no significant difference.

Phenomenon	English Response	Chinese Response
Self-dialogue	"Hmm, let me think..."	First, simplify $\sqrt{242}$.
Verification	"11 × 11 is 121... Yes!"	121 is a known square number...
Redundant check	"Let me double-check..."	No equivalent step
Alternative path	"What if I didn't start with 2..."	No alternative explored
Repeated confirmation	"I don't see any mistakes..."	$\sqrt{242}$ is $11\sqrt{2}$.

Table 10: Comparison of English and translated Chinese traces (Qwen 3, MATH500). English shows overthinking. Full output in Figure 5.