Do Thoughts Depth Affect Multilingual Reasoning?

Linjian Yang[1], **Xinyan Wang**[2], **Kunpeng Liu**[1]
[1]Clemson University, [2]Portland State University
{linjiay, kunpenl}@clemson.edu, xinyw@pdx.edu

Abstract

Chain-of-Thought (CoT) is commonly used to improve reasoning performance in large language models. We investigate its impact in multilingual contexts by systematically constraining reasoning steps across languages with varying resource levels. This study evaluates two models on two benchmarks with seven languages, comparing constrained CoT depth against zero-shot and free-CoT baselines. We demonstrate that increasing the number of reasoning steps does not consistently improve accuracy across various languages. While high-resource and mid-resource languages remain stable, low-resource languages often experience a decline in performance as the number of reasoning steps increases. We attribute this decline to error accumulation and reasoning noise, which are amplified under deeper reasoning in low-resource languages. These findings indicate that CoT is not inherently beneficial, but its effectiveness is significantly influenced by the interaction between reasoning steps and language resource availability.

1 Introduction

Existing reasoning-based large language models (LLMs) with Chain-of-Thought (CoT) capabilities (Wei et al., 2022) are developed on datasets that are predominantly English-centric (Xue et al., 2021; Huang et al., 2026). Consequently, these models suffer from a significant cross-lingual performance disparity (Wang et al., 2019; Qi et al., 2026; Li et al., 2025), i.e., when queried in non-English languages, their reasoning capabilities diminish drastically compared to their performance on identical English-language prompts.

The literature of multilingual CoT mainly focuses on the critical challenge of bridging the gap between high-resource and low-resource languages (Shi et al., 2022; Qi et al., 2025). Nevertheless, it is unclear how the depth of thoughts impacts multilingual reasoning. The existing studies (Zhou et al., 2022; Cox et al., 2025) are usually based on the assumption that extended intermediate reasoning chains could inherently improve outcomes. However, this assumption should not necessarily be true in terms of multilingual settings. This is because, if the multilingual large language models (MLLMs) conduct reasoning in a low-resource language that does not contain enough knowledge, then the large thought depth might increase the possibility of error accumulation (Wang et al., 2023). Thus, a critical research question is how to navigate the MLLMs to conduct reasoning via CoT across different languages in multilingual settings. Intuitively, for high-resource languages, extended intermediate reasoning is beneficial, but not for low-resource languages, which might require some additional constraints (Huang et al., 2021; Shi et al., 2022). We hypothesize that imposing explicit constraints on the reasoning steps may help mitigate the cross-lingual performance disparity issue. Thus, we aim to investigate whether constraining CoT reasoning steps could improve model accuracy across different languages.

To this end, we conducts a series of empirical analysis to examine the impact of reasoning thoughts depth in terms of various language resource levels. As shown in the Figure 1, CoT applied with explicit constraints in prompts on reasoning steps might be able to regulate the number of reasoning steps. Following the same principle, we compare multiple constrained CoT depths against free-CoT (Wei et al., 2022) and zero-shot (Kojima et al., 2022) baselines to isolate how reasoning thoughts depth impacts downstream task performance across various languages. This investigation is evaluated based on two MLLMs, DeepSeek (Bi et al., 2024) and Moonshot (Team et al., 2025), on two reasoning benchmarks, MGSM (Shi et al., 2022) in mathematical reasoning and X-CSQA (Lin et al., 2021) in cross-lingual commonsense reasoning. To ensure the

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 149–158

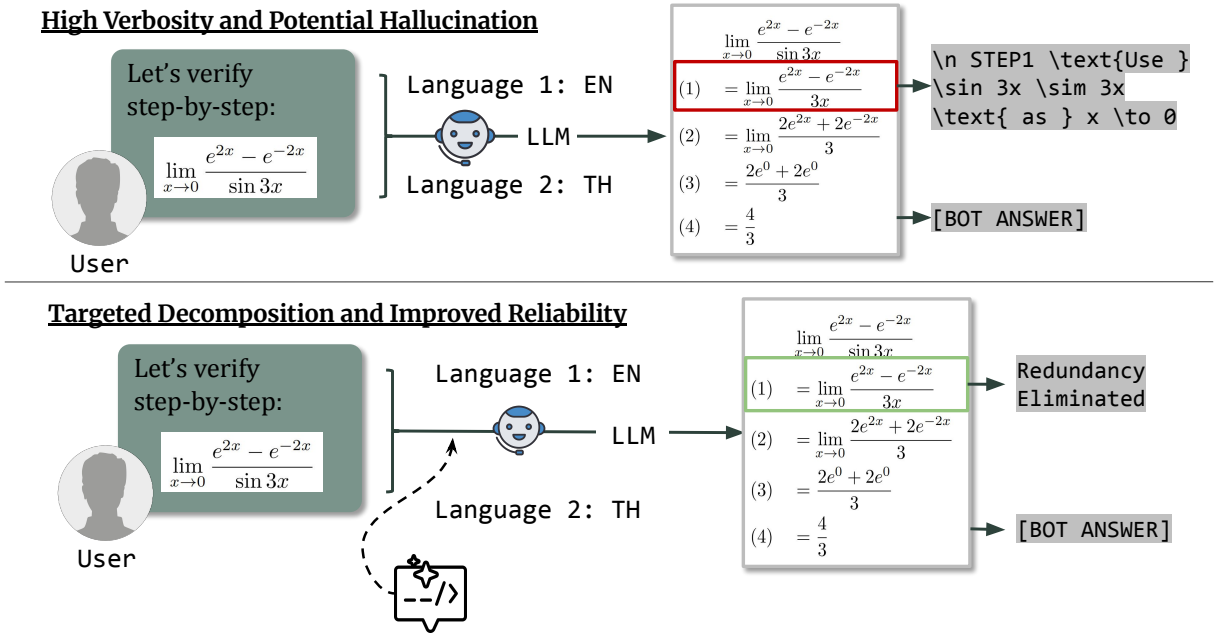

Figure 1: Unconstrained vs. Constrained Chain-of-Thought in Multilingual Reasoning. The prompt icon in the lower panel denotes the application of step-constrained prompting, which enforces targeted decomposition to eliminate redundant tokens.

reflection is not language-specific, the composition of low-resource languages differs across the two datasets. This intentional design enables us to assess whether the impact of CoT depth remains consistent across models, tasks, and language resource levels. Overall, we have the following key findings, which demonstrate that CoT is not a universally applicable mechanism in multilingual scenarios.

- **Resource-dependent Effectiveness:** We observe that the advantages of CoT are various by the availability of language resources.
- **Stability in High-Resource and Mid-Resource Languages:** The performance of multilingual reasoning exhibit stable performance improvements, even with extended reasoning steps.
- **Degradation in Low-Resource Languages:** In contrast, low-resource languages often exhibit minimal benefits and experience a decline in performance when reasoning steps are extended.

2 Related Work

2.1 General Reasoning in LLMs

Prompting with CoT has become a widely adopted approach for improving reasoning performance in LLM (Wei et al., 2022). Following the initial success of CoT prompting, subsequent research has extended this paradigm to include variants such as zero-shot CoT (Kojima et al., 2022), self-consistency decoding (Wang et al., 2022; Mo et al., 2026), and verification-based methods (Dhuliawala et al., 2024; Zhang et al., 2025a). Most recent work has further explored structured reasoning strategies, including decomposition-based (Khot et al., 2022) and search-based approaches (Mo et al., 2026; Huang et al., 2025; Zhang et al., 2025b) to highlight the importance of explicitly modeling the intermediate reasoning process. Collectively, these studies demonstrate that eliciting step-by-step reasoning can substantially improve performance on tasks requiring multi-step inference, such as mathematical problem-solving (Shi et al., 2022; Zhang et al., 2026) and commonsense question answering (Wei et al., 2022).

2.2 Multilingual Reasoning

Most existing optimization strategies for LLM reasoning are developed and evaluated in English. While recent cross-lingual benchmarks have expanded evaluation across a broader range of languages (Huang et al., 2026), the consistency of reasoning performance across languages remains underexplored (Wang et al., 2025; Qi et al., 2025).

In particular, CoT reasoning is often implicitly treated as universally effective, despite substantial disparities in language resources and representation quality. To study multilingual performance, prior work has introduced benchmark datasets, such as XTREME (Hu et al., 2020), MGSM (Shi et al., 2022), and X-CSQA (Lin et al., 2021), which consistently show that LLMs achieve higher accuracy in high-resource languages than in low-resource ones. Recent approaches (Qi et al., 2026) have explored cross-lingual reasoning strategies, such as a translation-based strategy from high-resource to low-resource languages (Ebing and Glavaš, 2024). However, these methods primarily focus on overall performance and do not systematically examine how reasoning strategies are impacted by language resource availability.

3 Methodology

3.1 Problem Setup

This study aims to examine the effect of reasoning depth on downstream task performance across various languages with different levels of resource availability. Formally, let $\mathcal{L} = \{l_1, \ldots, l_K\}$ represent a set of languages, and let $\rho(l)$ denote the resource level associated with language l. For each language $l \in \mathcal{L}$, we aim to evaluate the performance on top of a dataset $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is an input query and y_i is the corresponding ground-truth answer.

Given a language model f_θ, we evaluate its performance under CoT prompting with controlled reasoning steps. To this end, we introduce a reasoning step parameter $k \in \mathcal{K}$, which constrains the number of intermediate reasoning steps encouraged in the prompt. The model produces a prediction:

$$\hat{y} = f_\theta(x; k, l),$$

where k controls the explicit reasoning steps and l specifies the language of the input.

3.2 Multilingual CoT Prompting Design

To systematically investigate the impact of reasoning trajectory step on model performance across diverse languages, we define reasoning step as the precise number of explicitly intermediate reasoning steps mandated within the prompt structure.

In our approach, we employ a constrained prompting strategy to isolate the effect of reasoning steps from the model's inherent verbosity. We define multiple CoT depth conditions to capture the transition from shallow to more structured reasoning. In the Table 1, we interpret larger values of k as inducing deeper, more fine-grained reasoning, while smaller values encourage more concise, high-level reasoning.

Setting	Description
CoT = -1	Unconstrained reasoning (default behavior).
CoT = 0	Direct answer without explicit reasoning.
CoT = 2	Two-step reasoning (coarse-grained).
CoT = 4	Four-step reasoning (moderate granularity).
CoT = 6	Six-step reasoning (fine-grained).
CoT = 8	Eight-step reasoning (most fine-grained).

Table 1: Definition of CoT reasoning steps settings used to control reasoning granularity.

For $k \in \{2, 4, 6, 8\}$, the LLM is explicitly instructed to decompose its reasoning into exactly k concise steps before generating the final answer. To ensure consistency across languages, all prompts are manually translated into each target language while preserving both the original prompt structure and the imposed reasoning constraints. A unified prompt template is used across all experimental settings, as shown below:

```
<instruction>
Answer the question according to the specified
reasoning mode.
</instruction>

<mode>
CoT = −1: think step by step
CoT = 0: answer directly
CoT = k: exactly k concise reasoning steps,
where k ∈ {2, 4, 6, 8}
</mode>

<question>
{Question}
</question>

<choices>
A, B, C, D, E (if applicable)
</choices>

<answer>
Final Answer: [number / choice]
</answer>
```

4 Experiment Setup

4.1 Languages Coverage

As summarized in Table 2, we select languages to span different levels of resource availability,

Level	Language	Script	Morphology
High	English (en)	Latin	Analytic
	Chinese (zh)	CJK	Analytic
Medium	German (de)	Latin	Inflectional
	Japanese (ja)	Kana+Kanji	Agglutinative
Low	Swahili (sw)	Latin	Agglutinative
	Thai (th)	Thai	Analytic
	Urdu (ur)	Perso-Arabic	Inflectional

Table 2: Languages grouped by resource level, along with their script and morphological typology. CJK refers to the Chinese, Japanese, and Korean character system.

enabling a controlled comparison of model performance across high-resource level, medium-resource level, and low-resource level.

The *Script* column in Table 2 underscores the typological diversity in our dataset. These include Latin-alphabet scripts, logographic scripts such as Chinese and Japanese, and regional script such as Thai and the right-to-left (RTL) Perso-Arabic script used for Urdu. By leveraging these diverse scripts, we ensure that our findings regarding reasoning plateaus are robust across different tokenization densities and visual-linguistic representations.

The *Morphology* column in Table 2 characterizes how grammatical information is expressed in each language. Analytic languages rely on word order with minimal inflection, inflectional languages modify word forms to encode grammatical relations, and agglutinative languages concatenate multiple morphemes within a single word. These structural differences can influence tokenization patterns and may affect the stability of long reasoning chains in multilingual LLMs.

4.2 Backbone Language Models

The evaluation is conducted on top of two representative multilingual LLMs to investigate whether CoT prompting effects are unique to each model or generalizable across different model architectures.

- **DeepSeek-V1** (`deepseek-chat`) (Bi et al., 2024): A multilingual LLM developed by DeepSeek AI, chosen for its strong performance and extensive multilingual coverage.
- **Moonshot-v1** (`moonshot-v1-8k`) (Team et al., 2025): A multilingual LLM developed by Moonshot AI, utilized to evaluate the robustness of observed patterns across different systems.

This comparison enables the study to distinguish language-inherent effects from model-specific artifacts. For reproducibility, all models are accessed via their respective APIs, with temperature set to 0 for deterministic evaluation and a maximum output length of 1,000 tokens to accommodate extended reasoning chains.

Both models are instruction-tuned and may already exhibit implicit reasoning capabilities. Therefore, our manipulation of CoT depth should be interpreted as controlling the explicit structure of reasoning rather than introducing reasoning ability itself.

4.3 Evaluation Benchmarks

To evaluate cross-domain generalization, we utilize two distinct benchmarks that span structured and unstructured cognitive tasks:

- **MGSM** (Multilingual Grade School Math) (Shi et al., 2022): A multilingual mathematical reasoning benchmark consisting of 250 grade school math problems translated into ten different languages by human annotators. It requires multi-step arithmetic reasoning to derive a final numeric answer.
- **X-CSQA** (Cross-lingual Commonsense QA) (Lin et al., 2021): A multilingual commonsense reasoning benchmark derived from the CommonsenseQA dataset, covering 16 languages. The original test set contains 1,074 problems and aims to test the model's understanding of typical human knowledge extending beyond mathematics. This task involves unstructured, knowledge-based question answering.

The study uses two distinct task types: one containing deterministic mathematical logic, and another involving broader commonsense inference. Beyond evaluating task-specific performance, the primary purpose is to evaluate the dependency between CoT prompting and varying resource levels of languages across tasks.

4.4 Evaluation Protocol

Each configuration of model, language, dataset, and CoT prompting step is evaluated using prefixed 250 randomly sampled queries from the original test sets.

Accuracy is reported as the primary evaluation metric, defined as the percentage of correct responses under each language and CoT depth condition. Final answer extraction is task-specific:

- **MGSM:** The final numeric answer is extracted by parsing the numerical output from the model's response. Evaluation is conducted in English, Chinese, German, Japanese, Swahili, and Thai.
- **X-CSQA:** Rule-based parsing is used to extract the selected option (A to E) or the corresponding answer text. Responses that can not be mapped to a valid option are considered as incorrect. Evaluation is conducted in English, Chinese, German, Japanese, Swahili, and Urdu.

To improve robustness and avoid overfitting conclusions to specific languages, we intentionally vary the low-resource languages across datasets: Thai is used in MGSM and Urdu is used in X-CSQA. This cross-linguistic variation helps determine whether the observed CoT behavior are driven by general resource-level effects rather than specific linguistic features (e.g., script or morphology).

5 Results and Analysis

5.1 Full Experimental Results

We present the full evaluation results for DeepSeek-V1 and Moonshot-v1 across all selected benchmarks and languages. Tables 3 provides the accuracy for each CoT condition ($k \in \{-1, 0, 2, 4, 6, 8\}$). These results reveal distinct performance patterns across high-resource, medium-resource, and low-resource languages, forming the basis for our subsequent analysis of how reasoning step interacts with language resource availability.

5.2 CoT Depth Effects Across Languages in MGSM

DeepSeek Model's Performance on MGSM. All languages achieve over 85% accuracy across conditions (see Figure 2), with high-resource languages such as English showing minimal variance as reasoning step increases. Interestingly, some mid-resource languages such as German exhibit non-monotonic fluctuations, showing that increasing CoT depth does not consistently translate into improved performance. One possible explanation is

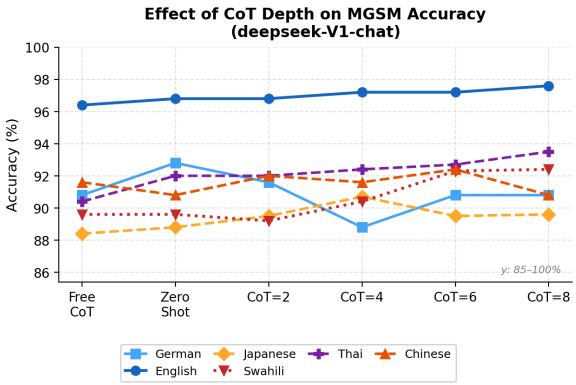

Figure 2: Deepseek-V1 accuracy matrix across all CoT conditions on MGSM

that extended reasoning chains introduce additional variability, potentially interacting with language-specific characteristics such as tokenization or representation quality. However, this mechanism is not directly examined in our study. In contrast, Thai demonstrates a steady, monotonic improvement with increased reasoning steps. Low-resource languages such as Swahili maintains a relatively high baseline accuracy approximately 90% but fluctuate without consistent gains from deeper reasoning, indicating a performance plateau where additional reasoning steps fails to yield further improvements.

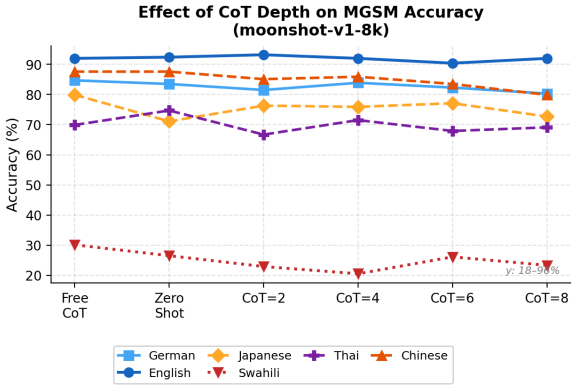

Figure 3: Moonshot-v1 Accuracy matrix across all CoT conditions on MGSM

Moonshot Model's Performance on MGSM. As illustrated in Figure 3, a divergent pattern emerges. While high-resource and mid-resource languages achieve accuracies ranging from 70% to 93%, the low-resource language Swahili consistently remains between 20% to 30% across all CoT conditions. Since increasing CoT depth does not lead to measurable improvements for Swahili. This reflects that prompting alone is insufficient to close

	MGSM						X-CSQA					
Language	**CoT$_{-1}$**	**CoT$_0$**	**CoT$_2$**	**CoT$_4$**	**CoT$_6$**	**CoT$_8$**	**CoT$_{-1}$**	**CoT$_0$**	**CoT$_2$**	**CoT$_4$**	**CoT$_6$**	**CoT$_8$**
DeepSeek-V1												
de	90.8%	**92.8%**	91.6%	88.8%	90.8%	90.8%	62.0%	**70.8%**	50.0%	42.8%	52.0%	54.4%
en	96.4%	96.8%	96.8%	97.2%	97.2%	**97.6%**	**58.8%**	55.6%	49.6%	52.0%	50.4%	50.2%
ja	88.4%	88.8%	89.5%	**90.7%**	89.5%	89.6%	35.5%	29.7%	34.2%	35.0%	36.8%	**39.4%**
sw	89.6%	89.6%	89.2%	90.4%	92.3%	**92.4%**	**33.2%**	30.5%	27.4%	26.4%	20.9%	24.3%
th	90.4%	92.0%	92.0%	92.4%	92.7%	**93.5%**	—	—	—	—	—	—
ur	—	—	—	—	—	—	32.2%	**35.8%**	27.0%	28.4%	23.9%	25.5%
zh	91.6%	90.8%	92.0%	91.6%	**92.4%**	90.8%	65.2%	61.2%	58.0%	63.6%	**67.1%**	65.6%
Moonshot-v1												
de	**84.7%**	83.5%	81.5%	83.9%	82.3%	80.3%	63.2%	60.4%	**66.4%**	61.2%	63.6%	57.6%
en	92.0%	92.4%	**93.2%**	92.0%	90.4%	92.0%	76.8%	71.6%	**77.2%**	76.0%	76.4%	74.0%
ja	**79.9%**	71.1%	76.3%	75.9%	77.1%	72.7%	**59.6%**	50.0%	53.8%	59.2%	58.0%	58.2%
sw	**30.1%**	26.5%	22.9%	20.5%	26.1%	23.3%	21.2%	**31.2%**	24.0%	26.4%	26.4%	20.0%
th	69.9%	**74.7%**	66.7%	71.5%	67.9%	69.1%	—	—	—	—	—	—
ur	—	—	—	—	—	—	32.4%	34.4%	**45.2%**	43.6%	42.8%	44.0%
zh	**87.6%**	**87.6%**	85.1%	85.9%	83.5%	79.9%	58.4%	44.8%	**61.6%**	59.2%	58.0%	59.6%

Table 3: Combined accuracy results for DeepSeek-V1 and Moonshot-v1 across the MGSM and X-CSQA benchmarks. Missing language-benchmark pairs are shown as dashes.

the performance gap and that it may be related to limitations in low-resource mathematical reasoning.

5.3 CoT Depth Effects Across Languages in X-CSQA

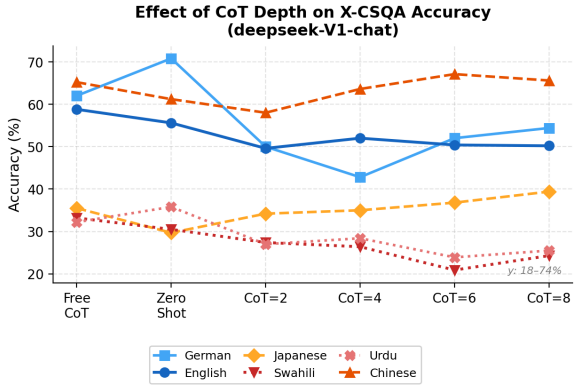

Figure 4: DeepSeek-V1 Accuracy matrix across all CoT conditions on X-CSQA

DeepSeek Model's Performance on X-CSQA. DeepSeek exhibits strong language-dependent sensitivity to CoT constraints. Chinese consistently outperform English (see Figure 4), exhibiting a U-shaped pattern in which accuracy dips at $k = 2$ before recovering at higher CoT depths $k = 6 - 8$. German shows high variance, it initially outperforms English but experiences a sharp decline under constrained CoT conditions. In contrast, Japanese, Swahili, and Urdu remain at lower performance level, and Swahili declines further as CoT depth increases, indicating instability in low-resource settings.

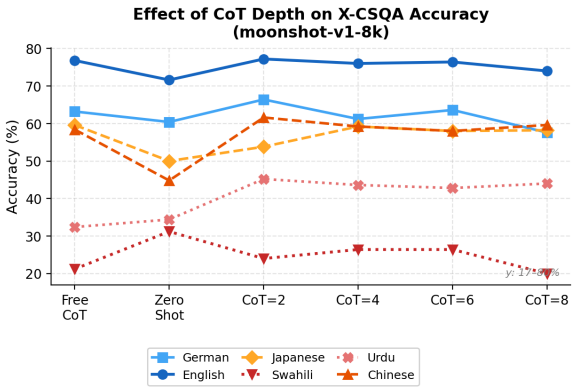

Figure 5: Moonshot-v1 Accuracy matrix across all CoT conditions on X-CSQA

Moonshot Model's Performance on X-CSQA. Unlike DeepSeek, Moonshot exhibits more stable cross-lingual performance, although substantial performance gap across resource levels remain. As shown in Figure 5, English achieves the highest accuracy among all languages. Urdu improves from Free-CoT to $k = 2$ and $k = 4$, suggesting that shallow reasoning scaffolding can be beneficial when native reasoning capabilities are limited. However, Swahili remains the lowest-performing language and declines further at $k = 8$.

Overall, CoT prompting with explicit step constraints provides limited benefits for low-resource languages. While shallow CoT may yield modest gains in certain cases, increasing the number of reasoning steps under this setting often leads to performance degradation.

5.4 Cross-Model Comparison

Figure 6 reveals two key trends. First, performance degradation across language resource levels varies by both task and model. Although MGSM achieves higher overall accuracy, the gap between high-resource and low-resource is strongly model-dependent: Moonshot experiences a severe collapse on low-resource mathematical reasoning, whereas DeepSeek remains relatively robust.

In contrast, X-CSQA shows a more consistent degradation across resource levels for both models. Second, model rankings reverse across tasks: DeepSeek consistently outperforms Moonshot on MGSM but underperforms on X-CSQA. This suggests that reasoning performance is task-dependent and does not transfer uniformly across domains.

5.5 Overall Implications

Across both models and tasks, we observe a consistent pattern: increasing CoT reasoning steps under step-constrained prompting does not reliably improve performance and often degrades accuracy in low-resource languages. This trend persists despite variation in linguistic typology and script, indicating that the effect is not driven by any specific language. Instead, it reflects a broader limitation associated with insufficient training data, which constrains the model's ability to benefit from extended reasoning. Overall, these findings show that the ineffectiveness of CoT in low-resource languages is more consistent with representation gaps than with a lack of reasoning capacity.

6 Discussion

6.1 Why Does Moderate CoT Depth Improve Performance?

Moderate CoT prompting can improve performance by decomposing complex problems into simpler intermediate steps. This decomposition reduces the effective search space and promotes locally coherent generation at each step. In high-resource languages, extensive training data allows this structure reasoning to better align with patterns learned during training, enabling the generation of coherent intermediate steps that support correct final answers.

This interpretation is consistent with our observations that English maintains stable performance across CoT depths, and Japanese shows consistent gains at moderate CoT depths across both models. However, increasing reasoning steps beyond a moderate range may introduce unnecessary inference redundancy, which can disrupt logical consistency and lead to compounding errors. Therefore, these results suggest that high-resource and mid-resource languages benefit structured reasoning primarily within an appropriate depth range.

6.2 Why Does Excessive CoT Degrade Accuracy?

The results show that the benefits of CoT prompting are not sustained as reasoning steps increase. Beyond a certain point, additional reasoning steps introduce several mechanisms that degrade performance:

- **Error accumulation:** Each reasoning step carries a probability of error that compounds across the reasoning chain, leading to cascading failures. Minor inaccuracies at early steps can amplify over time. For example, German's accuracy on DeepSeek declines monotonically with increasing CoT depth on both MGSM and X-CSQA, illustrating the cumulative effect of error propagation.

- **Overthinking:** Extending the reasoning process beyond what is necessary can lead to redundant or inconsistent intermediate steps. For example, Swahili maintains stable performance at moderate depths but degrades at higher CoT depths ($k = 6$ or $k = 8$) across both models, indicating that excessive reasoning introduces instability.

- **Spurious reasoning:** Longer reasoning chains may produce outputs that appear logically structured but are based on incorrect intermediate justifications. In such cases, additional reasoning does not correct errors but instead reinforces incorrect conclusions.

Overall, these findings show that excessive CoT depth amplifies noise rather than signal, leading to diminishing returns in reasoning performance.

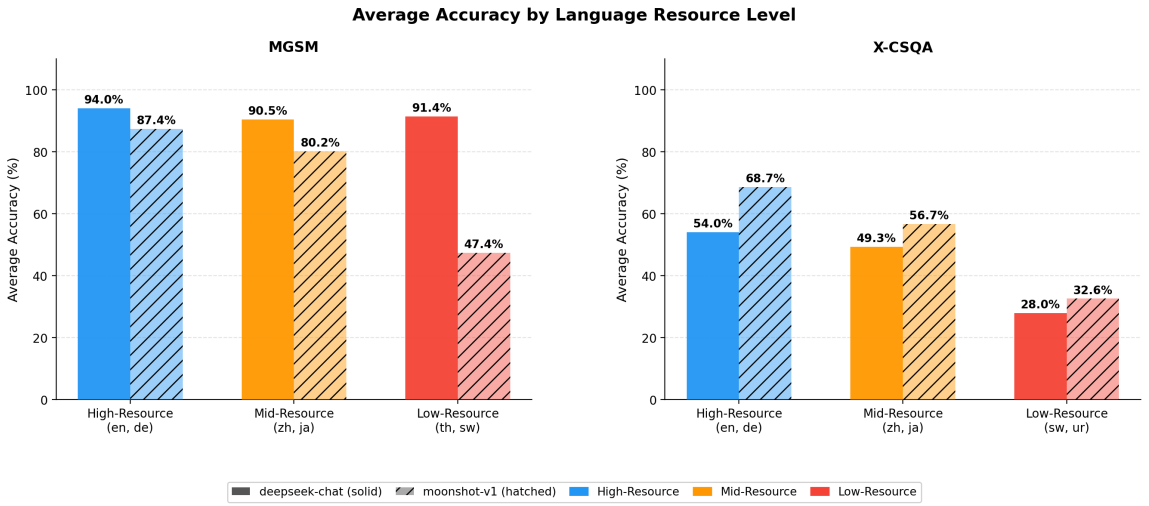

Figure 6: Average accuracy of models across varying language resource levels

6.3 Low-Resource Languages and Reasoning Noise

Limited training data leads to weaker and less reliable internal representations, increasing the likelihood of errors at each reasoning step. As reasoning steps increase, these errors can accumulate and amplify, resulting in degraded performance over longer inference chains. This pattern is reflected in the results for Swahili, which demonstrates consistently low performance with no improvement across CoT depths. Similarly, although Urdu exhibits modest gains under moderate CoT, it does not benefit from further increases in reasoning depth.

These findings indicate that poor performance in low-resource languages is less likely due to a lack of reasoning capacity and more consistent with an inability to suppress reasoning noise. Insufficient training data may limit the model's ability to maintain coherent multi-step reasoning, leading to error propagation that undermines inference.

7 Conclusion

This study investigates whether controlling the depth of CoT reasoning improves the accuracy of LLMs across languages with varying resource availability. We conduct experiments using two models (DeepSeek and Moonshot), two reasoning benchmarks (MGSM and X-CSQA), and seven languages spanning different resource levels. Our results show that increasing CoT depth does not consistently improve accuracy. While high-resource languages remain largely stable and low-resource languages exhibit only marginal gains, medium-resource languages show patterns similar to those of high-resource languages. Overall, the effectiveness of CoT depth is not directly correlated with performance, but instead depends on the interaction between language resource availability and model characteristics. These findings suggest that constraining CoT depth should be treated as a conditional strategy rather than a universally beneficial strategy.

8 Future Work

Future work can extend the analysis to a broader range of models to assess the generality of the observed patterns. Detailed error analysis could further distinguish between sources of failure, such as morphological ambiguity and hallucinations. Expanding evaluation to include more low-resource languages or diverse language families would also improve the robustness of these findings. Additionally, it is important to explore adaptive approaches that dynamically determine the optimal CoT depth for each language and task. Analyzing the role of pre-training data composition and tokenization efficiency may help clarify the reasoning noise observed in this study. Finally, developing intervention strategies, such as cross-lingual knowledge distillation or targeted reasoning fine-tuning, could help mitigate representation gaps and improve the effectiveness of CoT prompting in low-resource languages.

Acknowledgments

Dr. Kunpeng Liu is supported by the National Science Foundation (NSF) via the grant numbers 2550105, 2550106, and 2242812.

References

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, and 1 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Kyle Cox, Jiawei Xu, Yikun Han, Rong Xu, Tianhao Li, Chi-Yang Hsu, Tianlong Chen, Walter Gerych, and Ying Ding. 2025. Mapping from meaning: Addressing the miscalibration of prompt-sensitive language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23696–23703.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578.

Benedikt Ebing and Goran Glavaš. 2024. To translate or not to translate: A systematic investigation of translation-based cross-lingual transfer to low-resource languages. 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 5325–5344.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.

Chao Huang, Fengran Mo, Yufeng Chen, Changhao Guan, Zhenrui Yue, Xinyu Wang, Jinan Xu, and Kaiyu Huang. 2025. Boosting data utilization for multilingual dense retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12362–12378, Suzhou, China. Association for Computational Linguistics.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, and 1 others. 2026. A survey on large language models with multilingualism: Recent advances and new frontiers. *Artificial Intelligence Review*.

Kaiyu Huang, Keli Xiao, Fengran Mo, Bo Jin, Zhuang Liu, and Degen Huang. 2021. Domain-aware word segmentation for chinese language: A document-level context-aware model. *Transactions on Asian and Low-Resource Language Information Processing*, 21(2):1–16.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Hongliang Li, Jinan Xu, Gengping Cui, Changhao Guan, Fengran Mo, and Kaiyu Huang. 2025. Multilingual collaborative defense for large language models. *arXiv preprint arXiv:2505.11835*.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. *arXiv preprint arXiv:2106.06937*.

Fengran Mo, Zhan Su, Yuchen Hui, Jinghan Zhang, Jia Ao Sun, Zheyuan Liu, Chao Zhang, Tetsuya Sakai, and Jian-Yun Nie. 2026. Opendecoder: Open large language model decoding to incorporate document quality in rag. *arXiv preprint arXiv:2601.09028*.

Rui Qi, Zhibo Man, Yufeng Chen, Fengran Mo, Jinan Xu, and Kaiyu Huang. 2025. Sot: Structured-of-thought prompting guides multilingual reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11024–11039.

Rui Qi, Fengran Mo, Yufeng Chen, Xue Zhang, Shuo Wang, Hongliang Li, Jinan Xu, Meng Jiang, Jian-Yun Nie, and Kaiyu Huang. 2026. Language-coupled reinforcement learning for multilingual retrieval-augmented generation. *arXiv preprint arXiv:2601.14896*.

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*.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739.

Hao Wang, Pinzhi Huang, Jihan Yang, Saining Xie, and Daisuke Kawahara. 2025. Traveling across languages: Benchmarking cross-lingual consistency in multimodal llms. *arXiv preprint arXiv:2505.15075*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zihan Wang, Stephen Mayhew, Dan Roth, and 1 others. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 483–498.

Jinghan Zhang, Fengran Mo, Tharindu Cyril Weerasooriya, Ruimin Dai, Xiaoyan Han, Yanjie Fu, Dakuo Wang, and Kunpeng Liu. 2026. Starpo: Stability-augmented reinforcement policy optimization. *arXiv preprint arXiv:2604.08905*.

Jinghan Zhang, Xiting Wang, Fengran Mo, Yeyang Zhou, Wanfu Gao, and Kunpeng Liu. 2025a. Entropy-based exploration conduction for multi-step reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3895–3906.

Jinghan Zhang, Xiting Wang, Weijieying Ren, Lu Jiang, Dongjie Wang, and Kunpeng Liu. 2025b. Ratt: A thought structure for coherent and correct llm reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26733–26741.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.