A Comprehensive Evaluation of Multilingual Chain-of-Thought Reasoning: Performance, Consistency, and Faithfulness Across Languages

Anonymous ACL submission

Abstract

Large reasoning models (LRMs) increasingly rely on step-by-step Chain-of-Thought (CoT) reasoning to improve task performance, particularly in high-resource languages such as English. While recent work has examined final-answer accuracy in multilingual settings, the *thinking traces* themselves, i.e., the intermediate steps that lead to the final answer, remain underexplored. In this paper, we present the first comprehensive study of multilingual CoT reasoning, evaluating three key dimensions: *performance*, *consistency*, and *faithfulness*. We begin by measuring language compliance, answer accuracy, and answer consistency when LRMs are explicitly instructed or prompt-hacked to think in a target language, revealing strong language preferences and divergent performance across languages. Next, we assess *crosslingual consistency* of thinking traces by interchanging them between languages. We find that the quality and effectiveness of thinking traces vary substantially depending on the prompt language. Finally, we adapt perturbation-based techniques – i.e., *truncation* and *error injection* – to probe the *faithfulness* of thinking traces across languages, showing that models rely on traces to varying degrees. We will release our code and data to support future research.

1 Introduction

CoT prompting has emerged as a widely adopted technique for eliciting step-by-step *thinking traces* from LRMs (Wei et al., 2022; Kojima et al., 2022; Zhou et al., 2023). These traces have been shown to substantially improve model performance on complex reasoning tasks, while also offering an *interpretable* window for understanding the model's internal decision-making process (Grattafiori et al., 2024; OpenAI et al., 2024; Yang et al., 2025; DeepSeek-AI et al., 2025; Xu et al., 2025).

While most research on CoT reasoning has focused on English, the behavior of thinking traces in *multilingual* settings remains underexplored. A very recent line of studies has begun to examine LRM *performance* across languages, including scenarios where models are explicitly instructed or "forced" to reason in a specific language (Yong et al., 2025; Wang et al., 2025b; Qi et al., 2025). However, these efforts largely concentrate on final-answer accuracy, leaving open critical questions about the reasoning process itself, particularly: (1) *How* ***consistent*** *are the thinking traces across languages when answering semantically equivalent questions?* and (2) *To what extent are thinking traces* ***faithful*** *in languages other than English, especially the low-resource ones?*

To address these gaps, we conduct the first comprehensive evaluation of multilingual CoT reasoning across a diverse set of LRMs. Our study focuses on three core dimensions: *performance*, *consistency*, and *faithfulness*. In §4, we analyze language compliance, final-answer accuracy, and final-answer consistency when models are either explicitly instructed or prompt-hacked to think in a language aligned with the input. We find that LRMs exhibit strong language preferences during reasoning, and that performance varies substantially depending on the thinking language. To better understand these disparities, §5 introduces a novel method for *interchanging* thinking traces across languages. By substituting a thinking trace from one language into another, we assess whether reasoning is semantically aligned and transferable. Our results show that thinking traces are often inconsistent across languages, with quality varying largely by language. Surprisingly, we also find that final-answer accuracy is influenced not only by the thinking trace itself but also by the prompt language and thinking language. In §6, we evaluate the faithfulness of thinking traces. Extending prior monolingual work (Lanham et al., 2023), we apply perturbation-based interventions – such as truncation and error injection – and measure how these

changes impact model predictions. We find that models rely on their thinking traces to varying degrees across languages, suggesting that faithfulness is not uniformly preserved in multilingual contexts.

Overall, we make the following contributions: **(i)** We present the first comprehensive evaluation of multilingual CoT reasoning, covering three core dimensions – *performance*, *consistency*, and *faithfulness*. **(ii)** We propose a novel strategy: crosslingual thinking trace interchanging, to measure the semantic consistency of thinking traces across languages. **(iii)** We find that consistency of thinking traces varies across languages, and even with identical traces, accuracy is influenced by the language of the prompt. **(iv)** We show that languages other than English exhibit greater reliance on thinking traces, and that this reliance decreases as model scale increases. **(v)** We will release our code to facilitate future research on the evaluation of consistency and faithfulness in multilingual reasoning.

2 Related Work

Faithfulness in CoT Reasoning CoT prompting (Wei et al., 2022) has been shown to substantially improve the performance of LRMs across a variety of complex tasks (OpenAI et al., 2024; Snell et al., 2024; Muennighoff et al., 2025; DeepSeek-AI et al., 2025). Despite these gains, recent studies have raised concerns about the *faithfulness* of the generated thinking traces, i.e., whether the model's stated CoT truly reflects its internal decision-making process (Lyu et al., 2023; Turpin et al., 2023; Lanham et al., 2023; Tanneru et al., 2024; Arcuschin et al., 2025). One line of work evaluates faithfulness by introducing biases into the prompt, such as reordering multiple-choice options or injecting misleading arguments, and examining whether the model's answer changes accordingly (Turpin et al., 2023; Wang et al., 2024; Chua et al., 2025). Another line of work manipulates the thinking trace itself, e.g., by truncating it or inserting errors, and observes how such changes affect the model's prediction (Lanham et al., 2023; Yee et al., 2024; Xiong et al., 2025). These studies generally reveal that models may produce a thinking trace that is disconnected from the actual decision path leading to the answer. Our work builds on this latter line by extending it to *multilingual* settings. We manipulate thinking traces across languages, addressing the gap that existing studies evaluate faithfulness almost exclusively in English.

Evaluation of Multilingual Reasoning A growing body of work has evaluated CoT reasoning across languages (Shi et al., 2023a; Huang et al., 2023; Qin et al., 2023; Ahuja et al., 2023), showing that CoT prompting improves performance on a variety of multilingual tasks. More recent studies explore how manipulating the thinking trace, such as increasing the generation budget at test time or enforcing language-specific reasoning, can further affect model performance (Yong et al., 2025; Wang et al., 2025b; Qi et al., 2025). These works highlight that models often benefit from reasoning in high-resource languages like English or Chinese, or from being given more space to reason. However, existing multilingual reasoning evaluation studies almost exclusively focus on *performance*, overlooking whether models behave consistently across languages – that is, whether they produce correct answers consistently and whether the thinking traces themselves are semantically equivalent across languages. Such questions are especially important as LRMs are increasingly deployed in multilingual contexts (Ghosh et al., 2025). Our work moves beyond performance to offer a systematic evaluation of multilingual CoT reasoning along two additional dimensions: *consistency* and *faithfulness*, providing a complementary perspective on how reasoning behavior generalizes across languages.

3 Experimental Setup

3.1 Models

We evaluate a wide range of open-source LRMs of different model sizes. We consider the distilled versions of `DeepSeek-R1` (DeepSeek-AI et al., 2025): `DeepSeek-R1-Distill-Qwen-{1.5B, 7B, 14B, 32B}` whose base models are from `Qwen2.5` family (Qwen Team et al., 2025) and `DeepSeek-R1-Distill-Llama-{8B, 70B}` whose base models are from `Llama3` family (Grattafiori et al., 2024). Additionally, we consider two models from `Qwen3` family (Yang et al., 2025): `Qwen3-{8B, 32B}`.

3.2 Dataset

MMMLU Multilingual MMLU (Hendrycks et al., 2021) is a large-scale benchmark of general knowledge across various domains, such as Humanities and STEM, in a multiple-choice-question format, covering 15 typologically different languages.

MGSM Multilingual Grade School Math (Shi et al., 2023b) is a benchmark that contains 250

grade-school math problems from the GSM8K (Cobbe et al., 2021) (originally in English) that are manually translated into 10 additional languages.

3.3 Controlling Thinking Languages

Our motivation is that the language used for reasoning should match the language of the question, as users are very likely to prefer inspecting the reasoning process in the same language they use to pose the query. Accordingly, we consider two strategies for controlling the *thinking language*, i.e., the language used in the thinking trace (text generated between the special tokens <think> and </think>), to ensure it aligns with the *prompt language*, i.e., the language used in the original question.

Explicit Instruction The first strategy appends an *explicit* instruction to the prompt, directly asking the model to think in a particular language. For example, to elicit German reasoning, we insert the phrase "`Bitte denken Sie immer auf Deutsch.`" ["Please always think in German."] into the prompt. While intuitive, this approach seems less reliable: models may still default to their preferred thinking language, typically English, regardless of the instruction (Yong et al., 2025; Wang et al., 2025b).

Prompt Hacking The second strategy uses *prompt hacking*, a more targeted method to steer the model's language use (Schulhoff et al., 2023; Benjamin et al., 2024; Qi et al., 2025). Here, a short *prefix* in the desired language (e.g., "`Auf Anfrage werde ich anfangen, auf Deutsch zu denken.`" ["By Request, I will begin to think in German"] is inserted directly after the <think> token. The model is then expected to generate the remainder of the thinking trace, until the </think> token. This approach has been shown to be more effective than explicit instructions, often leading to language-consistent CoT generation that aligns with the prefix (Yong et al., 2025; Qi et al., 2025).

4 Language Compliance, Answer Accuracy, and Consistency

In this section, we evaluate the multilingual reasoning performance of LRMs under the two language control strategies introduced in §3.3. Using the metrics defined in §4.1, we assess each model's language compliance, final-answer accuracy, and crosslingual answer consistency. These results, presented and discussed in §4.2, allow us to examine the effectiveness of language control mechanisms and how the choice of thinking language influences model behavior. This multi-dimensional evaluation provides a foundation for our deeper analyses in later sections, particularly regarding reasoning consistency and faithfulness across languages.

4.1 Evaluation Metrics

Language Compliance Rate This metric measures the proportion of text within the thinking trace – i.e., between the special tokens <think> and </think> – that is generated in the intended target language (the prompt language). To compute this, we first split each thinking trace into individual sentences and then identify the language of each sentence using GlotLID (Kargaran et al., 2023). We then compute the overall proportion of reasoning content generated in the corresponding prompt language, following prior work (Yong et al., 2025; Wang et al., 2025b; Qi et al., 2025).[1]

Final Answer Accuracy This metric evaluates the correctness of the model's final prediction: $\mathrm{ACC}(l) = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} \mathbf{1}\left[\mathcal{M}(q_i^l) = o_i^l\right]$ where $\mathcal{D}$ is the dataset, $\mathcal{M}(q_i^l)$ the model prediction, and o_i^l the gold answer for question i. We compute accuracy *independently* for each language l.

Final Answer Consistency This metric quantifies the *crosslingual consistency* of model predictions. Given the same question posed in two languages, we evaluate whether the model produces the same and correct answer in both languages:

$$\mathrm{CO}(l_1, l_2) = \frac{\sum_{i=1}^{|\mathcal{D}|} \mathbf{1}[\mathcal{M}(q_i^{l_1})=o_i^{l_1} \wedge \mathcal{M}(q_i^{l_2})=o_i^{l_2}]}{\sum_{i=1}^{|\mathcal{D}|} \mathbf{1}[\mathcal{M}(q_i^{l_1})=o_i^{l_1} \vee \mathcal{M}(q_i^{l_2})=o_i^{l_2}]}$$

Consistency is widely used as a metric in knowledge probing, factual knowledge recall, and cultural awareness evaluation (Jiang et al., 2020; Wang et al., 2025a; Zhao et al., 2025b; Liu et al., 2025; Zhao et al., 2025a).

4.2 Results and Discussion

Table 1 reports the accuracy and language compliance rates of the evaluated LRMs on **MMMLU** across languages. Figure 1 illustrates the consistency across languages for R1-Qwen-32B and R1-Llama-70B (see §A.1 for additional results).

[1]In addition, we report the language usage distributions for English and Chinese across prompts in other languages, as well as token-level language distributions, in §A.1.

Method	Model	ar	bn	de	en	es	fr	hi	id	it	ja	ko	pt	sw	yo	zh
Explicit Instruction	R1-Qwen-1.5B	.25 (.07)	.24 (.03)	.35 (.24)	.47 (.95)	.32 (.89)	.35 (.43)	.28 (.05)	.21 (.20)	.30 (.51)	.24 (.14)	.32 (.02)	.29 (.81)	.19 (.40)	.21 (.12)	.37 (.85)
	R1-Qwen-7B	.24 (.69)	.34 (.20)	.41 (.95)	.58 (.97)	.42 (.96)	.42 (.91)	.31 (.78)	.52 (.92)	.44 (.90)	.34 (.83)	.38 (.17)	.41 (.97)	.21 (.07)	.26 (.16)	.54 (.87)
	R1-Qwen-14B	.66 (.78)	.56 (.02)	.66 (.17)	.76 (.97)	.67 (.48)	.62 (.17)	.47 (.04)	.67 (.17)	.68 (.24)	.65 (.32)	.66 (.05)	.71 (.12)	.32 (.08)	.31 (.09)	.67 (.85)
	R1-Qwen-32B	.65 (.70)	.54 (.17)	.70 (.87)	.78 (.96)	.71 (.38)	.73 (.05)	.53 (.19)	.72 (.05)	.73 (.50)	.68 (.37)	.70 (.15)	.77 (.08)	.38 (.40)	.28 (.20)	.74 (.88)
	Qwen-14B	.70 (.78)	.69 (.00)	.74 (.01)	.77 (.96)	.75 (.01)	.72 (.01)	.70 (.00)	.74 (.01)	.72 (.00)	.71 (.00)	.69 (.00)	.77 (.01)	.48 (.02)	.42 (.07)	.71 (.67)
	Qwen-32B	.63 (.70)	.51 (.00)	.76 (.01)	.81 (.95)	.63 (.01)	.68 (.01)	.58 (.00)	.58 (.02)	.66 (.01)	.64 (.00)	.64 (.00)	.64 (.02)	.43 (.02)	.33 (.05)	.78 (.81)
	R1-Llama-8B	.39 (.88)	.35 (.01)	.50 (.29)	.69 (.96)	.49 (.53)	.50 (.80)	.52 (.02)	.51 (.20)	.51 (.25)	.41 (.57)	.55 (.06)	.58 (.60)	.20 (.21)	.30 (.15)	.54 (.92)
	R1-Llama-70B	.76 (.83)	.51 (.01)	.78 (.05)	.84 (.95)	.68 (.09)	.76 (.10)	.65 (.04)	.66 (.12)	.72 (.02)	.66 (.29)	.74 (.03)	.74 (.04)	.63 (.06)	.40 (.09)	.76 (.86)
Prompt Hacking	R1-Qwen-1.5B	.08 (.75)	.14 (.97)	.23 (.82)	.40 (.97)	.30 (.90)	.25 (.96)	.15 (.86)	.22 (.69)	.31 (.94)	.24 (.56)	.10 (.40)	.31 (.65)	.05 (.66)	.09 (.73)	.36 (.89)
	R1-Qwen-7B	.29 (.66)	.27 (.92)	.37 (.95)	.58 (.97)	.48 (.96)	.48 (.93)	.29 (.84)	.54 (.91)	.42 (.96)	.34 (.74)	.23 (.61)	.43 (.97)	.04 (.76)	.08 (.97)	.55 (.91)
	R1-Qwen-14B	.61 (.73)	.34 (.94)	.60 (.94)	.72 (.97)	.65 (.95)	.68 (.97)	.36 (.77)	.58 (.98)	.68 (.97)	.61 (.93)	.65 (.97)	.65 (.98)	.27 (.95)	.23 (.75)	.65 (.92)
	R1-Qwen-32B	.66 (.78)	.49 (.91)	.71 (.97)	.80 (.96)	.74 (.96)	.75 (.97)	.52 (.80)	.73 (.98)	.70 (.96)	.67 (.73)	.70 (.97)	.73 (.98)	.35 (.94)	.25 (.94)	.75 (.85)
	Qwen3-14B	.62 (.73)	.52 (.91)	.64 (.89)	.71 (.96)	.68 (.96)	.64 (.94)	.54 (.76)	.64 (.97)	.67 (.96)	.63 (.86)	.62 (.98)	.67 (.97)	.24 (.96)	.20 (.91)	.70 (.90)
	Qwen3-32B	.62 (.78)	.65 (.85)	.67 (.59)	.80 (.97)	.55 (.42)	.55 (.50)	.73 (.67)	.70 (.80)	.68 (.40)	.63 (.73)	.71 (.86)	.77 (.61)	.40 (.91)	.31 (.91)	.74 (.88)
	R1-Llama-8B	.32 (.88)	.25 (.81)	.48 (.87)	.69 (.94)	.44 (.95)	.54 (.94)	.43 (.87)	.40 (.98)	.48 (.95)	.33 (.89)	.42 (.87)	.51 (.95)	.16 (.89)	.20 (.85)	.49 (.78)
	R1-Llama-70B	.70 (.83)	.62 (.91)	.76 (.81)	.86 (.95)	.79 (.78)	.78 (.93)	.75 (.73)	.71 (.94)	.74 (.88)	.70 (.85)	.74 (.77)	.80 (.94)	.62 (.90)	.36 (.94)	.76 (.87)

Table 1: Final-answer accuracy with sentence-level language compliance rates (in parentheses) for different LRMs across languages on the MMMLU task under two language-control strategies: *explicit instruction* and *prompt hacking*. Results for MGSM and token-level compliance rates are provided in §A.1.

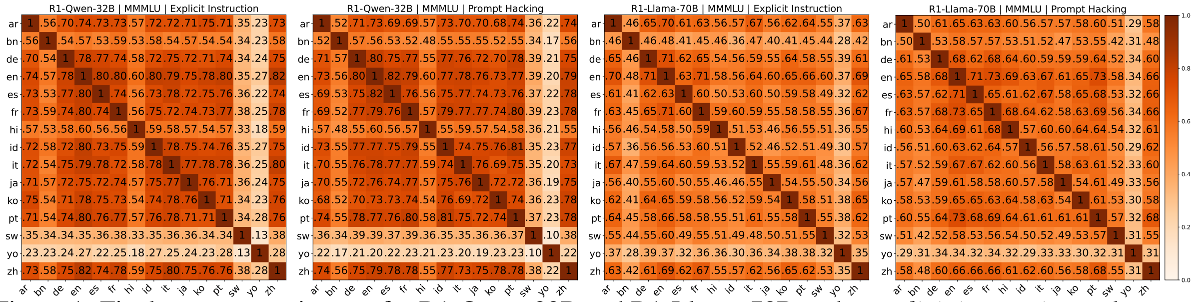

Figure 1: Final-answer consistency for R1-Qwen-32B and R1-Llama-70B under *explicit instruction* and *prompt hacking*. Similar language pairs, such as German and English, show higher consistency. Each cell shows the final-answer consistency between the language on the x-axis and the language on the y-axis.

Enforcing target-language reasoning improves compliance but may harm performance. When models are explicitly instructed to think in the same language as the prompt language, many fail to follow the instruction – especially in lower-resource languages such as Bengali (bn) and Yoruba (yo), which show low language compliance rates (less than 0.2 across all models).[2] Prompt hacking, by contrast, substantially improves language compliance, leading to very high alignment between the prompt and thinking language. However, this improved compliance often comes at the cost of reduced final-answer accuracy. For instance, while all models achieve remarkable compliance when forced to reason in Yoruba via prompt hacking, their accuracy drops substantially compared to explicit instruction. This reveals a trade-off between language control and task performance, consistent with findings in prior work (Qi et al., 2025).

Answer consistency reflects typological proximity across languages. Under both prompt hacking and explicit instruction setup, we observe that models exhibit high answer consistency across typologically similar languages. For instance, consistencies among Indo-European languages, e.g., English (en), German (de), and French (fr), tend to be high. To verify this, we compute average consistency for Indo-European language pairs versus mixed pairs (i.e., one Indo-European and one non–Indo-European), and find that the former is significantly higher (cf. Table 4 in §A.1.2). This suggests that *models reason similarly in these related languages*. In R1-Llama-70B, though the scaling improves the answer accuracy (cf. Table 1), the answer consistency is lower than that of its counterpart R1-Qwen-32B (cf. Figure 1), possibly due to different underlying base models. Nevertheless, the same trend holds: consistency is generally high among typologically similar languages.

Performance disparities persist across language control strategies, reflecting data exposure during training. Models consistently perform better on high-resource languages such as English and Chinese, which are overrepresented in pretraining and instruction tuning. In contrast, low-resource languages like Swahili and Yoruba yield lower accuracy in both explicit instruction and prompt hacking setups. These persistent gaps raise one core research question: *Why do models show different performance when the actual thinking languages vary, even with semantically equivalent prompts?* We hypothesize that this effect stems from inconsistencies in the quality and semantics of the thinking traces across languages – which we explore in §5.

[2]Models typically default to English reasoning. See §A.1 for the English proportion in different thinking traces.

Summary. Our analysis reveals three key patterns. First, models do not follow explicit instructions well, and while prompt hacking effectively enforces target-language reasoning, it often reduces accuracy. Second, consistency is high between similar languages. Finally, substantial performance gaps persist across languages regardless of control strategy, suggesting that inconsistencies in thinking trace quality may drive these gaps.

5 Consistency of Thinking Traces

We hypothesize that the disparities in final-answer accuracy and consistency observed in §4 stem primarily from the *quality* and *semantic inconsistency* of the generated thinking traces across languages. To investigate this, we introduce several novel substitution methods to evaluate the consistency of thinking traces between languages (§5.1). We further propose a new metric, *substitution consistency*, which quantifies how model predictions change before and after thinking trace substitution (§5.2). We then present and interpret our findings in §5.3.

5.1 Thinking Trace Interchanging

To better understand the disparities in thinking traces across languages, we consider three crosslingual substitution methods, each revealing different aspects of multilingual reasoning behavior. For the substitution, we reuse the thinking traces obtained for different languages in §4 and ask the model to directly generate the final answer based on the prompts and substituted thinking traces.

`BaseSub` In this setup, we interchange the thinking traces between languages l_1 and l_2 that are generated under *explicit instruction*. Since models often default to high-resource languages, even when instructed otherwise, many of these traces are in English regardless of the prompt language. This method allows us to understand why models present different performance even though the *thinking language* remains roughly flexible, but the *prompt language* varies.

`HackSub` Here, we interchange thinking traces generated under the *prompt hacking* setup. In this case, the thinking traces typically align with the prompt language due to the strong language control enforced by hacking prefixes. By interchanging these language-specific thinking traces between languages l_1 and l_2, we can examine how *consistent* the thinking traces are.

`TransSub` We first translate the thinking traces obtained under the *prompt hacking* setup into **English** using the Google Translate API.[3] We then interchange the translated English traces between language pairs l_1 and l_2. This setup removes the confounding variable of thinking language by standardizing all thinking traces to English. It provides a controlled environment to assess the quality of the generated thinking traces, independent of their original thinking languages.

5.2 Substitution Consistency

Beyond the final-answer accuracy defined in §4.1, we introduce *substitution consistency* to quantify how a model's predictions in language l change after its thinking trace is substituted with one from another language l'. Formally, let C_l denote the set of indices of questions for which the model produces a *correct* prediction in language l under the original thinking trace, and let $C_{l'\to l}$ denote the set of indices for which the model produces a correct prediction in l after the thinking trace from l' is substituted into l. We compute substitution consistency as the intersection-over-union (IoU) between these two sets: $CO(l', l) = \frac{|C_l \cap C_{l'\to l}|}{|C_l \cup C_{l'\to l}|}$. Intuitively, $CO(l', l)$ measures how stable the model's correct predictions in l remain after replacing its thinking trace with one from l'. Note that this metric is not symmetric – i.e., $CO(l', l) \neq CO(l, l')$ – because it specifically evaluates how the predictions in l change when thinking traces from another language are introduced.

5.3 Results and Discussion

Figure 2 and Figure 3 show the *final-answer accuracy* and *substitution consistency*, respectively, of R1-Qwen-14B under the three substitution strategies introduced in §5.1 for **MGSM**, where thinking traces are interchanged across languages.

Interchanging thinking traces substantially affects performance, revealing quality disparities across languages. We find that substituting thinking traces from one language into another often leads to large performance shifts. Low-resource languages generally benefit from substitution with high-resource thinking traces, while high-resource languages tend to suffer performance degradation when traces from low-resource languages are injected. For example, under the `HackSub` strategy, the accuracy of Chinese (zh) drops to 0.40 when

[3] https://cloud.google.com/translate

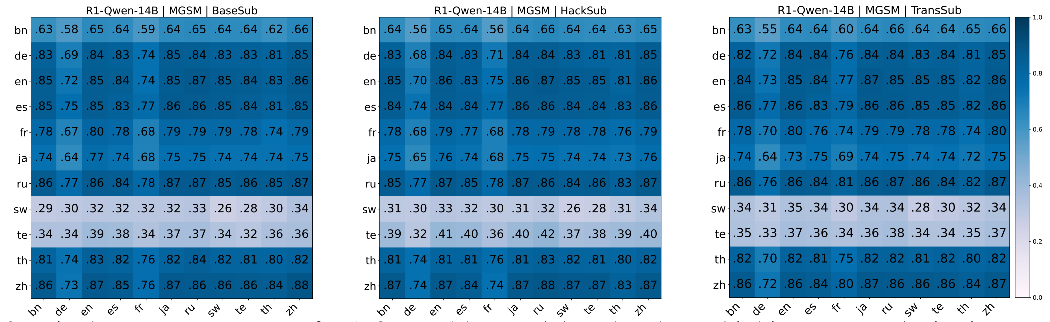

Figure 2: Final-answer accuracy of R1-Qwen-14B model under three thinking trace substitutions: `BaseSub`, `HackSub`, and `TransSub`. Each cell shows the accuracy when injecting thinking traces from a language on the y-axis into a language on the y-axis. Performance disparities indicate that thinking trace quality varies across languages.

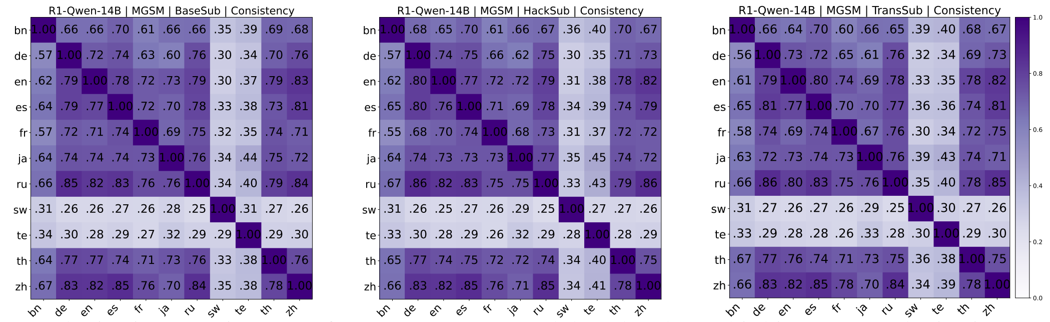

Figure 3: Substitution consistency of R1-Qwen-14B model under three thinking trace substitutions: `BaseSub`, `HackSub`, and `TransSub`. Each cell indicates the consistency between the original predictions in the language on the x-axis and the predictions after injecting thinking traces from the language on the y-axis. Higher consistency is observed when traces are substituted between similar languages.

thinking traces from Telugu (te) are used. Conversely, Telugu's accuracy rises to 0.87 when using traces from Chinese. This pattern is consistent across all three substitution strategies, suggesting that the quality of thinking traces varies dramatically across languages.

Substitution consistency is high between typologically-similar or resource-rich language pairs. We find that interchanging thinking traces between typologically-similar languages – such as English and German – yields relatively high substitution consistency. In contrast, substitution between more distant language pairs (e.g., Bengali and French) results in lower consistency (e.g., 0.61 in `BaseSub`). We further verify this by comparing the substitution consistency among different language pairs in Table 10 in §A.2. We also observe that language pairs where both languages are high-resource – i.e., well-represented in the model's pretraining data – tend to exhibit higher consistency. These findings suggest that the semantic consistency of thinking traces is easier to preserve when languages share language/geographic similarity or strong pretraining exposure.

Thinking traces alone do not fully determine final-answer accuracy. While thinking trace quality plays a major role in performance, it is not the only factor. We observe cases where models perform better in high-resource languages even when using identical thinking traces from low-resource languages. For example, when Swahili traces are injected into English prompts, the model achieves 0.33 accuracy – higher than Swahili's own original accuracy of 0.26 in `HackSub` (cf. Figure 2). This indicates that the *prompt language* also influences performance.

Models sometimes leverage English thinking traces better, even when semantically equivalent. An interesting pattern emerges when comparing `HackSub` and `TransSub` for Swahili accuracy (cf. Figure 2). When Swahili or Telugu thinking traces are first translated into English and then injected into other languages, the model usually achieves higher accuracy than when using the original Swahili or Telugu traces directly. This suggests that models are better at utilizing thinking traces expressed in English, even when the underlying semantics remain unchanged. This bias toward English may stem from both pretraining exposure and instruction tuning in English-heavy corpora.

Summary. Our analysis shows that the quality of thinking traces is highly uneven across languages,

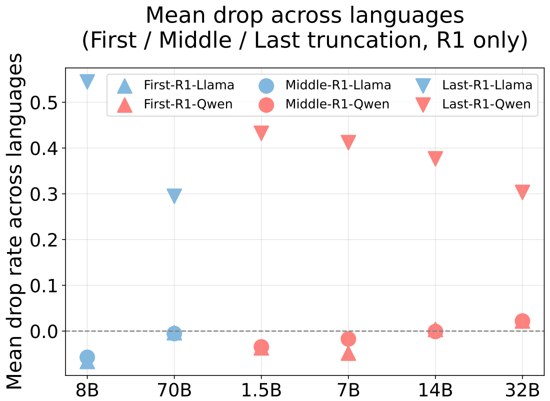

Figure 4: Mean accuracy drop (percentage) across languages for R1 distilled models under truncation of different parts of the thinking trace: first, middle, or last.

possibly shaped by both resource availability and inherent model biases. Semantic consistency of thinking traces across languages is also suboptimal, indicating that models do not generate equally aligned reasoning in different languages. Lastly, our results highlight that final-answer accuracy is jointly influenced by the *prompt language*, *thinking language*, and the *thinking trace*.

6 Faithfulness of Thinking Traces

In §5, we observed that thinking traces generated by LRMs are not consistent and not of the same quality across languages. In this section, we go one step further to explore the question: *Are thinking traces faithful across languages?* That is, do the generated reasoning steps reflect the actual reasoning process by which the model arrives at its final answer? Prior monolingual studies have shown that traces can be *unfaithful* (Lanham et al., 2023). However, whether this generalizes to other languages remains largely unexplored. To address this gap, we perturb the thinking traces and measure how these changes affect model predictions across languages. We describe the perturbation strategies in §6.1 and present results and discussion in §6.2.

6.1 Adding Perturbations to Thinking Traces

Following Lanham et al. (2023), we adopt two perturbation strategies to evaluate whether the model's final answer depends on its thinking traces. The more a model's predictions are influenced by changes to the thinking trace, the more faithfully it appears to use those traces during inference.

Trace Truncation In this setting, we truncate the thinking trace at different points and observe how the final answer changes. If the model's answer remains unchanged despite the removal of reasoning steps, this suggests that the original trace may have been post-hoc or ignored during inference. Concretely, for each generated trace, we segment the reasoning steps into *three* equal parts and perform targeted truncations: removing the *first* part, the *middle* part, or the *last* part. We then compare the model's predictions under each truncated trace to those obtained with the full thinking trace. This setup allows us to identify not only whether truncation affects predictions but also *which part* exerts the greatest influence across languages.

Error Injection In this setting, we introduce a small error into the *last sentence* of the thinking trace – by altering a number involved in the final computation step (e.g., changing it to another number). This design specifically targets the final stage of reasoning, where the model is expected to derive or summarize the correct answer. The goal is to assess whether the model relies on the correctness of the concluding reasoning step. If the model's answer changes in response to this minimal perturbation, it suggests that it is faithfully using its own reasoning. On the other hand, if the answer remains unchanged, despite the final reasoning step being incorrect, this may indicate that the model is ignoring its stated trace or relying on earlier steps, memorized patterns, or even contamination instead.

6.2 Results and Discussion

Table 2 reports the change in final-answer accuracy for each language on the **MGSM** dataset under two perturbation strategies – *last-part truncation* and *error injection* – in the HackSub setting. Figure 4 further visualizes the effect of truncating the first, middle, and last segments of the thinking trace across models in the DeepSeek-R1 distilled series.

Models show varying degrees of faithfulness across languages. We observe diverse sensitivity to perturbations across languages. For some low-resource languages like Swahili and Telugu, perturbations have little impact in R1 distilled models – largely because original performance is already low. In contrast, many languages experience substantial accuracy drops, suggesting that models do rely on their thinking traces to varying extents. This is further supported by our matching ratio results in Table 3, which measure how often the final predictions are influenced by the incorrect numbers injected into the trace. Notably, English consistently shows lower matching scores (e.g., 0.12 for

Operation	Model	de	en	es	fr	ja	ru	sw	th	zh	bn	te
Truncation (Last)	Qwen-14B	.68 (.75)	.75 (.79)	.69 (.79)	.64 (.77)	.69 (.82)	.68 (.75)	.40 (.84)	.65 (.72)	.67 (.79)	.66 (.83)	.63 (.87)
	Qwen-32B	.66 (.83)	.76 (.86)	.55 (.81)	.39 (.72)	.62 (.87)	.60 (.81)	.56 (.89)	.77 (.86)	.60 (.75)	.77 (.89)	.60 (.88)
	R1-Qwen-7B	.28 (.45)	.43 (.51)	.41 (.54)	.32 (.53)	.22 (.42)	.36 (.50)	.03 (.67)	.22 (.43)	.29 (.36)	.06 (.12)	.00 (.00)
	R1-Qwen-14B	.40 (.52)	.32 (.38)	.29 (.36)	.35 (.44)	.24 (.31)	.39 (.44)	.12 (.48)	.30 (.37)	.24 (.27)	.24 (.37)	.06 (.20)
	R1-Qwen-32B	.30 (.35)	.26 (.27)	.35 (.40)	.21 (.26)	.30 (.34)	.29 (.32)	.20 (.43)	.22 (.25)	.14 (.16)	.20 (.25)	-.22 (-1.22)
	R1-Llama-8B	.38 (.71)	.59 (.70)	.54 (.77)	.45 (.71)	.28 (.61)	.48 (.72)	.01 (.18)	.16 (.43)	.44 (.60)	.01 (.18)	.04 (.38)
	R1-Llama-70B	.36 (.44)	.22 (.23)	.38 (.42)	.35 (.41)	.28 (.34)	.39 (.43)	.32 (.37)	.26 (.31)	.29 (.33)	.11 (.15)	-.09 (-.19)
Error Inject.	Qwen-14B	.20 (.22)	.17 (.18)	.15 (.17)	.14 (.17)	.22 (.27)	.32 (.35)	.07 (.15)	.20 (.23)	.20 (.24)	.15 (.19)	.16 (.22)
	Qwen-32B	.10 (.12)	.22 (.25)	.03 (.04)	-.09 (-.16)	.08 (.12)	.10 (.13)	.08 (.13)	.13 (.14)	.18 (.23)	.17 (.20)	.32 (.47)
	R1-Qwen-7B	.56 (.92)	.74 (.87)	.68 (.89)	.54 (.89)	.44 (.83)	.66 (.92)	.03 (.58)	.43 (.82)	.69 (.86)	.40 (.85)	.09 (.59)
	R1-Qwen-14B	.61 (.80)	.22 (.26)	.66 (.80)	.64 (.82)	.58 (.77)	.56 (.63)	.20 (.81)	.67 (.82)	.59 (.67)	.47 (.71)	.08 (.30)
	R1-Qwen-32B	.57 (.65)	.16 (.16)	.60 (.68)	.60 (.73)	.62 (.72)	.57 (.63)	.33 (.73)	.67 (.75)	.64 (.71)	.56 (.70)	-.01 (-.07)
	R1-Llama-8B	.47 (.88)	.71 (.84)	.62 (.89)	.55 (.88)	.38 (.82)	.53 (.80)	.03 (.73)	.27 (.72)	.63 (.87)	-.02 (-.29)	.04 (.35)
	R1-Llama-70B	.48 (.59)	.41 (.44)	.61 (.68)	.56 (.65)	.41 (.49)	.38 (.42)	.59 (.70)	.65 (.76)	.12 (.13)	.42 (.59)	.24 (.54)

Table 2: Performance absolute drop after perturbation – *last-part truncation* and *error injection* – compared to original accuracy. Relative drops (in percentage) are shown in parentheses. Higher drops indicate greater sensitivity to the thinking trace perturbation, and therefore can be interpreted as stronger faithfulness.

Model	de	en	es	fr	ja	ru	sw	th	zh	bn	te
R1-Qwen-1.5B	.61	.46	.51	.58	.49	.53	.28	.16	.69	.38	.22
R1-Qwen-7B	.57	.56	.62	.46	.62	.69	.53	.59	.71	.62	.51
R1-Qwen-14B	.50	.26	.58	.54	.63	.53	.59	.62	.54	.53	.28
R1-Qwen-32B	.43	.12	.51	.46	.57	.54	.60	.61	.59	.56	.40
Qwen-14B	.06	.07	.06	.04	.08	.06	.10	.07	.03	.05	.06
Qwen-32B	.07	.02	.06	.06	.02	.05	.04	.03	.02	.04	.20
R1-Llama-8B	.48	.52	.50	.42	.65	.56	.33	.44	.63	.55	.38
R1-Llama-70B	.26	.24	.41	.34	.31	.33	.53	.56	.07	.38	.49

Table 3: Per-language matching ratio for each model, indicating the proportion of predictions that match the incorrect number injected into the final sentence of the thinking trace. Higher values suggest stronger reliance on the surface form of the reasoning.

R1-Qwen-32B) compared with other languages.

Truncating the final part is less disruptive than error injection, especially for R1 distilled models. Across all languages, we find that truncating the final segment of the thinking trace has less impact than injecting an incorrect value, particularly for the R1 distilled series.[4] This suggests that models may perform *latent-state reasoning* (Zhu et al., 2025), where inference continues internally even after the visible thinking trace is truncated. However, when a plausible but incorrect final value is injected, models often copy it directly, revealing that their outputs are sensitive to surface-level reasoning conclusions.

Model scale affects faithfulness behavior. As shown in Figure 4, model scale influences how different parts of the thinking trace affect predictions. Smaller models are more reliant on the final portion of the trace, aligning with their higher matching ratios (cf. Table 3), suggesting a higher degree of surface-level faithfulness. Larger models, by contrast, become less dependent on the final segments and more sensitive to earlier reasoning steps. This may indicate either (1) reduced faithfulness due to memorization or contamination, or (2) stronger latent-state reasoning capabilities that allow the model to recover from truncated traces or correct surface-level trace errors.

Summary. Our findings reveal that models vary in their faithfulness across languages and model scales. Languages other than English show stronger reliance on thinking traces. Truncating the final part of the trace is generally less disruptive than injecting incorrect information, especially for R1 distilled models, possibly suggesting latent-state reasoning. Larger models are less dependent on surface-level reasoning and more resilient to perturbations, though this may reflect either increased reasoning ability or memorization.

7 Conclusion

In this paper, we present the first comprehensive evaluation of multilingual CoT reasoning across a diverse set of LRMs. We examine three core dimensions, *performance*, *consistency*, and *faithfulness*, to provide a deeper understanding of how LRMs reason across languages. We show that LRMs exhibit strong language preferences in reasoning and that final-answer performance varies substantially across languages. Through our *crosslingual thinking trace interchanging* method, we show that thinking traces are often inconsistent across languages, with their quality strongly associated with the thinking language. Finally, our perturbation-based tests reveal that models rely on the traces to varying degrees, suggesting that reasoning faithfulness is uneven across languages. Our findings highlight the need for more robust and transparent evaluation of multilingual reasoning behavior.

[4] Interestingly, the Qwen3 models show the opposite trend. We hypothesize that this may be due to *data contamination*, as seen in Table 3, where these models rarely change their answers even when the final sentence is corrupted.

Limitations

While our work provides the first comprehensive study of multilingual CoT reasoning, we acknowledge that several limitations remain.

First, although we examine robustness through two perturbation strategies and show that robustness varies across languages, more sophisticated or adversarial perturbations (e.g., paraphrasing, distractor reasoning) remain unexplored and could be incorporated in future work.

Second, while we evaluate and analyze inconsistencies in multilingual reasoning, we do not provide a mechanistic explanation for why these inconsistencies arise. Future research could apply mechanistic interpretability methods to investigate model internals to better understand the sources of multilingual inconsistency and faithfulness.

Finally, due to resource constraints, our experiments are limited in the number of models and downstream tasks considered. Future work could extend our evaluation framework to a broader set of models, languages, and tasks.

Ethical Considerations

Use of AI Assistants The authors acknowledge the use of ChatGPT exclusively for grammar correction, improving the clarity and coherence of the draft, and assisting with code implementation.[5]

References

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

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. Chain-of-thought reasoning in the wild is not always faithful. *Preprint*, arXiv:2503.08679.

Victoria Benjamin, Emily Braca, Israel Carter, Hafsa Kanchwala, Nava Khojasteh, Charly Landow, Yi Luo, Caroline Ma, Anna Magarelli, Rachel Mirin, Avery Moyer, Kayla Simpson, Amelia Skawinski, and Thomas Heverin. 2024. Systematically analyzing prompt injection vulnerabilities in diverse llm architectures. *Preprint*, arXiv:2410.23308.

[5]https://chatgpt.com/

James Chua, Edward Rees, Hunar Batra, Samuel R. Bowman, Julian Michael, Ethan Perez, and Miles Turpin. 2025. Bias-augmented consistency training reduces biased reasoning in chain-of-thought. *Preprint*, arXiv:2403.05518.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. The multilingual mind : A survey of multilingual reasoning in language models. *Preprint*, arXiv:2502.09457.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *Preprint*, arXiv:2307.13702.

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Felicia Körner, Ercong Nie, Barbara Plank, François Yvon, and Hinrich Schütze. 2025. Tracing multilingual factual knowledge acquisition in pretraining. *Preprint*, arXiv:2505.14824.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *Preprint*, arXiv:2501.19393.

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 244 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

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. *Preprint*, arXiv:2505.22888.

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.

Qwen Team, :, 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, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Sander Schulhoff, Jeremy Pinto, Anaum Khan, Louis-François Bouchard, Chenglei Si, Svetlina Anati, Valen Tagliabue, Anson Kost, Christopher Carnahan, and Jordan Boyd-Graber. 2023. Ignore this title and HackAPrompt: Exposing systemic vulnerabilities of LLMs through a global prompt hacking competition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4945–4977, 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. 2023a. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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. 2023b. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. 2024. On the hardness of faithful chain-of-thought reasoning in large language models. *Preprint*, arXiv:2406.10625.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025a. Lost in multilinguality: Dissecting crosslingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5075–5094, Vienna, Austria. Association for Computational Linguistics.

Mingyang Wang, Lukas Lange, Heike Adel, Yunpu Ma, Jannik Strötgen, and Hinrich Schütze. 2025b.

Language mixing in reasoning language models: Patterns, impact, and internal causes. *Preprint*, arXiv:2505.14815.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Zidi Xiong, Shan Chen, Zhenting Qi, and Himabindu Lakkaraju. 2025. Measuring the faithfulness of thinking drafts in large reasoning models. *Preprint*, arXiv:2505.13774.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *Preprint*, arXiv:2501.09686.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Evelyn Yee, Alice Li, Chenyu Tang, Yeon Ho Jung, Ramamohan Paturi, and Leon Bergen. 2024. Dissociation of faithful and unfaithful reasoning in llms. *Preprint*, arXiv:2405.15092.

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. *Preprint*, arXiv:2505.05408.

Raoyuan Zhao, Beiduo Chen, Barbara Plank, and Michael A. Hedderich. 2025a. Makieval: A multilingual automatic wikidata-based framework for cultural awareness evaluation for llms. *Preprint*, arXiv:2505.21693.

Raoyuan Zhao, Abdullatif Köksal, Ali Modarressi, Michael A Hedderich, and Hinrich Schütze. 2025b. Do we know what llms don't know? a study of consistency in knowledge probing. *arXiv preprint arXiv:2505.21701*.

Metric	Group	Mean Value	P-Value
Without Low Resource Languages			
Final-Answer Consistency	Indo-European	0.565	0.034
	Non Indo-European	0.551	
With Low Resource Languages			
Final-Answer Consistency	Indo-European	0.5384	3.88e-20
	Non Indo-European	0.4894	

Table 4: Consistency comparison between Indo-European and non-Indo-European languages. Reported are mean consistency values for Final-Answer consistency metrics, with corresponding p-values (t-test). We also discard low-resouce languages sw and yo to conduct t-test. Indo-European languages generally achieve higher consistency, and the differences are statistically significant.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Rui-Jie Zhu, Tianhao Peng, Tianhao Cheng, Xingwei Qu, Jinfa Huang, Dawei Zhu, Hao Wang, Kaiwen Xue, Xuanliang Zhang, Yong Shan, Tianle Cai, Taylor Kergan, Assel Kembay, Andrew Smith, Chenghua Lin, Binh Nguyen, Yuqi Pan, Yuhong Chou, Zefan Cai, and 14 others. 2025. A survey on latent reasoning. *Preprint*, arXiv:2507.06203.

A Additional Results

A.1 Complete Results for Language Controlling

This section presents the complete compliance results for all languages, as well as additional results on MGSM, covering accuracy, consistency, and compliance.

A.1.1 Accuracy of Final-answer

Table 5 reports the final-answer accuracy of different LRMs on the MGSM task under explicit instruction and prompt hacking. The results confirm our findings: explicit instructions are often not strictly followed, and while prompt hacking improves language control, it generally comes at the cost of accuracy. In particular, accuracy drops are most pronounced in low-resource languages (e.g., Swahili, Telugu, Bengali), whereas high-resource languages (e.g., English, German, French, Chinese) show relatively stable performance across both control strategies.

Method	Model	de	en	es	fr	ja	ru	sw	th	zh	bn	te
Explicit Instruction	Qwen-14B	.91	.93	.91	.87	.82	.91	.62	.89	.83	.81	.73
	Qwen-32B	.89	.94	.92	.88	.86	.92	.80	.88	.87	.84	.79
	R1-Qwen-1.5B	.42	.78	.50	.47	.22	.47	.01	.03	.66	.13	.03
	R1-Qwen-7B	.69	.85	.75	.74	.56	.77	.10	.51	.80	.49	.26
	R1-Qwen-14B	.84	.93	.89	.86	.80	.93	.29	.85	.87	.72	.40
	R1-Qwen-32B	.89	.96	.91	.87	.86	.93	.44	.89	.87	.81	.40
	R1-Llama-8B	.57	.78	.65	.62	.45	.66	.04	.39	.63	.17	.18
	R1-Llama-70B	.88	.96	.92	.89	.85	.93	.85	.89	.88	.84	.55
Prompt Hacking	Qwen-14B	.91	.95	.88	.84	.84	.90	.48	.90	.85	.79	.72
	Qwen-32B	.80	.88	.68	.54	.72	.74	.63	.90	.80	.86	.69
	R1-Qwen-1.5B	.40	.79	.47	.47	.26	.44	.00	.01	.65	.06	.04
	R1-Qwen-7B	.62	.85	.76	.61	.53	.72	.05	.52	.80	.47	.15
	R1-Qwen-14B	.76	.84	.82	.78	.76	.88	.25	.82	.88	.67	.28
	R1-Qwen-32B	.90	.96	.90	.87	.86	.93	.49	.90	.87	.83	.45
	R1-Llama-8B	.56	.78	.64	.62	.46	.66	.07	.42	.62	.24	.21
	R1-Llama-70B	.86	.96	.86	.83	.80	.92	.62	.90	.84	.79	.47

Table 5: Final-answer accuracy for different LRMs across languages on the MGSM task under two language-control strategies: explicit instruction and prompt hacking.

A.1.2 Consistency of Final-answer

Figures 5 and 6 show the consistency results on MMMLU and MGSM of LRMs. Across both tasks, the scaling trend remains stable: larger models exhibit higher consistency. However, we also observe that applying prompt hacking tends to affect the internal language consistency of models, sometimes reducing alignment compared to explicit instructions.

A.1.3 Language Compliance

Table 6 and Table 7 report sentence-level and token-level compliance statistics on MMMLU and MGSM. Overall, prompt hacking improves alignment with the target language, with stronger effects for low-resource languages. Token-level compliance is consistently lower than sentence-level compliance, likely due to the difficulty of identifying language from individual tokens and the presence of borrowings or quoted words within sentences.

Table 8 and Table 9 present the proportions of Chinese and English content in the reasoning traces across different models and prompt languages. We observe that, under explicit instruction, models tend to default to English traces even when prompted in another language, with Chinese traces also appearing but less frequently. Prompt hacking can reduce this misalignment by shifting the distribution toward the target language.

A.2 Complete Results for Interchanging Thinking traces

Figure 7 shows the accuracy of R1-Qwen models on MGSM with interchanged reasoning traces. Injecting traces from low-resource languages into high-resource prompts lowers performance, while high-resource traces can boost low-resource prompts. This confirms the strong influence of reasoning trace quality on final accuracy. Figure 8 shows that consistency with original setup after substitution is generally higher between similar languages. HackSub follows the same trend as BaseSub but sometimes lowers cross-language consistency, especially for low-resource languages.

A.3 Complete Results for Faithfulness

Table 11 reports accuracy changes when truncating the first, middle, or last part of the reasoning traces. We observe that truncating the middle or the beginning generally has smaller impact, while removing the final part leads to larger performance drops, highlighting the importance of the concluding reasoning steps.

B Details of Datasets

B.1 Language Coverage

Table 12 summarizes the language coverage of the two datasets used in our experiments: MMMLU and MGSM.

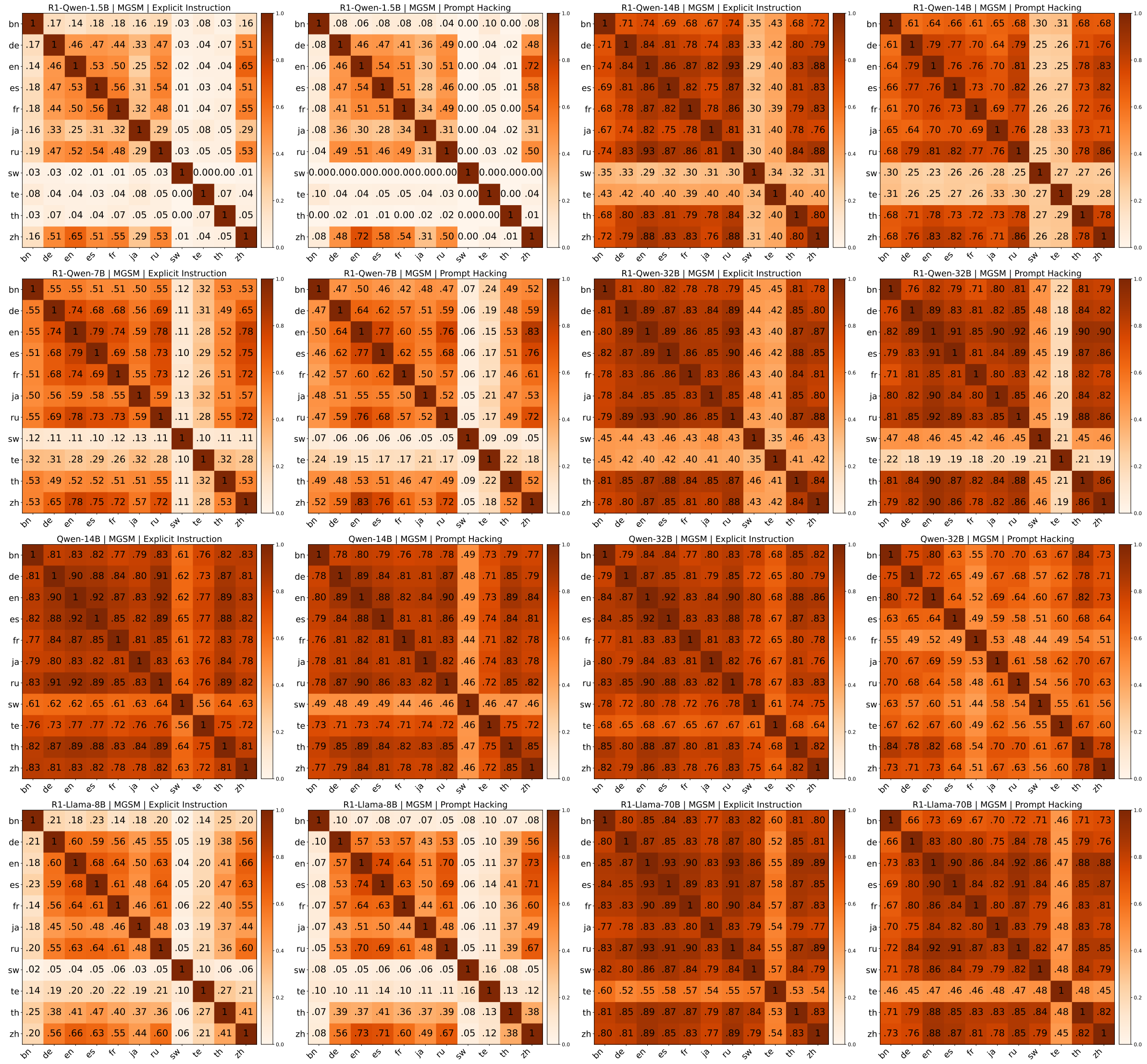

Figure 5: Final-answer consistency heatmaps on the MGSM dataset across different models under explicit instruction and prompt hacking.

Method	Model	bn	de	en	es	fr	ja	ru	sw	te	th	zh
Explicit Instruction	R1-Qwen-1.5B	.01/.00	.89/.29	.89/.38	.88/.27	.78/.12	.72/.43	.92/.42	.18/.05	.02/.00	.71/.82	.89/.56
	R1-Qwen-7B	.60/.04	.96/.32	.93/.37	.93/.28	.97/.15	.81/.56	.97/.47	.23/.07	.34/.03	.47/.34	.86/.55
	R1-Qwen-14B	.71/.06	.89/.29	.89/.38	.92/.28	.98/.14	.85/.73	.97/.48	.42/.12	.02/.00	.84/.50	.90/.63
	R1-Qwen-32B	.78/.05	.97/.31	.90/.38	.95/.28	.97/.15	.87/.70	.98/.48	.84/.28	.94/.08	.85/.52	.92/.58
	Qwen-14B	.01/.00	.02/.01	.87/.38	.02/.01	.02/.01	.01/.02	.97/.47	.04/.02	.02/.00	.00/.02	.81/.66
	Qwen-32B	.01/.00	.01/.01	.86/.38	.01/.01	.02/.00	.01/.02	.96/.47	.02/.01	.78/.07	.01/.01	.76/.63
	R1-Llama-8B	.45/.05	.93/.31	.93/.37	.93/.28	.97/.15	.85/.63	.97/.48	.92/.32	.02/.00	.84/.56	.90/.56
	R1-Llama-70B	.80/.06	.92/.31	.92/.38	.93/.28	.96/.14	.69/.53	.97/.49	.90/.24	.91/.07	.82/.52	.81/.67
Prompt Hacking	R1-Qwen-1.5B	.86/.06	.83/.32	.91/.39	.82/.28	.87/.13	.65/.13	.94/.42	.41/.03	.90/.08	.97/.58	.86/.56
	R1-Qwen-7B	.92/.07	.96/.36	.94/.39	.87/.30	.97/.15	.84/.27	.94/.43	.59/.34	.96/.08	.82/.45	.89/.61
	R1-Qwen-14B	.85/.07	.92/.33	.95/.40	.67/.21	.91/.14	.79/.63	.96/.47	.88/.30	.82/.07	.81/.52	.80/.58
	R1-Qwen-32B	.83/.06	.47/.16	.92/.39	.84/.27	.82/.13	.86/.68	.97/.47	.88/.31	.92/.10	.84/.56	.82/.64
	Qwen-14B	.89/.05	.03/.01	.90/.39	.82/.26	.78/.11	.87/.52	.96/.42	.87/.24	.92/.06	.88/.51	.88/.48
	Qwen-32B	.81/.04	.35/.20	.89/.38	.33/.14	.31/.06	.58/.35	.92/.43	.80/.23	.78/.07	.70/.47	.85/.40
	R1-Llama-8B	.75/.12	.90/.33	.91/.39	.88/.30	.93/.14	.83/.59	.89/.45	.67/.38	.46/.05	.88/.57	.83/.55
	R1-Llama-70B	.86/.06	.70/.29	.92/.39	.71/.26	.80/.13	.90/.59	.96/.49	.86/.26	.94/.07	.84/.53	.85/.61

Table 6: Sentence/Token Compliance Rate on **MGSM** across 11 languages.

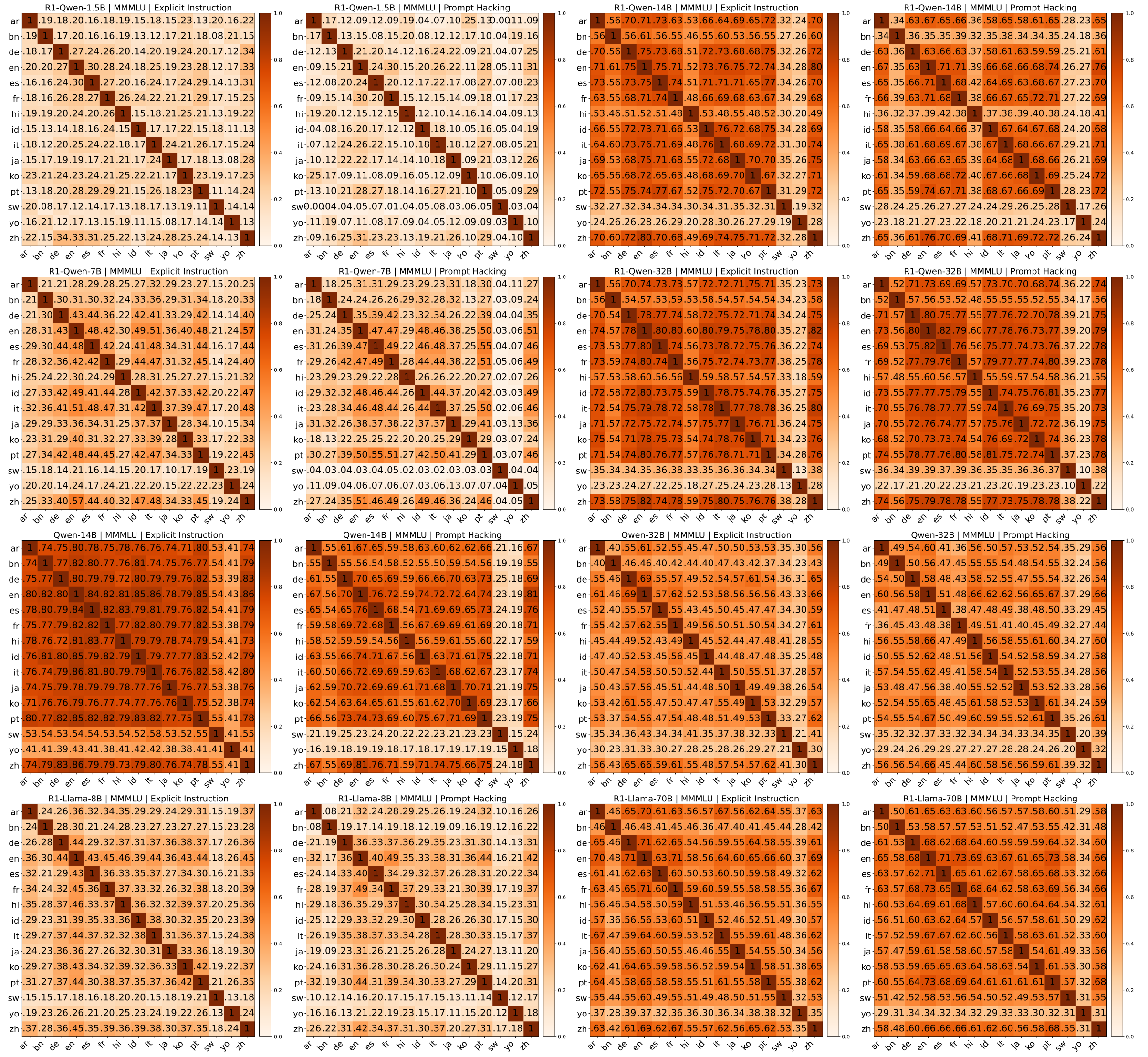

Figure 6: Final-answer consistency heatmaps on the MMLU dataset across different models under explicit instruction and prompt hacking.

Method	Model	ar	bn	de	es	fr	hi	id	it	ja	ko	pt	sw	yo	zh	en
Explicit Instruction	R1-Qwen-1.5B	.07/.05	.03/.01	.24/.08	.89/.29	.44/.09	.05/.06	.20/.12	.51/.28	.14/.16	.02/.03	.81/.32	.40/.34	.12/.02	.85/.63	.95/.45
	R1-Qwen-7B	.69/.38	.20/.04	.95/.37	.96/.30	.91/.16	.78/.36	.92/.32	.89/.44	.83/.26	.16/.10	.97/.38	.07/.03	.16/.03	.87/.67	.97/.46
	R1-Qwen-14B	.02/.01	.02/.01	.17/.07	.48/.18	.17/.03	.04/.14	.17/.05	.24/.12	.32/.39	.05/.04	.12/.05	.08/.02	.09/.02	.85/.69	.97/.46
	R1-Qwen-32B	.19/.13	.17/.05	.87/.38	.38/.15	.05/.01	.19/.27	.05/.01	.50/.26	.37/.37	.15/.11	.08/.04	.40/.17	.20/.04	.88/.68	.96/.45
	Qwen-14B	.00/.00	.00/.00	.01/.01	.01/.01	.01/.00	.00/.00	.01/.00	.00/.01	.00/.00	.00/.01	.01/.01	.02/.01	.07/.02	.67/.65	.96/.46
	Qwen-32B	.00/.00	.00/.00	.01/.01	.01/.01	.01/.00	.00/.00	.02/.01	.01/.01	.00/.00	.00/.01	.02/.01	.02/.01	.05/.01	.82/.69	.95/.46
	R1-Llama-8B	.52/.33	.01/.00	.29/.12	.53/.20	.80/.14	.02/.03	.20/.06	.25/.15	.57/.50	.06/.06	.60/.27	.21/.08	.15/.02	.92/.69	.96/.46
	R1-Llama-70B	.44/.27	.01/.01	.05/.02	.09/.04	.10/.02	.04/.11	.12/.03	.02/.02	.29/.20	.03/.02	.05/.02	.06/.01	.09/.01	.86/.68	.95/.46
Prompt Hacking	R1-Qwen-1.5B	.75/.38	.97/.07	.82/.34	.90/.29	.96/.16	.86/.35	.69/.33	.94/.47	.56/.31	.40/.35	.65/.38	.66/.58	.73/.11	.89/.56	.97/.47
	R1-Qwen-7B	.66/.37	.92/.07	.95/.39	.96/.32	.93/.16	.84/.36	.91/.32	.96/.48	.74/.25	.61/.32	.97/.41	.76/.56	.97/.10	.91/.61	.97/.47
	R1-Qwen-14B	.73/.43	.94/.08	.94/.40	.95/.32	.97/.17	.77/.35	.98/.33	.97/.49	.93/.74	.97/.86	.98/.41	.95/.36	.75/.16	.92/.62	.97/.47
	R1-Qwen-32B	.78/.44	.91/.07	.97/.40	.96/.32	.97/.17	.80/.35	.98/.32	.96/.49	.73/.68	.97/.85	.98/.42	.94/.40	.94/.26	.85/.62	.96/.46
	Qwen-14B	.78/.52	.90/.08	.89/.38	.96/.31	.93/.16	.76/.34	.97/.32	.96/.49	.86/.71	.98/.90	.97/.41	.96/.35	.91/.18	.90/.50	.96/.46
	Qwen-32B	.70/.49	.85/.06	.59/.34	.42/.18	.50/.10	.67/.34	.80/.27	.40/.25	.73/.61	.86/.80	.61/.34	.91/.33	.91/.19	.88/.58	.97/.48
	R1-Llama-8B	.88/.55	.81/.11	.87/.40	.95/.32	.94/.16	.87/.35	.98/.32	.95/.48	.89/.65	.87/.80	.95/.42	.89/.38	.85/.18	.78/.58	.94/.46
	R1-Llama-70B	.83/.55	.91/.07	.81/.38	.78/.28	.92/.17	.73/.35	.95/.28	.88/.46	.85/.61	.77/.69	.94/.43	.90/.33	.94/.18	.87/.59	.95/.47

Table 7: Compliance rates (sentence/token) for **MMMLU**, across 15 languages.

Model	Setting	bn	de	en	es	fr	ja	ru	sw	te	th	zh
R1-Qwen-1.5B	Explicit	.86/.37 | .00/.00	.03/.06 | .00/.00	.89/.38 | .00/.00	.00/.02 | .00/.00	.11/.09 | .00/.00	.02/.06 | .06/.05	.02/.04 | .00/.00	.73/.35 | .00/.00	.87/.38 | .00/.00	.20/.03 | .04/.00	.00/.00 | .89/.56
R1-Qwen-1.5B	Hacking	.03/.00 | .00/.00	.01/.05 | .00/.00	.91/.39 | .00/.00	.01/.03 | .00/.00	.00/.06 | .00/.00	.14/.40 | .05/.01	.04/.03 | .00/.00	.43/.08 | .00/.00	.01/.00 | .00/.00	.01/.01 | .00/.00	.00/.00 | .86/.56
R1-Qwen-7B	Explicit	.20/.04 | .00/.00	.00/.05 | .00/.00	.93/.37 | .00/.00	.00/.02 | .00/.00	.00/.05 | .00/.00	.05/.04 | .01/.01	.02/.02 | .00/.00	.58/.28 | .00/.00	.54/.20 | .00/.00	.36/.11 | .01/.01	.00/.00 | .86/.55
R1-Qwen-7B	Hacking	.00/.00 | .00/.00	.00/.05 | .00/.00	.94/.39 | .00/.00	.01/.02 | .00/.00	.00/.05 | .00/.00	.07/.12 | .01/.01	.01/.01 | .00/.00	.02/.04 | .01/.00	.00/.00 | .00/.00	.03/.02 | .00/.01	.00/.01 | .89/.61
R1-Qwen-14B	Explicit	.18/.02 | .00/.00	.08/.07 | .00/.00	.89/.38 | .00/.00	.00/.01 | .00/.00	.00/.05 | .00/.00	.05/.01 | .00/.01	.00/.00 | .00/.00	.44/.20 | .01/.01	.85/.36 | .00/.00	.00/.00 | .00/.00	.00/.00 | .90/.63
R1-Qwen-14B	Hacking	.00/.00 | .00/.00	.01/.04 | .00/.00	.95/.40 | .00/.00	.01/.01 | .21/.17	.00/.05 | .02/.02	.12/.05 | .01/.01	.00/.00 | .00/.00	.01/.00 | .00/.00	.03/.01 | .01/.01	.00/.00 | .00/.00	.00/.00 | .80/.58
R1-Qwen-32B	Explicit	.00/.00 | .00/.00	.00/.04 | .00/.00	.90/.38 | .00/.00	.00/.01 | .00/.00	.00/.05 | .00/.00	.00/.00 | .00/.01	.00/.00 | .00/.00	.01/.01 | .00/.00	.00/.00 | .00/.00	.01/.00 | .00/.00	.00/.00 | .92/.58
R1-Qwen-32B	Hacking	.01/.00 | .00/.00	.06/.05 | .34/.31	.92/.39 | .00/.00	.03/.02 | .02/.02	.03/.05 | .07/.07	.01/.01 | .01/.02	.00/.00 | .00/.00	.00/.01 | .00/.00	.00/.00 | .00/.00	.01/.00 | .01/.00	.00/.00 | .82/.64
Qwen-14B	Explicit	.87/.37 | .00/.00	.86/.36 | .00/.00	.87/.38 | .00/.00	.86/.36 | .00/.00	.86/.36 | .00/.00	.87/.38 | .00/.00	.00/.00 | .00/.00	.85/.37 | .00/.00	.85/.36 | .00/.00	.88/.37 | .00/.00	.00/.00 | .81/.66
Qwen-14B	Hacking	.00/.01 | .00/.00	.89/.38 | .00/.00	.90/.39 | .00/.00	.03/.03 | .00/.00	.15/.10 | .00/.00	.00/.03 | .02/.02	.00/.01 | .00/.00	.00/.01 | .00/.00	.01/.01 | .00/.00	.01/.01 | .00/.00	.00/.00 | .88/.48
Qwen-32B	Explicit	.86/.37 | .00/.00	.85/.36 | .00/.00	.86/.38 | .00/.00	.86/.36 | .00/.00	.84/.36 | .00/.00	.87/.37 | .00/.00	.01/.00 | .00/.00	.86/.37 | .00/.00	.12/.03 | .00/.00	.86/.37 | .00/.00	.00/.00 | .76/.63
Qwen-32B	Hacking	.00/.01 | .00/.00	.42/.18 | .00/.00	.89/.38 | .00/.00	.51/.20 | .00/.00	.55/.23 | .01/.01	.29/.15 | .01/.01	.02/.02 | .00/.00	.06/.04 | .00/.00	.12/.03 | .00/.00	.17/.05 | .00/.00	.01/.01 | .85/.40
R1-Llama-8B	Explicit	.30/.06 | .00/.00	.02/.04 | .00/.00	.93/.37 | .00/.00	.02/.02 | .00/.00	.00/.05 | .00/.00	.01/.00 | .03/.04	.01/.01 | .00/.00	.01/.01 | .00/.00	.90/.39 | .00/.00	.01/.00 | .02/.01	.00/.00 | .90/.56
R1-Llama-8B	Hacking	.06/.01 | .00/.00	.03/.05 | .00/.00	.91/.39 | .00/.00	.02/.02 | .00/.00	.02/.06 | .00/.00	.01/.01 | .02/.04	.00/.01 | .00/.00	.01/.01 | .01/.00	.37/.13 | .00/.00	.00/.00 | .00/.00	.00/.00 | .83/.55
R1-Llama-70B	Explicit	.01/.00 | .00/.00	.05/.05 | .00/.00	.92/.38 | .00/.00	.00/.02 | .00/.00	.02/.06 | .00/.00	.00/.00 | .15/.16	.00/.00 | .00/.00	.00/.00 | .00/.00	.02/.00 | .00/.00	.00/.00 | .02/.01	.00/.00 | .81/.67
R1-Llama-70B	Hacking	.01/.00 | .00/.00	.05/.05 | .13/.07	.92/.39 | .00/.00	.11/.05 | .05/.03	.11/.08 | .00/.00	.00/.02 | .01/.02	.00/.00 | .00/.00	.00/.01 | .01/.00	.01/.00 | .00/.00	.00/.01 | .01/.00	.00/.00 | .85/.61

Table 8: English and Chinese compliance rates (English-sentence-level / English-token-level | Chinese-sentence-level / Chinese-token-level) on **MGSM**. Explicit instruction vs. Prompt hacking.

Model	Setting	ar	bn	de	en	es	fr	hi	id	it	ja	ko	pt	sw	yo	zh
R1-Qwen-1.5B	Explicit	.42/.90 | .00/.00	.41/.93 | .00/.00	.36/.72 | .00/.00	.45/.95 | .00/.00	.05/.07 | .00/.00	.25/.38 | .00/.00	.38/.91 | .01/.00	.32/.38 | .00/.00	.20/.43 | .00/.00	.27/.68 | .12/.10	.47/.94 | .00/.00	.12/.15 | .00/.00	.19/.47 | .00/.00	.40/.82 | .00/.00	.63/.85
R1-Qwen-1.5B	Hacking	.12/.11 | .01/.00	.00/.00 | .00/.00	.05/.01 | .00/.00	.47/.97 | .00/.00	.04/.01 | .00/.00	.09/.00 | .00/.00	.00/.00 | .00/.00	.02/.00 | .00/.00	.02/.01 | .00/.00	.21/.32 | .03/.05	.29/.25 | .00/.01	.03/.00 | .00/.00	.00/.00 | .00/.00	.01/.00 | .00/.00	.56/.89
R1-Qwen-7B	Explicit	.12/.17 | .02/.01	.16/.76 | .00/.00	.08/.01 | .00/.00	.46/.97 | .00/.00	.06/.00 | .00/.00	.11/.06 | .00/.00	.00/.03 | .00/.00	.05/.06 | .00/.00	.03/.04 | .00/.00	.35/.05 | .02/.02	.36/.69 | .02/.01	.03/.00 | .00/.00	.40/.67 | .00/.00	.40/.79 | .00/.00	.67/.87
R1-Qwen-7B	Hacking	.12/.18 | .02/.01	.00/.00 | .00/.00	.06/.02 | .00/.00	.47/.97 | .00/.00	.03/.00 | .00/.00	.08/.01 | .00/.00	.00/.00 | .00/.00	.02/.00 | .00/.00	.02/.00 | .00/.00	.29/.16 | .05/.01	.28/.19 | .02/.01	.03/.00 | .00/.00	.01/.00 | .00/.01	.00/.00 | .00/.00	.61/.91
R1-Qwen-14B	Explicit	.46/.95 | .00/.00	.42/.96 | .00/.00	.40/.80 | .00/.00	.46/.97 | .00/.00	.23/.50 | .00/.00	.39/.81 | .00/.00	.28/.92 | .01/.00	.40/.82 | .00/.00	.35/.72 | .00/.00	.20/.63 | .05/.02	.45/.92 | .00/.00	.41/.86 | .01/.00	.44/.90 | .00/.00	.41/.85 | .00/.00	.69/.85
R1-Qwen-14B	Hacking	.03/.04 | .09/.09	.00/.00 | .00/.00	.06/.02 | .00/.00	.47/.97 | .00/.00	.02/.00 | .00/.00	.08/.00 | .00/.00	.00/.01 | .00/.00	.02/.00 | .00/.00	.01/.00 | .00/.00	.00/.00 | .01/.01	.00/.00 | .02/.01	.02/.00 | .00/.00	.01/.01 | .00/.00	.10/.17 | .00/.00	.62/.92
R1-Qwen-32B	Explicit	.36/.75 | .01/.00	.15/.80 | .00/.00	.09/.09 | .00/.00	.45/.96 | .00/.00	.27/.60 | .00/.00	.44/.93 | .00/.00	.11/.75 | .00/.00	.44/.92 | .00/.00	.22/.47 | .00/.00	.13/.45 | .14/.12	.41/.82 | .01/.00	.43/.89 | .00/.00	.27/.57 | .00/.00	.36/.73 | .00/.00	.68/.88
R1-Qwen-32B	Hacking	.03/.05 | .05/.03	.00/.00 | .00/.00	.05/.00 | .00/.00	.46/.96 | .00/.00	.02/.00 | .00/.00	.08/.00 | .00/.00	.00/.00 | .00/.00	.02/.00 | .00/.00	.01/.00 | .00/.00	.01/.00 | .02/.13	.00/.00 | .03/.02	.02/.00 | .00/.00	.01/.00 | .00/.00	.01/.00 | .00/.00	.62/.85
Qwen-14B	Explicit	.45/.96 | .00/.00	.45/.97 | .00/.00	.44/.95 | .00/.00	.46/.96 | .00/.00	.45/.95 | .00/.00	.45/.95 | .00/.00	.45/.96 | .00/.00	.46/.95 | .00/.00	.45/.96 | .00/.00	.46/.95 | .00/.00	.45/.96 | .00/.00	.45/.95 | .00/.00	.44/.94 | .00/.00	.41/.89 | .00/.00	.65/.67
Qwen-14B	Hacking	.04/.10 | .02/.01	.00/.00 | .00/.00	.09/.09 | .00/.00	.46/.96 | .00/.00	.02/.00 | .00/.00	.09/.05 | .00/.00	.00/.00 | .00/.00	.02/.01 | .00/.00	.02/.02 | .00/.00	.01/.00 | .02/.01	.00/.00 | .00/.00	.02/.00 | .00/.00	.00/.00 | .00/.00	.00/.00 | .00/.00	.50/.90
Qwen-32B	Explicit	.47/.98 | .00/.00	.47/.97 | .00/.00	.45/.95 | .00/.00	.46/.95 | .00/.00	.45/.96 | .00/.00	.45/.94 | .00/.00	.47/.97 | .00/.00	.46/.95 | .00/.00	.46/.96 | .00/.00	.47/.96 | .00/.00	.47/.96 | .00/.00	.46/.95 | .00/.00	.47/.93 | .00/.00	.46/.92 | .00/.00	.69/.82
Qwen-32B	Hacking	.04/.08 | .02/.02	.01/.04 | .00/.00	.11/.16 | .00/.00	.48/.97 | .00/.00	.20/.41 | .01/.01	.20/.40 | .01/.01	.01/.08 | .00/.04	.04/.03 | .00/.00	.23/.50 | .00/.00	.06/.16 | .02/.01	.04/.04 | .02/.02	.11/.21 | .00/.00	.02/.01 | .00/.01	.01/.00 | .00/.00	.58/.88
R1-Llama-8B	Explicit	.19/.40 | .03/.01	.46/.97 | .00/.00	.34/.68 | .00/.00	.46/.96 | .00/.00	.20/.44 | .00/.00	.14/.18 | .00/.00	.41/.94 | .04/.02	.40/.79 | .00/.00	.34/.72 | .00/.00	.09/.30 | .08/.06	.40/.88 | .07/.04	.19/.38 | .00/.00	.38/.76 | .00/.00	.42/.81 | .00/.00	.69/.92
R1-Llama-8B	Hacking	.01/.01 | .01/.00	.00/.10 | .00/.00	.05/.02 | .00/.00	.46/.94 | .00/.00	.02/.00 | .00/.00	.09/.03 | .00/.00	.00/.00 | .00/.00	.02/.00 | .00/.00	.02/.01 | .00/.00	.01/.01 | .04/.02	.02/.02 | .01/.01	.03/.00 | .00/.00	.01/.00 | .00/.01	.08/.13 | .00/.00	.58/.78
R1-Llama-70B	Explicit	.22/.45 | .06/.03	.44/.96 | .00/.00	.44/.92 | .00/.00	.46/.95 | .00/.00	.42/.89 | .00/.00	.43/.87 | .00/.00	.33/.94 | .00/.00	.43/.86 | .00/.00	.46/.95 | .00/.00	.03/.07 | .48/.51	.46/.94 | .02/.01	.45/.92 | .00/.00	.46/.90 | .00/.00	.43/.89 | .00/.00	.68/.86
R1-Llama-70B	Hacking	.02/.02 | .02/.02	.00/.00 | .00/.00	.06/.03 | .02/.06	.47/.95 | .00/.00	.06/.10 | .02/.02	.09/.02 | .00/.00	.00/.00 | .00/.00	.02/.00 | .00/.00	.03/.04 | .01/.00	.01/.00 | .05/.02	.00/.00 | .13/.16	.03/.00 | .00/.00	.01/.00 | .00/.01	.01/.01 | .00/.00	.59/.87

Table 9: English and Chinese compliance rates (English-sentence-level / English-token-level | Chinese-sentence-level / Chinese-token-level) on **MMMLU**. Explicit instruction vs. Prompt hacking.

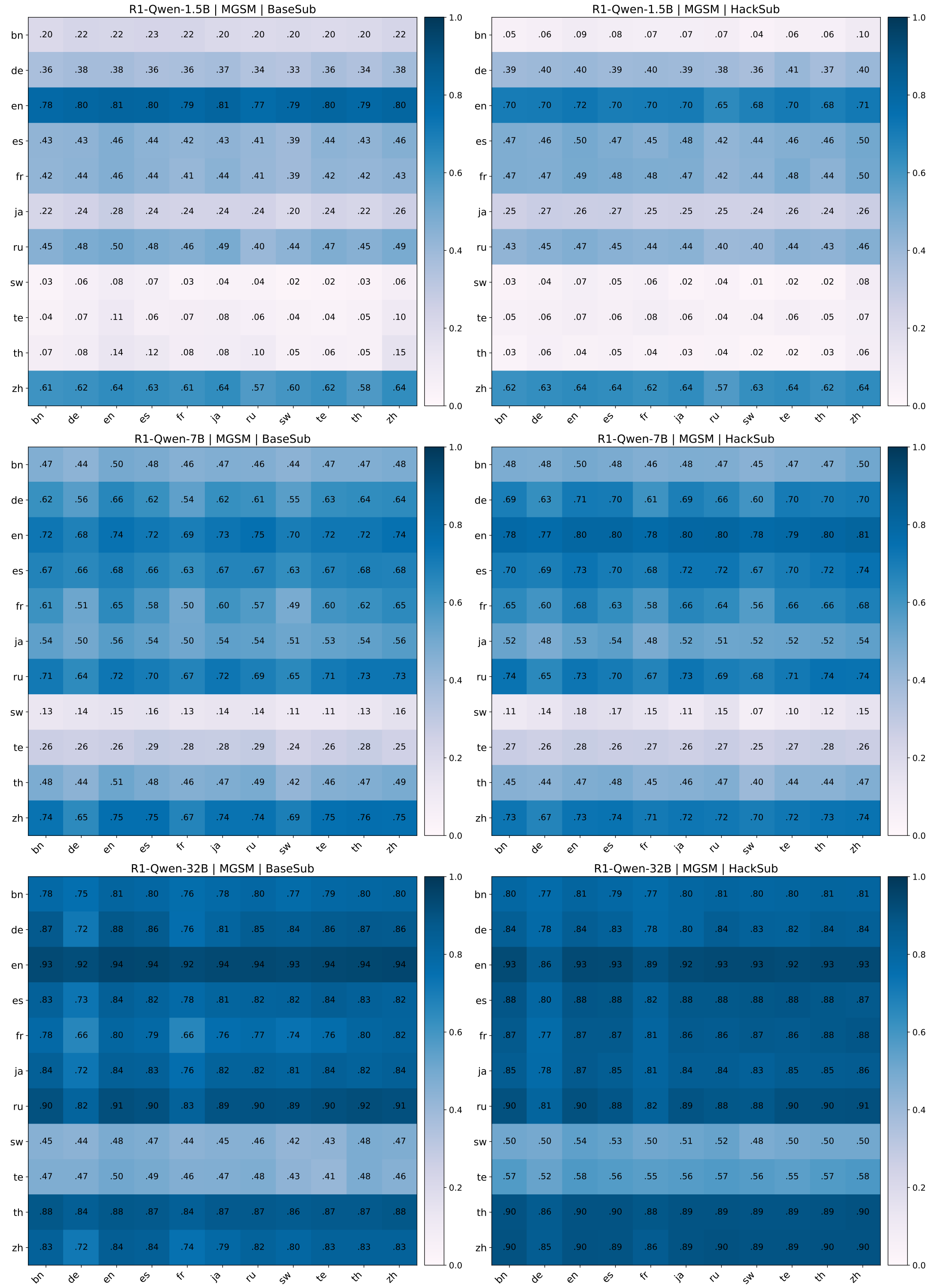

Figure 7: Final-answer accuracy of LRMs under two thinking trace substitutions: BaseSub, HackSub. Each cell shows the accuracy when injecting thinking traces from a language on the y-axis into a language on the y-axis. Performance disparities indicate that thinking trace quality varies across languages.

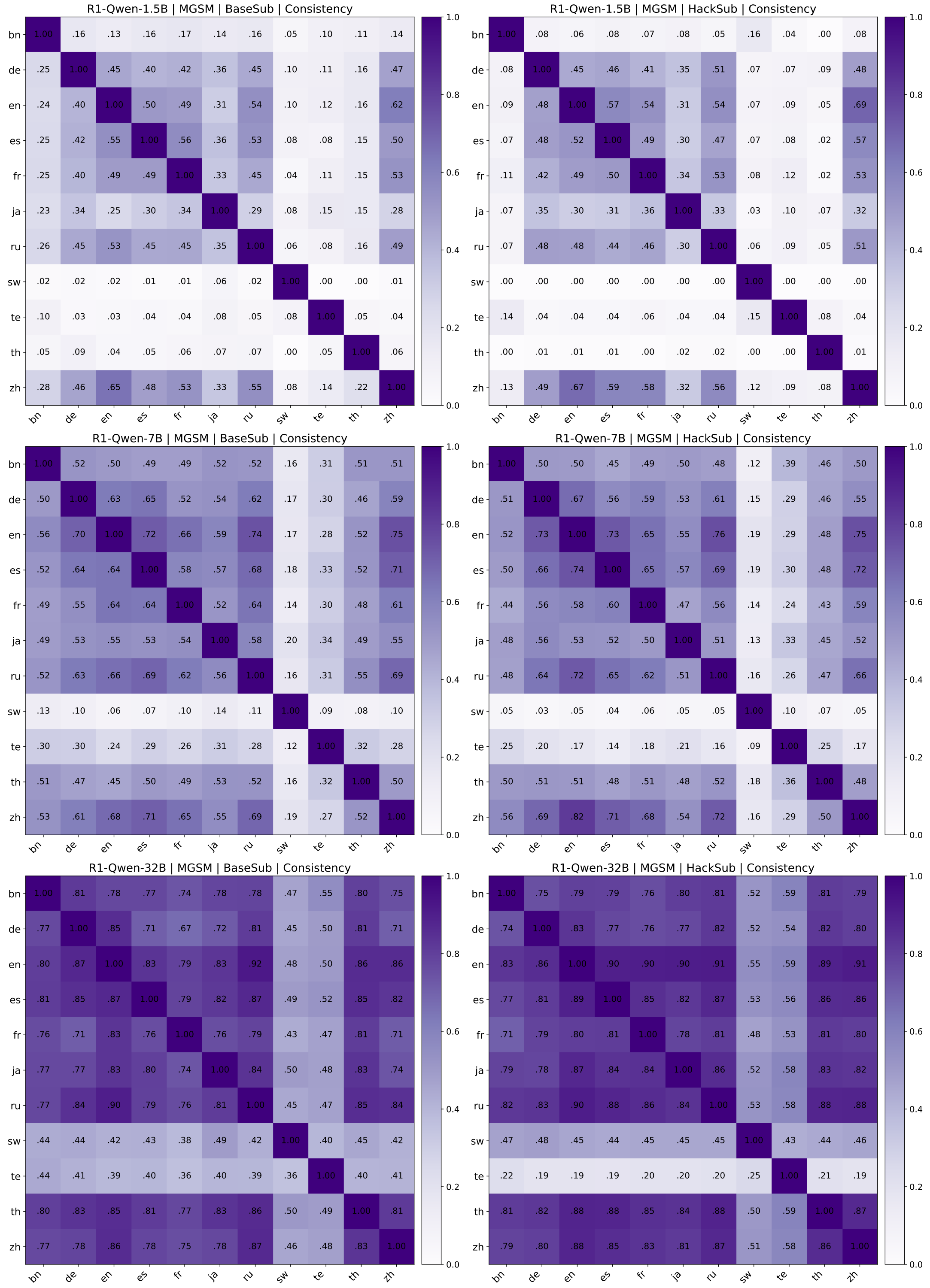

Figure 8: Substitution consistency of LRMs under two thinking trace substitutions: BaseSub, HackSub. Each cell indicates the consistency between the original predictions in the language on the x-axis and the predictions after injecting thinking traces from the language on the y-axis. Higher consistency is observed when traces are substituted between similar languages.

Metric	Group	Mean Value	P-Value
Without Low Resource Languages			
Trace- Substitution Consistency	Indo-European	0.7208	9.05e-03
	Non Indo-European	0.7108	
With Low Resource Languages			
Trace- Substitution Consistency	Indo-European	0.7108	1.99e-52
	Non Indo-European	0.5452	

Table 10: Consistency comparison between Indo-European and non-Indo-European languages. Reported are mean consistency values for Consistency of Trace-Substitution metrics, with corresponding p-values (t-test). We also discard low-resouce languages sw and te to conduct t-test. Indo-European languages generally achieve higher consistency, and the differences are statistically significant.

B.2 Test Instances

To reduce computational costs, we limit each dataset to a maximum of 250 test instances per language. For MMMLU, we randomly sample 250 examples from the full 14K test set. This sampling is applied consistently across all parallel language versions to ensure comparability. For MGSM, we use the official test set, which consists of 250 parallel examples available across all supported languages.

B.3 Prompt Templates

Table 9 provides an overview of the prompts designed for model instruction. However, we observed that, even when explicit prompts were provided, the model frequently conducted its intermediate reasoning in a language different from that of the input prompt. To mitigate this inconsistency, we appended the language-specific instructions listed in Table 13 after the query and the <think> tag. This strategy effectively enforces the model to align its reasoning language with the prompt language. To avoid redundancy, we present prompt templates for the MGSM task in Table 9 only, while templates for MMMLU share the same structural format and are available in the released code.

C Experimental Environment and Hyperparameters

We set the maximum generation length to 8192 tokens for all models. We use the recommended configurations provided on HuggingFace for all models. Specifically, we set the temperature to 0.6 and top-p to 0.95 for distilled versions of `DeepSeek R1`. For Qwen3 models, we set the temperature to 0.6, top-p to 0.95, and top-k to 20.

Experiments are primarily conducted on NVIDIA A100 GPUs. For larger models, such as `DeepSeek-R1-Distill-Llama-70B`, we use NVIDIA H200 GPUs for inference.

To evaluate final-answer correctness, we adopt an exact matching strategy. Following prior work (Qi et al., 2025), we prompt the model to wrap its final answer in `\boxed{}`, and extract the boxed content for comparison against the gold answer.

Lang	Instruction
ar	فكر دائمًا بالعربية. حل مسألة الرياضيات التالية خطوة بخطوة. اكتب تفكيرك ضمن <think>...</think>. وأخيرًا ضع النتيجة النهائية داخل \boxed{}.
bn	অনুগ্রহ করে সবসময় বাংলায় ভাবুন। ধাপে ধাপে নিচের গণিত সমস্যা সমাধান করুন। <think>...</think> এ যুক্তি লিখুন এবং চূড়ান্ত ফলাফল \boxed{}এ দিন।
de	Bitte denken Sie immer auf Deutsch. Lösen Sie das folgende Mathematikproblem Schritt für Schritt. Schreiben Sie Ihre Begründung in <think>...</think>. Geben Sie schließlich das Ergebnis in \boxed{} an.
en	Always think in English. Solve the following math problem step by step. Write your reasoning in <think>...</think>. Finally, provide the final result enclosed in \boxed{}.
es	Piense siempre en español. Resuelva el siguiente problema de matemáticas paso a paso. Escriba su razonamiento en <think>...</think> y encierre el resultado final en \boxed{}.
fr	Veuillez toujours réfléchir en français. Résolvez le problème mathématique suivant étape par étape. Écrivez le raisonnement dans <think>...</think>. Enfin, encadrez le résultat final dans \boxed{}.
hi	कृपया हमेशा हिंदी में सोचें। नीचे दिए गए गणित प्रश्न को चरणबद्ध तरीके से हल करें। तर्क <think>...</think> में लिखें और अंतिम परिणाम \boxed{}में दें।
id	Selalu berpikir dalam bahasa Indonesia. Selesaikan soal matematika berikut langkah demi langkah. Tulis penalaran di <think>...</think> dan berikan hasil akhir dalam \boxed{}.
it	Pensa sempre in italiano. Risolvi il seguente problema matematico passo dopo passo. Scrivi il ragionamento in <think>...</think>. Infine racchiudi il risultato in \boxed{}.
ja	常に日本語で考えてください。 以下の数学問題を段階的に解いてください。 推論は <think>...</think> に記述し，最終結果を \boxed{}に示してください。
ko	항상 한국어로 사고하세요. 다음 수학 문제를 단계별로 풀이하세요. 추론은 <think>...</think>에 쓰고 최종 결과를 제시하세요.
pt	A pedido, pense sempre em português. Resolva o problema de matemática a seguir paso a paso. Escreva o raciocínio em <think>...</think> e coloque o resultado final em \boxed{}.
ru	Всегда рассуждай по-русски. Решай следующую математическую задачу шаг за шагом. Пиши рассуждения внутри <think>...</think>. В конце помести итог внутри \boxed{}.
sw	Tafadhali kila mara fikiria kwa Kiswahili. Tatua tatizo lifuatalo la hisabati hatua kwa hatua. Andika hoja kwenye <think>...</think> na weka matokeo ya mwisho ndani ya \boxed{}.
te	దయచేసి ఎల్లప్పుడూ తెలుగులో ఆలోచించండి. క్రింది గణిత సమస్యను దశలవారీగా పరిష్కరించండి. మీ తర్కాన్ని <think>...</think> లో వ్రాయండి. చివరగా తుది ఫలితాన్ని \boxed{}లో ఇవ్వండి.
th	โปรดคิดเป็นภาษาไทยเสมอ แก้ปัญหาคณิตต่อไปนี้แบบเป็นขั้นตอน เขียนเหตุผลไว้ใน <think>...</think> และใส่ผลลัพธ์สุดท้ายใน \boxed{}.
yo	Jọ̀wọ́ máa rò ní Yorùbá. Ṣe iṣoro iṣirò yìí ní ìgbésẹ̀-nípẹ̀yà. Kọ ìrònú sínú <think>...</think> kí o sì fi esi ikẹhin sínú \boxed{}.
zh	请始终用中文思考。 逐步解决以下数学问题。 每一步将推理写在 <think>...</think> 中。最后，请将最终结果放在 \boxed{}中。

Figure 9: Prompts for **MGSM** task in different languages.

Operation	Model	de	en	es	fr	ja	ru	sw	th	zh	bn	te
Truncation (First)	Qwen-14B	.46 (.51)	.54 (.57)	.40 (.46)	.32 (.39)	.46 (.55)	.32 (.36)	.28 (.60)	.25 (.28)	.56 (.67)	.51 (.64)	.36 (.49)
	Qwen-32B	.53 (.66)	.62 (.71)	.43 (.64)	.31 (.58)	.46 (.64)	.37 (.50)	.51 (.80)	.60 (.67)	.38 (.48)	.46 (.53)	.26 (.38)
	R1-Qwen-1.5B	.05 (.13)	.08 (.11)	.04 (.08)	-.00 (-.01)	.03 (.12)	.02 (.05)	-.00 (NaN)	-.02 (-2.00)	.03 (.04)	-.02 (-.29)	-.02 (-.56)
	R1-Qwen-7B	.08 (.12)	.06 (.07)	.06 (.07)	.06 (.09)	.03 (.05)	.05 (.07)	-.02 (-.50)	.09 (.18)	.08 (.10)	.00 (.00)	-.12 (-.78)
	R1-Qwen-14B	.08 (.10)	-.02 (-.02)	-.02 (-.02)	.06 (.07)	.03 (.04)	.02 (.02)	.00 (.02)	.04 (.05)	.02 (.02)	.04 (.06)	-.08 (-.30)
	R1-Qwen-32B	.11 (.12)	.02 (.03)	.01 (.01)	-.01 (-.01)	.02 (.02)	.01 (.01)	.00 (.01)	.01 (.01)	.01 (.01)	.01 (.01)	-.36 (-1.98)
	R1-Llama-8B	.10 (.20)	.13 (.16)	.11 (.15)	.16 (.26)	.04 (.09)	.04 (.06)	-.02 (-.55)	-.02 (-.05)	.04 (.05)	-.01 (-.18)	-.10 (-.92)
	R1-Llama-70B	.08 (.10)	.03 (.03)	.07 (.08)	.07 (.08)	.06 (.07)	.06 (.07)	.04 (.04)	.03 (.04)	.17 (.19)	-.02 (-.02)	-.32 (-.72)
Truncation (Middle)	Qwen-14B	.46 (.51)	.56 (.58)	.46 (.52)	.39 (.47)	.42 (.51)	.30 (.34)	.29 (.61)	.33 (.37)	.52 (.61)	.56 (.71)	.38 (.53)
	Qwen-32B	.54 (.67)	.64 (.73)	.44 (.65)	.31 (.58)	.48 (.67)	.36 (.48)	.50 (.78)	.61 (.68)	.39 (.49)	.53 (.62)	.36 (.53)
	R1-Qwen-1.5B	.03 (.07)	.09 (.11)	.04 (.08)	.02 (.03)	.02 (.08)	.03 (.07)	-.00 (NaN)	-.01 (-1.50)	.01 (.02)	-.01 (-.21)	-.02 (-.56)
	R1-Qwen-7B	.09 (.15)	.06 (.07)	.06 (.08)	.08 (.13)	.02 (.04)	.06 (.08)	-.01 (-.17)	.07 (.13)	.07 (.09)	-.00 (-.01)	-.12 (-.78)
	R1-Qwen-14B	.08 (.11)	.00 (.00)	-.02 (-.02)	.05 (.06)	.04 (.05)	.02 (.03)	-.02 (-.06)	.04 (.04)	.02 (.02)	.04 (.06)	-.08 (-.30)
	R1-Qwen-32B	.10 (.12)	.04 (.04)	.02 (.03)	-.02 (-.03)	.02 (.03)	.02 (.02)	-.01 (-.03)	.01 (.01)	.01 (.01)	.02 (.02)	-.36 (-1.93)
	R1-Llama-8B	.09 (.17)	.13 (.15)	.13 (.18)	.15 (.24)	.04 (.08)	.04 (.05)	-.02 (-.36)	-.03 (-.09)	.03 (.04)	-.02 (-.24)	-.09 (-.85)
	R1-Llama-70B	.10 (.12)	.02 (.02)	.09 (.10)	.08 (.10)	.06 (.07)	.07 (.08)	.04 (.05)	.03 (.04)	.13 (.15)	-.04 (-.06)	-.33 (-.73)
Truncation (Last)	Qwen-14B	.68 (.75)	.75 (.79)	.69 (.79)	.64 (.77)	.69 (.82)	.68 (.75)	.40 (.84)	.65 (.72)	.67 (.79)	.66 (.83)	.63 (.87)
	Qwen-32B	.66 (.83)	.76 (.86)	.55 (.81)	.39 (.72)	.62 (.87)	.60 (.81)	.56 (.89)	.77 (.86)	.60 (.75)	.77 (.89)	.60 (.88)
	R1-Qwen-1.5B	.22 (.55)	.51 (.65)	.31 (.67)	.23 (.50)	.14 (.56)	.28 (.64)	-.01 (NaN)	-.02 (-2.50)	.26 (.40)	.01 (.14)	-.01 (-.22)
	R1-Qwen-7B	.28 (.45)	.43 (.51)	.41 (.54)	.32 (.53)	.22 (.42)	.36 (.50)	.03 (.67)	.22 (.43)	.29 (.36)	.06 (.12)	.00 (.00)
	R1-Qwen-14B	.40 (.52)	.32 (.38)	.29 (.36)	.35 (.44)	.24 (.31)	.39 (.44)	.12 (.48)	.30 (.37)	.24 (.27)	.24 (.37)	.06 (.20)
	R1-Qwen-32B	.30 (.35)	.26 (.27)	.35 (.40)	.21 (.26)	.30 (.34)	.29 (.32)	.20 (.43)	.22 (.25)	.14 (.16)	.20 (.25)	-.22 (-1.22)
	R1-Llama-8B	.38 (.71)	.59 (.70)	.54 (.77)	.45 (.71)	.28 (.61)	.48 (.72)	.01 (.18)	.16 (.43)	.44 (.60)	.01 (.18)	.04 (.38)
	R1-Llama-70B	.36 (.44)	.22 (.23)	.38 (.42)	.35 (.41)	.28 (.34)	.39 (.43)	.32 (.37)	.26 (.31)	.29 (.33)	.11 (.15)	-.09 (-.19)

Table 11: Absolute drop and relative drop rate compared to baseline (First, Middle and Last truncation).

Dataset	Languages
MMMLU	Arabic, Bengali, German, Spanish, French, Hindi, Indonesian, Italian, Japanese, Korean, Portuguese, Swahili, Yoruba, Chinese
MGSM	Bengali, Chinese, French, German, Japanese, Russian, Spanish, Swahili, Telugu, Thai

Table 12: Languages covered in MMMLU and MGSM.

Lang	Hacking_Prefix
ar	بناءً على الطلب، سأبدأ في التفكير باللغة العربية.
bn	অনুরোধে, আমি বাংলায় চিন্তা করা শুরু করব.
de	Auf Anfrage werde ich anfangen, auf Deutsch zu denken.
en	By request, I will start thinking in English.
es	Por solicitud, comenzaré a pensar en español.
fr	À la demande, je vais commencer à penser en français.
hi	अनुरोध पर, मैं हिंदी में सोचना शुरू करूँगा।
id	Atas permintaan, saya akan mulai berpikir dalam bahasa Indonesia.
it	Su richiesta, inizierò a pensare in italiano.
ja	リクエストに応じて、日本語で考え始めます。
ko	요청에 따라 한국어로 생각하기 시작하겠습니다.
pt	A pedido, começarei a pensar em português.
ru	По просьбе я начну рассуждать по-русски.
sw	Kwa ombi, nitaanza kufikiri kwa Kiswahili.
te	అభ్యర్థన మేరకు, నేను తెలుగులో ఆలోచించడం ప్రారంభిస్తాను.
th	ตามคำขอ ฉันจะเริ่มคิดเป็นภาษาไทย
yo	Lórí ìbéèrè, màá bẹ̀rẹ̀ sí rò ní èdè Yorùbá.
zh	应要求，我将开始用中文思考.

Table 13: Language-specific answer instructions.