Large Reasoning Models Are (Not Yet) Multilingual Latent Reasoners

Yihong Liu*, **Raoyuan Zhao***, **Hinrich Schütze**†, **and Michael A. Hedderich**†

Center for Information and Language Processing, LMU Munich
Munich Center for Machine Learning (MCML)
{yihong, rzhao, hedderich}@cis.lmu.de

Abstract

Large reasoning models (LRMs) achieve strong performance on mathematical reasoning tasks, often attributed to their capability to generate explicit chain-of-thought (CoT) explanations. However, recent work shows that LRMs often arrive at the correct answer before completing these textual reasoning steps, indicating the presence of *latent reasoning* – internal, non-verbal computation encoded in hidden states. While this phenomenon has been explored in English, its multilingual behavior remains largely unknown. In this paper, we conduct a systematic investigation of multilingual latent reasoning in LRMs across 11 languages. Using a truncation-based strategy, we examine how the correct answer emerges as the model is given only partial reasoning traces, allowing us to measure stepwise latent prediction formation. Our results reveal clear evidence of multilingual latent reasoning, though unevenly: strong in resource-rich languages, weaker in low-resource ones, and broadly less observable on harder benchmarks. To understand whether these differences reflect distinct internal mechanisms, we further perform representational analyses. Despite surface-level disparities, we find that the internal evolution of predictions is highly consistent across languages and broadly aligns with English – a pattern suggesting an English-centered latent reasoning pathway.[1]

1 Introduction

Recent large reasoning models (LRMs) (OpenAI et al., 2024; Yang et al., 2025; DeepSeek-AI et al., 2025) have rapidly advanced the state of the art on many challenging tasks, such as coding, mathematical reasoning, and logical reasoning (Li et al., 2025). This is largely thought to be due to their capacity to generate explicit CoT explanations (Wei et al., 2022) that scaffold multi-step problem solving, especially through test-time scaling, where enough computation budget is given to allow the model to generate longer reasoning traces (Snell et al., 2024; Muennighoff et al., 2025).

Despite this reliance on explicit CoT explanations, emerging evidence shows that models often engage in *latent reasoning* – computing intermediate or final answers within hidden states. Such latent behavior has been observed in multi-hop factual knowledge recall (Yang et al., 2024; Biran et al., 2024) and, in the context of CoT, in models that internally form solutions well before they articulate the answer in their reasoning (Lanham et al., 2023; Pfau et al., 2024; Mao et al., 2025). This phenomenon aligns with recent findings that LLMs can "think ahead" by predicting future tokens directly from intermediate hidden states (Pal et al., 2023; Wu et al., 2024; Cai et al., 2024). Together, these observations indicate that explicit CoT generation is not the sole mechanism through which LRMs solve problems and that reasoning may be occurring within the model's latent space.

However, existing studies of latent reasoning focus almost exclusively on English, leaving open how these latent reasoning processes behave across languages. At the explicit reasoning level, multilingual performance is already known to be uneven: models trained on English-centric corpora often struggle with underrepresented languages due to limited multilingual reasoning training data (Wang et al., 2025a; Huang et al., 2025), weaker language understanding ability (Yoon et al., 2024; Kang et al., 2025), and lower-quality reasoning trace generation (Yong et al., 2025; Zhao et al., 2026). These findings raise a natural question: if explicit reasoning varies across languages, does latent reasoning exhibit similar disparities, or does it follow a language-independent mechanism? This motivates our two research questions: (**RQ1**) *Do LRMs exhibit latent reasoning across languages,*

*Equal contribution.
†Equal advising.
[1]We make our code publicly available at: https://github.com/cisnlp/multilingual-latent-reasoner

Findings of the Association for Computational Linguistics: ACL 2026, pages 22330–22358

and how does the strength vary? and (**RQ2**) *Do languages follow different internal latent reasoning pathways, or do they share a common mechanism?*

To answer these questions, we conduct a systematic investigation of multilingual latent reasoning in LRMs using two mathematical reasoning benchmarks of different difficulty across 11 languages. To address **RQ1**, we quantify how strongly LRMs rely on explicit reasoning traces by eliciting and evaluating their stepwise early predictions in truncated traces, and we propose novel aggregate metrics capturing different dimensions of latent reasoning (cf. §4). To address **RQ2**, we analyze the internal evolution of answer formation using the logit lens approach (Nostalgebraist, 2020), examining when the correct answer becomes probable across layers in each language, and we compare hidden-state similarity trajectories across languages (cf. §5). We further disentangle latent reasoning from potential memorization effects (cf. §7).

Our key findings are as follows: (**i**) Latent reasoning exists across languages. However, resource-rich languages show strong early-emergent correctness, while low-resource ones display weaker latent reasoning signals. (**ii**) Latent reasoning is less pronounced under increased task difficulty. On harder benchmarks, early answer formation largely disappears across all languages and model sizes. (**iii**) Internal latent reasoning dynamics are shared across languages. Such dynamics converge to an English-centered pathway, especially for high-resource languages and correctly solved instances. (**iv**) While models show partial memorization, latent reasoning remains evident for high-resource languages, and this capability scales with model size.

2 Related Work

Multilingual Reasoning Multilingual reasoning remains challenging due to the strong language bias of most models (Ghosh et al., 2025). Since models often rely on English as a pivot, translate-then-solve strategies are frequently effective for under-resourced languages (Qin et al., 2023; Huang et al., 2023; Zhu et al., 2024). Recent work shows that post-training on multilingual reasoning data can substantially improve crosslingual performance (Chen et al., 2024; Huang et al., 2025). At inference time, model behavior is highly sensitive to the language used in the reasoning process (Wang et al., 2025c; Qi et al., 2025; Yong et al., 2025; Tam et al., 2025). These performance gaps have been attributed to disparities in the quality of language-specific reasoning traces (Zhao et al., 2026) and to failures in basic understanding for low-resource inputs (Kang et al., 2025; Bafna et al., 2025). However, existing studies focus almost exclusively on *explicit* multilingual reasoning behavior. Our work aims to investigate *latent* reasoning across languages systematically.

Implicit Latent Reasoning Unlike *explicit* reasoning, where models produce step-by-step textual explanations, implicit or *latent* reasoning refers to the internal computation that occurs in the model's hidden representations (Cheng and Durme, 2024; Li et al., 2025; Chen et al., 2025a). Prior work shows that LLMs may pursue multiple latent reasoning paths in parallel, gradually increasing confidence in a particular solution as explicit reasoning unfolds (Prystawski et al., 2023; Dutta et al., 2024; Qian et al., 2025). Yet even when the model has internally formed the correct answer, it may continue generating unnecessary reasoning steps, referred to as "overthinking" (Chen et al., 2025b; Sui et al., 2025). Motivated by these observations, there have been new approaches aiming to train models to reason directly in latent space without producing full textual traces (Deng et al., 2024; Hao et al., 2025; Lin et al., 2025; Saunshi et al., 2025; Xu et al., 2025a). However, this line of work focuses almost exclusively on English, leaving open whether latent reasoning behaviors emerge across languages. Our work addresses this gap by systematically evaluating and comparing latent reasoning dynamics in a multilingual setting.

3 Experimental Setup

3.1 Models

We use three distilled variants of `DeepSeek-R1` (DeepSeek-AI et al., 2025), `DeepSeek-R1-Distill-Qwen-{7B, 14B, 32B}`, whose backbone models are based on the `Qwen2.5` family (Qwen Team et al., 2025). These models are selected because they exhibit strong reasoning performance while providing multiple sizes, enabling us to analyze how multilingual latent reasoning varies with model capacity.

3.2 Datasets and Languages

MGSM Multilingual Grade School Math dataset (Shi et al., 2023) contains 250 grade-school math problems sourced from GSM8K (Cobbe et al., 2021), originally written in English and manually

translated into 10 additional languages: French (FR), German (DE), Chinese (ZH), Japanese (JA), Russian (RU), Spanish (ES), Swahili (SW), Bengali (BN), Telugu (TE), and Thai (TH). Since the underlying mathematical problems are identical across languages, it is well-suited for studying multilingual (latent) reasoning dynamics.

Multilingual AIME The Multilingual American Invitational Mathematics Examination (AIME) datasets are translated versions of AIME2024 and AIME2025 introduced by Qi et al. (2025), covering the same 11 languages as MGSM. These datasets contain substantially more challenging, competition-level math problems, enabling us to examine how increased problem difficulty influences multilingual latent reasoning dynamics.

We categorize the languages considered in this study into **high-resource** (EN, ES, DE, FR, RU, ZH), **mid-resource** (BN, JA, TH), and **low-resource** (SW, TE) groups. This categorization is based on the relative availability of large-scale training resources and the degree of language coverage in contemporary multilingual LLMs (Joshi et al., 2020; Blasi et al., 2022; Xu et al., 2025b)

3.3 Language Control

LLMs may generate explicit reasoning traces in a language different from that of the prompt (Wang et al., 2025c; Qi et al., 2025), which is undesirable for crosslingual analysis of latent reasoning. To ensure that explicit reasoning is produced in the same language as the input, we employ a *prompt-hacking strategy* (Qi et al., 2025; Zhao et al., 2026) that inserts a language-specific prefix immediately after the `<think>` token, reliably steering the reasoning trace to the target language (see §F.1 for details).

4 Latent Reasoning Identification

To address **RQ1**: *Do LRMs exhibit latent reasoning across languages, and how does the strength vary?*, we analyze the model's early predictions under reasoning-trace truncation. This protocol connects the *explicit* reasoning process with the model's *internal* answer construction: if the model already knows the answer early in the trace, it should often answer correctly even when only a small portion of the reasoning is visible. This method is similar to concurrent work on early stopping and stepwise answer prediction (Mao et al., 2025; Wang et al., 2025d; Zhao et al., 2026; Chen et al., 2026), but we leverage such truncation to identify latent reasoning and complement it with novel metrics that quantify latent reasoning capability across languages.

4.1 Truncating Reasoning Traces

Let x denote a math problem and let the model produce a full reasoning trace $c = (t_1, t_2, \ldots, t_T)$ in the target language, followed by a final answer, where t_i indicates the i-th reasoning step.[2] We then consider a set of truncation ratios $\mathcal{R} = \{r_1, r_2, \ldots, r_M\} \subset [0, 1]$, where each $r \in \mathcal{R}$ specifies the fraction of the reasoning trace that is retained (e.g., 10%). For a ratio r, we define the truncation index as $m(r) = \lfloor r \cdot T \rfloor$ and the truncated reasoning trace as $c_{\leq r} = (t_1, t_2, \ldots, t_{m(r)})$. We then ask the model to directly produce a numerical prediction based on the original math problem x and the truncated reasoning trace $c_{\leq r}$.[3]

4.2 Evaluation Metrics

We evaluate performance over a set of truncation ratios $r \in \mathcal{R}$ using the following metrics.[4]

Truncated Pass@k. This metric estimates the probability that at least one correct answer appears among the top-k attempts for a given problem (Kulal et al., 2019; Chen et al., 2021). Let $a_k(r)$ denote the pass@k accuracy at truncation ratio r, i.e.,

$$a_k(r) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\Big[\exists j \leq k : \hat{y}_j^{(i)}(r) = y^{(i)\star} \Big]$$

where N is the number of problems, $\hat{y}_j^{(i)}(r)$ is the j-th sampled prediction for problem i based on the truncated reasoning trace $c_{\leq r}^{(i)}$, and $y^{(i)\star}$ is the gold answer. This metric measures performance under partial reasoning: if $a_k(r)$ is high even for small r, the model may not rely on fully explicit reasoning traces and instead perform latent reasoning.

Gold-in-Trace Rate. There are cases where the model explicitly articulates the answer in early reasoning steps, and then continues to refine it or explore additional paths in later steps. In such cases, correct predictions may depend on the explicitly written answer. To distinguish these cases, we additionally track whether the gold answer already

[2] The reasoning trace is regarded as the tokens between special thinking markers, e.g., `<think>` and `</think>`. We view each individual sentence as a reasoning step.

[3] This is achieved by adding `</think>` right after the truncated reasoning trace and then appending a short prefix to elicit the numerical answer prediction (see §F.1 for details).

[4] For MGSM, we consider every 10%, i.e., $\mathcal{R} = \{0\%, 10\%, 20\%, \ldots, 100\%\}$. For Multilingual AIME, we consider every 5%, i.e., $\mathcal{R} = \{0\%, 5\%, 10\%, \ldots, 100\%\}$. The choice is based on a preliminary analysis of the average number of steps across languages (see §A for details).

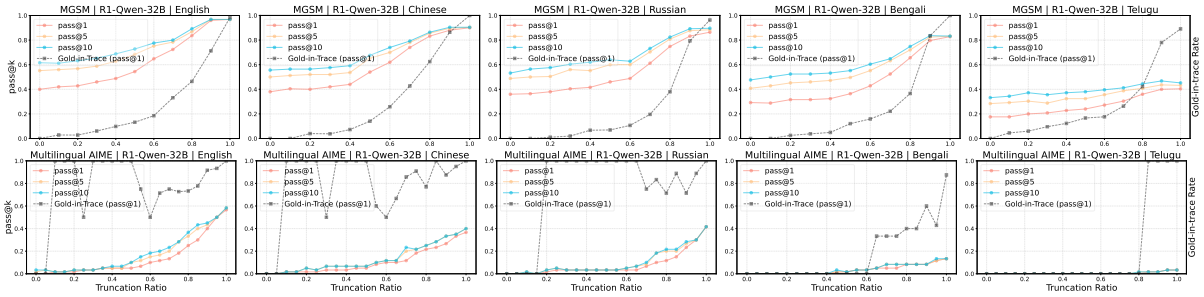

Figure 1: Pass@k accuracy ($k = 1, 5, 10$) and gold-in-trace rate under reasoning-trace truncation for R1-Qwen-32B. High accuracy with a low gold-in-trace rate indicates latent reasoning. The model shows strong evidence of latent reasoning in high-resource languages (e.g., English) on MGSM, but it is less detectable on Multilingual AIME.

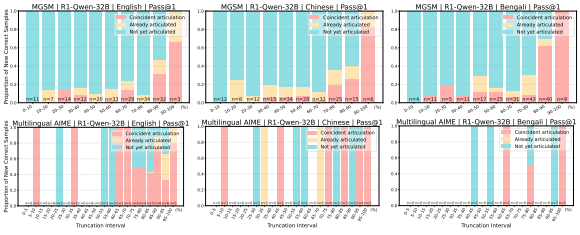

Figure 2: Causal decomposition of newly correct predictions across truncation intervals. Each bar partitions gains into three cases: (**i**) the gold answer is first articulated in the newly added reasoning steps, (**ii**) it was already articulated in earlier steps, or (**iii**) it has not yet appeared in the visible truncated trace. On MGSM, performance improvements at early and intermediate truncation ratios are dominated by case (iii), indicating that many gains arise from latent reasoning.

appears in the visible reasoning prefix $c_{\leq r}$. We define the gold-in-trace rate at truncation ratio r as

$$g_k(r) = \frac{1}{|\mathcal{C}_k(r)|} \sum_{i \in \mathcal{C}_k(r)} \mathbf{1}\Big[\, y^{(i)\star} \text{ appears in } c^{(i)}_{\leq r} \,\Big]$$

Here, $\mathcal{C}_k(r)$ denotes **the set of correctly solved instances** under truncation ratio r according to pass@k. Importantly, a high gold-in-trace rate is expected at large truncation ratios (e.g., $r \approx 1$), where the full reasoning trace should usually contain the final answer. Thus, gold-in-trace is primarily informative at *small* truncation ratios: a high value early in the trace suggests that correctness may be driven by explicit answer articulation, whereas a low value indicates that correct predictions are more likely supported by latent reasoning.

Area Under the Truncation Accuracy Curve (AUTC). We define the **AUTC** as

$$\mathrm{AUTC}_k = \int_0^1 a_k(r)\, dr.$$

A model that reaches high accuracy early (i.e., needs only a short prefix of the trace) will yield a larger AUTC than a model whose accuracy only improves near $r \approx 1$. AUTC is thus a measure of *how early and robustly* correct predictions emerge as more reasoning is revealed.

Area Under the Gold-in-Trace Curve (AUGC). Analogously, we define the **AUGC** as

$$\mathrm{AUGC}_k = \int_0^1 g_k(r)\, dr$$

A high AUGC indicates that, when the model is correct, the gold answer tends to be articulated early, while a low AUGC indicates that the gold answer usually appears near the end of the trace.

Latent Reasoning Score (LRS). To focus on correctness that is not trivially attributable to copying the answer from the trace, we define **LRS** as

$$\mathrm{LRS}_k = \int_0^1 a_k(r)\,\big(1 - g_k(r)\big)\, dr.$$

Intuitively, we weight performance at each truncation ratio by the complement of the gold-in-trace rate: correctness that occurs *after* the answer is already articulated in the trace (high $g_k(r)$) is downweighted, while correctness that occurs *before* the answer is visible (low $g_k(r)$) is upweighted. Therefore, LRS can be regarded as a proxy measure for the model's *latent reasoning capability*.

We approximate AUTC, AUGC, and LRS numerically using the trapezoidal rule (Hildebrand, 1987) over all considered truncation ratios $r \in \mathcal{R}$.

4.3 Results and Discussion

Figure 1 presents truncation curves for R1-Qwen-32B across 5 languages on two benchmarks (see §C for full results). Table 1 summarizes the corresponding AUTC, AUGC, and LRS scores across all 11 languages, models, and datasets. Finally, Figure 2 breaks down newly correct predictions in each truncation ratio interval by whether their gold answers are articulated in the newly added reasoning steps, already appear in earlier steps, or do not appear in the current truncated trace at all.

The model often knows the answer even before any reasoning is articulated. Across many high-resource languages – most notably English,

Dataset	Model	Metric	DE	EN	ES	FR	RU	ZH	BN	JA	TH	SW	TE
MGSM	R1-Qwen-7B	AUTC	0.45	0.52	0.45	0.43	0.46	0.53	0.38	0.38	0.37	0.10	0.24
		AUGC	0.22	0.21	0.16	0.22	0.24	0.27	0.27	0.24	0.33	0.19	0.32
		LRS	0.32	0.38	0.35	0.32	0.31	0.34	0.25	0.26	0.22	0.08	0.15
	R1-Qwen-14B	AUTC	0.54	0.59	0.59	0.55	0.58	0.62	0.51	0.55	0.57	0.22	0.28
		AUGC	0.20	0.19	0.20	0.18	0.22	0.26	0.27	0.24	0.27	0.25	0.26
		LRS	0.40	0.44	0.44	0.42	0.41	0.41	0.33	0.39	0.36	0.16	0.20
	R1-Qwen-32B	AUTC	0.67	0.75	0.69	0.64	0.68	0.70	0.61	0.63	0.69	0.38	0.39
		AUGC	0.20	0.25	0.20	0.17	0.21	0.30	0.23	0.21	0.28	0.20	0.23
		LRS	0.51	0.53	0.52	0.51	0.51	0.45	0.44	0.47	0.46	0.30	0.30
Multilingual AIME	R1-Qwen-7B	AUTC	0.07	0.10	0.06	0.05	0.06	0.09	0.04	0.02	0.02	0.00	0.01
		AUGC	0.52	0.51	0.19	0.23	0.55	0.60	0.57	0.12	0.17	0.00	0.00
		LRS	0.02	0.03	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0.00	0.01
	R1-Qwen-14B	AUTC	0.05	0.12	0.07	0.07	0.05	0.08	0.08	0.02	0.05	0.00	0.04
		AUGC	0.66	0.44	0.52	0.25	0.41	0.79	0.70	0.06	0.29	0.00	0.08
		LRS	0.02	0.04	0.03	0.04	0.02	0.01	0.01	0.02	0.02	0.00	0.04
	R1-Qwen-32B	AUTC	0.06	0.18	0.08	0.09	0.10	0.13	0.04	0.04	0.07	0.01	0.01
		AUGC	0.29	0.61	0.32	0.72	0.66	0.75	0.18	0.74	0.82	0.05	0.17
		LRS	0.03	0.06	0.04	0.02	0.03	0.03	0.02	0.01	0.02	0.00	0.00

Table 1: Truncation-based metrics (AUTC, AUGC, LRS) across models and benchmarks. Latent reasoning capability scales with model size and language resource availability, but emerges primarily on the simpler MGSM benchmark and is largely undetectable on the more challenging benchmark, Multilingual AIME.

French, and Chinese, the pass@1 accuracy at *zero* reasoning steps is already nontrivial (around 0.2). This suggests that for MGSM, the model can frequently compute the answer *directly in its latent representations*, without requiring explicit step-by-step CoT generation. As the truncation ratio increases, accuracy rises steadily in all languages, accompanied by a growing gold-in-trace rate that typically approaches 1.0 once the full trace is revealed. Figure 2 further supports this observation: most early correct predictions do not depend on the articulation of the gold answer in the visible trace. Together, these findings suggest that explicit chain-of-thought primarily serves to *surface* an answer that has already been internally computed: latent reasoning precedes explicit verbal reasoning.

Latent reasoning is substantially stronger in high-resource languages. Comparisons across languages using AUTC and LRS reveal clear multilingual disparities. For MGSM, high-resource languages such as English and Chinese obtain both high AUTC and high LRS. For example, English achieves AUTC 0.52 and LRS 0.38 with R1-Qwen-7B, indicating that a large fraction of early accuracy cannot be explained by explicit answer articulation. In contrast, low-resource languages such as Swahili show much lower AUTC and LRS, meaning that the model struggles to produce correct answers under truncation and relies more heavily on fully articulated reasoning traces. Increasing model size from 7B to 32B seems to improve AUTC and LRS in all languages, but does *not* eliminate the gap: latent reasoning remains markedly less effective in low-resource languages. Overall, latent reasoning is a strongly resource-dependent phenomenon.

Latent reasoning is less pronounced on more challenging benchmarks. On Multilingual AIME, both AUTC and LRS drop sharply across languages and model sizes compared to MGSM. For example, LRS decreases from about 0.38 on MGSM to 0.03 on Multilingual AIME for English with R1-Qwen-7B, with similar trends in other languages and larger models. This pattern indicates that for problems requiring longer, more complex reasoning, models rarely form correct predictions early, prior to explicit answer articulation, and instead rely more heavily on extended explicit reasoning.

5 Latent State Dynamics

In §4, we observed that models exhibit clear signs of latent reasoning across languages – particularly on lower-complexity tasks – but with substantial crosslingual variation: high-resource languages tend to reach correct predictions earlier and more reliably. To better understand the origins of these differences, we must go beyond surface-level outputs and examine the model's internal representations. We therefore turn to **RQ2**: *Do different languages rely on different internal latent reasoning mechanisms?* To answer this question, we analyze both the layer-wise evolution of the model's implicit predictions (§5.1) and the similarity of hidden states across languages (§5.2).

5.1 Dynamic of Ranking Across Layers

To investigate whether different languages rely on distinct *internal* latent reasoning mechanisms, we analyze how evidence for the correct answer emerges across model layers using the logit lens (Nostalgebraist, 2020). While the logit lens is not a perfect probe of intermediate representations, particularly due to residual stream entanglement

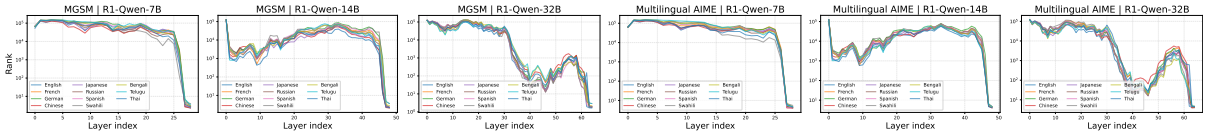

Figure 3: Layer-wise rank of the gold answer obtained via logit lens across languages on MGSM (left three panels) and Multilingual AIME (right three panels). Rank trajectories exhibit highly similar trends across languages, suggesting that latent reasoning progresses through comparable layer-wise transformations regardless of language.

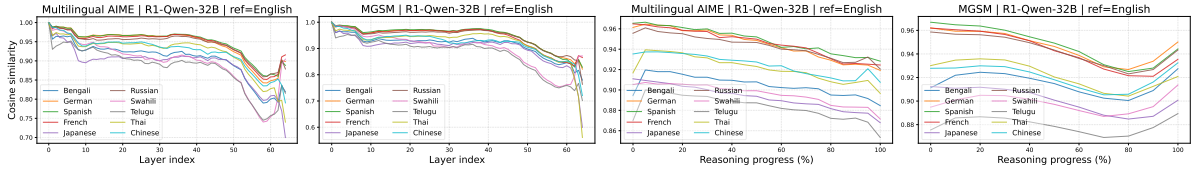

Figure 4: Aggregated cosine similarity between hidden states in each language and English (reference), averaged over both reasoning steps and layers, for R1-Qwen-32B. High-resource languages show consistently higher similarity to English, suggesting convergence toward an English-centered latent reasoning pathway.

(Belrose et al., 2025), it remains a useful diagnostic tool for tracking *relative* changes in answer salience across layers when applied consistently within the same model (Wendler et al., 2024; Wang et al., 2025b). Concretely, at each layer, we project the hidden state (i.e., residual stream activation) through the model's layer normalization and unembedding matrix and record the rank of the gold answer.[5] By comparing these *rank trajectories* across languages for a fixed model, we can assess whether layers play comparable functional roles in latent reasoning across different languages.

Figure 3 shows rank trajectories across languages, models, and datasets. A striking observation is that all languages exhibit highly similar ranking curves for a fixed model, suggesting that the internal mechanism used to form the solution is largely *language-invariant*. Despite differences in surface language realization and accuracy, **the underlying latent computation appears to follow the same structural progression across layers**.

At the same time, we observe distinct patterns across model sizes. These differences suggest that model capacity can shape latent reasoning dynamics, consistent with prior work showing that larger models exhibit qualitatively different intermediate-layer behavior, such as stronger representation compression, compared to smaller models (Skean et al., 2025). In particular, larger models appear to distribute reasoning more evenly across depth, allowing intermediate representations to encode increasingly informative abstractions. Notably, this pattern aligns with recent findings that multilingual models maintain a largely language-independent conceptual space in their middle layers (Wang et al., 2025b; Lu et al., 2025). The emergence of intermediate answer salience in larger models may therefore reflect a greater capacity to exploit this shared space, enabling earlier and more stable accumulation of evidence toward the correct solution.

[5]The gold answer is always a numeric value and is identical across languages. We track the rank of the first token, as generating this token is a necessary condition for producing the correct final answer. This practice is widely adopted in prior work (Hernandez et al., 2024; Kargaran et al., 2025).

5.2 Hidden State Similarity

We showed that a model presents consistently similar rank trajectories across layers across languages in §5.1. We further hypothesize that such consistency may reflect an *English-centered* latent reasoning process, in which reasoning in other languages implicitly aligns with the pathway used for English.

To test this hypothesis, we compute cosine similarity between the hidden states of each target language and those of English. For each example in a target language, at each truncation ratio, we extract the hidden state of the final token of the reasoning trace and measure its similarity to the corresponding hidden state of its English counterpart. We aggregate similarities in two ways: (i) averaging over layers and (ii) averaging over reasoning steps.

Figure 4 summarizes these results. Overall, we observe consistently higher similarity with English for high-resource languages, including those using non-Latin scripts such as Chinese and Russian. In contrast, mid-resource languages with distinct scripts (e.g., Japanese) and low-resource languages (e.g., Telugu) exhibit lower similarity to English. This pattern is stable across layers and reasoning steps, suggesting that **reasoning in high-resource languages may be processed in a representational space more closely aligned with English,**

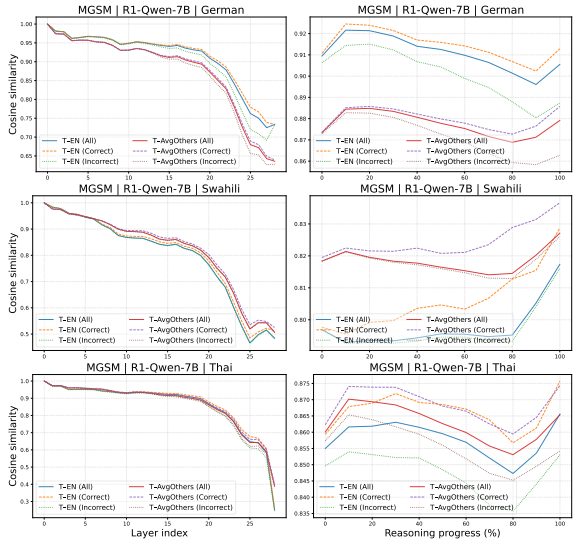

Figure 5: Comparison of cosine similarity with English versus average similarity with other languages, shown separately for correctly and incorrectly solved examples. High-resource languages show stronger alignment with English, whereas low- and mid-resource languages show weaker or correctness-dependent alignment.

whereas mid- and low-resource languages deviate more substantially.[6] However, similarity alone does not distinguish whether this alignment reflects an English-centered reasoning process or merely arises from *shared correct answers*.

To disentangle these effects, for each language, we group MGSM examples into *correct* and *incorrect* sets based on pass@10 accuracy, and compare the (i) similarity with English and (ii) average similarity with other languages (excluding itself and English). Figure 5 presents the results over layers and reasoning steps (see full results in §D.2).

Language resource level modulates alignment with English in latent reasoning. High-resource languages (e.g., German) exhibit consistently strong alignment with English that is largely *independent of correctness*: incorrect instances remain nearly as similar to their English counterparts as correct ones, particularly in early reasoning stages. Although a modest gap emerges as the reasoning trace unfolds, this appears to reflect increasing commitment to an (incorrect) solution rather than a shift away from English-aligned latent trajectories. The absence of a substantial correct–incorrect gap overall indicates that, for high-resource languages, alignment with English reflects a stable latent reasoning trajectory rather than a byproduct of successful solution formation. In contrast, low-resource languages (e.g., Swahili) show weaker alignment with English across both correct and incorrect examples, while exhibiting relatively higher similarity to other languages. This pattern suggests a more autonomous subspace that is less shaped by English-centric post-training and more influenced by language-specific representations formed during pretraining (Chang et al., 2022; Liu et al., 2024). Mid-resource languages (e.g., Thai) occupy an intermediate regime: alignment with English is more pronounced for correct instances than for incorrect ones, suggesting that convergence toward English-like latent trajectories occurs primarily when reasoning is successful.

[6] Crosslingual similarity may also be influenced by linguistic and typological relatedness between languages, which could partially contribute to the observed alignment patterns.

6 Is Latent Reasoning Behavior Specific to Distilled Models?

A potential concern is *whether the observed latent reasoning behavior is specific to distilled reasoning models*, or *whether it generalizes to models with different post-training paradigms*. Prior work suggests that reinforcement learning (RL) and distillation can lead to different reasoning behaviors (Yue et al., 2025; Kim et al., 2025; Baek and Tegmark, 2025). To investigate this, we extend our latent reasoning identification (§4) to additional models.

6.1 Experimental Setup

Model Selection. In addition to the distilled reasoning model `R1-Qwen-32B` used in §4, we consider two additional models derived from the same base model: (1) `Qwen2.5-32B-Instruct` (Qwen Team et al., 2025), an instruction-tuned model without explicit reasoning distillation, and (2) `Nemotron-32B` (Moshkov et al., 2025), a reasoning model trained with RL-based post-training.

Evaluation Metrics. We adopt the same truncation-based evaluation as in the main experiments (cf. §4.2), including **AUTC** (Area Under the Truncation Accuracy Curve), **AUGC** (Area Under the Gold-in-Trace Curve), and **LRS** (Latent Reasoning Score). These metrics jointly characterize when correctness emerges during reasoning and whether it depends on explicit answer articulation.

6.2 Results and Discussion

Table 2 reports our results. Overall, the main findings remain consistent across all models. In particular, we observe high AUTC across models, indicating that correct answers can often be produced

Dataset	Model	Metric	DE	EN	ES	FR	RU	ZH	BN	JA	TH	SW	TE
MGSM	Qwen2.5-32B	AUTC	0.62	0.70	0.67	0.63	0.65	0.66	0.66	0.63	0.63	0.41	0.49
		AUGC	0.15	0.16	0.15	0.14	0.16	0.20	0.17	0.19	0.18	0.17	0.17
		LRS	0.50	0.57	0.55	0.53	0.52	0.50	0.52	0.48	0.49	0.33	0.39
	R1-Qwen-32B	AUTC	0.67	0.75	0.69	0.64	0.68	0.70	0.61	0.63	0.69	0.38	0.39
		AUGC	0.20	0.25	0.20	0.17	0.21	0.30	0.23	0.21	0.28	0.20	0.23
		LRS	0.51	0.53	0.52	0.51	0.51	0.45	0.44	0.47	0.46	0.30	0.30
	Nemotron-32B	AUTC	0.84	0.88	0.86	0.80	0.84	0.83	0.70	0.78	0.77	0.34	0.26
		AUGC	0.86	0.82	0.84	0.81	0.77	0.81	0.81	0.82	0.77	0.45	0.57
		LRS	0.09	0.11	0.11	0.11	0.15	0.11	0.11	0.11	0.14	0.18	0.11
Multilingual AIME	Qwen2.5-32B	AUTC	0.03	0.04	0.03	0.03	0.04	0.02	0.01	0.01	0.03	0.02	0.00
		AUGC	0.07	0.04	0.32	0.48	0.16	0.09	0.03	0.33	0.62	0.02	0.00
		LRS	0.02	0.04	0.02	0.02	0.03	0.02	0.01	0.00	0.01	0.02	0.00
	R1-Qwen-32B	AUTC	0.06	0.18	0.08	0.09	0.10	0.13	0.04	0.04	0.07	0.01	0.01
		AUGC	0.29	0.61	0.32	0.72	0.66	0.75	0.18	0.74	0.82	0.05	0.17
		LRS	0.03	0.06	0.04	0.02	0.03	0.03	0.02	0.01	0.02	0.00	0.00
	Nemotron-32B	AUTC	0.38	0.43	0.41	0.38	0.36	0.32	0.30	0.37	0.30	0.05	0.13
		AUGC	0.79	0.88	0.84	0.75	0.73	0.73	0.93	0.81	0.79	0.26	0.56
		LRS	0.03	0.02	0.03	0.04	0.03	0.03	0.02	0.03	0.03	0.03	0.04

Table 2: Truncation-based metrics across different post-training paradigms. While RL-trained models (Nemotron-32B) achieve stronger overall performance (higher AUTC), multilingual asymmetry persists. Higher AUGC in RL models indicates earlier answer articulation, often followed by extended reasoning ("overthinking").

from partial reasoning traces. However, latent reasoning remains uneven across languages for both instruction-tuned and RL-based models, suggesting that this disparity is not specific to distilled models.

Early answer articulation and overthinking in RL models. Nemotron-32B achieves substantially higher AUTC than the other models, indicating stronger overall reasoning capability. However, it also exhibits much higher AUGC and relatively low LRS, suggesting that correctness often relies on early answer articulation. Inspection of reasoning traces reveals that the model frequently produces the correct answer within the first $\sim$20% of steps, followed by extended reflective reasoning. This behavior is consistent with *overthinking* (Chen et al., 2025b; Sui et al., 2025), where the model continues generation after already reaching the answer.

Task difficulty governs latent reasoning dynamics. Despite differences in post-training, all models exhibit similar trends across benchmarks. On MGSM, models can often produce correct answers with minimal or even no reasoning trace, indicating the presence of latent reasoning. In contrast, on the more challenging Multilingual AIME benchmark, correct predictions typically require explicit answer articulation, resulting in lower AUTC/LRS and higher AUGC. This further confirms that task difficulty plays a central role in shaping latent reasoning dynamics.

7 Complementary Analysis: Memorization or Latent Reasoning

In §4, we observed that for MGSM, models can sometimes predict the correct answer even when no reasoning trace is provided (i.e., truncation ratio 0%). While this behavior may suggest that the model has already implicitly computed the answer and thus exhibits latent reasoning, an alternative explanation is that the model has memorized the solution due to exposure during pre-training or post-training, a phenomenon commonly referred to as *data contamination* or *benchmark leakage* (Xu et al., 2024; Balloccu et al., 2024). Under such circumstances, correct predictions may arise from direct recall rather than genuine latent reasoning.

To disentangle memorization from latent reasoning, we conduct a complementary analysis that probes the model's sensitivity to controlled question perturbations. The intuition is as follows: if a model relies on memorization, small but *meaning-altering* edits should not substantially change its predictions, as the original answer may still be recalled. In contrast, a reasoning-based model should adapt its prediction to such changes. Conversely, when the underlying meaning of a question is preserved but surface form is altered via *paraphrasing*, a reasoning model should remain robust and continue to produce the correct answer, whereas a memorization-based model may fail.

7.1 Method

We focus on MGSM and restrict our attention to **questions that are answered correctly under the pass@10 when no reasoning trace is provided** (i.e., truncation ratio $= 0\%$), for each language and each model independently, as they are particularly ambiguous cases where memorization and latent reasoning are difficult to distinguish. For each question, we apply the following editing strategies.[7]

[7] See §E for the details of altering the original problem.

Edit Method	Model	Setup	DE	EN	ES	FR	RU	ZH	BN	JA	TH	SW	TE
NumEdit (↓)	R1-Qwen-7B	w/o Trace	0.40	0.31	0.29	0.34	0.36	0.33	0.38	0.51	0.33	0.52	0.40
		w/ Trace	0.27	0.25	0.21	0.22	0.24	0.16	0.21	0.32	0.30	0.19	0.17
	R1-Qwen-14B	w/o Trace	0.33	0.27	0.29	0.35	0.35	0.29	0.30	0.43	0.29	0.47	0.39
		w/ Trace	0.19	0.16	0.21	0.20	0.19	0.13	0.15	0.28	0.15	0.22	0.12
	R1-Qwen-32B	w/o Trace	0.31	0.32	0.29	0.27	0.35	0.28	0.32	0.47	0.31	0.37	0.33
		w/ Trace	0.18	0.20	0.20	0.19	0.19	0.19	0.16	0.26	0.18	0.22	0.18
Paraphrase (↑)	R1-Qwen-7B	w/o Trace	0.66	0.70	0.73	0.72	0.71	0.74	0.67	0.64	0.57	0.45	0.58
		w/ Trace	1.00	1.00	0.96	0.99	0.99	0.96	0.95	0.90	0.96	0.35	0.85
	R1-Qwen-14B	w/o Trace	0.79	0.86	0.73	0.77	0.85	0.72	0.79	0.75	0.81	0.55	0.63
		w/ Trace	0.97	1.00	0.98	0.96	0.99	0.95	0.95	0.94	0.99	0.69	0.54
	R1-Qwen-32B	w/o Trace	0.86	0.81	0.86	0.83	0.90	0.91	0.90	0.85	0.85	0.83	0.76
		w/ Trace	0.99	0.99	0.98	0.96	0.98	0.96	0.99	0.98	0.96	0.92	0.78

Table 3: Pass@10 results on edited MGSM questions across 11 languages. For **NumEdit** (↓), values report the *matching ratio* with the original gold answer after a single-number perturbation (lower is better). For **Paraphrase** (↑), values report *accuracy*, as the gold answer is unchanged. "w/o Trace" denotes inference without a reasoning trace (empty `<think></think>` block), while "w/ Trace" allows a newly generated trace.

NumEdit We modify exactly one numerical value in the original question while keeping the rest of the problem unchanged. The edit is chosen such that it alters the solution, and therefore, the original gold answer is no longer correct. Accordingly, we evaluate NumEdit using the *matching ratio* with the original gold answer, where lower values indicate better sensitivity to the perturbation.

Paraphrase We paraphrase and reorder the question text while preserving all numerical values, mathematical expressions, and the overall semantics. The paraphrased question is logically *equivalent* to the original, and thus the gold answer is unchanged. In this case, we evaluate performance using standard *accuracy* (the higher the better).

7.2 Results and Discussions

Models exhibit partial memorization, but latent reasoning remains evident. Table 3 shows that under NumEdit, models still match the original gold answer in a non-trivial fraction of cases, with matching ratios typically around 30% across languages, with high-resource languages (e.g., English) generally showing a lower matching ratio than low-resource languages (e.g., Swahili), under the *w/o Trace* setting. However, the matching ratio consistently decreases when models are allowed to generate a new reasoning trace, often dropping below 25% for smaller models and below 20% for the 32B model in most languages.[8] Taken together, these results indicate that while memorization is present, models largely recompute solutions rather than merely recalling memorized answers, providing evidence in favor of latent reasoning.

[8]We use `Gemini-2.5-Flash` to validate NumEdit; around 10% of the edited questions retain the same gold answer (see §E). As a result, the reported matching ratios should be interpreted as an upper bound on the extent of memorization.

Robustness to paraphrasing argues against pattern-matching memorization. In Paraphrase, pass@10 accuracy under the *w/o Trace* setting is typically above 70%, and increases further when allowing the model to generate a new reasoning trace across languages. For instance, R1-Qwen-32B reaches near-perfect accuracy in high-resource languages such as English and German under the *w/ Trace* setting. Although performance is lower for under-resourced languages (e.g., Swahili), the same trend holds. Additionally, accuracy consistently improves with model scale, and explicit reasoning traces further amplify this effect. These results suggest that the models do not rely solely on surface-level pattern matching to the original question wording. Instead, their robustness to paraphrasing provides converging evidence that the models engage in genuine reasoning processes.

8 Conclusion

We present a systematic study of multilingual latent reasoning in LRMs. Using truncation-based analyses, we show that LRMs can perform latent reasoning, but this capability is highly uneven: it is strong in resource-rich languages on easier tasks, weak in low-resource languages, and is largely undetectable on more challenging benchmarks. Our representational analyses reveal highly consistent layer-wise dynamics across languages, with latent reasoning converging toward an English-centered pathway, particularly for high-resource languages and correctly solved instances. Finally, we demonstrate that these behaviors cannot be explained by surface-level memorization alone. Together, our findings suggest that current LRMs exhibit real but fragile multilingual latent reasoning, shaped by English-centric post-training and task complexity.

Limitations

While this work offers a systematic analysis of multilingual latent reasoning in LRMs, it is subject to several limitations.

First, we implement reasoning-trace truncation at the *step* level rather than the *token* level, following previous work. While step-based truncation aligns with sentence-level CoT structure and improves interpretability, finer-grained token-level truncation may reveal more precise dynamics of latent answer formation and is left to future work.

Second, our Gold-in-Trace metric relies on string matching to detect whether the gold answer appears in the reasoning trace, which may yield false positives when intermediate values happen to match the final answer. However, as discussed in §B, this effect is limited in our setting due to the prevalence of multi-digit answers and conditioning on correct predictions. More sophisticated and robust matching strategies can be explored in future work.

Third, our experiments focus on mathematical reasoning benchmarks. While this choice enables precise measurement of latent reasoning through clear numeric answers and structured multi-step reasoning, it may limit generalizability to other reasoning domains, such as commonsense reasoning. Extending our analysis to such tasks is an important direction for future work.

Fourth, due to computational constraints, our experiments are limited to models up to 32B parameters. While we include models with different post-training paradigms and observe consistent trends, extending the analysis to larger-scale models can be an interesting direction for future work.

Finally, our work is primarily diagnostic and does not propose a new inference method. Instead, we focus on analyzing whether latent reasoning exists across languages and how its behavior varies with language and task difficulty. While our findings provide insights into multilingual reasoning (e.g., asymmetry across languages and task-dependent dynamics), future work can translate these insights into concrete improvements for multilingual inference, such as transferring reasoning capabilities from high- to low-resource languages.

Ethical Considerations

Use of AI Assistants. The authors acknowledge the use of `ChatGPT 5.2` for language editing (grammar, clarity, and coherence) and limited assistance with code implementation;[9] all technical content and experimental decisions were made by the authors.

Acknowledgments

This research was supported by the Munich Center for Machine Learning (MCML) and German Research Foundation (DFG, grant SCHU 2246/14-1).

References

David D. Baek and Max Tegmark. 2025. Towards understanding distilled reasoning models: A representational approach. *Preprint*, arXiv:2503.03730.

Niyati Bafna, Tianjian Li, Kenton Murray, David R. Mortensen, David Yarowsky, Hale Sirin, and Daniel Khashabi. 2025. The translation barrier hypothesis: Multilingual generation with large language models suffers from implicit translation failure. *Preprint*, arXiv:2506.22724.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2025. Eliciting latent predictions from transformers with the tuned lens. *Preprint*, arXiv:2303.08112.

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.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. 2024. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14113–14130, Miami, Florida, USA. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

[9] `https://chatgpt.com/`

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. 2024. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Beiduo Chen, Tiancheng Hu, Caiqi Zhang, Robert Litschko, Anna Korhonen, and Barbara Plank. 2026. Decoupling the effect of chain-of-thought reasoning: A human label variation perspective. *Preprint*, arXiv:2601.03154.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. 2025a. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. *Preprint*, arXiv:2505.16782.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025b. Do not think that much for 2+3=? on the overthinking of o1-like llms. *Preprint*, arXiv:2412.21187.

Jeffrey Cheng and Benjamin Van Durme. 2024. Compressed chain of thought: Efficient reasoning through dense representations. *Preprint*, arXiv:2412.13171.

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.

Yuntian Deng, Yejin Choi, and Stuart Shieber. 2024. From explicit cot to implicit cot: Learning to internalize cot step by step. *Preprint*, arXiv:2405.14838.

Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. 2024. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *Trans. Mach. Learn. Res.*, 2024.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. A survey of multilingual reasoning in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8920–8936, Suzhou, China. Association for Computational Linguistics.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2025. Training large language models to reason in a continuous latent space. *Preprint*, arXiv:2412.06769.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Francis Begnaud Hildebrand. 1987. *Introduction to numerical analysis*. Courier Corporation.

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.

Shulin Huang, Yiran Ding, Junshu Pan, and Yue Zhang. 2025. Beyond english-centric training: How reinforcement learning improves cross-lingual reasoning in llms. *Preprint*, arXiv:2509.23657.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Deokhyung Kang, Seonjeong Hwang, Daehui Kim, Hyounghun Kim, and Gary Geunbae Lee. 2025. Why do multilingual reasoning gaps emerge in reasoning language models? *Preprint*, arXiv:2510.27269.

Amir Hossein Kargaran, Yihong Liu, François Yvon, and Hinrich Schuetze. 2025. How programming concepts and neurons are shared in code language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26905–26917, Vienna, Austria. Association for Computational Linguistics.

Minwu Kim, Anubhav Shrestha, Safal Shrestha, Aadim Nepal, and Keith Ross. 2025. Reinforcement learning vs. distillation: Understanding accuracy and capability in llm reasoning. *Preprint*, arXiv:2505.14216.

Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy Liang. 2019. Spoc: Search-based pseudocode to code. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11883–11894.

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.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-Long Bi, Ling-Rui Mei, Junfeng Fang, Xiao Liang, Zhijiang Guo, and 2 others. 2025. From system 1 to system 2: A survey of reasoning large language models. *Preprint*, arXiv:2502.17419.

Tianhe Lin, Jian Xie, Siyu Yuan, and Deqing Yang. 2025. Implicit reasoning in transformers is reasoning through shortcuts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9470–9487, Vienna, Austria. Association for Computational Linguistics.

Yihong Liu, Chunlan Ma, Haotian Ye, and Hinrich Schuetze. 2024. TransliCo: A contrastive learning framework to address the script barrier in multilingual pretrained language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2476–2499, Bangkok, Thailand. Association for Computational Linguistics.

Meng Lu, Ruochen Zhang, Carsten Eickhoff, and Ellie Pavlick. 2025. Paths not taken: Understanding and mending the multilingual factual recall pipeline. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15077–15107, Suzhou, China. Association for Computational Linguistics.

Minjia Mao, Bowen Yin, Yu Zhu, and Xiao Fang. 2025. Early stopping chain-of-thoughts in large language models. *Preprint*, arXiv:2509.14004.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. 2025. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *Preprint*, arXiv:2504.16891.

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.

Nostalgebraist. 2020. Interpreting GPT: the logit lens. Blog post.

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.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron Wallace, and David Bau. 2023. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 548–560, Singapore. Association for Computational Linguistics.

Jacob Pfau, William Merrill, and Samuel R. Bowman. 2024. Let's think dot by dot: Hidden computation in transformer language models. *Preprint*, arXiv:2404.15758.

Ben Prystawski, Michael Li, and Noah D. Goodman. 2023. Why think step by step? reasoning emerges from the locality of experience. 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*.

Jirui Qi, Shan Chen, Zidi Xiong, Raquel Fernández, Danielle Bitterman, and Arianna Bisazza. 2025. When models reason in your language: Controlling thinking language comes at the cost of accuracy. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20279–20296, Suzhou, China. Association for Computational Linguistics.

Chen Qian, Dongrui Liu, Haochen Wen, Zhen Bai, Yong Liu, and Jing Shao. 2025. Demystifying reasoning dynamics with mutual information: Thinking tokens are information peaks in llm reasoning. *Preprint*, arXiv:2506.02867.

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.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. 2025. Reasoning with latent thoughts: On the power of looped transformers. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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. 2023. 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.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. 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.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *Trans. Mach. Learn. Res.*, 2025.

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

Jiangkuo Wang, Suyv Ma, and Mingpeng Wei. 2025a. Enhancing multilingual reasoning in LLMs: Insights from cross-linguistic correlations and optimal data proportions.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025b. Lost in multilinguality: Dissecting cross-lingual 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 Schuetze. 2025c. Language mixing in reasoning language models: Patterns, impact, and internal causes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2637–2665, Suzhou, China. Association for Computational Linguistics.

Xinpeng Wang, Nitish Joshi, Barbara Plank, Rico Angell, and He He. 2025d. Is it thinking or cheating? detecting implicit reward hacking by measuring reasoning effort. *Preprint*, arXiv:2510.01367.

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*.

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

Wilson Wu, John X. Morris, and Lionel Levine. 2024. Do language models plan ahead for future tokens? *Preprint*, arXiv:2404.00859.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *Preprint*, arXiv:2404.18824.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025a. SoftCoT: Soft chain-of-thought for efficient reasoning with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23336–23351, Vienna, Austria. Association for Computational Linguistics.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025b. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

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.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.

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.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. LangBridge: Multilingual reasoning without multilingual supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7502–7522, Bangkok, Thailand. Association for Computational Linguistics.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Preprint*, arXiv:2504.13837.

Raoyuan Zhao, Yihong Liu, Hinrich Schuetze, and Michael A. Hedderich. 2026. A comprehensive evaluation of multilingual chain-of-thought reasoning: Performance, consistency, and faithfulness across languages. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5223–5247, Rabat, Morocco. Association for Computational Linguistics.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

A Reasoning Trace Statistics

Table 4 reports the average and median number of reasoning steps for MGSM and Multilingual AIME across languages.

For MGSM, we observe that most languages exhibit an average of approximately 10 reasoning steps. Accordingly, we adopt a truncation granularity of 10%, which roughly corresponds to removing one reasoning step at a time.

In contrast, Multilingual AIME displays substantially longer reasoning traces on average, while the median number of steps is markedly smaller. This discrepancy is primarily driven by a small number of outlier instances, which is expected given the limited dataset size (60 problems). To better account for crosslingual variation in reasoning length while controlling computational cost, we therefore use a finer truncation granularity of 5% for Multilingual AIME.

B Potential False Positives in Gold-in-Trace Matching

A potential limitation of the **Gold-in-Trace** metric is that it relies on string matching to detect whether the gold answer appears in the truncated reasoning trace. In principle, this may introduce false positives when intermediate values happen to match the final answer. For example, if the correct answer is "5", an intermediate step such as "5 apples remain after step 2" would be counted as a match, even if it is unrelated to the final solution. However, we expect such cases to be rare in our experimental setting for the following reasons.

First, in both MGSM and Multilingual AIME, gold answers are typically multi-digit integers. The probability that an unrelated intermediate value exactly matches the final answer is therefore low. To quantify this, we report the proportion of single-digit answers in each dataset in Table 5.

Second, Gold-in-Trace and AUGC are computed *only over correctly solved instances*. If the model produces the correct final answer and the same value appears earlier in the reasoning trace, this is more likely to reflect early internal convergence rather than random coincidence.

While false positives cannot be completely ruled out, the combination of multi-digit answer distributions and conditioning on correct predictions substantially reduces their likelihood. We therefore consider Gold-in-Trace to be a reasonable proxy for early answer articulation in our setting.

C Complete Truncation Results

C.1 Truncation Curves

Figures 6, 7, 8 and their Multilingual AIME counterparts (Figure 9, 10, 11) visualize pass@k accuracy ($k = 1, 5, 10$) and the gold-in-trace rate as a function of reasoning-trace truncation ratio across languages and model sizes. Across models, MGSM exhibits a characteristic pattern in which accuracy increases well before the gold answer is explicitly articulated, whereas on Multilingual AIME accuracy typically remains low until late

Dataset	Model	EN	FR	DE	ZH	JA	RU	ES	SW	BN	TE	TH
MGSM	R1-Qwen-7B	9.5 (9.0)	11.4 (9.0)	10.9 (9.5)	10.8 (9.0)	13.1 (8.0)	9.4 (8.0)	9.6 (8.0)	21.6 (9.0)	16.1 (14.0)	23.9 (20.0)	18.9 (9.0)
	R1-Qwen-14B	11.2 (10.0)	13.6 (13.0)	14.7 (13.0)	16.3 (13.0)	11.1 (10.0)	13.8 (12.0)	13.8 (11.0)	32.6 (18.0)	15.8 (13.0)	22.6 (17.0)	12.4 (11.0)
	R1-Qwen-32B	23.2 (19.0)	9.2 (8.0)	11.7 (11.0)	20.2 (18.0)	10.2 (9.0)	10.1 (9.0)	10.3 (8.0)	20.3 (16.0)	14.8 (13.0)	18.2 (15.5)	12.6 (11.0)
Multilingual AIME	R1-Qwen-7B	234.9 (168.0)	197.2 (132.0)	206.3 (124.5)	263.9 (181.0)	169.0 (47.0)	143.8 (54.0)	226.1 (152.5)	92.5 (10.0)	114.4 (66.5)	117.5 (125.5)	95.6 (15.5)
	R1-Qwen-14B	231.8 (169.0)	183.4 (120.5)	162.4 (107.0)	267.2 (178.5)	144.6 (67.5)	217.8 (157.0)	205.3 (139.0)	95.7 (66.0)	141.5 (108.0)	122.3 (112.0)	99.3 (62.5)
	R1-Qwen-32B	282.4 (207.0)	123.5 (102.5)	213.7 (152.0)	279.6 (215.0)	411.4 (89.0)	216.5 (119.0)	166.0 (114.0)	112.2 (43.5)	138.6 (121.5)	122.5 (126.0)	159.3 (116.5)

Table 4: Reasoning-step statistics across languages on MGSM and Multilingual AIME. Each cell reports the average number of reasoning steps, with the median shown in parentheses.

Dataset	#Questions	#Single-Digit	%
MGSM	250	38	15.2%
AIME-2024 (Multilingual)	30	0	0.0%
AIME-2025 (Multilingual)	30	0	0.0%

Table 5: Proportion of single-digit gold answers.

truncation ratios. The figures also highlight substantial crosslingual variation, with high-resource languages generally achieving earlier and more stable gains than mid- and low-resource languages.

C.2 Causal Decompositions

Figures 12, 13, 14 (MGSM) and Figures 15, 16, 17 (Multilingual AIME) further decompose accuracy gains between consecutive truncation intervals into three cases: (i) the gold answer is newly articulated in the added reasoning steps, (ii) it was already present earlier in the visible trace, or (iii) it has not yet appeared in the truncated trace. This analysis disentangles improvements driven by explicit answer articulation from those arising prior to any verbalized solution. On MGSM, gains at early and intermediate truncation ratios are largely attributed to case (iii), indicating that correct predictions often emerge before the answer is explicitly articulated. In contrast, on Multilingual AIME, gains are sparser and increasingly dominated by cases (i) and (ii), reflecting a stronger dependence on explicit reasoning and a reduced role for early latent solution formation.

D Complete Similarity results

D.1 Similarity with English

We conduct a focused analysis of crosslingual representational alignment by computing the cosine similarity between hidden states in each target language and those of English. For each instance, and at each truncation ratio, we extract the hidden state corresponding to the final token of the partial reasoning trace and compare it with the hidden state obtained from the English version of the same problem. To summarize these similarities, we aggregate them both across layers and across reasoning steps. Figure 18 and Figure 19 present the resulting trends for MGSM and Multilingual AIME, respectively.

Across both benchmarks, we observe systematic differences in representational alignment with English that seem to correlate with language resource levels: high-resource languages exhibit consistently stronger alignment with English representations, whereas mid- and low-resource languages show reduced alignment.

D.2 Similarity vs. Correctness

We examine whether answer correctness affects crosslingual alignment by grouping MGSM examples into *correct* and *incorrect* sets and comparing their similarity to English and to other languages. Results for R1-Qwen-{7B,14B,32B} are shown in Figures 20, 21, 22. Across models, alignment with English seems to vary systematically with language resource level. High-resource languages show consistently strong similarity to English regardless of correctness, whereas low-resource languages (e.g., Swahili) are more similar to other non-English languages, suggesting a more autonomous latent subspace. Mid-resource languages exhibit an intermediate pattern, with stronger English alignment primarily for correctly solved instances.

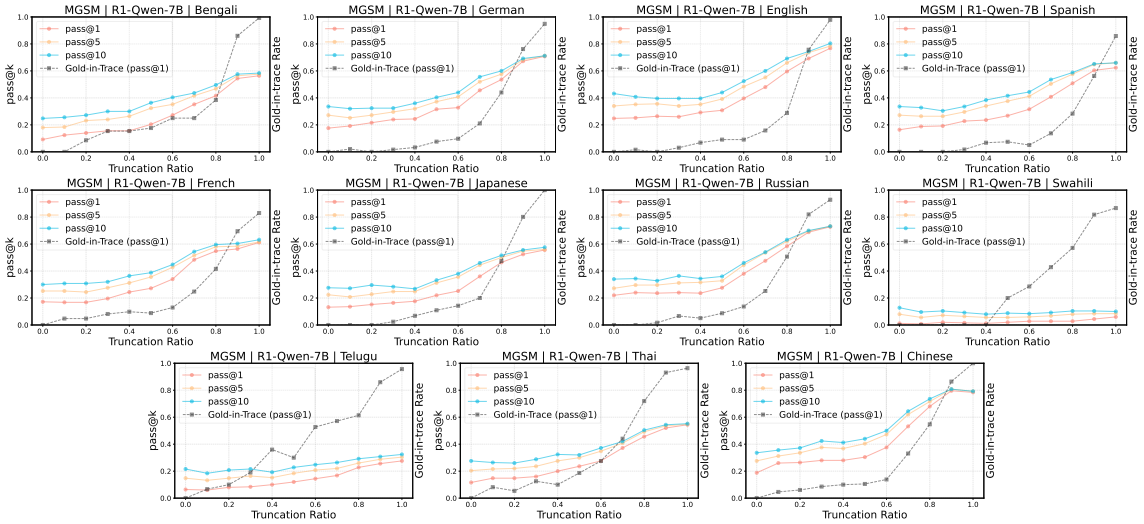

Figure 6: Pass@k accuracy ($k = 1, 5, 10$) and gold-in-trace rate under reasoning-trace truncation for **R1-Qwen-7B** on **MGSM**. The model shows stronger latent reasoning in high-resource languages (e.g., English).

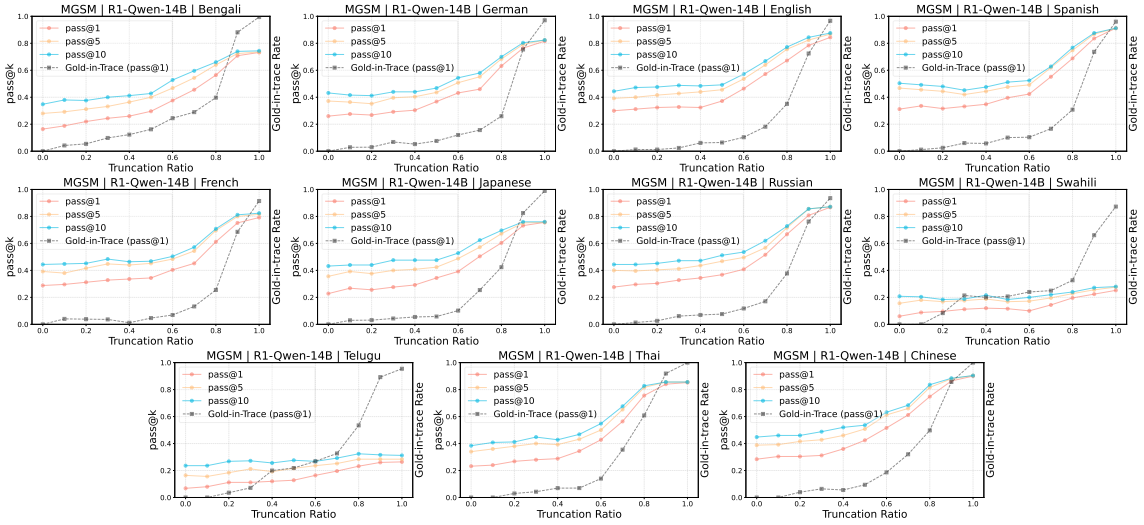

Figure 7: Pass@k accuracy ($k = 1, 5, 10$) and gold-in-trace rate under reasoning-trace truncation for **R1-Qwen-14B** on **MGSM**. The model shows stronger latent reasoning in high-resource languages (e.g., English).

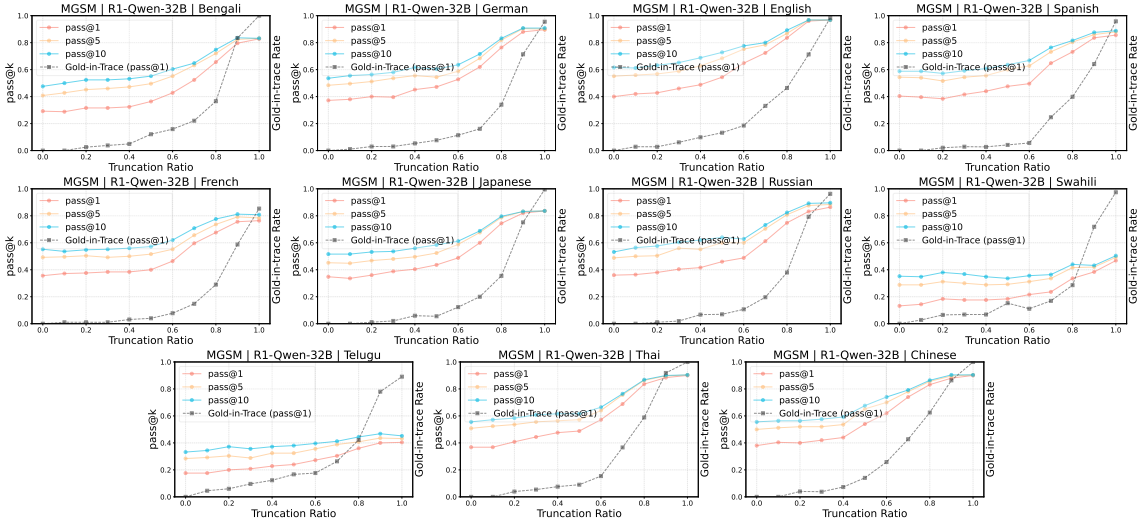

Figure 8: Pass@k accuracy ($k = 1, 5, 10$) and gold-in-trace rate under reasoning-trace truncation for **R1-Qwen-32B** on **MGSM**. The model shows stronger latent reasoning in high-resource languages (e.g., English).

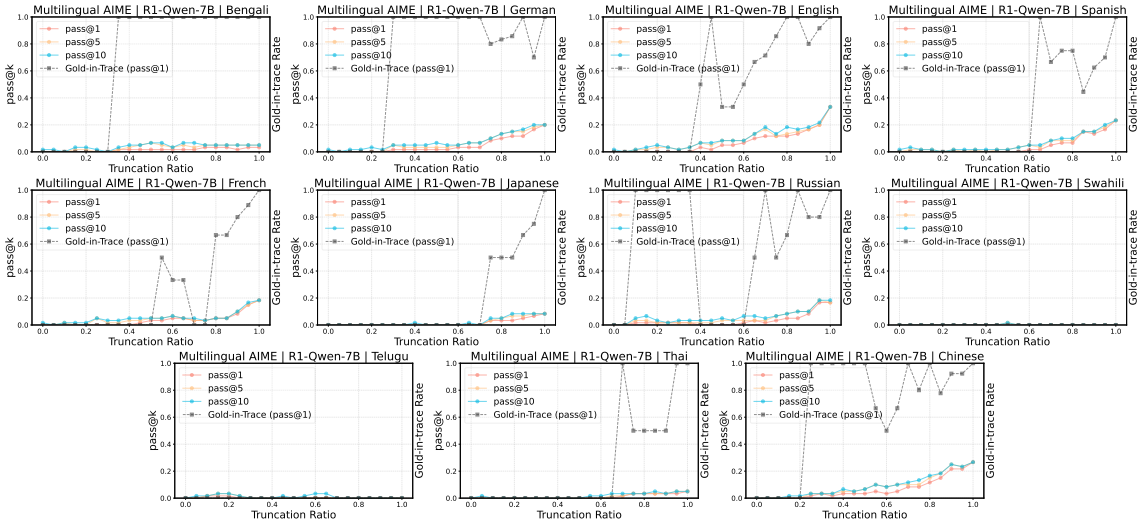

Figure 9: Pass@k accuracy ($k = 1, 5, 10$) and gold-in-trace rate under reasoning-trace truncation for **R1-Qwen-7B** on **Multilingual AIME**. Latent reasoning is less pronounced compared to MGSM.

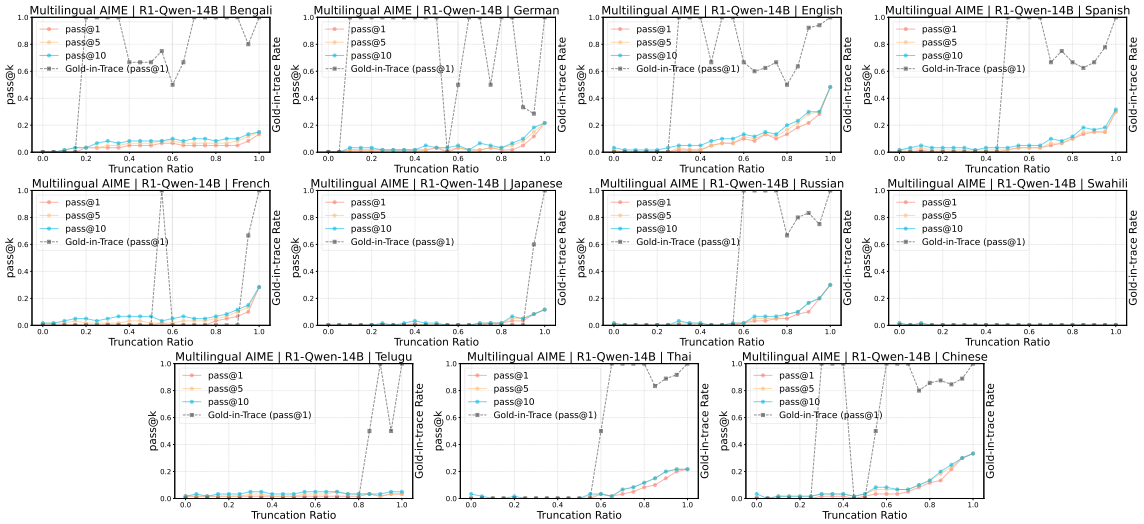

Figure 10: Pass@k accuracy ($k = 1, 5, 10$) and gold-in-trace rate under reasoning-trace truncation for **R1-Qwen-14B** on **Multilingual AIME**. Latent reasoning is less pronounced compared to MGSM.

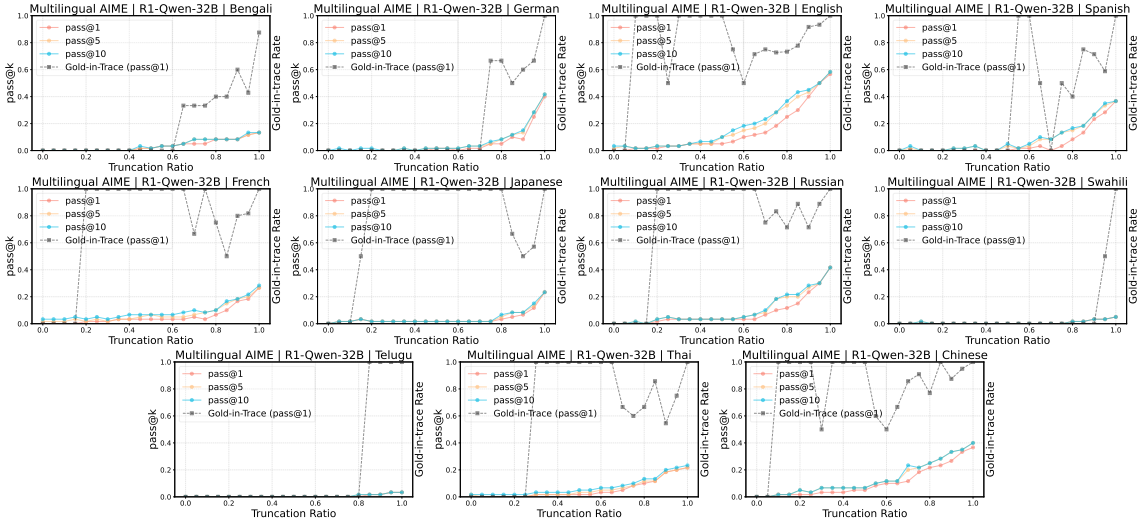

Figure 11: Pass@k accuracy ($k = 1, 5, 10$) and gold-in-trace rate under reasoning-trace truncation for **R1-Qwen-32B** on **Multilingual AIME**. Latent reasoning is less pronounced compared to MGSM.

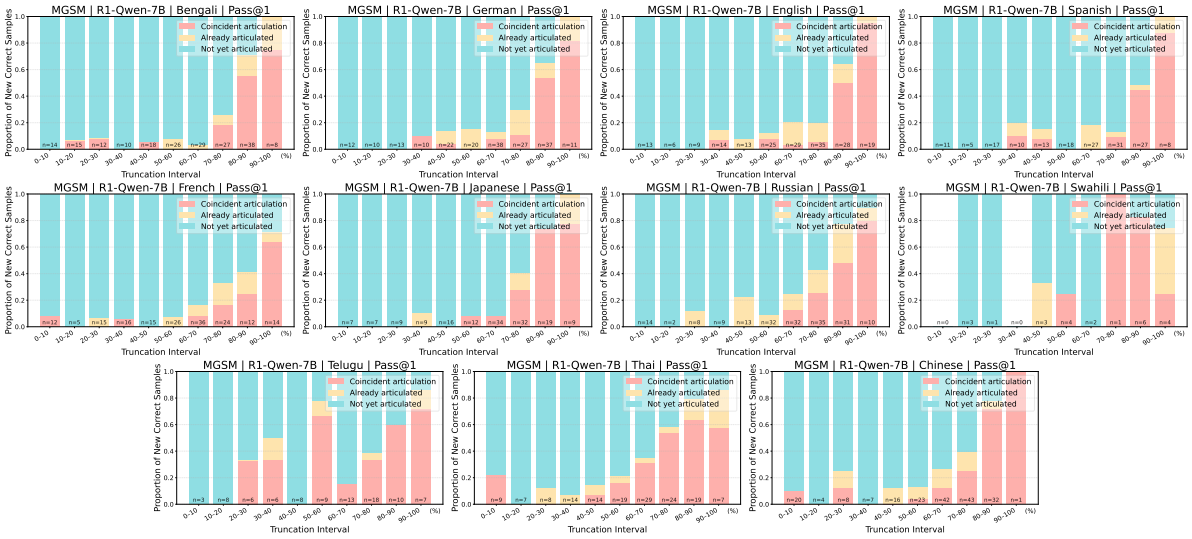

Figure 12: Causal decomposition of newly correct predictions across truncation intervals on **MGSM** with **R1-Qwen-7B**. Each bar partitions gains into three cases: (**i**) the gold answer is first articulated in the newly added reasoning steps, (**ii**) it was already articulated in earlier steps, or (**iii**) it has not yet appeared in the visible truncated trace. Early and intermediate gains are dominated by case (**iii**), indicating latent reasoning.

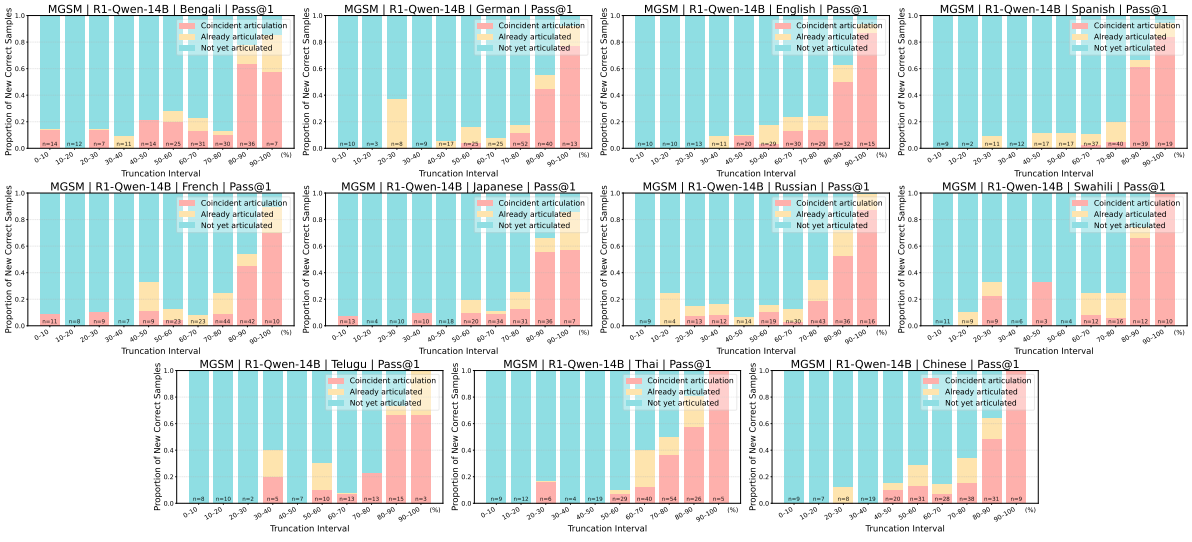

Figure 13: Causal decomposition of newly correct predictions across truncation intervals on **MGSM** with **R1-Qwen-14B**. Each bar partitions gains into three cases: (**i**) the gold answer is first articulated in the newly added reasoning steps, (**ii**) it was already articulated in earlier steps, or (**iii**) it has not yet appeared in the visible truncated trace. Early and intermediate gains are dominated by case (**iii**), indicating latent reasoning.

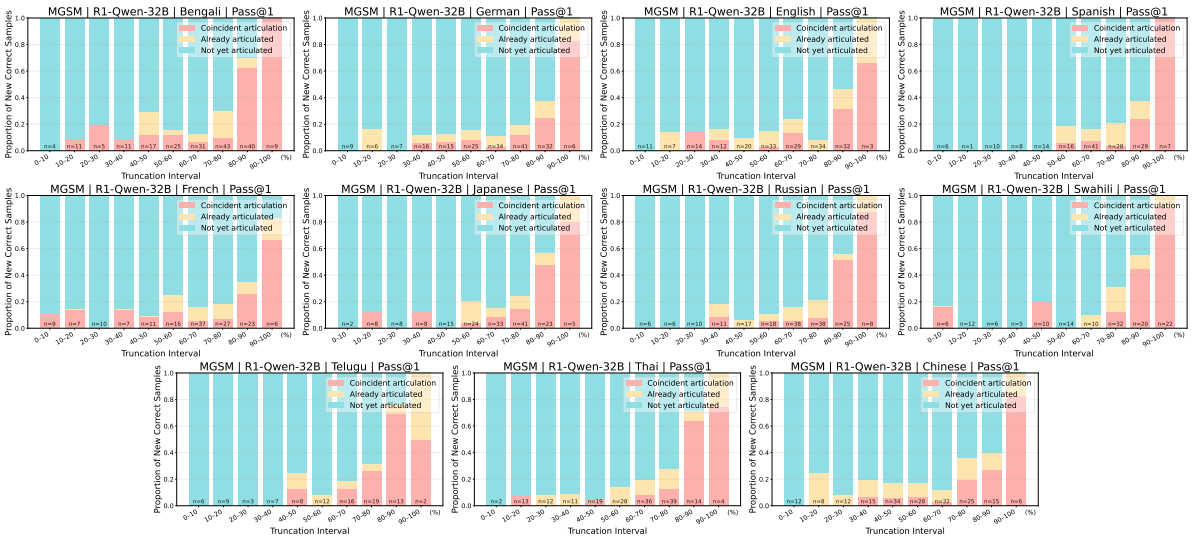

Figure 14: Causal decomposition of newly correct predictions across truncation intervals on **MGSM** with **R1-Qwen-32B**. Each bar partitions gains into three cases: (**i**) the gold answer is first articulated in the newly added reasoning steps, (**ii**) it was already articulated in earlier steps, or (**iii**) it has not yet appeared in the visible truncated trace. Early and intermediate gains are dominated by case (**iii**), indicating latent reasoning.

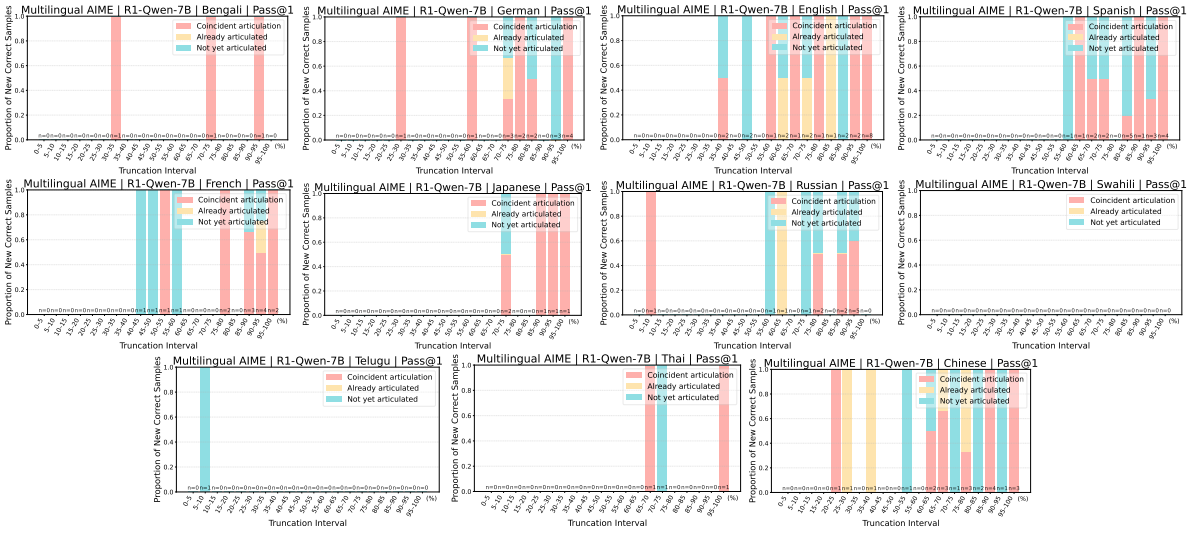

Figure 15: Causal decomposition of newly correct predictions across truncation intervals on **Multilingual AIME** with **R1-Qwen-7B**. Each bar partitions gains into three cases: (**i**) the gold answer is first articulated in the newly added reasoning steps, (**ii**) it was already articulated in earlier steps, or (**iii**) it has not yet appeared in the visible truncated trace. Compared to MGSM, gains are sparser and less dominated by latent reasoning.

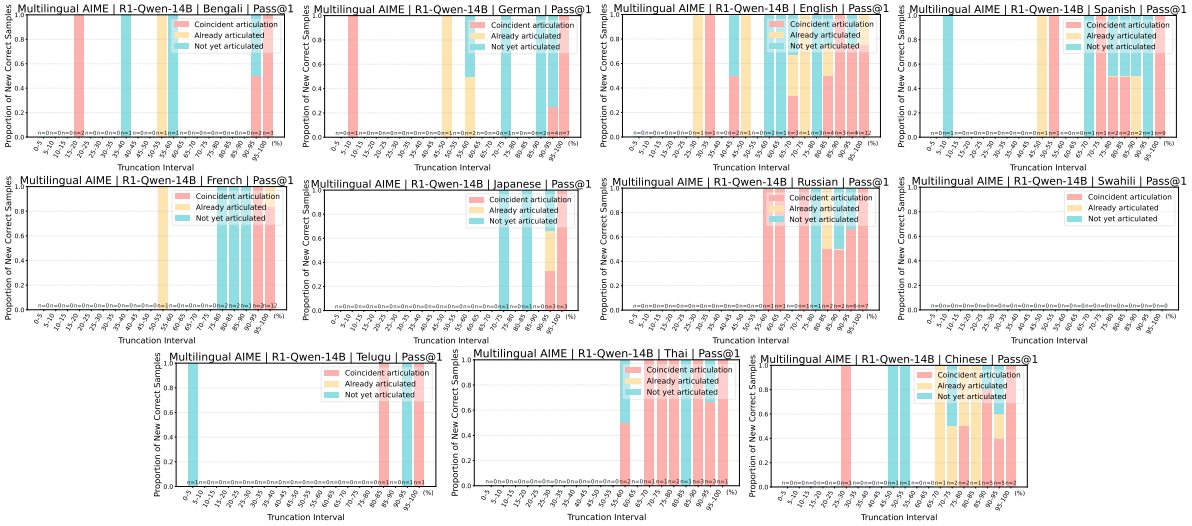

Figure 16: Causal decomposition of newly correct predictions across truncation intervals on **Multilingual AIME** with **R1-Qwen-14B**. Each bar partitions gains into three cases: (**i**) the gold answer is first articulated in the newly added reasoning steps, (**ii**) it was already articulated in earlier steps, or (**iii**) it has not yet appeared in the visible truncated trace. Compared to MGSM, gains are sparser and less dominated by latent reasoning.

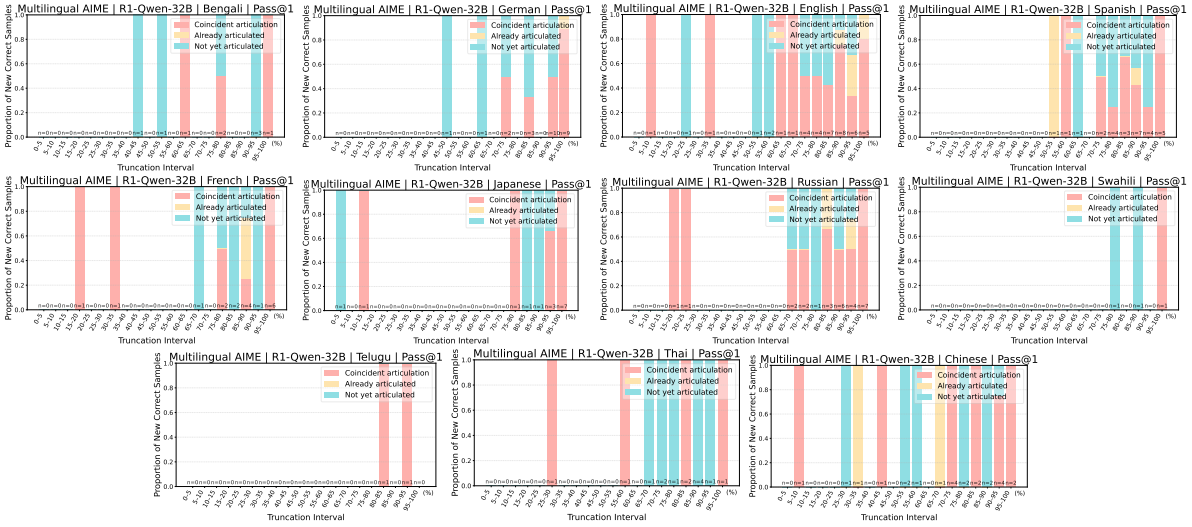

Figure 17: Causal decomposition of newly correct predictions across truncation intervals on **Multilingual AIME** with **R1-Qwen-32B**. Each bar partitions gains into three cases: (**i**) the gold answer is first articulated in the newly added reasoning steps, (**ii**) it was already articulated in earlier steps, or (**iii**) it has not yet appeared in the visible truncated trace. Compared to MGSM, gains are sparser and less dominated by latent reasoning.

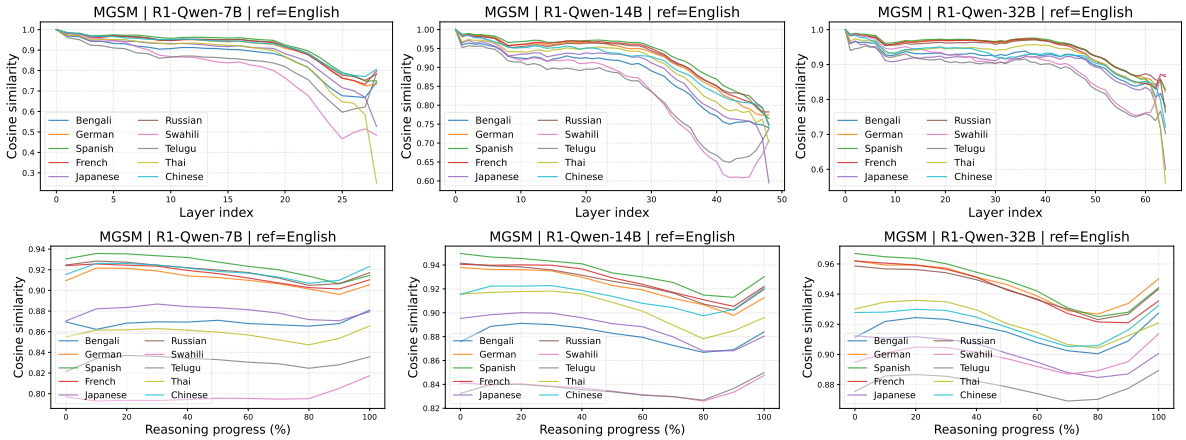

Figure 18: Aggregated cosine similarity on **MGSM** between hidden states in each language and English (reference), averaged over both reasoning steps and layers. High-resource languages show consistently higher similarity to English, suggesting convergence toward an English-centered latent reasoning pathway.

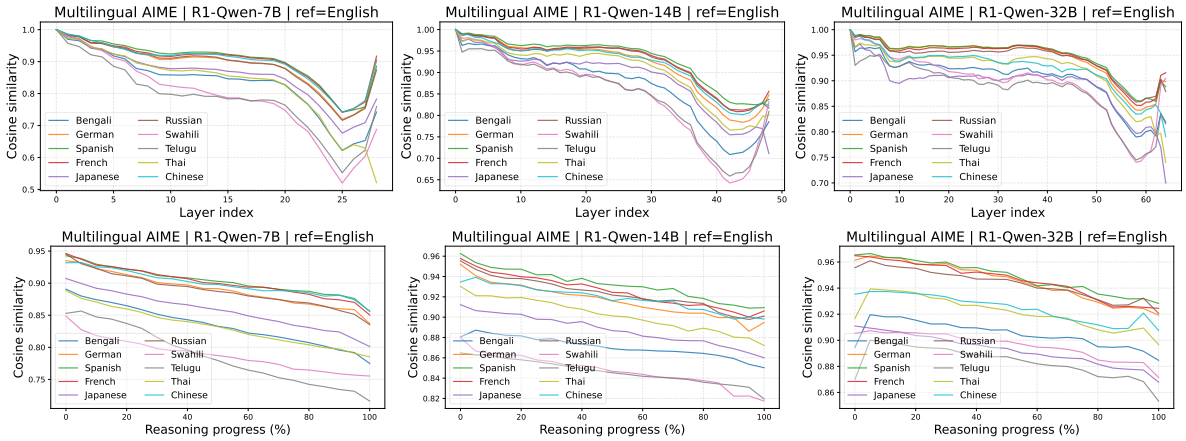

Figure 19: Aggregated cosine similarity on **Multilingual AIME** between hidden states in each language and English (reference), averaged over both reasoning steps and layers. High-resource languages show consistently higher similarity to English, suggesting convergence toward an English-centered latent reasoning pathway.

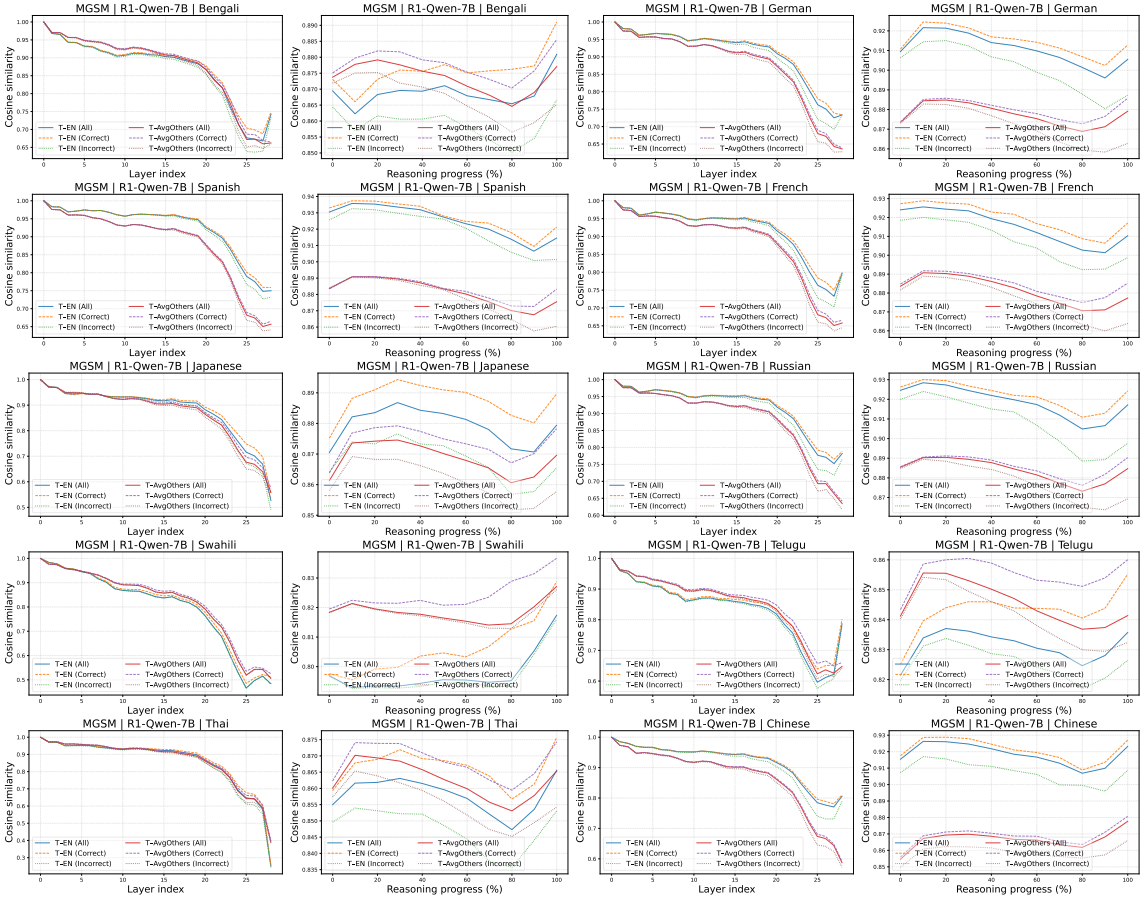

Figure 20: Comparison of cosine similarity with English versus average similarity with other languages, shown separately for correctly and incorrectly solved examples in **MGSM** with **R1-Qwen-7B**.

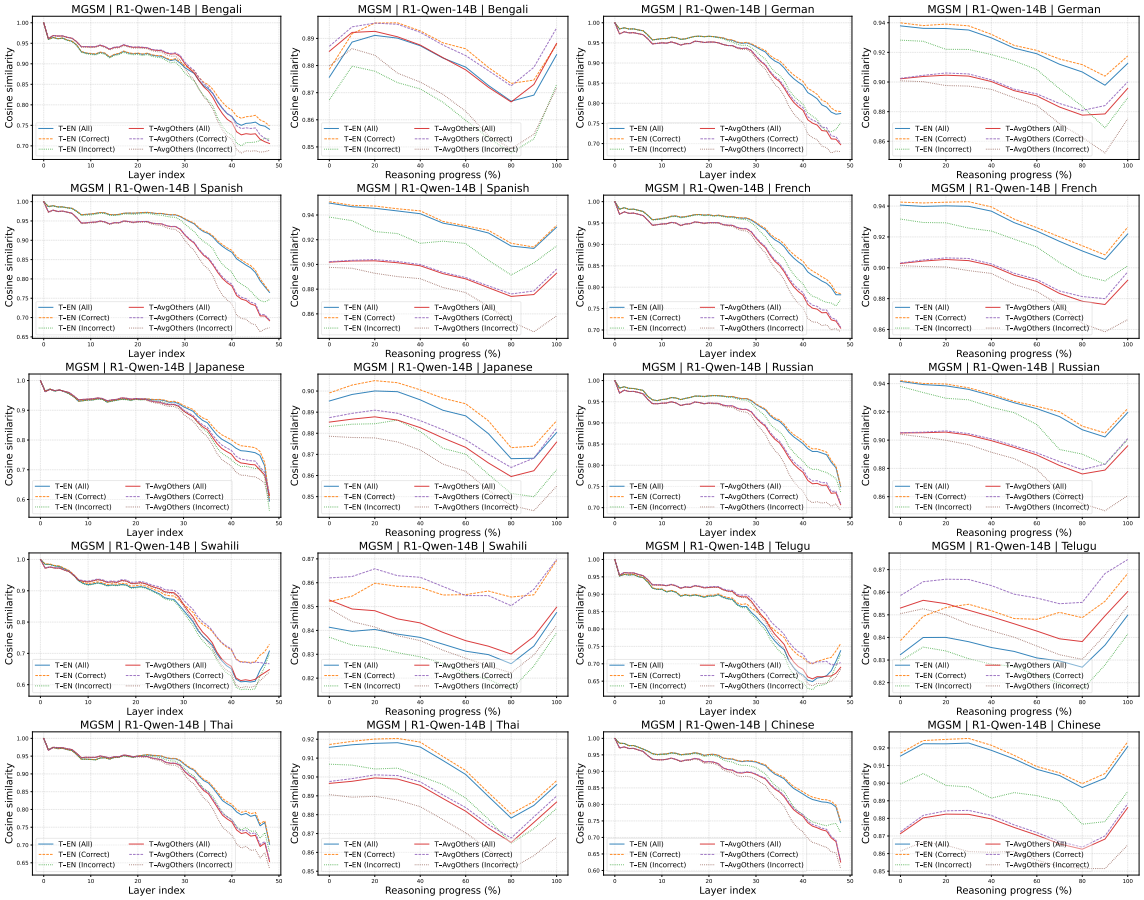

Figure 21: Comparison of cosine similarity with English versus average similarity with other languages, shown separately for correctly and incorrectly solved examples in **MGSM** with **R1-Qwen-14B**.

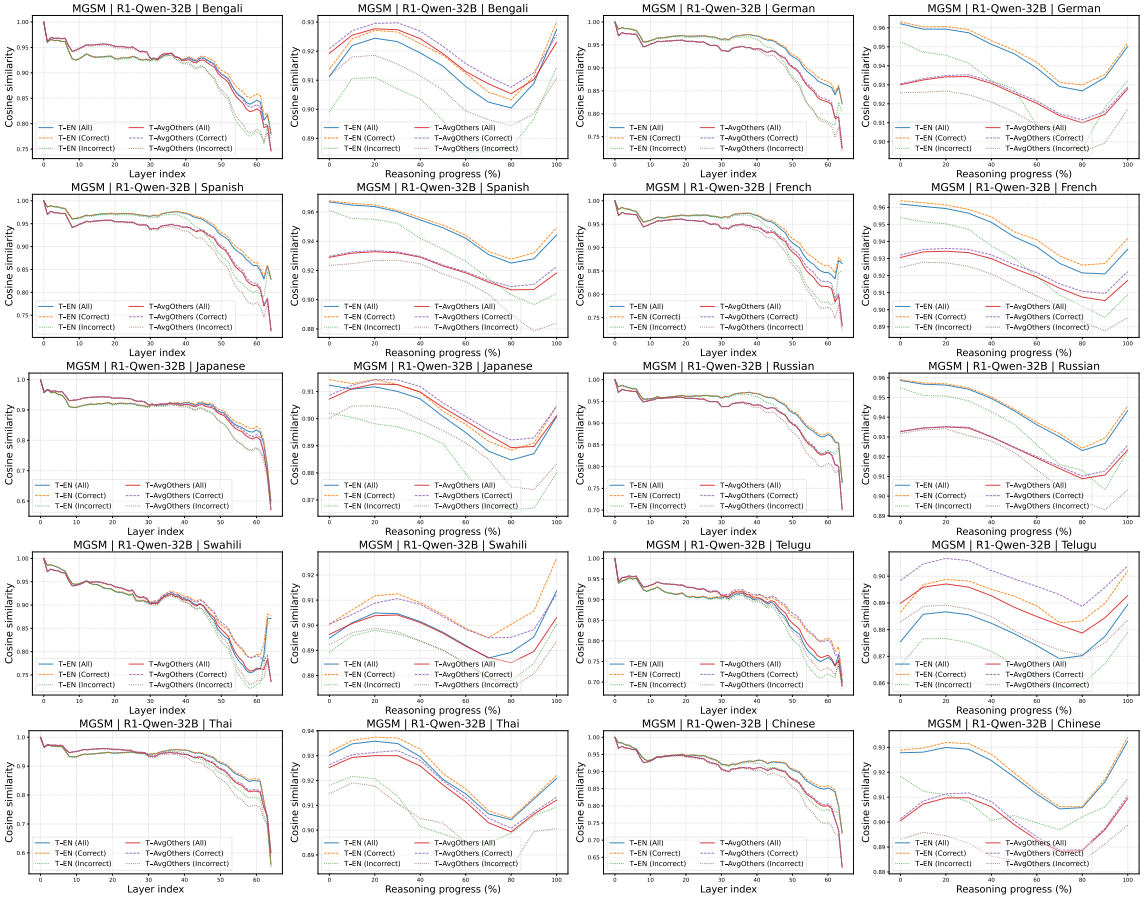

Figure 22: Comparison of cosine similarity with English versus average similarity with other languages, shown separately for correctly and incorrectly solved examples in **MGSM** with **R1-Qwen-32B**.

E Perturbation Details

We used the two perturbation methods in §7.1 to probe memorization versus (latent) reasoning. Both methods operate *only* on the subset of MGSM instances that the model answers correctly under **pass@10** when the truncation ratio is **0%** (i.e., with an empty `<think></think>` block). For each selected instance, we produce two edited variants: **NumEdit** (meaning-altering) and **Paraphrase** (meaning-preserving).

E.1 NumEdit

NumEdit creates a minimally changed but *meaning-altering* counterfactual by perturbing exactly *one* number in the question. The expected behavior differs depending on whether the model relies on memorization or reasoning: a memorization-driven model may continue to output the original answer, whereas a reasoning-driven model should adapt its answer to the changed quantity.

Numeric span detection. We identify candidate numbers using a conservative regular expression that matches standalone numeric tokens (including optional negative sign and decimal part), while avoiding matches that are embedded in words or common formats that are likely to break semantics:

- **Exclude years:** numbers matching `19xx` or `20xx` are skipped.
- **Exclude ordinals:** tokens followed by `st/nd/rd/th` are skipped.
- **Exclude fractions:** occurrences adjacent to `/` (e.g., `1/2`) are skipped.

Perturbation rule. Among the remaining candidates, we perturb *exactly one* numeric span with a small additive change:

- If the token is an integer:
  - For $\{0, 1, 2\}$, add $+1$.
  - Otherwise add a small $\Delta \in \{1, 2\}$ (seeded randomness).
- If the token is a float, add a small fixed delta depending on magnitude: $+0.1$ for $|x| < 1$, $+0.5$ for $|x| < 10$, otherwise $+1.0$.

E.2 Paraphrase

Paraphrase produces a meaning-preserving rewrite intended to reduce lexical overlap with the original prompt while keeping the problem logically equivalent. Unlike NumEdit, the gold answer remains the same. Thus, high performance on paraphrased questions supports generalization beyond surface-form memorization.

LLM-based Question Paraphrasing. We leverage `Gemini-2.5-Flash` to paraphrase each question. We have the following constraints in the prompt to ensure that the paraphrased question is equivalent to the original one:

- Preserve all numbers exactly.
- Preserve all LaTeX math segments (anything inside `$...$`) exactly as-is.
- Maintain logical equivalence and ask for the same final quantity.
- Do not add/remove constraints, entities, or units.

To prevent crosslingual drift, we additionally specify that the paraphrase must be written in the same language as the input question. The prompt template is shown in Figure 23.

Automatic validation. Each paraphrase is validated before acceptance: the multiset of numeric tokens in the paraphrase must match the original. If not, the same prompt will be applied again until the paraphrased question is valid.

E.3 Perturbation Solvability

To verify that the generated counterfactual questions remain *solvable* and do not introduce unintended artifacts, we conduct an auxiliary evaluation using a strong commercial model again (`Gemini-2.5-Flash`). This analysis serves as a sanity check that the perturbations preserve mathematical well-formedness while modifying surface form or numerical content as intended.

Setup. For each selected MGSM problem, we query the model on three inputs: the original question, its NumEdit variant, and its Paraphrase variant (for each input, we only query once). The model is instructed to optionally generate intermediate steps but to output a single final answer in a strict, parseable format (see prompt in Figure 24). We then compare Gemini's predictions across variants. Specifically, we report: (i) **Orig Acc**, Gemini's accuracy on the original question; (ii) **NumEdit Match**, the proportion of NumEdit predictions

identical to the original prediction (lower is better, as the gold answer is expected to be changed); and (iii) **Paraphrase Match**, the proportion of Paraphrase predictions identical to the original prediction (higher is better, as the gold answer is expected to be preserved).

Results and Discussion. Table 6 shows that Gemini achieves consistently high accuracy on the original questions across models and languages, indicating that the selected problem subset is reliably solvable. For NumEdit, the matching ratio remains low (typically below 10%), confirming that the numerical perturbations effectively alter the solution and are not trivially ignored. In contrast, Paraphrase variants exhibit very high matching ratios (often above 95%), demonstrating that meaning-preserving rewrites retain solvability and solution consistency. These trends are stable across model sizes and languages, including lower-resource settings. Together, these results confirm that both perturbation methods produce mathematically valid and solvable questions. NumEdit reliably changes the target answer, while Paraphrase preserves it, validating their use as controlled probes for disentangling memorization from latent reasoning.

F Experimental Details

F.1 Language Control

Following Qi et al. (2025); Zhao et al. (2026), we adopt a set of complementary strategies to ensure that the model's explicit reasoning trace is produced in the same language as the input prompt.

Prompt Formation We prepend each input with an explicit language-specific instruction that explicitly specifies the target language for reasoning. This instruction is embedded into a standardized prompt template, shown in Figure 25, which is used consistently across all languages.

Prompt Hacking Even with explicit language instructions in the prompt, LLMs may still generate reasoning traces in a language different from that of the input, a phenomenon observed in prior work on multilingual reasoning and language mixing (Wang et al., 2025c; Qi et al., 2025; Zhao et al., 2026). Such behavior is undesirable in our setting, as it confounds cross-lingual comparisons of latent reasoning by introducing variation at the level of explicit verbalization. To mitigate this issue, we adopt a *prompt-hacking* strategy (Schulhoff et al., 2023; Benjamin et al., 2024) that reinforces the language constraint at inference time. Concretely, following prior work (Qi et al., 2025; Zhao et al., 2026), we insert a language-specific prefix immediately after the opening `<think>` tag (e.g., "`By request, I will begin to think in English`"). This prefix explicitly restates the target language at the onset of the reasoning trace and reliably steers the model to produce explicit reasoning in the same language as the prompt until the closing `</think>` tag is reached. The complete set of prompt-hacking prefixes used in our experiments is listed in Figure 26.

Answer Elicitation To analyze early answer formation during truncated reasoning, we aim to elicit the model's prediction immediately after the visible reasoning prefix, without allowing further reasoning steps. Accordingly, we append a language-specific answer-elicitation prefix directly after the closing `</think>` tag. This prefix prompts the model to produce only the final numerical answer, preventing additional reasoning or thought continuation beyond the truncated trace. The answer-elicitation prefixes used in our experiments are shown in Figure 27.

F.2 Pass@k Evaluation

For each question, we generate 10 independent samples and evaluate each prediction separately. Correctness is assessed via exact matching. Following prior work (Qi et al., 2025; Zhao et al., 2026), models are instructed to enclose their final answers in `\boxed{}`, from which the boxed content is extracted and compared against the gold answer using mathematical equivalence rather than raw string matching.

G Environment and Hyperparameters

We set the maximum generation length to 4K tokens for the MGSM benchmark and 16K tokens for Multilingual AIME across all evaluated models. Unless stated otherwise, we adopt the recommended generation configurations provided on HuggingFace.[10] In particular, we use a temperature of 0.6 and top-p sampling with $p = 0.95$. For reproducibility, the random seed is fixed to 42.

Experiments on identifying multilingual latent reasoning (cf. §4) and on memorization versus latent reasoning (cf. §7) are conducted using the vLLM framework,[11] while experiments analyzing

[10] https://huggingface.co

[11] https://vllm.ai/

Original Model	Metric	EN	FR	DE	ZH	JA	RU	ES	SW	BN	TE	TH
R1-Qwen-7B	**Orig Acc** ↑	0.98	0.93	0.95	0.94	0.88	1.00	0.96	0.94	0.93	0.91	0.96
	NumEdit Match ↓	0.07	0.05	0.08	0.06	0.16	0.07	0.07	0.16	0.05	0.08	0.09
	Paraphrase Match ↑	0.97	0.93	0.99	0.95	0.97	1.00	0.95	0.97	0.93	0.87	0.91
R1-Qwen-14B	**Orig Acc** ↑	0.97	0.96	0.94	0.94	0.88	0.95	0.94	0.86	0.99	0.89	0.96
	NumEdit Match ↓	0.07	0.06	0.07	0.09	0.16	0.08	0.12	0.06	0.07	0.04	0.08
	Paraphrase Match ↑	0.97	0.94	0.97	0.96	0.93	0.95	0.97	0.88	0.97	0.93	0.94
R1-Qwen-32B	**Orig Acc** ↑	0.98	0.97	0.93	0.93	0.94	0.98	0.97	0.95	0.96	0.89	0.96
	NumEdit Match ↓	0.10	0.06	0.09	0.11	0.15	0.11	0.08	0.09	0.11	0.13	0.09
	Paraphrase Match ↑	0.94	0.95	0.97	0.98	0.95	0.99	0.97	0.95	0.97	0.95	0.95

Table 6: Gemini performance on original and counterfactual MGSM questions. **Orig Acc** compares Gemini's prediction on the original question to the gold answer. **NumEdit Match** measures the fraction of NumEdit predictions identical to the original prediction (lower is better). **Paraphrase Match** measures the fraction of Paraphrase predictions identical to the original prediction (higher is better).

Paraphrase prompt template

```
You are rewriting a math problem.
Language constraint (MUST follow):
- The paraphrase MUST be written in the SAME language as the original question.
- The original question language is: {language_name}. Do NOT translate to any other language.
Hard constraints:
1) Preserve ALL numbers exactly (character-for-character).
2) Preserve ALL LaTeX math exactly as-is (anything inside $...$ must appear unchanged).
3) Keep the question asking for the same final quantity; the problem must be logically equivalent.
4) Reduce lexical overlap by paraphrasing and reordering sentences outside math mode.
5) Do NOT include any solution steps, explanations, or the final answer.
6) Do NOT add or remove any facts, entities, units, or constraints.

Return ONLY valid JSON with exactly these keys:
{"paraphrase": "...", "changes": "..."}
Original problem:
{problem}
```

Figure 23: Prompt used to generate meaning-preserving paraphrases using `Gemini-2.5-Flash`. Placeholders `{language_name}` and `{problem}` are substituted per instance.

latent dynamics (cf. §5) are performed using the HuggingFace `Transformers` library.[12]

All experiments are run on NVIDIA A100 GPUs (80 GB) and NVIDIA RTX A6000 GPUs (48 GB).

[12] https://huggingface.co/docs/transformers

Solvability evaluation prompt

```
You are given a grade-school math word problem.

Language constraint:
- Write your solution in the SAME language as the problem.
- The problem language is: {language_name}. Do not translate.

You may write intermediate steps.
Hard requirement:
- End your response with a SINGLE final line in the following exact format:
FINAL_ANSWER: <answer>

Rules for <answer>:
- Provide only the final numeric value (or a simplified number).
- Do not wrap it in LaTeX, do not add units, and do not add extra words.
- Do not output anything after the FINAL_ANSWER line.

Problem:
{problem}
```

Figure 24: Prompt used to evaluate the solvability of original and counterfactual MGSM questions using `Gemini-2.5-Flash`. Placeholders `{language_name}` and `{problem}` are substituted per instance.

Lang	Prompt Template
BN	দয়া করে সর্বদা বাংলায় চিন্তা করুন।\n\n নিচের গাণিতিক সমস্যাটি ধাপে ধাপে সমাধান করুন। শেষে, আপনার চূড়ান্ত উত্তরটি\boxed{}-এর মধ্যে প্রদান করুন।\n সমস্যা: {}\n\n
DE	Bitte denken Sie immer auf Deutsch.\n\n Lösen Sie das folgende Mathematikproblem Schritt für Schritt. Am Ende geben Sie Ihre endgültige Antwort in \boxed{} ein.\n Problem: {}\n\n
EN	Please always think in English.\n\n Solve the following mathematics problem step by step. At the end, provide your final answer enclosed in \boxed{}.\n Problem: {}\n\n
ES	Por favor, siempre piensa en español.\n\n Resuelve el siguiente problema matemático paso a paso. Al final, proporciona tu respuesta final encerrada en \boxed{}.\n Problema: {}\n\n
FR	Veuillez toujours réfléchir en français.\n\n Résolvez le problème mathématique suivant étape par étape. À la fin, fournissez votre réponse finale entourée de \boxed{}.\n Problème : {}\n\n
JA	常に日本語で考えてください。\n\n 以下の数学問題をステップバイステップで解いてください。最後に、最終的な答えを\boxed{}で囲んで提供してください。\n 問題：{}\n\n
RU	Пожалуйста, всегда думайте по-русски.\n\n Решите следующую математическую задачу шаг за шагом. В конце предоставьте свой окончательный ответ, заключенный в \boxed{}.\n Задача: {}\n\n
SW	Tafadhali daima fikiria kwa Kiswahili.\n\n Tatuwa tatizo la hisabati lifuatalo hatua kwa hatua. Mwishoni, toa jibu lako la mwisho lililozungukwa na \boxed{}.\n Tatizo: {}\n\n
TE	దయచేసి ఎప్పుడూ తెలుగు లో ఆలోచించండి.\n\n క్రింది గణిత సమస్యను దశదిశగా పరిష్కరించండి. చివరగా, మీ తుదిపరిష్కారాన్ని \boxed{} లో రాంచు.\n సమస్య: {}\n\n
TH	กรุณาคิดเป็นภาษาไทยเสมอ.\n\n แก้ปัญหาทางคณิตศาสตร์ต่อไปนี้ทีละขั้นตอน สุดท้ายให้ระบุคำตอบสุดท้ายของคุณใน \boxed{}.\n ปัญหา: {}\n\n
ZH	请始终用中文思考。\n\n逐步解决以下数学问题。最后，将您的最终答案放在\boxed{}中。\n问题：{}\n\n

Figure 25: Language-specific prompt templates (containing the explicit language instruction) used for controlling the reasoning language.

Lang	Prompt Hacking Prefix
BN	অনুরোধ করলে, আমি বাংলায় চিন্তা করা শুরু করব।
DE	Auf Anfrage werde ich anfangen, in Deutsch zu denken.
EN	By request, I will start thinking in English.
ES	A petición, empezaré a pensar en español.
FR	Sur demande, je commencerai à penser en français.
JA	要求があれば、日本語で考え始めます。
RU	По запросу я начну думать на русском.
SW	Kwa ombi, nitaanza kufikiria kwa Kiswahili.
TE	అభ్యర్థన మేరకు, నేను తెలుగులో ఆలోచించడం ప్రారంభిస్తాను.
TH	ตามคำขอ ฉันจะเริ่มคิดเป็นภาษาไทย
ZH	应要求，我将开始用中文思考。

Figure 26: Language-specific prompt-hacking prefixes used to reinforce language control. Each prefix is inserted immediately after the <think> tag to steer the model's explicit reasoning trace to match the prompt language.

Lang	Answer Elicitation Prefix
BN	উত্তর হল: \boxed{
DE	Die Antwort ist: \boxed{
EN	The answer is: \boxed{
ES	La respuesta es: \boxed{
FR	La réponse est : \boxed{
JA	答えは： \boxed{
RU	Ответ: \boxed{
SW	Jibu ni: \boxed{
TE	సమాధానం: \boxed{
TH	คำตอบคือ: \boxed{
ZH	答案是： \boxed{

Figure 27: Language-specific answer-elicitation prefixes used to directly prompt the model for a final numerical answer. Each prefix is appended immediately after the </think> tag to elicit the prediction.