Reasoning as State Transition: A Representational Analysis of Reasoning Evolution in Large Language Models

Siyuan Zhang [1] **Jialian Li** [1] **Yichi Zhang** [1] **Xiao Yang** [1] **Yinpeng Dong** [2] **Hang Su** [1]

Abstract

Large Language Models have achieved remarkable performance on reasoning tasks, motivating research into how this ability evolves during training. Prior work has primarily analyzed this evolution via explicit generation outcomes, treating the reasoning process as a black box and obscuring internal changes. To address this opacity, we introduce a representational perspective to investigate the dynamics of the model's internal states. Through comprehensive experiments across models at various training stages, we discover that post-training yields only limited improvement in static initial representation quality. Furthermore, distinct from non-reasoning tasks, reasoning involves a significant continuous distributional shift in representations during generation. Comparative analysis indicates that post-training empowers models to drive this transition toward a better distribution for task solving. To clarify the relationship between internal states and external outputs, statistical analysis confirms a high correlation between generation correctness and the final representations; while counterfactual experiments identify the semantics of the generated tokens, rather than additional computation during inference or intrinsic parameter differences, as the key driver of the transition. Collectively, we offer a novel understanding of the reasoning process and the effect of training on reasoning enhancement, providing valuable insights for future model analysis and optimization.

[1]Dept. of Comp. Sci. and Tech., Institute for AI, Tsinghua-Bosch Joint ML Center, THBI Lab, BNRist Center, Tsinghua University, Beijing 100084, China [2]College of AI, Tsinghua University, Beijing 100084, China. Correspondence to: Yinpeng Dong <dongyinpeng@mail.tsinghua.edu.cn>, Hang Su <suhangss@mail.tsinghua.edu.cn>.

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

1. Introduction

Large Language Models (LLMs) (Jaech et al., 2024; Guo et al., 2025; Team et al., 2025) have demonstrated remarkable performance on reasoning tasks requiring multi-hop analysis and complex calculations (Yue et al., 2025; Wu et al., 2025a). Through targeted post-training techniques, specifically Reinforcement Learning (RL) (Shao et al., 2024; Zeng et al., 2025; Zheng et al., 2025) and distillation (Guo et al., 2025; Face, 2025; Tian et al., 2025), these models acquire the capacity to generate long Chain-of-Thought (CoT) sequences (Wei et al., 2022). This capability enables models to effectively decompose and solve problems that remain challenging for base models (Li et al., 2025; Wu et al., 2025b). Such advancements have motivated research into the underlying mechanisms of how reasoning evolves during training (Zhao et al., 2025; Matsutani et al., 2025).

Prior research has primarily analyzed this evolution from the perspective of the explicit *generation* process, focusing on the improvement in response accuracy (Yue et al., 2025; Shao et al., 2025; Zhang et al., 2025b) or the emergence of special CoT patterns (Li et al., 2025; Matsutani et al., 2025; Wang et al., 2025a). However, by relying solely on external behaviors, these approaches treat reasoning as a black box, overlooking the dynamics of the model's internal states. Consequently, they fail to provide a deeper and mechanistic explanation of how reasoning functions and how reasoning ability fundamentally develops through training.

To address this limitation and enhance interpretability, we extend our investigation beyond explicit generation to the model's latent *representations*, which serve as a proxy for its internal states. We quantify representation quality using the probing technique (Ye et al., 2024), which measures how accurately the correct answer can be predicted directly from the high-dimensional representation vector, thereby reflecting whether the model internally encodes the solution to a task. As illustrated in Figure 1, our analysis framework examines both the static *initial* quality prior to generation and the dynamic quality variation throughout the generation process. Comprehensive experiments on models at different training stages, spanning diverse representative tasks, LLM families, and model sizes, reveal several unexpected yet universal findings:

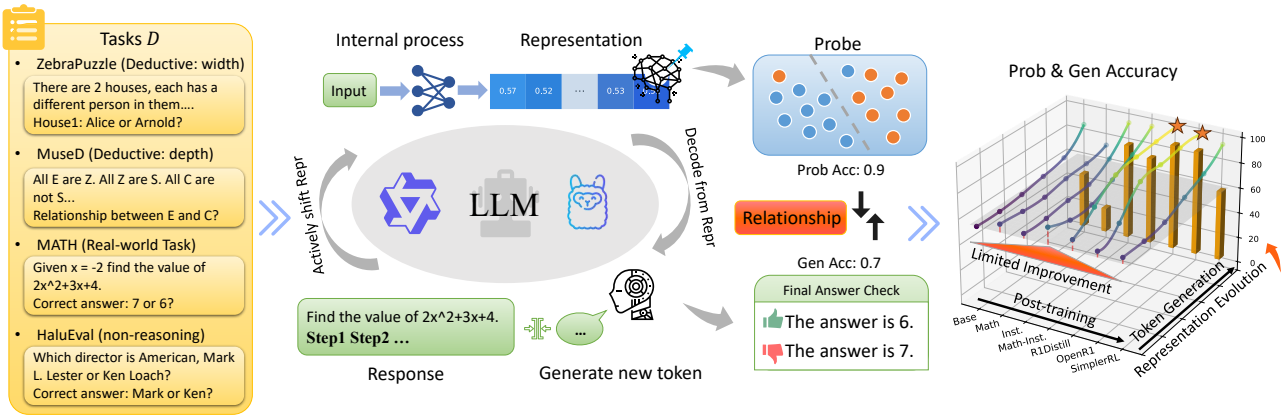

Figure 1. **Overview of the dual-perspective analysis framework.** We conduct a comprehensive study across four representative tasks, measuring representation quality via probing alongside explicit generation accuracy. Our results indicate that post-training significantly improves generation accuracy but yields limited enhancement to static initial representation quality. Furthermore, tracking representation dynamics reveals that the reasoning process involves a significant distributional shift, where post-training empowers models to drive representations toward a higher-quality state. Finally, we investigate the relationship between explicit generation and implicit representation, quantifying their alignment and identifying the key factor driving this state transition.

As a foundational starting point for the reasoning process, we investigate the evolution of initial representation quality during post-training. We find that reasoning ability is already acquired during the pre-training phase. LLMs exhibit substantial representation quality on reasoning tasks, with probing accuracy far exceeding random guessing. Notably, even without token generation, this initial representation quality surpasses explicit generation accuracy in over 60% of tasks, indicating a strong latent reasoning ability. However, differing from the substantial gains in generation accuracy, post-training improves initial representation quality to only a limited extent, with probing accuracy increasing by less than 5% across nearly all training stages. This demonstrates that the improvement in generation accuracy does not simply stem from a better reasoning ability available immediately upon encountering a question.

We further observe that strong reasoning models can achieve higher generation accuracy that exceeds their initial representation quality on difficult tasks. Since this cannot be explained by the negligible improvement in static initial quality, we investigate the dynamics of the representation during generation. Our experiments reveal that, while representation quality remains relatively stable on non-reasoning tasks, reasoning tasks involve a continuous and significant representation change throughout the generation process. Post-training serves to enhance the model's capacity to drive this transition, enabling it to attain a higher *final* quality after extensive CoT. Notably, although explicit generation accuracy can surpass the initial representation quality, it generally lags behind the final quality, indicating untapped potential within the model's internal states.

While we have observed internal changes during reasoning, the precise relationship between explicit generation and implicit representations remains underexplored. To this end, we first statistically investigate their alignment, finding a high correlation between generation correctness and the final representations. Conversely, the low correlation with initial representations suggests that the model does not merely verbalize its initial latent thoughts. Furthermore, the sharp divergence between initial and final representations confirms a fundamental distributional shift, and the extensive reasoning process facilitates this state transition. Second, we identify the primary driver of this transition. Counterfactual analysis demonstrates that the semantic content of the CoT is the key factor, whereas solely increasing computation during inference is insufficient to improve representation quality, and intrinsic parameter differences affect the shift results minimally.

Through comprehensive and in-depth analysis, our work offers a novel perspective to understand the reasoning process and the evolution of reasoning during training, elucidating the relationship between the model's internal representations and external generation. These findings provide valuable insights for further reasoning interpretability and highlight promising directions for model optimization from the viewpoint of internal states and signals.

2. Related Work

Analysis of LLM reasoning evolution. Following the success of reasoning models (Jaech et al., 2024; Guo et al., 2025; Team et al., 2025), extensive research has investigated how reasoning ability evolves during post-training. One line of work (Li et al., 2025; Wang et al., 2025a; Gandhi et al., 2025; Vassoyan et al., 2025) analyzes training data, identifying that specific patterns, such as reasoning struc-

tures and high-entropy tokens, are crucial for optimization results. Another stream of research (Havrilla et al., 2024; Yue et al., 2025; Liu et al., 2025; Wu et al., 2025a; Zhang et al., 2025b; Karan & Du, 2025; Matsutani et al., 2025) compares reasoning boundaries before and after training, revealing that while RL may not extend reasoning capability beyond the base model, distillation can effectively transfer new abilities. However, these studies mainly analyze reasoning through explicit generation outcomes, treating the model as a black box and overlooking changes in the model's internal states. To address this limitation, we introduce another perspective, from implicit representations, to directly observe the model's latent judgments and reasoning processes. This approach uncovers meaningful findings through a distinct viewpoint, providing insights into both reasoning interpretability and the underlying mechanisms of reasoning evolution.

Analysis of LLM representation. Interpretability research suggests that internal representations encode meaningful knowledge (Hewitt & Manning, 2019; Mohebbi et al., 2021; Ghandeharioun et al., 2024; Gurnee & Tegmark, 2024; Ataishiyev et al., 2025; Yusupov et al., 2025), enabling the prediction of specific properties via probing techniques (Belinkov, 2022; Zhao et al., 2024). Prior studies have leveraged these representations to detect output correctness (Orgad et al., 2025; Zhang et al., 2025d; Cencerrado et al., 2025; Zhang et al., 2025a; Liang et al., 2025), uncertainty (Wang et al., 2025b), and quality (Yusupov et al., 2025). Distinct from predicting external generation performance, our analysis focuses on determining whether the model has internally discerned the correct solution for a task. While recent studies (Yan et al., 2024; Ye et al., 2024) reveal that models possess awareness of correct answers via latent reasoning (Yang et al., 2024c; Biran et al., 2024) prior to generation, we extend this investigation to complex reasoning tasks and analyze the dynamic evolution of this awareness throughout the CoT. Most closely related to our work, recent studies (Kudo et al., 2024; Afzal et al., 2025; Wang et al., 2025c) have also examined representation shifts during generation. However, we distinguish our contribution by investigating how these shifts evolve across different training stages and by offering a deeper analysis of the relationship between the internal representations and external generation.

3. Preliminaries

In this section, we introduce the preliminaries for our experiments and analysis, detailing the formulation of core metrics and the experimental setup.

3.1. Formulation

We define two core metrics in our analysis: representation quality and generation accuracy. For clarity, we first formalize the LLM working process. We consider a Transformer-based model (Vaswani et al., 2017) M_θ, which can be functionally decomposed into two components: a backbone function f_θ that maps the input to a latent representation and a decoding function g_θ that computes the next-token distribution from this representation. Following previous work (Huh et al., 2024; Yan et al., 2024; Zhang et al., 2025c), we define the representation c as the hidden state of the last token at the final layer. Given a prompt x, the auto-regressive generation proceeds through discrete time steps $t = 1, \ldots, T$. At each step t, the input $(x, y_{1:t-1})$ is first transformed into the representation $c_t := f_\theta(x, y_{1:t-1}) \in \mathbb{R}^m$, where m denotes the hidden dimension. We adopt the representation c_t as a proxy for the model's full internal state S_t at step t. Subsequently, the next token is sampled according to $y_t \sim g_\theta(\cdot|c_t)$. We denote the complete generation process of a sequence $y_{1:T}$ given a prompt x as $y_{1:T} \sim M_\theta(\cdot|x)$.

Representation Quality. We quantify representation quality using the widely-adopted probing technique (Belinkov et al., 2017; Belinkov, 2022; Azaria & Mitchell, 2023; Sky et al., 2024). Given a dataset $\mathcal{D}^{\text{repr}} = \{(x^{(i)}, z^{(i)}\}$ for an s-class task, where the label $z \in \{0, \ldots, s-1\}$, standard linear-probing (Hewitt & Manning, 2019; Li et al., 2023b; Orgad et al., 2025) involves freezing the LLM backbone and training a linear classifier $h \in \mathbb{R}^{m \times s}$ on the training split. This classifier maps the representation c to a predicted probability distribution $p := \text{softmax}(h(c))$. The classifier is then optimized via cross-entropy loss:

$$\mathcal{L}_{\text{CE}}(\mathcal{D}^{\text{repr}}_{\text{train}}) = \mathbb{E}_{\{x^{(i)}, z^{(i)}\} \in \mathcal{D}^{\text{repr}}_{\text{train}}}[-\log(p(z^{(i)}|x^{(i)}))]. \quad (1)$$

The **representation quality** $Acc_{\text{repr}}(\mathcal{D}^{\text{repr}}_{\text{test}})$ is then defined as the classification accuracy on the test split:

$$\mathbb{E}_{\{x^{(i)}, z^{(i)}\} \in \mathcal{D}^{\text{repr}}_{\text{test}}}[\mathbf{1}(\arg\max_z(p(z|x^{(i)})) = z^{(i)})]. \quad (2)$$

However, standard probing is insufficient for tasks where the target label is not fixed but is conditioned on multiple variables, so we adopt the V-probing technique (Ye et al., 2024), which encapsulates relevant variables within special tokens and trains their embeddings via a low-rank update to the embedding layer. Formally, the training and calculation procedure is presented in Algorithm 1.

Under this definition, representation quality functions as a static metric for a specific task and context. Specifically, when the input comprises only the problem description x prior to any token generation, the measured accuracy constitutes the **initial representation quality**, which characterizes the model's capacity for pure latent reasoning (Yang et al., 2024c) derived from a single forward pass.

Although V-probing is nearly-linear, we posit that it remains a faithful indicator of the model's intrinsic representation

quality. In Appendix C.3, we theoretically analyze that the limited budget of trainable parameters is insufficient to independently achieve high probing accuracy on tasks of sufficient complexity. Empirical results in Table 1 verify this conclusion, showing that V-probing on a randomly initialized model fails to outperform majority guessing. The substantial performance gap between the pre-trained and randomly initialized models confirms that the observed accuracy derives primarily from the knowledge encoded within the model's existing parameters rather than the probe's capacity to fully fit the data distribution.

Generation Accuracy. **Generation accuracy** quantifies the model's ability to explicitly generate correct solutions. We evaluate this metric $Acc_{\text{gen}}(\mathcal{D}_{\text{test}}^{\text{gen}})$ on a test set $\mathcal{D}_{\text{test}}^{\text{gen}}$ containing identical problem instances to those in the probing dataset $\mathcal{D}_{\text{test}}^{\text{repr}}$:

$$\mathbb{E}_{\{x^{(i)},z^{(i)}\}\in\mathcal{D}_{\text{test}}^{\text{gen}},y_{1:T}^{(i)}\sim M_\theta(\cdot|x^{(i)})}[\mathbf{1}(\text{Ext}(x^{(i)}+y_{1:T}^{(i)})=z^{(i)})], \quad (3)$$

where $\text{Ext}(\cdot)$ denotes a function that extracts the final answer from the model response and maps it to the target label. In contrast to the static nature of representation quality, generation accuracy reflects the dynamic outcome of the entire auto-regressive reasoning process for a given task.

3.2. Experimental Setup

We select four representative tasks to comprehensively analyze reasoning ability. We adopt the deductive tasks Zebra-Puzzle (Stojanovski et al., 2025) and MuseD (Li et al., 2024) to evaluate reasoning breadth and depth, respectively, following Allen-Zhu (2025). **ZebraPuzzle** presents constraint satisfaction problems requiring the simultaneous resolution of multiple relational dependencies, while **MuseD** demands long-range, multi-hop deductive reductions. Both synthetic tasks are stratified into three difficulty levels (Low, Med, High). In addition to the deductive tasks, we select **MATH** (Hendrycks et al., 2021) to evaluate unstructured real-world reasoning. Finally, **HaluEval** (Li et al., 2023a) serves as a control task for factuality, with minimal reasoning requirements. To satisfy the probing constraints, we construct True-False (TF) and Multiple-Choice (MC) variants for each task. Detailed specifications are provided in Appendix A.

For V-probing, we employ LoRA (Hu et al., 2022) training with the rank $r = 4$, the learning rate $\gamma = 1e^{-4}$, and the batch size $b = 32$. For generation, we adopt a context window of $32,768$ tokens to accommodate long CoT, consistent with recent evaluations (Guo et al., 2025; Team, 2025), and utilize a rule-based function for answer extraction $\text{Ext}(\cdot)$. More implementation details are provided in Appendix C.1 and Appendix D.

We conduct our primary analysis on the Qwen2.5-7B series (Team, 2024; Yang et al., 2024b), evaluating seven models

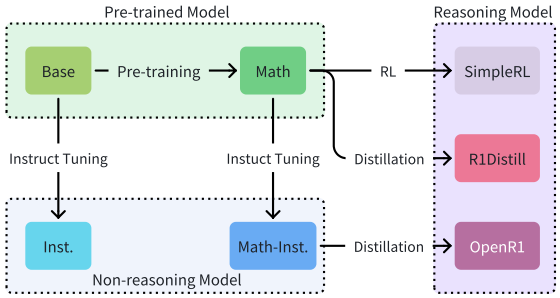

Figure 2. **Training relationships among the Qwen2.5-7B series.**

across distinct training stages. As illustrated in Figure 2, we categorize these models into three classes: Pre-trained models (Base, Math); Non-reasoning models with standard instruct-tuning (Instruct, Math-Instruct); and Reasoning models developed via distillation or RL (R1Distill (Guo et al., 2025), OpenR1 (Face, 2025), and SimpleRL (Zeng et al., 2025)). Detailed model specifications are provided in Appendix B. To verify the universality of our findings across different model families and parameter scales, we replicate representative experiments on the Llama3.1-8B series (Dubey et al., 2024) and the smaller Qwen2.5-1.5B series, with these results detailed in Appendix E.

4. Representational Analysis of Reasoning Evolution

Having established the metrics for representation quality and generation accuracy, we now present our primary experimental findings in this section.

4.1. Development of Initial Representation Quality

To analyze the evolution of reasoning ability in LLMs, we first examine the development of initial representation quality across various training stages.

Table 1. **Probing accuracy** (%) of the Pre-trained (PT) model across all tasks, compared against majority guessing and randomly initialized model probing baselines.

	Zebra-TF			Zebra-MC		
	High	Med	Low	High	Med	Low
Majority guessing	50.00	50.00	50.00	22.22	28.57	40.00
PT model probing	**60.25**	**63.65**	**80.65**	**37.00**	**49.55**	**78.90**
Random model probing	52.20	53.80	51.20	25.00	30.95	43.25
	MuseD-TF			**MuseD-MC**		
	High	Med	Low	High	Med	Low
Majority guessing	50.00	50.00	50.00	25.00	25.00	25.00
PT model probing	**92.44**	**100.00**	**100.00**	**89.08**	**100.00**	**100.00**
Random model probing	51.72	51.36	52.68	35.28	55.00	57.24
	MATH			**HaluEval**		
	TF	MC	-	TF	MC	-
Majority guessing	50.00	50.00	-	50.00	50.00	-
PT model probing	**64.90**	**60.95**	-	**60.02**	**60.81**	-
Random model probing	53.70	51.45	-	52.31	51.13	-

As a preliminary validation, we assess the reliability of the

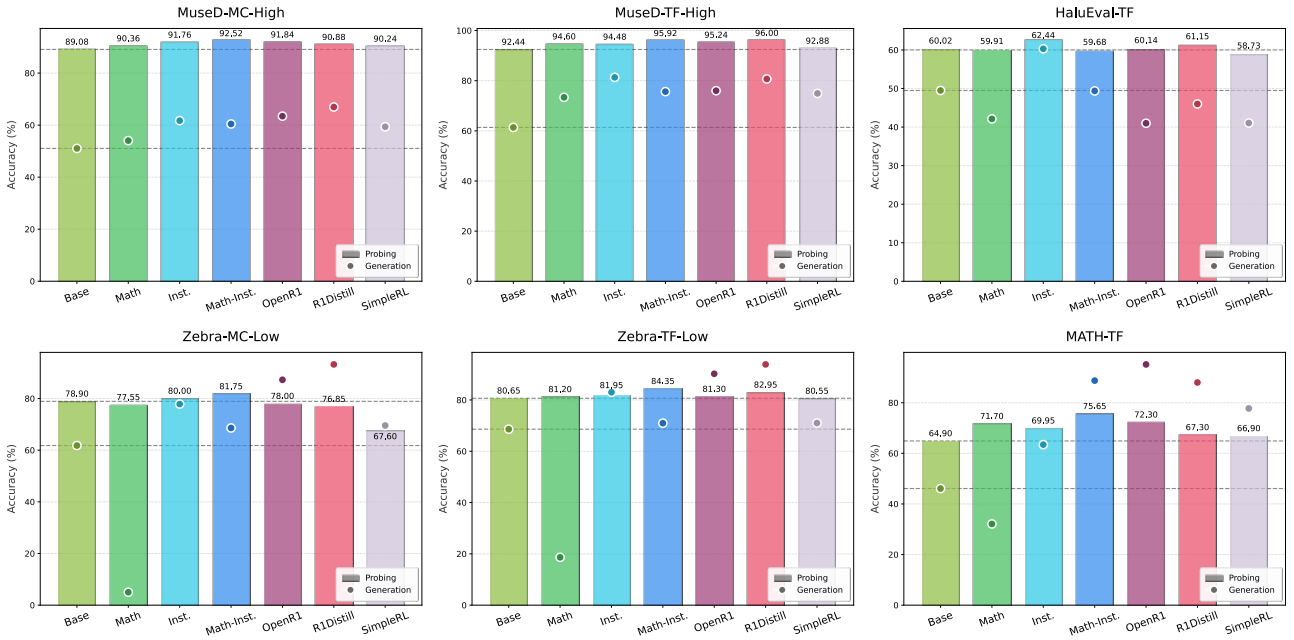

Figure 3. **Development of initial representation quality and generation accuracy.** The gray dashed lines indicate the baseline performance of the Base model for probing and generation. See Appendix E.1 for full results.

probing methodology by comparing the accuracy of the pre-trained model against a randomly initialized baseline. The results in Table 1 reveal that the pre-trained model significantly outperforms the random model. Crucially, the random baseline's accuracy aligns closely with majority guessing, which always predicts the most frequent label. This confirms that the tasks are sufficiently complex, and the observed high accuracy derives from the model's internal parameters rather than the probe's trainable components. Furthermore, these results yield a notable finding that even in the pre-training phase, prior to any alignment, the model already exhibits significant latent reasoning capacity, enabling it to discern correct solutions before generation.

Table 2. **The maximum performance gain** (Max Δ) represents the highest accuracy improvement (%) achieved by any other model over the Base model. The ratio indicates the relative scale of probing gains compared to generation gains.

	Zebra-TF			Zebra-MC		
Max Δ	High	Med	Low	High	Med	Low
Probing gain	2.05	5.30	3.70	2.95	1.60	2.85
Generation gain	30.05	25.75	25.25	12.75	33.20	31.40
Ratio (Prob / Gen)	**0.07**	**0.21**	**0.15**	**0.23**	**0.05**	**0.09**
	MuseD-TF			MuseD-MC		
Max Δ	High	Med	Low	High	Med	Low
Probing gain	3.56	0	0	3.44	0	0
Generation gain	19.96	12.60	11.16	15.92	14.44	14.24
Ratio (Prob / Gen)	**0.18**	**0.00**	**0.00**	**0.22**	**0.00**	**0.00**
	MATH			HaluEval		
Max Δ	TF	MC	-	TF	MC	-
Probing gain	10.75	12.05	-	2.42	4.22	-
Generation gain	49.05	54.25	-	10.81	32.15	-
Ratio (Prob / Gen)	**0.22**	**0.22**	-	**0.22**	**0.13**	-

We then analyze the development of initial representation quality and generation accuracy across tasks, as shown in Figure 3. We observe that post-training yields only marginal gains in representation quality, with improvement typically remaining under 5% across training stages. This contrasts sharply with the substantial gains in generation accuracy. Specifically, as shown in Table 2, the maximum performance gain in probing is less than 25% of the corresponding generation gain. These findings demonstrate that the development of reasoning does not stem primarily from the improvement of the initial internal states formed upon processing a question. Moreover, among reasoning models, those trained via distillation consistently outperform the pure RL variant in both representation quality and generation accuracy, suggesting that distillation may offer a more effective strategy for reasoning enhancement. Notably, initial representation quality frequently surpasses generation accuracy (observed in 10 out of 16 tasks, $\approx 60\%$). This indicates that models possess a strong latent reasoning capacity that is often not fully realized in their generation outcomes.

4.2. Representation Quality Dynamics during Generation

Despite the general trend of strong initial representations, we observe that strong reasoning models can achieve generation accuracy that surpasses their initial representation quality, particularly on complex tasks such as Zebra and MATH. This gap motivates us to investigate whether CoT enables the model to progressively improve its internal judgments. Therefore, we track probing accuracy across dif-

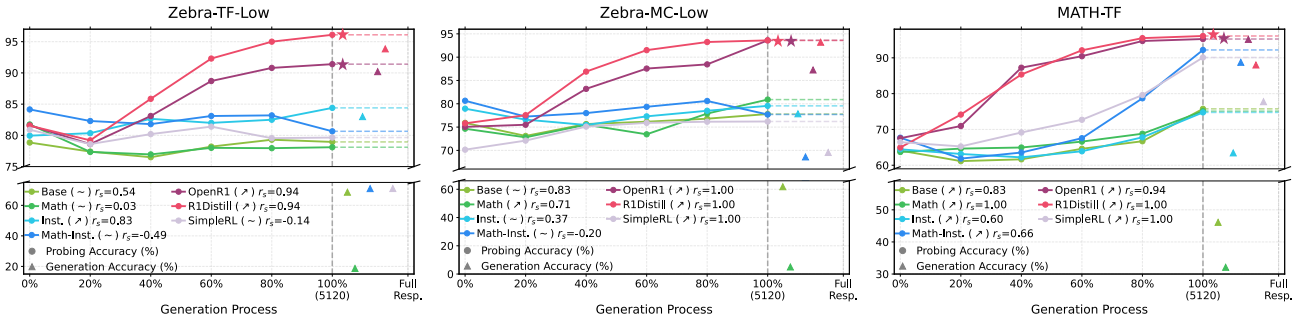

Figure 4. **Representation quality dynamics during generation.** Trends are analyzed using linear regression and Spearman rank correlation. We mark the final probing accuracy of the strong reasoning models with ★. See Appendix E.2 for full results.

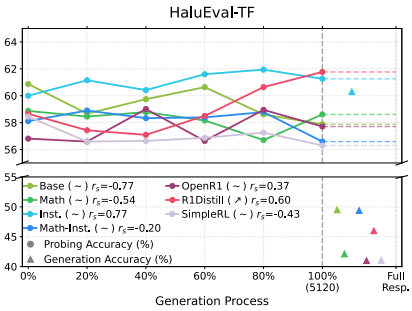

Figure 5. **Representation quality remains relatively stable** during generation on the factuality task.

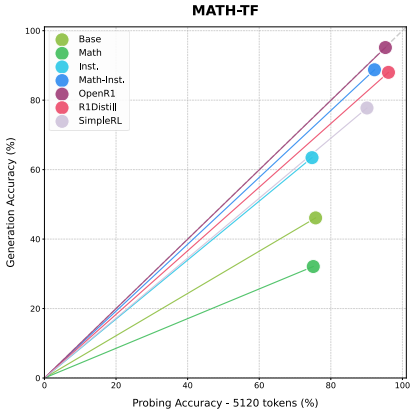

Figure 6. **Generation accuracy versus probing accuracy.**

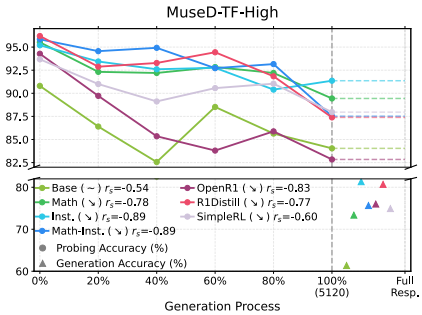

Figure 7. **Representation quality exhibits a decreasing trend** during generation on a simple reasoning task.

ferent stages of generation on representative tasks where generation accuracy exceeds the initial representation quality. For comparison, we conduct identical experiments on HaluEval. We split the model's response by "\n" and construct progressive probing datasets by appending different portions of CoT. Due to GPU memory constraints, we truncate contexts to the first $5,120$ tokens.

The main results are presented in Figure 4 and Figure 5, where we quantify probing accuracy trends using linear regression and Spearman rank correlation (Spearman, 1904). For linear regression, we determine the trend (↗, ↘) based on the slope if the p-value < 0.1; otherwise, the trend is classified as fluctuating ($\sim$). Additionally, we employ the Spearman correlation coefficient r_s to assess monotonic relationships without relying on linear assumptions.

On the factuality task, which demands minimal reasoning, representation quality remains relatively stable, exhibiting only minor fluctuations throughout the CoT. In contrast, on reasoning tasks, representation quality evolves continuously during generation. We observe that models achieve improved representations following CoT, indicating that the latent reasoning capacity of a single forward pass is often insufficient for complex problems; instead, CoT enables models to iteratively refine internal states. Crucially, advanced reasoning models distinguish themselves by leveraging more effective CoT to attain significantly higher final representation quality, a capability acquired during post-training. Furthermore, as illustrated in Figure 6, post-training can also enhance the model's ability to convert these internal representations into generation accuracy. However, a gap persists: while explicit generation accuracy can surpass initial quality, it frequently falls short of the representation quality of the better final state.

We also conduct an experiment on a simple task (MuseD) where the initial representation is already nearly-perfect. As shown in Figure 7, we find that the reasoning process can be a double-edged sword: the representation quality generally exhibits a downward trend, with degradation being particularly significant in strong reasoning models.

Table 3. **Generation-representation alignment.** Significance levels P_{MWU} of Mann-Whitney U test: NS denotes non-significant ($> 5e^{-2}$), * ($< 5e^{-2}$), and *** ($< 1e^{-10}$). **Bold**: higher alignment with CoT; Blue: R1Distill alignment is lower than Inst. Full results are provided in Appendix E.3.

	Inst.				R1Distill			
High-Difficulty Task	Trend	$r_s\uparrow$	ROC-AUC↑	P_{MWU}	Trend	$r_s\uparrow$	ROC-AUC↑	P_{MWU}
MusED-TF	↗	0.59	0.57	*	∼	0.07	0.53	NS
MusED-TF+CoT	↗	**0.99**	**0.85**	***	↗	**0.95**	**0.88**	***
Zebra-TF	↗	0.88	0.55	*	∼	0.07	0.51	NS
Zebra-TF+CoT	↗	**1.00**	**0.86**	***	↗	**0.76**	**0.62**	***
MATH-TF	↗	0.88	0.59	*	∼	0.45	0.57	*
MATH-TF+CoT	↗	**0.98**	**0.70**	***	↗	**0.52**	**0.90**	***

5. Deep Analysis

While previous experiments establish the dynamic evolution of representation quality, the underlying relationship between explicit token generation and latent representation transitions remains unclear, which we specifically investigate in this section.

5.1. Statistical Analysis of Alignment

We first examine the alignment between internal representations and external generation outcomes. We select Inst. and R1Distill as representative non-reasoning and reasoning models, respectively, given their superior performance among post-trained models. To quantify the alignment, we define two variables: the probing probability of the correct label $p \in [0, 1]$ and the generation correctness indicator $\delta \in \{0, 1\}$. We partition samples into ten equal-width buckets based on p, calculate the average accuracy within each bucket, and derive the linear regression trend and Spearman rank correlation coefficient r_s. Additionally, we report the sample-level ROC-AUC score and the p-value of the Mann-Whitney U test (Wilcoxon, 1945; Mann & Whitney, 1947) to assess the degree and significance of the divergence in p distributions between correct and incorrect generations.

As presented in Table 3, generation correctness exhibits a weak correlation with the initial representations, as evidenced by low ROC-AUC scores consistently remaining below 0.6. This demonstrates that the generation process does not merely articulate the model's initial thought, and that the final answer in reasoning scenarios is largely independent of the initial representations. Conversely, generation correlates significantly more strongly with the final representations following CoT. However, the degree of alignment for R1Distill is not higher than that of Inst., indicating that reasoning models do not possess superior efficiency in converting internal representation quality into generation accuracy compared to non-reasoning models.

Additionally, we investigate the relationship between initial and final representations to quantify the extent of the state transition. The results are shown in Table 4, reporting the Pearson correlation coefficient r_p (Pearson, 1896) and linear regression r^2, alongside the Spearman rank correlation coefficient r_s. Across all metrics, the initial and final representations exhibit extremely low correlation, which confirms that the transition is not simply a minor refinement of the initial states, but rather a fundamental distributional reshaping. Notably, R1Distill consistently exhibits lower correlations than Inst. across all reasoning tasks, indicating that reasoning facilitates a more intense state transition.

Table 4. **Alignment between representations before and after CoT**. r_s: Spearman rank correlation coefficient; r_p: Pearson correlation coefficient; r^2 is adopted as the primary metric of linear regression, as p-values tend to become uninformative in large datasets. Blue: R1Distill alignment is lower than Inst.

	Inst.			R1Distill		
	$r_s\uparrow$	$r_p\uparrow$	$r^2\uparrow$	$r_s\uparrow$	$r_p\uparrow$	$r^2\uparrow$
MusED-TF-High	0.16	0.13	0.02	0.10	0.05	0.00
Zebra-TF-High	0.30	0.30	0.09	0.29	0.29	0.08
MATH-TF	0.28	0.29	0.08	0.09	0.05	0.00

5.2. Counterfactual Analysis of Transition Causes

Previous sections show that the representations undergo a distributional transition during generation. However, the driver of this transition remains unclear. To this end, we conduct counterfactual experiments to disentangle the effects of the generated content's semantics, the additional computation introduced by CoT inference, and the parameter differences among models at different stages.

Specifically, we measure the representation quality of the Inst. conditioned on CoT sequences from different sources: Base, Inst. itself, and R1Distill. To control for computation and context length, we set three baselines: gibberish dot sequences, repeated prompts (which prior work (Xu et al., 2024; Pfau et al., 2024) suggests may aid reasoning), and CoT on irrelevant problems. For the gibberish and irrelevant baselines, we keep the token length the same as the original Inst. CoT. For the repeated prompt baseline, we repeat the

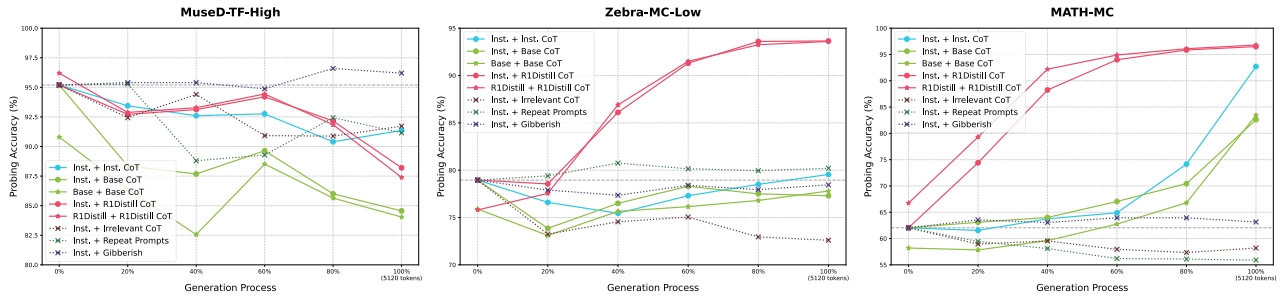

Figure 8. **Inst. representation quality dynamics across different CoT sources.** The dashed lines indicate the initial probing accuracy.

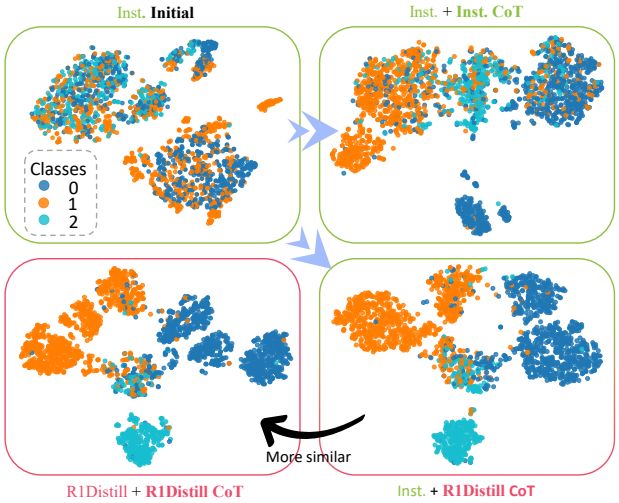

Figure 9. **Visualization of the distribution shift in Inst. representations** induced by Inst. versus R1Distill CoT on Zebra-MC-Low.

problem statement up to five times. We conduct this analysis on three representative tasks where the Inst. exhibits downward, fluctuating, and rising trends.

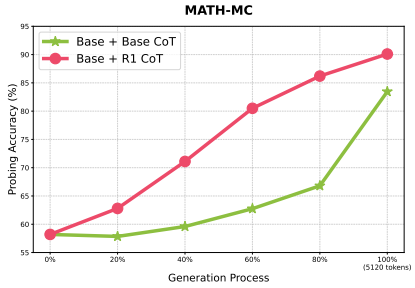

Figure 10. **Dynamics of Base model representation quality** induced by Base versus R1 CoT.

The results in Figure 8 demonstrate that the semantics of the generated content, compared with the other two factors, is the key driver of the observed transitions. Merely increasing computation through meaningless and semantically irrelevant content fails to enhance representation quality, and repeated prompts also prove ineffective in the absence of generated responses. Moreover, when Inst. is conditioned on external CoT, its representation quality aligns more closely with that of the source model than with its self-generated CoT. This underscores the dominance of generation semantics over intrinsic parameter differences in shaping representations, suggesting that the model's fundamental representational capacity remains relatively stable throughout post-training. We provide further evidence by visualizing the representation distributions in Figure 9 using t-SNE (Maaten & Hinton, 2008), where the distribution of Inst. conditioned on R1Distill CoT overlaps more significantly with that of R1Distill. Additionally, we apply DeepSeek-R1 (Guo et al., 2025) CoT to the Base model, and the results in Figure 10 reveal that high-quality reasoning paths can elicit superior representations even in pre-trained models. This highlights the primacy of semantics and also explains why distillation can effectively empower weak models with strong reasoning capability.

6. Conclusion and Discussion

We investigate the evolution of model reasoning ability from the perspectives of generation and representation. Through comprehensive experiments, we discover that post-training improves the static initial representation quality only to a limited extent. More importantly, the generation process does not simply verbalize the model's internal thoughts but continuously shifts representations into a new distribution; post-training enables the model to leverage this process more effectively to attain higher final representation quality. Furthermore, we explore the deep relationship between internal representations and external generation through statistical analysis and identify the main driver of the representation transition via counterfactual experiments. These findings offer valuable insights into reasoning interpretability and highlight promising directions for model enhancement, including: utilizing internal signals for self-training to align generation with representation; adjusting thinking budgets based on the trend of representation transition; controlling model behaviors at test time to prevent representation degradation; and developing optimized generation strategies to attain high-quality representation efficiently, etc.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Limitations

Model Coverage. Our primary analysis focuses on the Qwen2.5-7B model family across four training stages, supplemented by validation on Llama-3.1-8B and Qwen2.5-1.5B. While the core findings are consistent across these models and scales, we acknowledge that broader validation across more diverse architectures (e.g., Mixtral (Jiang et al., 2024), Gemma (Team et al., 2024)) and larger scales (e.g., 70B+) would further strengthen the generalizability of our conclusions. However, due to limited computational resources, we only conduct validations on small-scale models and the most widely used models.

Task Format. To enable controlled probing with a classification head, we convert tasks into TF or MC formats. Although this design is necessary for quantitative measurement of representation quality, it constrains our analysis to structured answer spaces. Open-ended generation scenarios (e.g., free-form mathematical derivation, code generation) may exhibit different representation dynamics, and extending our framework to such settings remains an important direction.

From Analysis to Application. Based on our findings, we outline several promising directions for model improvement, such as utilizing internal representation signals for self-training, adaptive thinking budget allocation, and test-time behavior control. However, as the primary contribution of this work lies in the analytical framework and empirical understanding of reasoning dynamics, we leave the implementation and validation of these proposed methods to future work.

References

Afzal, A., Matthes, F., Chechik, G., and Ziser, Y. Knowing before saying: LLM representations encode information about chain-of-thought success before completion. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 12791–12806. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-acl.662/.

Allen-Zhu, Z. Physics of language models: Part 4.1, architecture design and the magic of canon layers. *arXiv preprint arXiv:2512.17351*, 2025.

Atakishiyev, S., Babiker, H. K., Dai, J., Farruque, N., Hayashi, T., Hriti, N. S., Rahman, M. A., Smith, I., Kim, M.-Y., Zaïane, O. R., et al. Explainability of large language models: Opportunities and challenges toward generating trustworthy explanations. *arXiv preprint arXiv:2510.17256*, 2025.

Azaria, A. and Mitchell, T. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Belinkov, Y., Màrquez, L., Sajjad, H., Durrani, N., Dalvi, F., and Glass, J. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1–10, 2017.

Biran, E., Gottesman, D., Yang, S., Geva, M., and Globerson, A. 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*, pp. 14113–14130, 2024.

Cencerrado, I. V. M., Masdemont, A. P., Hawthorne, A. G., Africa, D. D., and Pacchiardi, L. No answer needed: Predicting llm answer accuracy from question-only linear probes. *arXiv preprint arXiv:2509.10625*, 2025.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Face, H. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Gandhi, K., Chakravarthy, A., Singh, A., Lile, N., and Goodman, N. D. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

Ghandeharioun, A., Caciularu, A., Pearce, A., Dixon, L., and Geva, M. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *International Conference on Machine Learning*, pp. 15466–15490. PMLR, 2024.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J., Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Gurnee, W. and Tegmark, M. Language models represent space and time. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Havrilla, A., Du, Y., Raparthy, S. C., Nalmpantis, C., Dwivedi-Yu, J., Zhuravinskyi, M., Hambro, E., Sukhbaatar, S., and Raileanu, R. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*, 2024.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

Hewitt, J. and Manning, C. D. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, 2019.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Huh, M., Cheung, B., Wang, T., and Isola, P. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=BH8TYy0r6u.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Karan, A. and Du, Y. Reasoning with sampling: Your base model is smarter than you think. *arXiv preprint arXiv:2510.14901*, 2025.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Kudo, K., Aoki, Y., Kuribayashi, T., Sone, S., Taniguchi, M., Brassard, A., Sakaguchi, K., and Inui, K. Think-to-talk or talk-to-think? when llms come up with an answer in multi-step reasoning. *arXiv e-prints*, pp. arXiv–2412, 2024.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Li, D., Cao, S., Griggs, T., Liu, S., Mo, X., Tang, E., Hegde, S., Hakhamaneshi, K., Patil, S. G., Zaharia, M., et al. Llms can easily learn to reason from demonstrations structure, not content, is what matters! *arXiv preprint arXiv:2502.07374*, 2025.

Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., and Wen, J.-R. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, 2023a.

Li, J., Zhang, Y., Shen, W., Yan, Y., Xie, J., and Yan, D. Boosting deductive reasoning with step signals in rlhf, 2024. URL https://arxiv.org/abs/2410.09528.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023b.

Liang, Z., Li, R., Zhou, Y., Song, L., Yu, D., Du, X., Mi, H., and Yu, D. Clue: Non-parametric verification from experience via hidden-state clustering. *arXiv preprint arXiv:2510.01591*, 2025.

Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S., and Lin, M. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov): 2579–2605, 2008.

Mann, H. B. and Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pp. 50–60, 1947.

Matsutani, K., Takashiro, S., Minegishi, G., Kojima, T., Iwasawa, Y., and Matsuo, Y. Rl squeezes, sft expands: A comparative study of reasoning llms. *arXiv preprint arXiv:2509.21128*, 2025.

Mohebbi, H., Modarressi, A., and Pilehvar, M. T. Exploring the role of bert token representations to explain sentence probing results. In *The 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 792–806. Association for Computational Linguistics, 2021.

Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., and Belinkov, Y. Llms know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Pearson, K. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 187:253–318, 1896.

Pfau, J., Merrill, W., and Bowman, S. R. Let's think dot by dot: Hidden computation in transformer language models. *arXiv preprint arXiv:2404.15758*, 2024.

Shao, R., Li, S. S., Xin, R., Geng, S., Wang, Y., Oh, S., Du, S. S., Lambert, N., Min, S., Krishna, R., et al. Spurious rewards: Rethinking training signals in rlvr. *arXiv preprint arXiv:2506.10947*, 2025.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Sky, C.-W., Van Durme, B., Eisner, J., and Kedzie, C. Do androids know they're only dreaming of electric sheep? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4401–4420, 2024.

Spearman, C. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Stojanovski, Z., Stanley, O., Sharratt, J., Jones, R., Adefioye, A., Kaddour, J., and Köpf, A. Reasoning gym: Reasoning environments for reinforcement learning with verifiable rewards, 2025. URL https://arxiv.org/abs/2505.24760.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Team, K., Du, A., Gao, B., Xing, B., Jiang, C., Chen, C., Li, C., Xiao, C., Du, C., Liao, C., et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Team, Q. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Team, Q. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

Tian, X., Zhao, S., Wang, H., Chen, S., Peng, Y., Ji, Y., Zhao, H., and Li, X. Deepdistill: Enhancing llm reasoning capabilities via large-scale difficulty-graded data training. *arXiv preprint arXiv:2504.17565*, 2025.

Vassoyan, J., Beau, N., and Plaud, R. Ignore the kl penalty! boosting exploration on critical tokens to enhance rl fine-tuning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6108–6118, 2025.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, S., Yu, L., Gao, C., Zheng, C., Liu, S., Lu, R., Dang, K., Chen, X., Yang, J., Zhang, Z., et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025a.

Wang, Y., Wang, Y., Zhou, X., and Shen, Z. Response uncertainty and probe modeling: Two sides of the same coin in llm interpretability? *arXiv preprint arXiv:2505.18575*, 2025b.

Wang, Z., Zeng, X., Liu, W., Wang, Y., Li, L., Wang, Y., Shang, L., Jiang, X., Liu, Q., and Wong, K.-F. Chain-of-probe: Examining the necessity and accuracy of cot step-by-step. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2586–2606, 2025c.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

Wu, F., Xuan, W., Lu, X., Harchaoui, Z., and Choi, Y. The invisible leash: Why rlvr may not escape its origin. *arXiv e-prints*, pp. arXiv–2507, 2025a.

Wu, J., Xiong, Y., Li, X., Hu, Z., Yu, T., Wang, R., Chen, X., Shang, J., and McAuley, J. Ctrls: Chain-of-thought reasoning via latent state-transition. *arXiv preprint arXiv:2507.08182*, 2025b.

Xu, X., Tao, C., Shen, T., Xu, C., Xu, H., Long, G., Lou, J.-G., and Ma, S. Re-reading improves reasoning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15549–15575, 2024.

Yan, Y., Li, J., Zhang, Y., and Yan, D. Exploring the llm journey from cognition to expression with linear representations. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 55778–55796, 2024.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.

Yang, A., Zhang, B., Hui, B., Gao, B., Yu, B., Li, C., Liu, D., Tu, J., Zhou, J., Lin, J., Lu, K., Xue, M., Lin, R., Liu, T., Ren, X., and Zhang, Z. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024b.

Yang, S., Gribovskaya, E., Kassner, N., Geva, M., and Riedel, S. 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)*, pp. 10210–10229, 2024c.

Ye, T., Xu, Z., Li, Y., and Allen-Zhu, Z. Physics of language models: Part 2.1, grade-school math and the hidden reasoning process. In *ICLR 2025: International Conference on Learning Representations*, 2024.

Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Song, S., and Huang, G. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Yusupov, V., Maksimov, D., Alaeva, A., Vasileva, A., Antipina, A., Zaitseva, T., Ermilova, A., Burnaev, E., and Shvetsov, E. From internal representations to text quality: A geometric approach to llm evaluation. *arXiv preprint arXiv:2509.25359*, 2025.

Zeng, W., Huang, Y., Liu, Q., Liu, W., He, K., Ma, Z., and He, J. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*, 2025.

Zhang, A., Chen, Y., Pan, J., Zhao, C., Panda, A., Li, J., and He, H. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*, 2025a.

Zhang, C., Neubig, G., and Yue, X. On the interplay of pre-training, mid-training, and rl on reasoning language models. *arXiv preprint arXiv:2512.07783*, 2025b.

Zhang, S., Zhang, Y., Dong, Y., and Su, H. Exploring the generalizability of factual hallucination mitigation via enhancing precise knowledge utilization. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3936–3968, Suzhou, China, November 2025c. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.211. URL https://aclanthology.org/2025.findings-emnlp.211/.

Zhang, Z., Hu, X., Zhang, H., Zhang, J., and Wan, X. Icr probe: Tracking hidden state dynamics for reliable hallucination detection in llms. In *Proceedings of the 63rd*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 17986–18002, 2025d.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., and Du, M. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.

Zhao, R., Meterez, A., Kakade, S., Pehlevan, C., Jelassi, S., and Malach, E. Echo chamber: Rl post-training amplifies behaviors learned in pretraining. *arXiv preprint arXiv:2504.07912*, 2025.

Zheng, C., Liu, S., Li, M., Chen, X.-H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

A. Experimental Task Introduction

In this section, we detail the settings of the experimental tasks. We describe the construction of both the probing dataset $\mathcal{D}^{\text{repr}}$ and the generation dataset $\mathcal{D}^{\text{gen}}$, and provide illustrative examples for each task.

ZebraPuzzle (Stojanovski et al., 2025) is a classic deductive logic puzzle designed to evaluate a model's ability to solve constraint satisfaction problems. Concretely, the problem consists of n houses occupied by n different residents, where each resident is characterized by m attributes. Given a set of constraints, the model must deduce the specific resident associated with each house.

Since the solution requires the simultaneous resolution of all dependencies, this task effectively serves as a metric for reasoning **breadth**. Therefore, we stratify the dataset into three difficulty levels based on n and m: Low for $n, m \in \{2, 3\}$, Med (Medium) for $n, m \in \{3, 4\}$, and High for $n, m \in \{4, 5\}$. For the TF version, the model must verify whether a specific resident occupies a given house. The dataset is balanced with an equal distribution of correct matches and randomly sampled incorrect matches; For the MC version, the model is required to identify the correct resident of a given house from the list of all possible names. For each difficulty level, we generate a training set of $20,000$ samples and a test set of $2,000$ samples, ensuring a balanced distribution across (n, m) combinations. Illustrative examples are shown in Figure 11.

Illustrative examples of ZebraPuzzle

Problem description:
This is a logic puzzle. There are 5 houses (numbered 1 on the left, 5 on the right), from the perspective of someone standing across the street from them. Each has a different person in them. They have different characteristics:\n - Each person has a unique name: arnold, alice, david, carol, bob\n - Everyone has a different favorite cigar: red eye, dunhill, blue master, pall mall, prince\n - They all have a different favorite flower: lilies, daffodils, iris, carnations, orchid\n - Everyone has something different for lunch: pizza, spaghetti, stir fry, grilled cheese, soup\n - The people keep different animals: cat, dog, horse, fish, bird\n\n1. The cat lover is the person who loves a bouquet of daffodils.\n2. The person who loves the soup is the person who smokes Blue Master.\n3. Carol and the person who smokes Red Eye are next to each other.\n4. The person who loves the boquet of orchid is directly left of the person who loves a bouquet of daffodils.\n5. The person who loves stir fry is in the second house.\n6. The person who loves the boquet of lilies is Alice.\n7. The Dunhill smoker is the person who keeps horses.\n8. Bob is directly left of Carol.\n9. Alice is the Dunhill smoker.\n10. The dog owner is the person who smokes Red Eye.\n11. The bird keeper is in the fourth house.\n12. The person who loves the boquet of iris is the person who loves the spaghetti eater.\n13. Arnold is in the second house.\n14. The person who loves eating grilled cheese is directly left of the person who smokes Blue Master.\n15. The person who loves a bouquet of daffodils is the Prince smoker.

TF version:
Probe: <|START|>House 1<|MID|>alice<|END|> *(0 for False and 1 for True.)*
Question: Decide whether alice is the name of the person who lives in House 1.
Correct answer: 1 / True *(alice live in House 1)*

MC version:
Probe: <|START|>House 1<|END|> *(0, 1, 2, 3, 4 for alice, arnold, bob, carol, david, respectively.)*
Question: What is the name of the person who lives in House 1?
Correct answer: 0 / alice *(alice live in House 1)*

Figure 11. Illustrative examples of **ZebraPuzzle**.

MuseD (Li et al., 2024) is a deductive reasoning task designed to evaluate a model's ability to solve multi-hop syllogistic reasoning. Concretely, the problem presents a set of *premises* defining relationships between various entities, utilizing the four classical Aristotelian categorical propositions, where S and P denote the *Subject* and *Predicate*:

- Type 1 (*A*): All S are P.
- Type 2 (*E*): All S are not P.

- Type 3 (*I*): There exists one S that is P.
- Type 4 (*O*): There exists one S that is not P.

Solving the problem requires chaining valid syllogisms (e.g., All M are P + All S are M $\implies$ All S are P) to derive the final relationship between a target subject and predicate.

Since the solution requires a sequence of d deductive steps, this task effectively serves as a metric for reasoning **depth**. Therefore, we stratify the dataset into three difficulty levels based on the information structure: Low, where premises are all necessary and presented in the optimal deductive order; Med, where premises are all necessary but permuted randomly; High, where premises are permuted randomly and include 20% distraction noise (irrelevant relationships). For the TF version, the model must verify the correctness of a specific relationship type between two entities. The dataset is balanced with an equal distribution of correct relationships and incorrect matches (Type A $\leftrightarrow$ Type O, Type E $\leftrightarrow$ Type I). For the MC version, the model is required to identify the best relationship type for a given pair of entities. For each difficulty level, we generate a training set of $20,000$ samples and a test set of $2,500$ samples, covering deduction depths $d \in [3, 25]$ and ensuring a balanced distribution across four relationship types. Illustrative examples are shown in Figure 12.

Illustrative examples of MuseD

Problem description:
Given:\nAll Q are not N.\nThere exists one Gamma that is W.\nAll S are Q.\nAll W are Rho.\nAll Rho are S.

TF version:
Probe: <|START|>W<|MID1|>N<|MID2|>2<|END|> *(0 for False and 1 for True.)*
Question: Decide whether the statement is true or false: All W are not N.
Correct answer: 1 / True *(The relationship type is E)*

MC version:
Probe: <|START|>W<|MID|>N<|END|> *(0, 1, 2, 3 for A, E, I, O, respectively.)*
Question: Decide the relationship between W and N.\n\nType 1: All W are N.\nType 2: All W are not N.\nType 3: There exists one W that is N.\nType 4: There exists one W that is not N.\n\nChoose the single best option that describes the relationship between W and N.
Correct answer: 1 / Type 2 *(The relationship type is E)*

Figure 12. Illustrative examples of **MuseD**.

MATH (Hendrycks et al., 2021) evaluates a model's reasoning ability on **unstructured** and **real-world** mathematical problems. Unlike the limited output spaces of the previous two tasks, mathematical questions typically feature open-ended numerical or algebraic solutions, which are incompatible with fixed-class probing. To adapt the dataset for our framework, we implement a rule-based algorithm to generate plausible incorrect distractors, converting the task into answer verification.

For the TF version, the model must verify whether a provided candidate answer is correct. The dataset is balanced with an equal distribution of ground truth answers and generated incorrect distractors. For the MC version, the model is required to identify the correct answer from a pair of candidates. We construct the training set using the $7,500$ samples from the original MATH training split and create a test set of $2,000$ samples randomly drawn from the original test set. Illustrative examples are shown in Figure 13.

HaluEval (Li et al., 2023a) is a factuality benchmark to evaluate *hallucination* detection in general knowledge queries. It serves as a control task that requires little reasoning ability, relying primarily on knowledge retrieval. We initially randomly split the dataset into a training set of $8,000$ samples and a test set of $2,000$ samples. However, we observe a significant format bias in the original dataset: ground truth answers are typically concise entities, whereas hallucinated answers are complete, lengthy sentences. To prevent models from exploiting answer length as a spurious shortcut, we employ GPT-4o (Hurst et al., 2024) to rewrite the hallucinated answers into the same concise format as the ground truth answers. After filtering out instances where the rewritten hallucinated answer becomes semantically the same as the correct answer, $7,037$ training samples and $1,776$ test samples are retained.

Illustrative examples of MATH

Problem description:
Given $x = -2$ find the value of $2x^2 + 3x + 4$.

TF version:
Probe: <|ANS|>7<|END|> *(0 for False and 1 for True.)*
Question: I want you act as an answer judge. Given a math question and a candidate answer, your objective is to determine if the provided answer is correct or not.\nQuestion: Given $x = -2$ find the value of $2x^2 + 3x + 4$.\nCandidate answer: 7
Correct answer: 0 / False *(The correct answer is 6)*

MC version:
Probe: <|CA1|>7<|CA2|>6<|END|> *(0, 1 for the first and the second candidate answers, respectively.)*
Question: A. 7\nB.6\n\nWhich answer is correct?
Correct answer: 1 / B *(The correct answer is 6)*

Figure 13. Illustrative examples of **MATH**.

Similar to MATH, for the TF version, the model must verify whether a provided candidate answer is factually correct. The dataset is balanced with an equal distribution of correct answers and hallucinated answers. For the MC version, the model is required to identify the correct answer from a pair of candidates. Illustrative examples are shown in Figure 14.

Illustrative examples of HaluEval

Problem description:
Dua Lipa, an English singer, songwriter and model, the album spawned the number-one single N̈ew Rulesïs a song by English singer Dua Lipa from her eponymous debut studio album, released in what year?

TF version:
Probe: <|ANS|>2018<|END|> *(0 for False and 1 for True.)*
Question: *HaluEval template* + Question: Dua Lipa, an English singer, songwriter and model, the album spawned the number-one single N̈ew Rulesïs a song by English singer Dua Lipa from her eponymous debut studio album, released in what year?\nAnswer: 2018\nYour Judgement:
Correct answer: 0 / False *(The correct answer is 2017)*

MC version:
Probe: <|CA1|>2017<|CA2|>2018<|END|> *(0, 1 for the first and the second candidate answers, respectively.)*
Question: A. 2017\nB.2018\n\nWhich answer is correct?
Correct answer: 0 / A *(The correct answer is 2017)*

Figure 14. Illustrative examples of **HaluEval**.

B. Model Introduction

In this section, we detail the training relationships within our experimental models, including Qwen2.5-7B, -1.5B series (Team, 2024; Yang et al., 2024b) and Llama3.1-8B series (Dubey et al., 2024). We primarily conduct experiments and analysis on the Qwen2.5-7B series, as its diverse variants spanning multiple training stages provide an ideal testbed for comprehensively analyzing the evolution of reasoning ability.

Qwen2.5-7B. In this work, we evaluate a suite of seven different models from the Qwen2.5-7B series, spanning three distinct

developmental stages: pre-training, standard instruct tuning, and reasoning-focused training.

- Pre-trained models: Qwen2.5-7B-Base, the foundational general purpose model. Building on this, Qwen2.5-7B-Math undergoes continuous pre-training on a specialized mathematical corpus and synthetic data from Qwen2-Math-Instruct (Yang et al., 2024a). Pre-trained models represent the base capability of the model series prior to any alignment.
- Instruction-tuned models: Derived from the Base and Math models, Qwen2.5-7B-Instruct and Qwen2.5-7B-Math-Instruct, respectively, undergo a standard post-training pipeline that includes Supervised Fine-Tuning (SFT) and RL to enhance instruction following ability. While we also refer to these as "non-reasoning" models in the main body to distinguish them from the reasoning models, they naturally retain reasoning ability to some extent.
- Reasoning-optimized models: We include three specialized models designed to maximize reasoning performance. Qwen2.5-7B-OpenR1 (Face, 2025) and Qwen2.5-7B-R1Distill (Guo et al., 2025) are distilled models based on Math-Instruct and Math models, respectively. They are fine-tuned (SFT) on reasoning responses generated by DeepSeek-R1 (Guo et al., 2025) without additional RL. Conversely, Qwen2.5-7B-SimpleRL-Zoo (Zeng et al., 2025) applies RL (specifically GRPO (Shao et al., 2024)) directly to the Math model. These models typically exhibit superior performance on tasks necessitating multi-hop reasoning and analysis compared to their predecessors.

Qwen2.5-1.5B. The Qwen2.5-1.5B series follows a training pipeline identical to that of the 7B series, with the exception of the OpenR1 variant, which does not exist at this scale. Therefore, we include the remaining six corresponding models in our experimental analysis.

Llama3.1-8B. We evaluate four different variants of the Llama3.1-8B series. In addition to the Base model, we include the Instruct, R1Distill, and SimpleRL variants, which are derived from the Base model via instruction tuning, distillation, and RL, respectively.

C. Probing Settings

In this section, we detail the implementation of the V-probing technique (Ye et al., 2024) and specify the experimental hyperparameters and settings to ensure reproducibility. Furthermore, we provide an information-theoretic discussion suggesting that, for tasks of sufficient complexity, the additional trainable parameters in V-probing alone lack the capacity to solve the task.

C.1. Training Settings

We implement V-probing using Low-Rank Adaptation (LoRA) (Hu et al., 2022). Specifically, we adopt the configuration with a small rank of $r = 4$, a scaling factor of $\alpha = 16$, a dropout ratio of $p = 0.1$, and no bias terms. Optimization is performed using the standard Adam optimizer (Kingma & Ba, 2015) with a learning rate of $\gamma = 1e^{-4}$ and a batch size of $b = 32$ by default. All models are initialized in bfloat16 precision.

For the evaluation of initial representation quality in section 4.1, we train the model for a fixed budget of $N = 10$ epochs to ensure sufficient training ($N = 30$ epochs for MATH and HaluEval tasks that have less training data). The input x is constructed by concatenating the probe trigger after the problem description, and the label is predicted using the hidden state of the last token at the final layer. To account for training variance (where the standard deviation across three runs with different random seeds on the same task can reach up to 5%), we conduct four independent runs for each experiment and report the highest value. When evaluating the development of representation quality during the CoT process in section 4.2, we dynamically adjust the training epochs based on the convergence speed observed in the initial experiments. In this experimental setting, we introduce another special token <|Reasoning|> to separate the problem statement and the CoT, where the input x is formatted as follows: `problem description` + <|Reasoning|> + `CoT` + `probe trigger`.

For the comparative analysis in section 5.2, we establish three baseline conditions to evaluate the necessity of meaningful CoT content: gibberish sequences, repeated prompts, and irrelevant CoT. In each experimental setting, we replace the original `CoT` with a corresponding baseline sequence. For gibberish sequences, we substitute the `CoT` part with a sequence of repeated dots (" .") of the same length; For repeated prompts, we replace the `CoT` part with the probe trigger appended with the problem description, repeated 1 to 5 times; For irrelevant CoT, we substitute the `CoT` part with a CoT generated for an unrelated decryption task, truncated to match the original token length.

C.2. Probing Algorithm

The pseudocode for the V-probing implementation is presented in Algorithm 1.

Algorithm 1 Representation Quality Quantification via V-probing

Input: Dataset $\mathcal{D}^{\text{repr}}$ split into $\mathcal{D}^{\text{repr}}_{\text{train}}$ and $\mathcal{D}^{\text{repr}}_{\text{test}}$, number of special tokens k, number of classes s; model component f_θ with hidden size m; learning rate γ, LoRA rank r, number of epochs N.

Training:
Freeze backbone parameters θ
Initialize trainable parameters $\Phi = \{E_{\text{sp}}, \Delta\theta_{\text{emb}}, h\}$, where:
  $E_{\text{sp}} \in \mathbb{R}^{k\times m}$ (Special token embeddings)
  $\Delta\theta_{\text{emb}}$ (LoRA adapter for embeddings, rank r)
  $h \in \mathbb{R}^{m\times s}$ (Linear probe head)
for epoch $= 1$ **to** N **do**
  for each batch $\mathcal{B} = \{(x^{(i)}, z^{(i)})\} \subset \mathcal{D}^{\text{repr}}_{\text{train}}$ **do**
    Compute representations: $c^{(i)} = f_{\theta,E_{\text{sp}},\Delta\theta_{\text{emb}}}(x^{(i)})$
    Estimate distributions: $p_\Phi(z|x^{(i)}) = \text{softmax}(h(c^{(i)}))$
    Compute Loss: $\mathcal{L}_{\text{CE}} = \frac{1}{|\mathcal{B}|}\sum_{(x^{(i)},z^{(i)})\in\mathcal{B}} -\log p_\Phi(z^{(i)}|x^{(i)})$
    Update $\Phi \leftarrow \Phi - \gamma\nabla_\Phi\mathcal{L}_{\text{CE}}$
  end for
end for

Evaluation:
Initialize correct count $C \leftarrow 0$
for each batch $\mathcal{B} = \{(x^{(i)}, z^{(i)})\} \subset \mathcal{D}^{\text{repr}}_{\text{test}}$ **do**
  Compute representations: $c^{(i)} = f_{\theta,E_{\text{sp}},\Delta\theta_{\text{emb}}}(x^{(i)})$
  Estimate distributions: $p_\Phi(z|x^{(i)}) = \text{softmax}(h(c^{(i)}))$
  Predict label: $\hat{z}^{(i)} = \arg\max_z p_\Phi(z|x^{(i)})$
  $C \leftarrow C + \sum_{(x^{(i)},z^{(i)})\in\mathcal{B}} \mathbf{1}(\hat{z}^{(i)} = z^{(i)})$
end for
Return $Acc_{\text{repr}} = C/|\mathcal{D}^{\text{repr}}_{\text{test}}|$

C.3. Information-Theoretic Analysis of V-probing

In this section, we analyze the capacity of the V-probing parameters to fit the task data. We establish an upper bound on the achievable training accuracy based on information theory, under the assumption that the model's pre-existing parameters possess no knowledge of the target labels. We demonstrate that, for tasks of sufficient complexity, the probe's limited parameter budget lacks the capacity to reach high accuracy independently. The following derivation focuses on the simplest binary classification setting.

Assumption 1 (Random Labels). Let the target labels $Y^N = (Y_1, \ldots, Y_N) \in \{0,1\}^N$ be independent and identically distributed (*i.i.d.*) random variables following a Bernoulli distribution with $p = 0.5$, where N denotes the dataset size. Consequently, the total label entropy is $H(Y^N) = N$ bits. This assumption represents the scenario where the task is sufficiently complex, containing N bits of incompressible entropy, and the model's existing parameters provide no informative signal.

Assumption 2 (Limited Capacity). Let Θ denote the trainable parameters of the probe. We bound the effective information capacity of these parameters by P_{eff} bits, such that $H(\theta) \leq P_{eff}$.

Proposition C.1 (Capacity-Constrained Accuracy Bound)**.** *Let Acc be the random variable representing the training accuracy of the probe with parameter capacity P_{eff} bits on a balanced binary classification dataset of size N. Under the two assumptions, the expected accuracy is upper bounded by:*

$$\mathbb{E}[Acc] \leq \frac{1}{2} + \sqrt{\frac{\ln 2}{2} \cdot \frac{P_{eff}}{N}} \tag{4}$$

Proof

1. **Global Information Constraint:** Let $\hat{Y}^N$ denote the sequence of predictions. The mutual information between the true labels and the predictions is bounded by the capacity of the probe parameters Θ:
$$I(Y^N; \hat{Y}^N) \leq I(Y^N; \Theta) \leq H(\Theta) \leq P_{eff} \tag{5}$$

2. **Conditional Entropy Lower Bound:** Since $I(Y^N; \hat{Y}^N) = H(Y^N) - H(Y^N|\hat{Y}^N)$ and $H(Y^N) = N$, we have:
$$H(Y^N|\hat{Y}^N) = N - I(Y^N; \hat{Y}^N) \geq N - P_{eff} \tag{6}$$

3. **Decomposition and Fano's Inequality:** The joint conditional entropy is upper bounded by the sum of marginal conditional entropies:
$$H(Y^N|\hat{Y}^N) \leq \sum_{i=1}^{N} H(Y_i|\hat{Y}_i) \tag{7}$$
Let $P_{e,i} = \mathbb{P}(\hat{Y}_i \neq Y_i)$ be the marginal error probability for the i-th sample. By Fano's inequality for binary variables, $H(Y_i|\hat{Y}_i) \leq h_2(P_{e,i})$, where $h_2(\cdot)$ is the binary entropy function. Thus:
$$H(Y^N|\hat{Y}^N) \leq \sum_{i=1}^{N} h_2(P_{e,i}) \tag{8}$$

4. **Applying Jensen's Inequality:** Let the expected error rate be $\bar{P}_e = \frac{1}{N}\sum_{i=1}^{N} P_{e,i} = 1 - \mathbb{E}[Acc]$. Since the binary entropy function $h_2(p)$ is concave, Jensen's Inequality implies that the average of the function values is less than or equal to the function of the average:
$$\frac{1}{N}\sum_{i=1}^{N} h_2(P_{e,i}) \leq h_2\left(\frac{1}{N}\sum_{i=1}^{N} P_{e,i}\right) = h_2(\bar{P}_e) \tag{9}$$
Combining Equation (6), Equation (8), and Equation (9):
$$N - P_{eff} \leq N \cdot h_2(\bar{P}_e) \implies h_2(1 - \mathbb{E}[Acc]) \geq 1 - \frac{P_{eff}}{N} \tag{10}$$
Note that $h_2(1-x) = h_2(x)$, so we have $h_2(\mathbb{E}[Acc]) \geq 1 - \frac{P_{eff}}{N}$.

5. **Pinsker's Inequality Bound:** We relate the binary entropy to the classification accuracy deviation using Pinsker's inequality. The KL divergence between a Bernoulli variable with parameter p and the uniform distribution is $1 - h_2(p)$. Pinsker's inequality states:
$$1 - h_2(p) \geq \frac{2}{\ln 2}(p - 0.5)^2 \tag{11}$$
Substituting $p = \mathbb{E}[Acc]$ and rearranging:
$$h_2(\mathbb{E}[Acc]) \leq 1 - \frac{2}{\ln 2}(\mathbb{E}[Acc] - 0.5)^2 \tag{12}$$

6. **Final Derivation:** Substituting Equation (12) into Equation (10):
$$1 - \frac{P_{eff}}{N} \leq 1 - \frac{2}{\ln 2}(\mathbb{E}[Acc] - 0.5)^2 \tag{13}$$
$$\frac{2}{\ln 2}(\mathbb{E}[Acc] - 0.5)^2 \leq \frac{P_{eff}}{N} \tag{14}$$
Solving for $\mathbb{E}[Acc]$ gives the result:
$$\mathbb{E}[Acc] \leq 0.5 + \sqrt{\frac{\ln 2}{2} \cdot \frac{P_{eff}}{N}} \tag{15}$$
So far, we have proven the Proposition C.1. □

The term $\sqrt{\frac{P_{eff}}{N}}$ decays as the dataset size N increases relative to the probe's capacity P_{eff}. Therefore, if the model's existing parameters do not contain any knowledge about the complex task and the sample size is sufficiently large, then the V-probing accuracy will converge to random guessing. Conversely, observing a high training accuracy implies that the hypothesis is false, i.e., the model's existing parameters must contain significant information about the target labels.

D. Generation Settings

In this section, we detail the generation settings used to evaluate generation accuracy.

We employ vLLM (Kwon et al., 2023) for generation, loading all models in bfloat16 precision to align with the probing settings. A context window of $32,768$ tokens is allocated to ensure sufficient space for long reasoning chains. For decoding configurations, we apply the default temperature for pre-trained models and instruction-tuned (non-reasoning) models. For reasoning models, we adopt the widely utilized configuration of temperature 0.6 and top-p 0.95 following Guo et al. (2025); Face (2025). To construct the input, we append the specific question after the problem description. To facilitate automated answer parsing, we prompt the model to enclose its final answer in "\\boxed{}". While most models utilize their default chat templates, for models lacking a native chat template (Llama3.1-8B-Base and Llama3.1-8B-SimpleRL) or those with limited instruction-following ability (Qwen2.5-1.5B pre-trained and SimpleRL variants), we adopt the simple chat template proposed by Zeng et al. (2025).

Since models after post-training employ stochastic decoding with a moderate temperature, unlike the greedy decoding used in pre-trained models, the variance introduced by random sampling may affect the reliability of reported generation accuracies. To assess this, we evaluate generation stability by running each experiment three times with different random seeds on two representative models: Inst. and R1Distill, across three tasks spanning different categories: Zebra-MC-High, MuseD-TF-High, and MATH-MC.

Table 5. **Generation stability across three independent runs** with different random seeds. We report accuracy (%) for each run, the mean, and the standard deviation (Std).

MODEL	TASK	RUN 1	RUN 2	RUN 3	MEAN	STD
INST.	ZEBRA-MC-HIGH	35.10	37.55	35.00	35.88	1.18
	MUSED-TF-HIGH	80.88	79.68	81.16	80.57	0.64
	MATH-MC	89.05	89.60	88.25	88.97	0.55
R1DISTILL	ZEBRA-MC-HIGH	20.80	18.90	18.55	19.42	0.99
	MUSED-TF-HIGH	82.64	82.52	81.64	82.27	0.45
	MATH-MC	96.30	96.65	96.50	96.48	0.14

As shown in Table 5, the standard deviations across three runs remain consistently low, ranging from 0.14% to 1.18% across all model-task combinations. This indicates that stochastic decoding introduces negligible variance in generation accuracy, and the reported results are robust and reproducible.

E. Additional Experimental Results

In this section, we present supplementary experimental results to support the main analysis. We provide the complete results for Qwen2.5-7B series, as well as the results from the replication experiments conducted on the Qwen2.5-1.5B and Llama3.1-8B series.

E.1. Development of Initial Representation Quality

In this section, we present the complete experimental results of section 4.1 (Figure 15 and Figure 16). Additionally, we detail the replication experiments conducted on representative tasks for Llama3.1-8B series (Figure 17, Table 6) and Qwen2.5-1.5B series (Figure 18, Table 7).

The additional results from the Llama3.1-8B and Qwen2.5-1.5B series demonstrate findings consistent with the main body. First, both model families confirm the existence of latent reasoning ability in the pre-trained models, with probing accuracies significantly above the majority guessing baseline. Second, the improvement in representation quality after post-training remains marginal, typically staying below 5% across training stages. Notably, reasoning models even exhibit a degradation in representation quality after training on complex tasks (e.g., MATH and ZebraPuzzle) in Qwen2.5-7B series. Compared

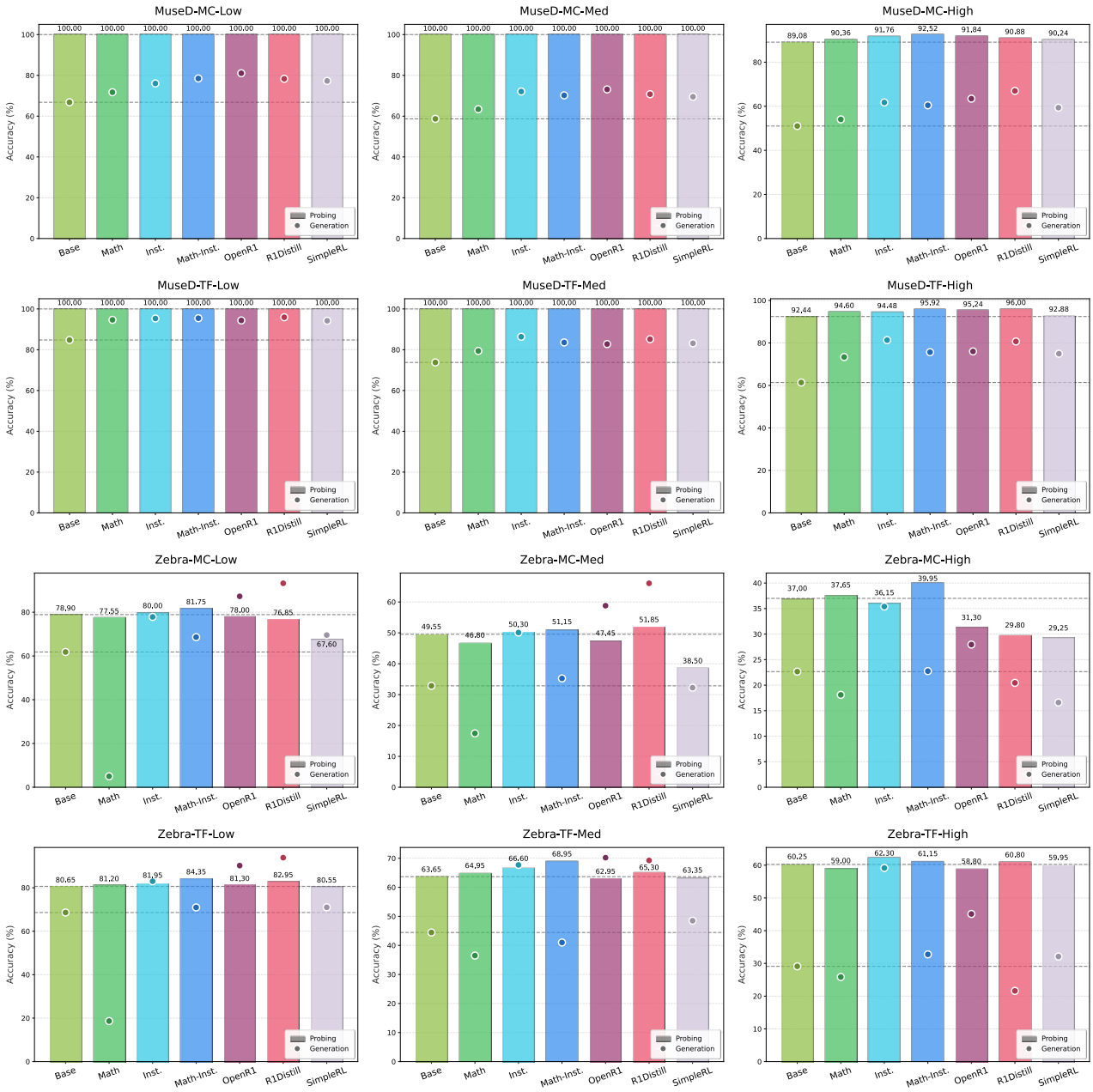

Figure 15. **Qwen2.5-7B results (Part 1): Development of initial representation quality and generation accuracy across all tasks.** Gray dashed lines indicate the baseline performance of the Base model for probing and generation.

Table 6. **Llama3.1-8B results: Maximum performance gain** (Max Δ) represents the highest accuracy improvement (%) achieved by any other model over the Base model. The ratio indicates the relative scale of probing gains compared to generation gains.

MAX Δ	ZEBRA-TF-LOW	ZEBRA-MC-HIGH	ZEBRA-MC-LOW	MUSED-TF-HIGH	MATH-TF	MATH-MC	HALUEVAL-TF
PROBING GAIN	1.85	5.25	1.25	0.58	4.15	1.85	1.69
GENERATION GAIN	46.35	30.10	51.35	40.96	57.40	46.30	13.68
RATIO (PROB / GEN)	**0.04**	**0.17**	**0.02**	**0.01**	**0.07**	**0.04**	**0.12**

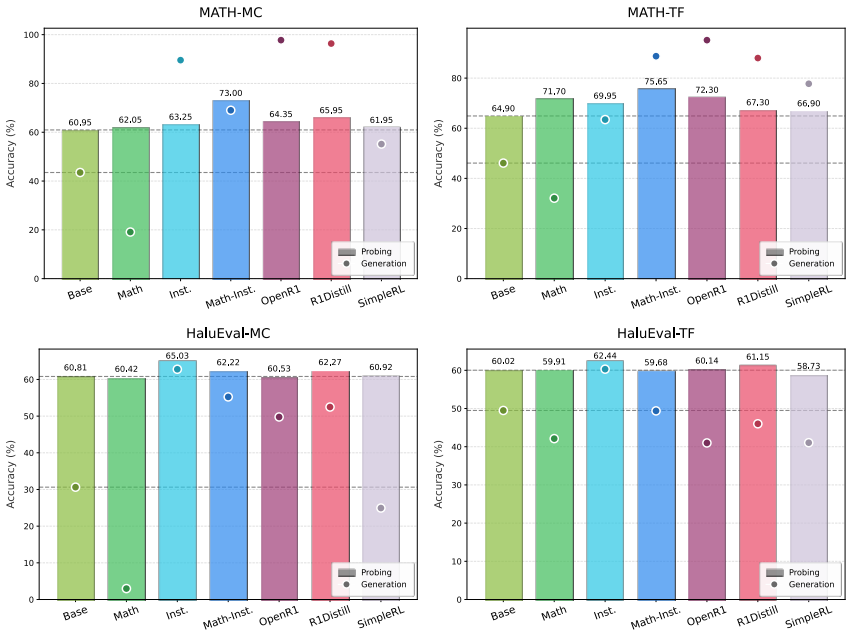

Figure 16. **Qwen2.5-7B results (Part 2): Development of initial representation quality and generation accuracy across all tasks.** Gray dashed lines indicate the baseline performance of the Base model for probing and generation.

Table 7. **Qwen2.5-1.5B results: Maximum performance gain** (Max Δ) represents the highest accuracy improvement (%) achieved by any other model over the Base model. The ratio indicates the relative scale of probing gains compared to generation gains.

MAX Δ	ZEBRA-TF-LOW	ZEBRA-MC-HIGH	ZEBRA-MC-LOW	MUSED-TF-HIGH	MATH-TF	MATH-MC	HALUEVAL-TF
PROBING GAIN	5.90	0.00	4.10	3.56	4.50	1.90	0.73
GENERATION GAIN	12.20	4.35	25.65	12.24	29.00	42.15	3.83
RATIO (PROB / GEN)	**0.48**	**0.00**	**0.16**	**0.29**	**0.16**	**0.05**	**0.19**

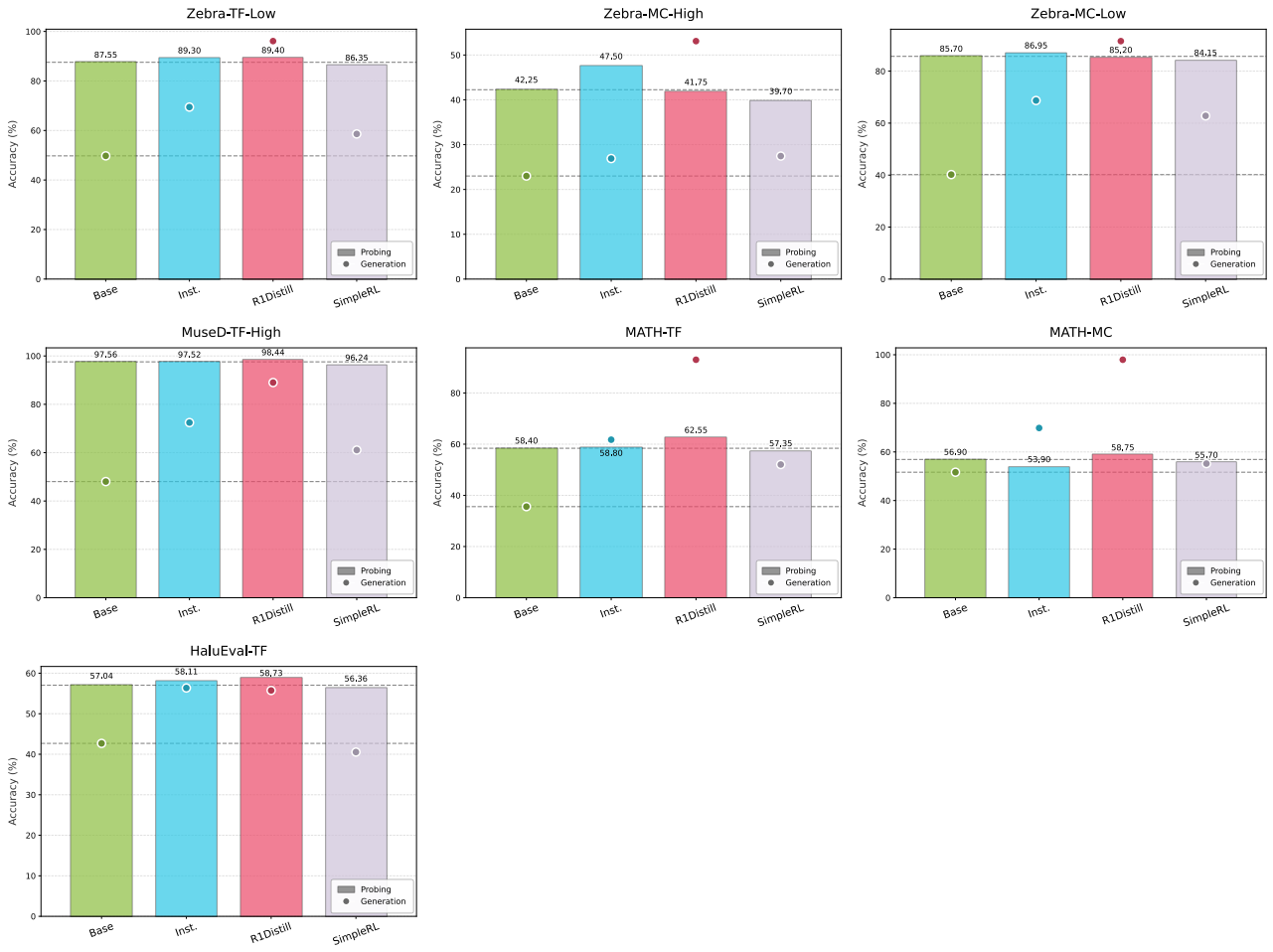

Figure 17. **Llama3.1-8B results: Development of initial representation quality and generation accuracy on representative tasks.** Gray dashed lines indicate the baseline performance of the Base model for probing and generation.

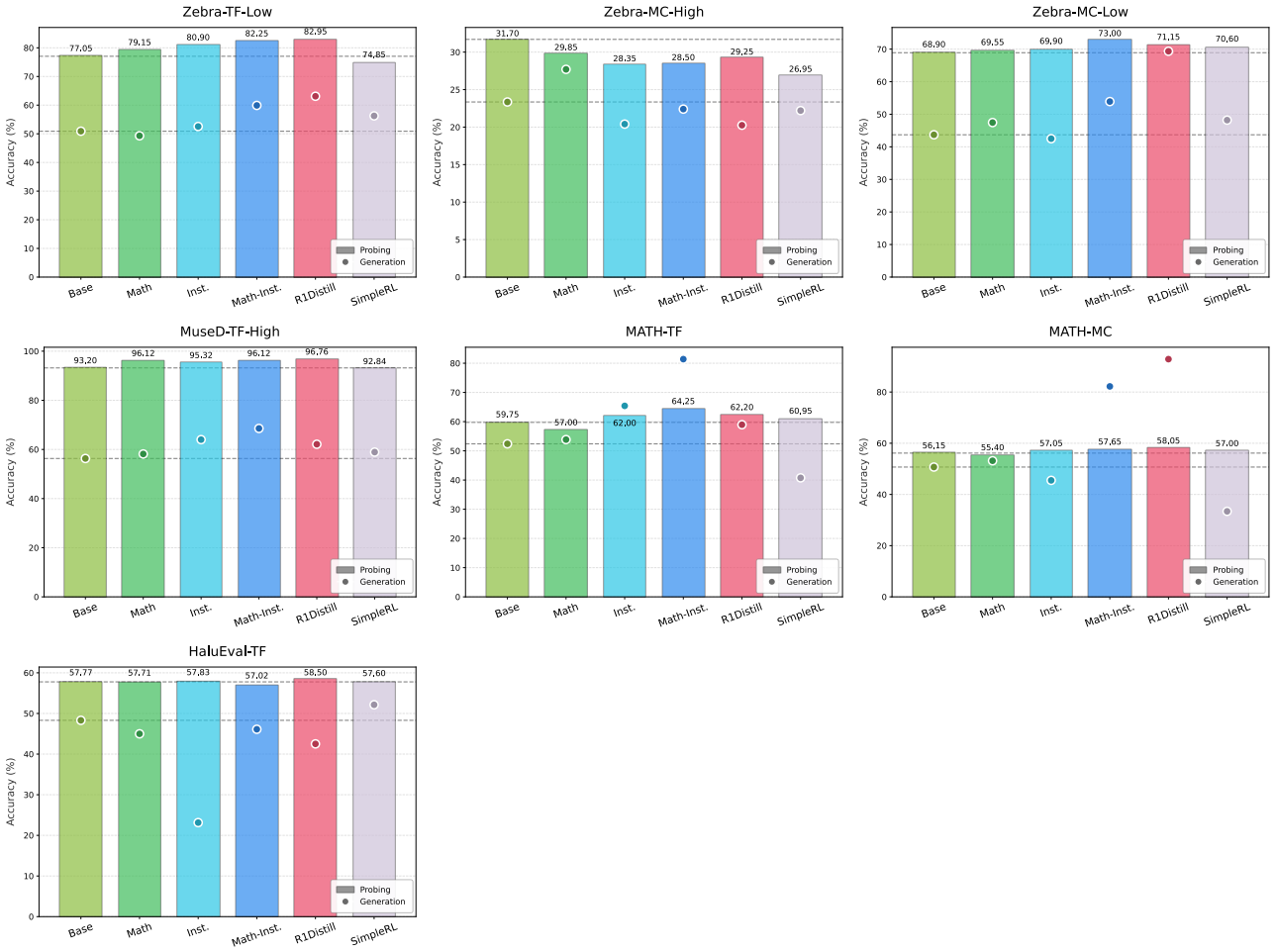

Figure 18. **Qwen2.5-1.5B results: Development of initial representation quality and generation accuracy on representative tasks.** Gray dashed lines indicate the baseline performance of the Base model for probing and generation.

with maximum generation gains, the maximum improvement in probing is also extremely low, especially for Llama3.1-8B series. Nevertheless, probing accuracy exceeds generation accuracy in the majority of cases, except for highly optimized reasoning models. Finally, consistent with our main results, reasoning models trained via distillation consistently outperform the pure RL variants in both representation quality and generation accuracy.

E.2. Representation Quality Dynamics during Generation

For this experiment, we select one high-difficulty subset for each task, as well as specific sub-tasks where the model's generation accuracy exceeds its initial probing accuracy. The complete results for the Qwen2.5-7B series are presented in Figure 19.

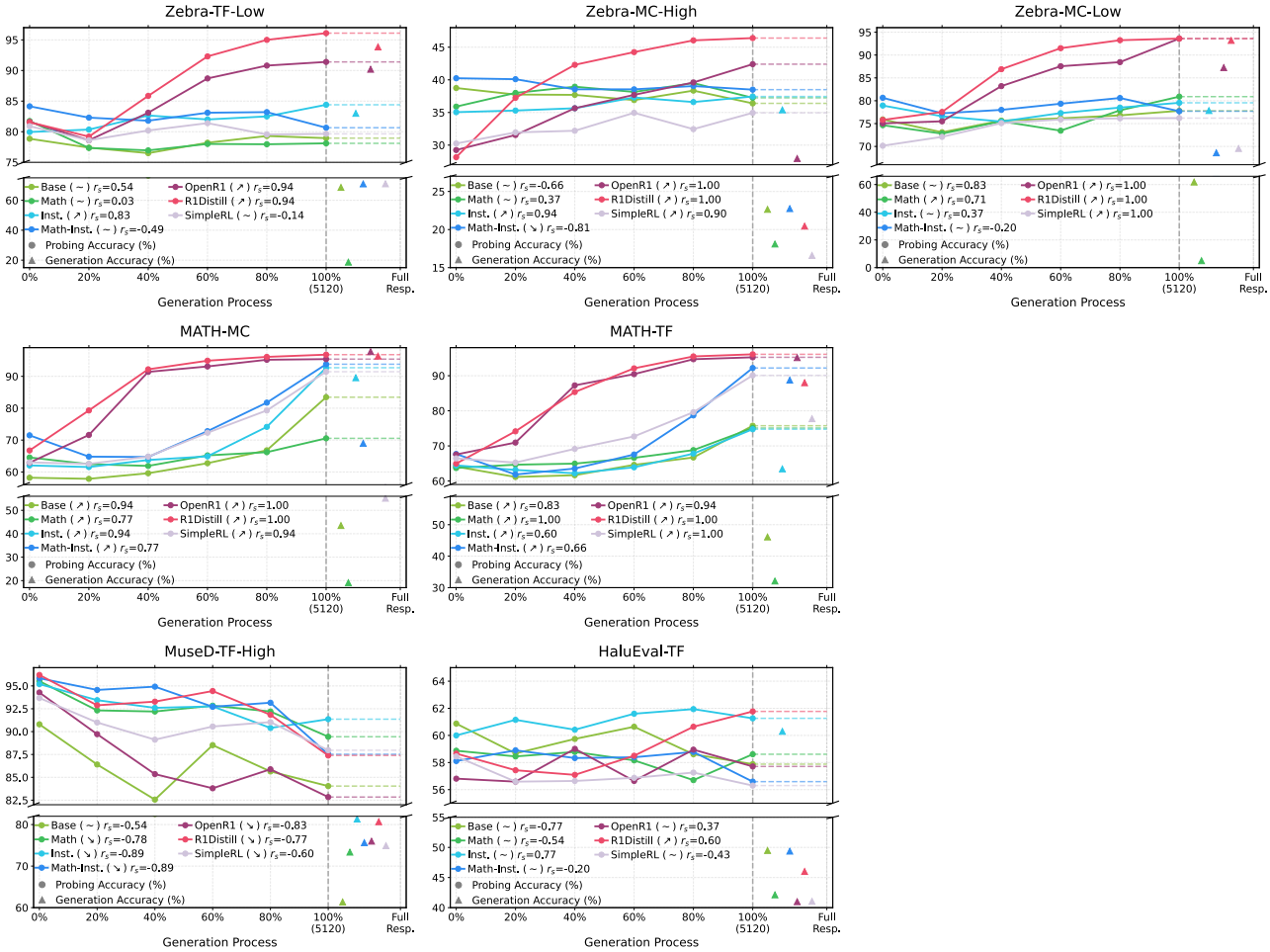

Figure 19. **Qwen2.5-7B results: Representation quality dynamics during generation on all selected tasks.** Trends are analyzed using linear regression and Spearman rank correlation.

With the exception of the factual task (HaluEval) and the task where the model already possesses substantial initial representation quality (MuseD), strong reasoning models (OpenR1 and R1Distill) demonstrate a consistent, significant upward trend in representation quality during generation, ultimately achieving the highest values among all models. While non-reasoning models also show gradual improvement via CoT, the gain is markedly smaller compared to reasoning models.

We select three typical tasks where the Inst. model exhibits downward, fluctuating, and rising trends (MuseD-TF-High, Zebra-MC-Low, and MATH-MC, respectively), and replicate the experiments on the Llama3.1-8B series and Qwen2.5-1.5B series. As shown in Figure 20 and Figure 21, the observed trends remain consistent across different model series. Notably, R1Distill demonstrates the most significant upward trend and achieves the highest final representation quality on the latter

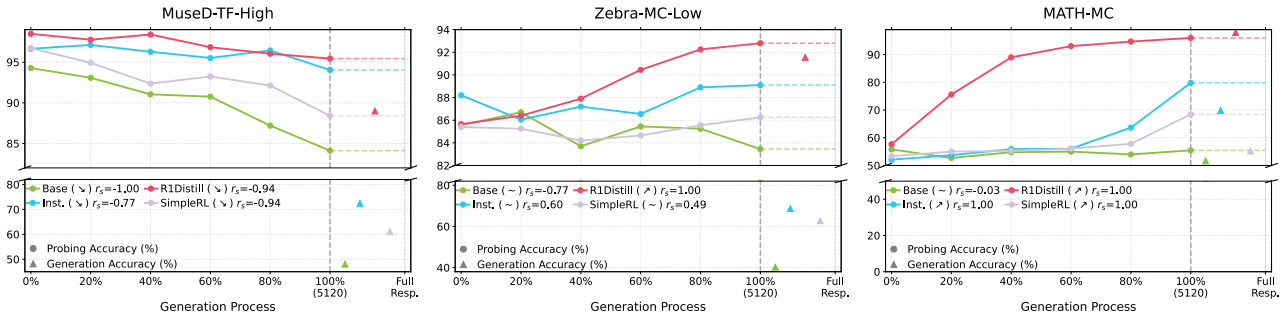

Figure 20. **Llama3.1-8B results: Representation quality dynamics during generation on representative tasks.** Trends are analyzed using linear regression and Spearman rank correlation.

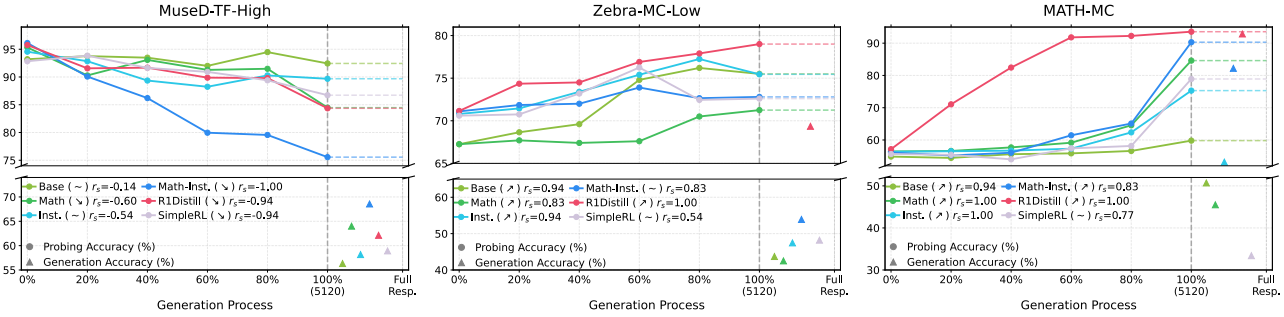

Figure 21. **Qwen2.5-1.5B results: Representation quality dynamics during generation on representative tasks.** Trends are analyzed using linear regression and Spearman rank correlation.

two tasks.

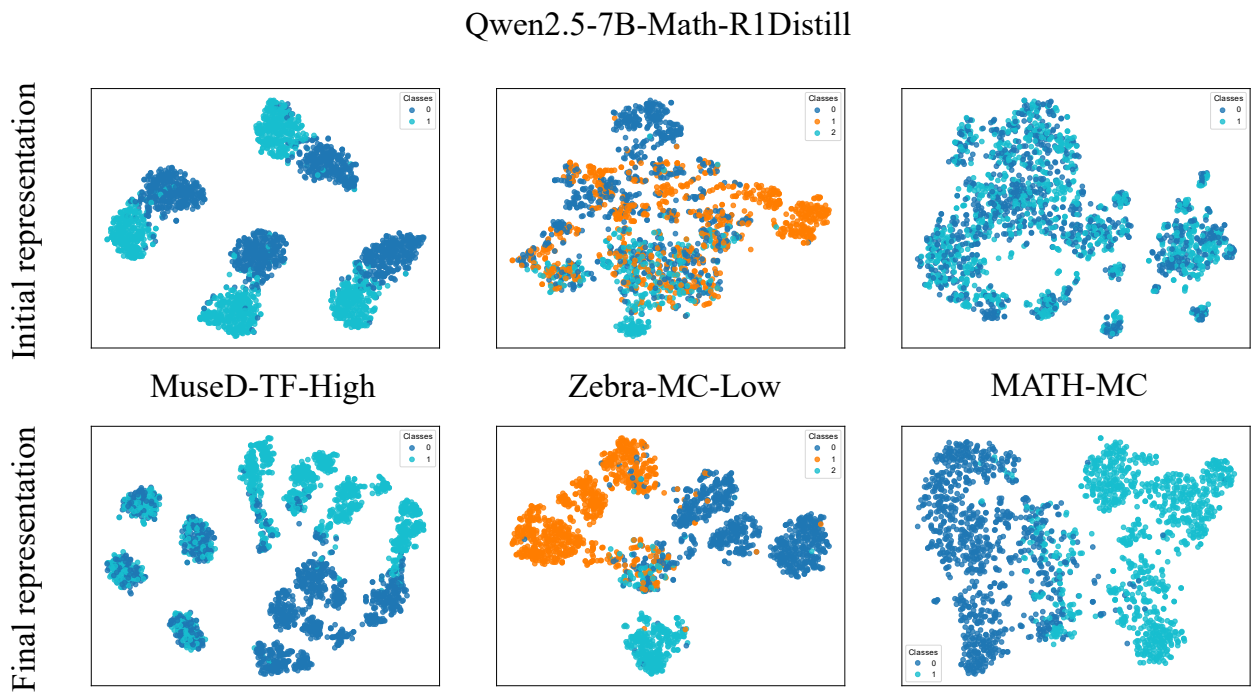

Figure 22. **Visualization of Representation Shifts.** Comparison of the R1Distill model's representation distributions before (Initial) and after (Final) the CoT across three tasks.

To intuitively visualize how the model's internal representations change during generation, we compare the initial and final representation distributions of R1Distill using t-SNE (Maaten & Hinton, 2008). As shown in Figure 22, for MuseD, the initial representations already exhibit strong class separability, and the separation diminishes after CoT, indicating that the reasoning process can negatively impact internal judgment. In contrast, for Zebra and MATH, where initial representations show significant overlap between classes, the generation process yields a marked improvement in cluster separation, confirming that reasoning effectively enables the model to distinguish between different labels.

Finally, we compare explicit generation accuracy with latent probing accuracy, as shown in Figure 23, Figure 24, and Figure 25. Across all models and tasks, generation accuracy consistently fails to exceed or match probing accuracy, even though probing is restricted to the first 5,120 tokens rather than the full response. This indicates that the model's latent potential remains underutilized. Post-training generally enhances the model's ability to realize this potential, as evidenced by steeper slopes compared to pre-trained models.

E.3. Statistical Analysis of Alignment

In this section, we provide the full statistical results for all reasoning tasks (High-difficulty level) in Table 8 and Table 9. Additionally, we visualize the relationship between the initial and final probing probabilities p in Figure 26 and Figure 27. When calculating the trend based on linear regression and the Spearman rank correlation coefficient r_s, we exclude buckets containing fewer than 10 data samples to ensure statistical robustness.

Consistent with the results in section 5.1, generation correctness exhibits minimal correlation with the initial representation but significantly higher correlation with the final representation. Notably, R1Distill generally shows weaker alignment than Inst. Furthermore, the correlation between initial and final representations is low for both models, with R1Distill demonstrating an even lower degree of alignment.

The visualizations in Figure 26 and Figure 27 further demonstrate the drastic changes occurring between the initial and final representations.

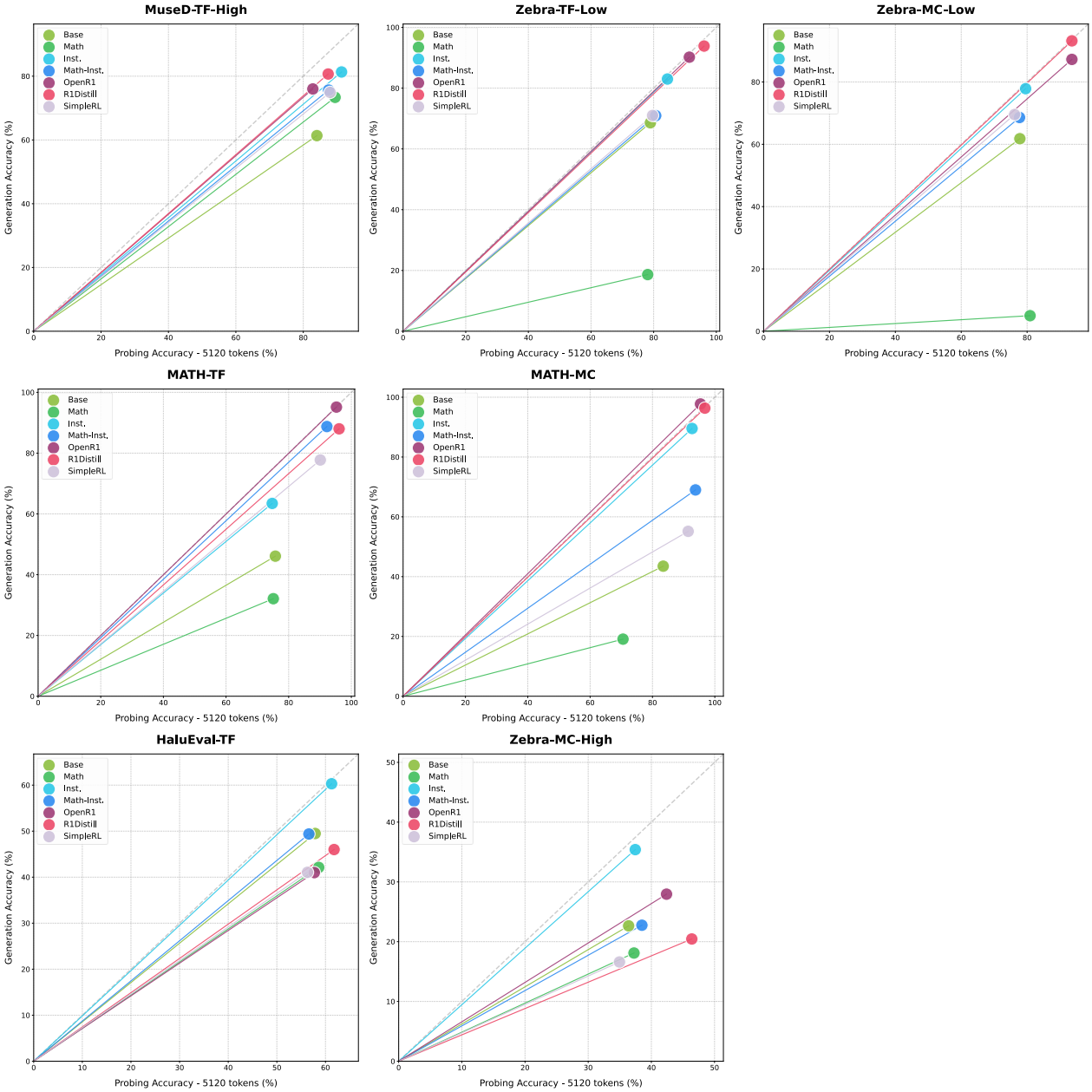

Figure 23. **Qwen2.5-7B results: Generation accuracy versus probing accuracy on all selected tasks.**

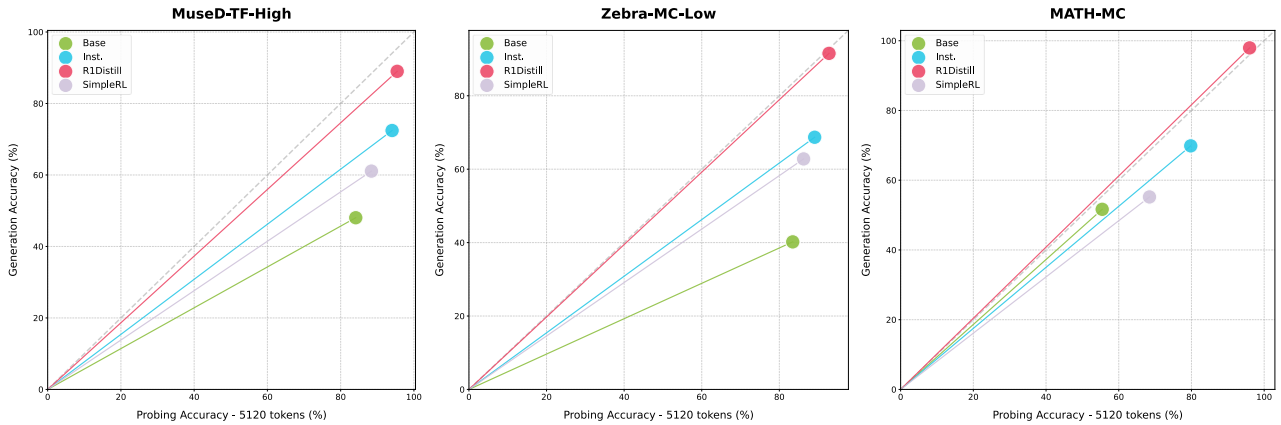

Figure 24. **Llama3.1-8B results: Generation accuracy versus probing accuracy on representative tasks.**

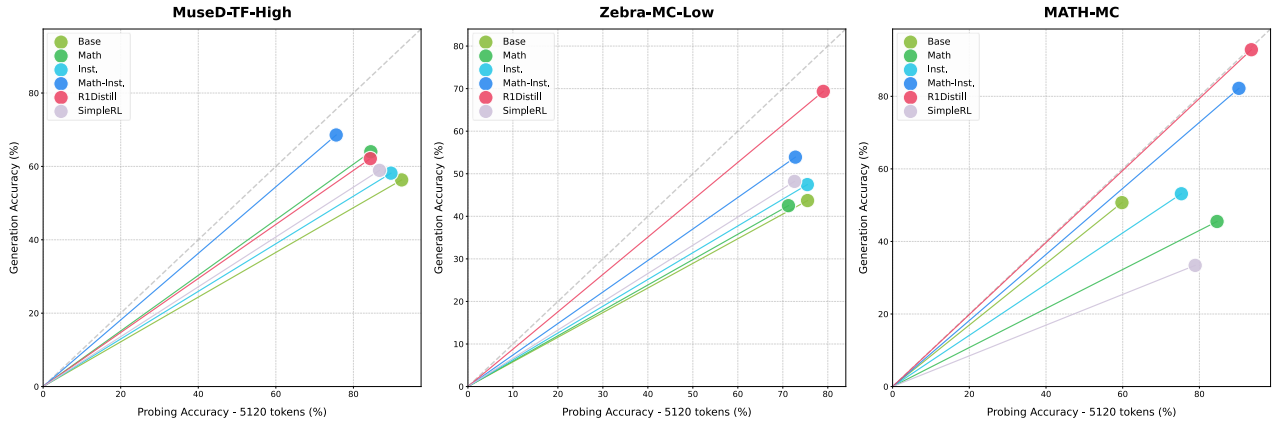

Figure 25. **Qwen2.5-1.5B results: Generation accuracy versus probing accuracy on representative tasks.**

Table 8. **Generation-representation alignment.** Significance levels P_{MWU} of Mann-Whitney U test: NS denotes non-significant ($> 5e^{-2}$), * ($< 5e^{-2}$), and *** ($< 1e^{-10}$). **Bold**: higher alignment with CoT; Blue: R1Distill alignment is lower than Inst.

	Inst.				R1Distill			
High-Difficulty Task	Trend	$r_s\uparrow$	ROC-AUC↑	P_{MWU}	Trend	$r_s\uparrow$	ROC-AUC↑	P_{MWU}
MusED-TF	↗	0.59	0.57	*	∼	0.07	0.53	NS
MusED-TF+CoT	↗	**0.99**	**0.85**	***	↗	**0.95**	**0.88**	***
MusED-MC	∼	0.49	0.57	*	∼	0.28	0.57	*
MusED-MC+CoT	↗	**0.90**	**0.78**	***	↗	**0.96**	**0.78**	***
Zebra-TF	↗	0.88	0.55	*	∼	0.07	0.51	NS
Zebra-TF+CoT	↗	**1.00**	**0.86**	***	↗	**0.76**	**0.62**	***
Zebra-MC	↗	1.00	0.61	***	↗	1.00	0.60	*
Zebra-MC+CoT	↗	0.99	**0.94**	***	↗	0.98	**0.76**	***
HaluEval-TF	∼	0.45	0.56	*	∼	0.21	0.50	NS
MATH-TF	↗	0.88	0.59	*	∼	0.45	0.57	*
MATH-TF+CoT	↗	**0.98**	**0.70**	***	↗	**0.52**	**0.90**	***
MATH-MC	∼	-0.16	0.50	NS	↘	-0.60	0.46	NS
MATH-MC+CoT	↗	**0.95**	**0.95**	***	∼	**0.43**	**0.85**	***

Table 9. **Alignment between representations before and after CoT**. r_s: Spearman rank correlation coefficient; r_p: Pearson correlation coefficient; r^2 is adopted as the primary metric of linear regression, as p-values tend to become uninformative in large datasets. Blue: R1Distill alignment is lower than Inst.

	Inst.			R1Distill		
	$r_s\uparrow$	$r_p\uparrow$	$r^2\uparrow$	$r_s\uparrow$	$r_p\uparrow$	$r^2\uparrow$
MusED-TF-High	0.16	0.13	0.02	0.10	0.05	0.00
MusED-MC-High	0.36	0.30	0.09	0.25	0.21	0.04
Zebra-TF-High	0.30	0.30	0.09	0.29	0.29	0.08
Zebra-MC-High	0.23	0.32	0.10	0.18	0.15	0.02
MATH-TF	0.28	0.29	0.08	0.09	0.05	0.00
MATH-MC	0.13	0.07	0.01	0.08	0.01	0.00

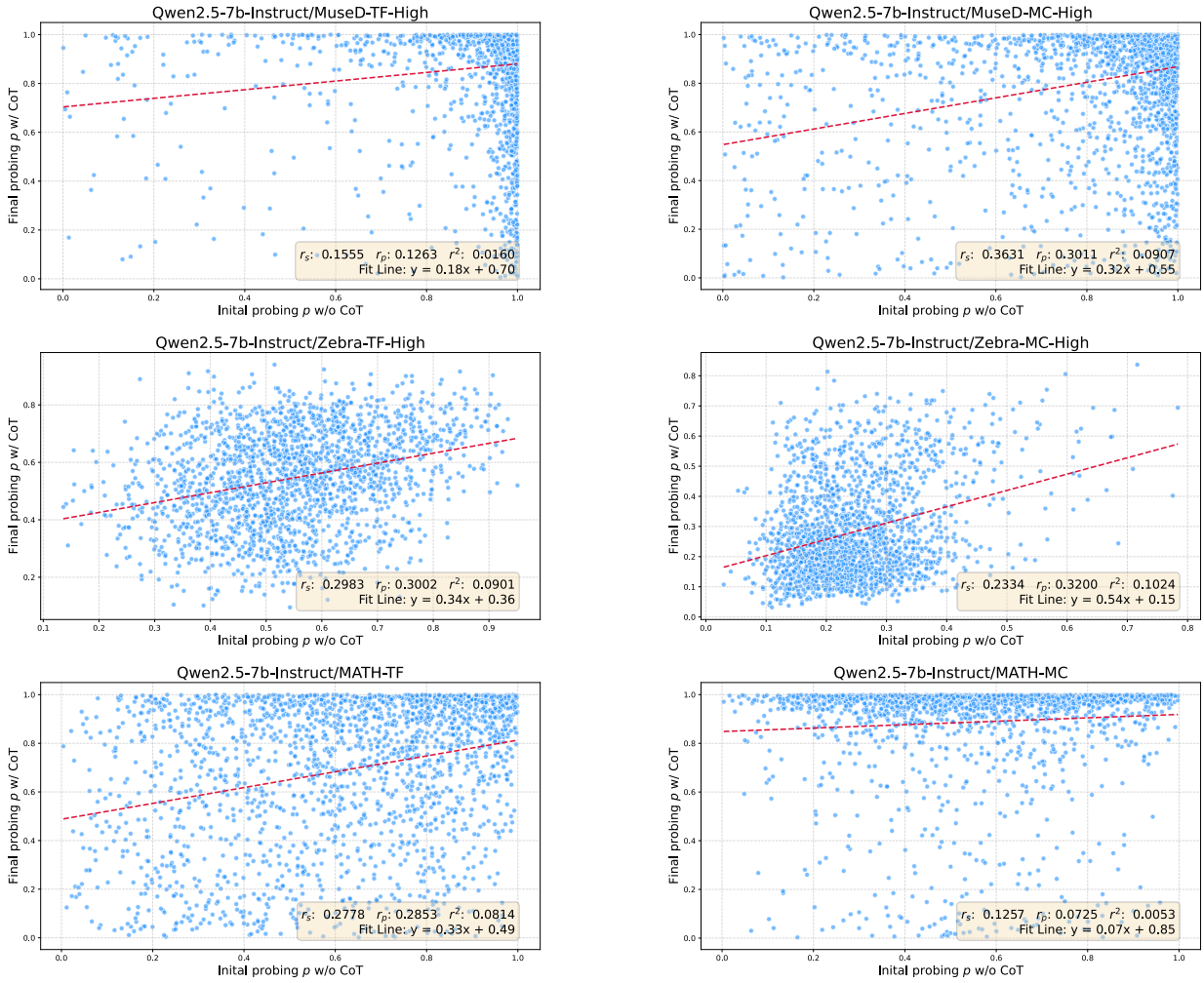

Figure 26. **Qwen2.5-7B-Instruct: Visualization of the relationship between initial and final probing probabilities p.**

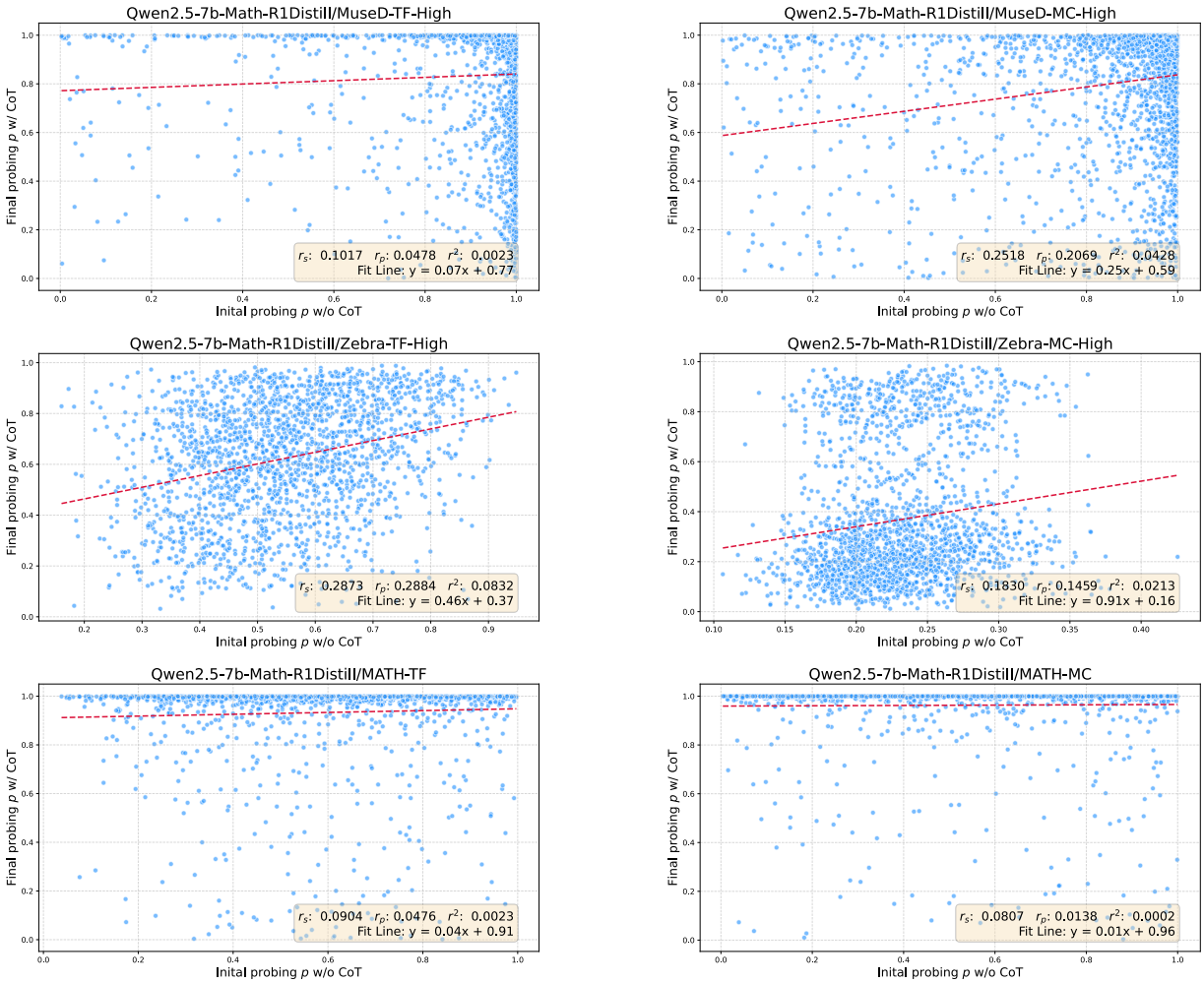

Figure 27. **Qwen2.5-7B-Math-R1Distill: Visualization of the relationship between initial and final probing probabilities p.**

E.4. *Generation Accuracy* Dynamics during Generation

While previous experiments have established that representation quality evolves alongside the CoT process, this quality reflects the model's *potential* to generate the correct answer, rather than guaranteeing its successful elicitation, as we also demonstrate in Figure 23. Therefore, we further investigate how the progression of CoT influences the actual probability of the model outputting the correct answer. The *generation accuracy* evaluated in this section differs from its definition in the main text. Specifically, we force the model to yield a final answer by appending a strict answering trigger, "Therefore, the answer is \\boxed{", to a truncated CoT prefix. This forced generation accuracy is not strictly equivalent to that in natural generation scenarios: During natural decoding, a model might remain in a low-confidence state on complex problems, failing to produce an answer within limited inference steps or falling into repetitive reasoning loops. In contrast, the appended trigger effectively compels the model to output a definitive answer.

Specifically, aligning with the settings of representation quality dynamics experiments, we segment the generated CoT (truncated to a maximum of 5,120 tokens) into discrete progressive stages. At each stage, we append the aforementioned trigger and allow the model to continue generating. We then employ the same extraction function $Ext()$ to parse the response and verify its correctness against the ground truth. Experiments are conducted on three representative tasks (Zebra-MC-Low, MuseD-TF-High, and MATH-MC) using two representative models: Inst. and R1Distill for non-reasoning model and strong reasoning model, respectively. The results are presented in Figure 28.

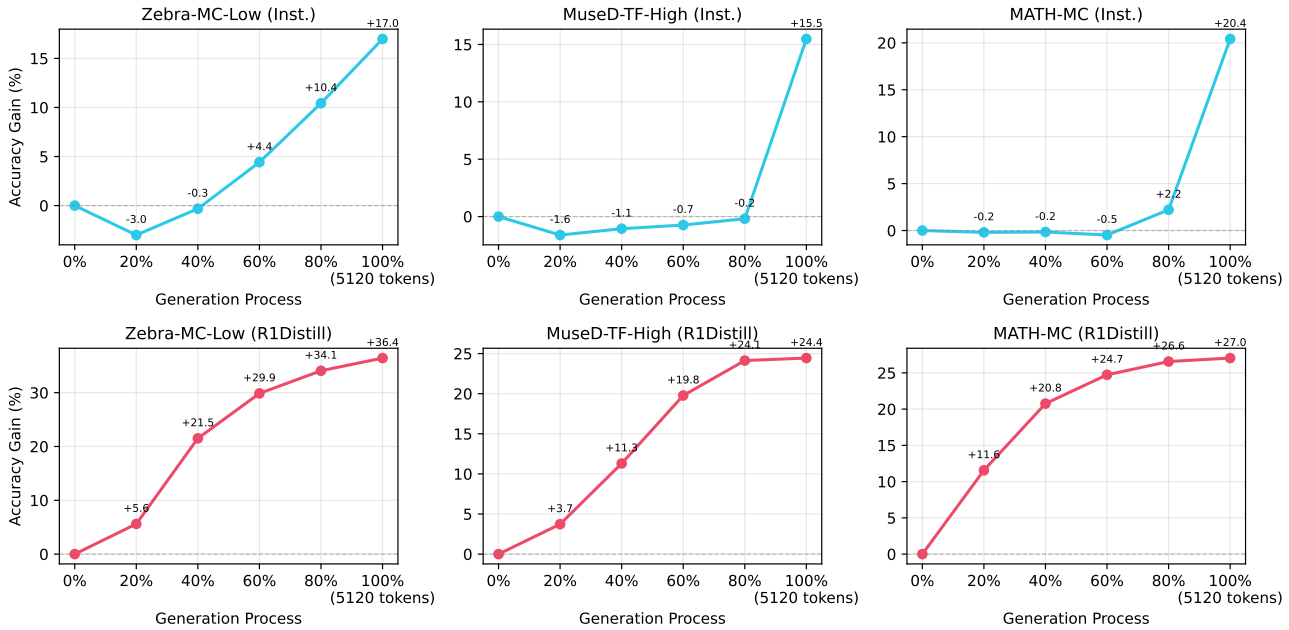

Figure 28. ***Generation accuracy*** **dynamics during generation**. We report the gains compared to the 0% baseline.

The results indicate a general upward trend in *generation accuracy* as the CoT progresses across all three tasks for both models. However, the transition patterns of the accuracy diverge significantly, highlighting a fundamental difference in the reasoning paradigms between reasoning and non-reasoning models. For R1Distill, the accuracy surges rapidly during the early stages of CoT and subsequently plateaus, forming a concave curve. Conversely, the Inst. model exhibits a convex curve, where substantial accuracy gains only materialize in the final stages of generation. This suggests that the strong reasoning model engages in a highly progressive and effective deduction process, rapidly accumulating confidence and converging towards the correct answer early on. In contrast, the intermediate "thoughts" of the non-reasoning model are less effective at accumulating early certainty and lead to less improvement in *generation accuracy*. Notably, these contrasting concave and convex transitions largely align with the dynamics observed in their representation qualities.

Nevertheless, the *generation accuracy* dynamics on the MuseD task reveal that the theoretical potential encapsulated in representation quality does not always strictly align with actual elicited generation performance. While previous analyses show that the representation quality for MuseD slightly degrades as CoT progresses, its forced *generation accuracy* still

exhibits an upward trend. This discrepancy implies a trade-off: Although the linear separability of the internal representations might slightly diminish, the generated intermediate tokens provide a more explicit contextual foundation. This rich textual context ultimately compensates for the latent representation degradation, translating into enhanced expressive power and actual decoding accuracy during the final generation phase.