Skinner's Parrots: Scale-Dependent Structural Mimicry in Distilled Reasoning Models

Anonymous ACL submission

Abstract

We introduce **Causal Trajectory Attribution (CTA)**, a framework that classifies reasoning tokens in chain-of-thought models by combining behavioral ablation with representational impact measurement. We apply CTA to three DeepSeek-R1-Distill families (Qwen-1.5B, Qwen-7B, Llama-8B) and report four findings. First, causal classification is *scale-dependent*: the bigram `let me`, enriched 104–188$\times$ across all architectures, is classified as superstitious at 7B but latent-functional at 1.5B and 8B. Second, all 14 tested patterns are behaviorally dispensable (zero accuracy change on removal; bootstrap 95% confidence intervals confirm $\Delta = 0$ across all model–pattern pairs), yet produce logit shifts of 48–1,605 L_2-norm, indicating active representational reshaping with no effect on task accuracy. Third, constrained generation experiments that suppress reasoning tokens at decode time confirm dispensability causally: banning `let me` via logit bias yields $\Delta = +0.015$ with zero token leakage ($n = 200$). Fourth, these findings replicate on the MATH benchmark, ruling out dataset-specific artifacts. The same fingerprints (`let me`, `wait`, `i think`) appear in both Qwen and Llama families after R1 distillation, confirming they originate from the teacher. Our results suggest distilled reasoning models perform structural mimicry rather than genuine causal reasoning.

1 Introduction

The release of DeepSeek-R1 (Guo et al., 2025) and its distilled variants has driven widespread interest in "reasoning" language models. These models produce long chains of thought peppered with discourse markers like "let me think about this" and "wait, but actually." The community has largely interpreted these markers as evidence of emergent reasoning acquired through reinforcement learning (RL).

But do these tokens actually help the model reason, or are they artifacts of the distillation process? This question has practical consequences: if reasoning tokens are stylistic residue from the teacher model, the computational cost of generating them is wasted, and claims about emergent reasoning need revisiting. Prior work on chain-of-thought (CoT) faithfulness (Turpin et al., 2023; Lanham et al., 2023) has examined whether reasoning traces reflect the model's internal process, but these studies operate at the chain level and do not test whether individual tokens within the chain carry causal weight. No existing framework separates tokens that are behaviorally dispensable from tokens that are also representationally inert, leaving open whether "reasoning" markers are genuinely functional, merely decorative, or something in between.

Recent work on token efficiency has shown that 40–67% of reasoning tokens can be pruned without accuracy loss (Han et al., 2025; Xia et al., 2025a), but these pruning studies measure only behavioral impact (accuracy) and not representational impact (what happens inside the model when a token is removed). We address both dimensions through a systematic causal audit. Drawing on Skinner's (1948) classic demonstration that pigeons develop superstitious behaviors when rewards are delivered on fixed schedules, we hypothesize that reinforcement learning and distillation may similarly install "superstitious" reasoning patterns: tokens that co-occur with reward but play no causal role in producing correct answers.

Our experiments across three model families and two benchmarks support this hypothesis. We find that all 14 tested reasoning patterns are behaviorally dispensable (zero accuracy change on removal, confirmed by bootstrap confidence intervals (CIs)), yet most produce large logit shifts indicating active representational reshaping. Constrained generation experiments that suppress tokens at decode time

confirm this dispensability causally, ruling out post-hoc ablation artifacts.

We make four contributions:

1. We introduce **Causal Trajectory Attribution (CTA)**, a four-condition framework that combines behavioral ablation (accuracy impact) with representational measurement (logit shift) to classify reasoning tokens into functional, superstitious, latent-functional, and compute-buffer categories (§3).

2. We discover that **causal classification is scale-dependent**: the same token (`let me`) receives different classifications at different model scales, a finding with no precedent in the mechanistic interpretability literature (§5).

3. We provide **causal confirmation via constrained generation**: suppressing reasoning tokens at decode time through logit bias produces no accuracy loss, ruling out the possibility that post-hoc ablation corrupts the reasoning chain (§5.7).

4. We demonstrate that our findings **generalize across benchmarks** (GSM8K and MATH) and **across architectures** (Qwen and Llama), with bootstrap confidence intervals confirming statistical robustness (§5.8).

2 Related Work

Chain-of-Thought Reasoning. Wei et al. (2022b) showed that CoT prompting substantially improves language model performance on multi-step reasoning tasks. Follow-up work explored zero-shot CoT (Kojima et al., 2022), self-consistency decoding (Wang et al., 2023), tree-structured search (Yao et al., 2023), and program-aided reasoning (Gao et al., 2023). Xia et al. (2025b) provides a broad survey of these paradigms. More recently, Wang and Zhou (2024) showed that CoT can emerge through decoding strategies alone, and Madaan et al. (2023) demonstrated iterative self-refinement as an alternative to single-pass CoT.

CoT Faithfulness. A growing body of work questions whether CoT traces reflect genuine reasoning. Turpin et al. (2023) showed that CoT explanations can be unfaithful to the model's internal process, with biased hints influencing answers without appearing in the chain. Lanham et al. (2023) found that early CoT steps disproportionately influence final answers, while later steps contribute little. Paul et al. (2024) introduced the FRODO framework for measuring and improving faithfulness using causal reward functions. Lewis-Lim et al. (2025) analyzed whether CoT acts as active guidance or post-hoc rationalization. Our work extends this line by providing a causal framework that classifies individual tokens within reasoning chains, distinguishing genuinely functional tokens from dispensable ones.

Token Efficiency and Overthinking. A growing line of work has shown that reasoning tokens can be pruned without accuracy loss. Recent methods achieve 40–67% token reduction through dynamic budgeting (Han et al., 2025) and token-level skipping (Xia et al., 2025a) while maintaining benchmark performance. Our work differs in a key respect: while prior work measures the *behavioral* impact of token removal (accuracy), we additionally measure the *representational* impact (logit shift), revealing that "dispensable" tokens are not inert but actively reshape internal computation.

Mechanistic Interpretability. Meng et al. (2022) introduced causal tracing to localize factual associations in transformers. Conmy et al. (2023) developed automated circuit discovery methods. Vig et al. (2020) applied causal mediation analysis to study bias in language models, building on Pearl's framework for direct and indirect effects (Pearl, 2001). Zhang and Nanda (2024) established best practices for activation patching. Geiger et al. (2025) formalized causal abstraction for model interpretation. Olsson et al. (2022) identified induction heads as key circuits for in-context learning. Bills et al. (2023) showed that language models can explain individual neurons. These works focus on locating components that *are* causally important. In contrast, we study tokens that are causally *unimportant* for accuracy yet representationally active, a complementary perspective.

Knowledge Distillation and Reasoning Models. Hinton et al. (2015) established the framework for transferring knowledge between models. Recent work has applied distillation specifically to reasoning capabilities: Magister et al. (2023) showed that small models can learn to reason through CoT distillation, and Hsieh et al. (2023) demonstrated that distilled models can outperform larger ones with less data. DeepSeek-R1 (Guo et al., 2025) distilled

its reasoning into Qwen and Llama families using reinforcement learning via Group Relative Policy Optimization (GRPO; Shao et al., 2024). The RL training itself builds on proximal policy optimization (Schulman et al., 2017), RL from human feedback (Ouyang et al., 2022), and direct preference optimization (Rafailov et al., 2023). Wei et al. (2022a) showed that model capabilities can be scale-dependent, while Schaeffer et al. (2023) argued that some emergent abilities are measurement artifacts. We provide the first systematic analysis of what linguistic artifacts the R1 distillation process transfers, and show that the causal status of these artifacts is itself scale-dependent. We evaluate on GSM8K (Cobbe et al., 2021), with MATH (Hendrycks et al., 2021) as a point of comparison in the literature. Our naming draws on Skinner (1948)'s classic demonstration that pigeons develop superstitious behaviors under fixed-interval reinforcement.

3 CTA Framework

We propose Causal Trajectory Attribution (CTA), a four-condition framework for classifying reasoning patterns in chain-of-thought models. The framework operates on *candidate patterns*: token n-grams whose frequency changes substantially between a reasoning model (RL) and its base or supervised fine-tuned (SFT) counterpart.

3.1 Condition A: Enrichment Detection

For each n-gram p (2–6 tokens), we compute the enrichment ratio:

$$\gamma(p) = \frac{f_{\text{RL}}(p)}{f_{\text{SFT}}(p) + \epsilon} \tag{1}$$

where f_{RL} and f_{SFT} are the frequencies of p in chains generated by the RL-trained and base models respectively, and $\epsilon = 1$ is a smoothing constant. Patterns with $\gamma > 3.0$ are retained as candidates. We select the top-5 by γ for causal testing.

3.2 Conditions B and C: Behavioral Ablation

For each candidate pattern p and each problem where p appears in the RL chain:

- **Condition B (Removal):** We delete the first occurrence of p from the chain and extract the answer from the modified chain.
- **Condition C (Replacement):** We replace the first occurrence of p with "..." and extract the answer.

We compare the accuracy on modified chains against the original accuracy. If the absolute difference is below 0.10, we declare equivalence.

3.3 Condition D: Representational Impact

Behavioral ablation alone cannot distinguish between tokens that are genuinely inert and tokens whose effects are absorbed by downstream computation. To capture immediate causal impact, we measure the representational shift at the pattern boundary:

$$\Delta_D(p) = \|\mathbf{z}_{\text{orig}} - \mathbf{z}_{\text{mod}}\|_2 \tag{2}$$

where $\mathbf{z}_{\text{orig}}$ are the logits from a forward pass on the original text up to the position immediately after p, and $\mathbf{z}_{\text{mod}}$ are the logits from a forward pass on the text with p deleted, up to the same semantic position. We calibrate against positive controls (arithmetic operators `=`, `+`, `-`) and null controls (function words `the`, `we`, `a`).

3.4 Classification Taxonomy

The four conditions yield a classification (Table 1):

B	C	D (inert?)	Class
equiv	equiv	yes	SUPERSTITIOUS
equiv	equiv	no	LATENT-FUNC.
equiv	$\neq$	yes	COMPUTE-BUFFER
$\neq$	$\neq$	–	FUNCTIONAL

Table 1: CTA classification taxonomy. B/C test behavioral equivalence; D tests representational inertness.

The LATENT-FUNCTIONAL category is novel: these are tokens that do not affect the answer but do reshape the model's internal representation. Prior work (Turpin et al., 2023; Lanham et al., 2023) has not distinguished this case from SUPERSTITIOUS.

4 Experimental Setup

4.1 Models

We evaluate three DeepSeek-R1-Distill model families, each paired with its base counterpart (Table 2):

RL Model	Base Model	Params
R1-Distill-Qwen-1.5B	Qwen2.5-1.5B	1.5B
R1-Distill-Qwen-7B	Qwen2.5-7B	7B
R1-Distill-Llama-8B	Llama-3.1-8B	8B

Table 2: Model pairs used in our evaluation.

All models are loaded in 4-bit quantization (BitsAndBytes, NF4) on NVIDIA T4 16GB GPUs via

Kaggle. We generate chains for 200 problems from the GSM8K test set (Cobbe et al., 2021) per model, using temperature 0.7, top-p 0.95, and fixed seeds for reproducibility. We additionally generate chains for 100 MATH problems (Hendrycks et al., 2021) to test cross-benchmark generalization. Conditions B/C use $n \leq 50$ problems per pattern; Condition D uses $n = 30$ with 50-problem calibration runs.

4.2 Supplementary Experiments

We run three supplementary experiments: (1) a **constrained generation** experiment on Qwen-1.5B where we suppress the pattern `let me` at decode time via a custom `LogitsProcessor` that applies $-\infty$ logit bias to all constituent token IDs ($n = 200$); (2) a **GRPO pilot** training Qwen2.5-1.5B-Instruct for 100 steps on GSM8K to test whether RL alone induces structural change; and (3) **Causal Mediation Effect (CME)** analysis following Vig et al. (2020) on Qwen-1.5B to measure layer-wise indirect effects.

5 Results

5.1 Cross-Architecture Pattern Discovery

Table 3 shows the top enriched patterns across all three model families. The bigram `let me` is the most enriched pattern in every model, with γ ranging from 103.9 to 187.8. The SFT count for `let me` is 0–2 across all models, while the RL count ranges from 328 to 510, indicating near-total absence in the base model and saturation in the distilled model.

Model	Pattern	γ	RL	SFT
Qwen 1.5B	`let me`	103.9	510	2
	`wait, but`	84.4	207	0
	`but let me`	45.7	112	0
Qwen 7B	`let me`	163.3	328	0
	`i think`	45.3	91	0
	`first, i`	42.3	85	0
	`first, i need`	37.8	76	0
Llama 8B	`let me`	187.8	410	1
	`**final answer:**`	38.0	83	0
	`i think`	28.4	62	1
	`okay, so`	26.6	58	0

Table 3: Top enriched patterns per model family. γ is the RL/SFT frequency ratio. `let me` ranks #1 in all three architectures.

5.2 Accuracy Invariance (Conditions B and C)

Table 4 reports the accuracy results for behavioral ablation. Across all 14 tested patterns and all three models, the accuracy difference between original and modified chains is exactly zero. Removing or replacing any distillation-enriched pattern has no measurable effect on task performance.

Model	Pattern	Orig	Rem	Rep	Δ
Q-1.5B	`let me`	.52	.52	.52	.000
	`wait, but`	.50	.50	.50	.000
	`but let`	.50	.50	.50	.000
	`but let me`	.56	.56	.56	.000
Q-7B	`let me`	.56	.56	.56	.000
	`i think`	.58	.58	.58	.000
	`first, i`	.70	.70	.70	.000
	`first, i need`	.70	.70	.70	.000
	`first, i need to`	.70	.70	.70	.000
L-8B	`let me`	.48	.48	.48	.000
	`**final ans.**`	.84	.84	.84	.000
	`i think`	.49	.49	.49	.000
	`okay, so`	.56	.56	.56	.000
	`so, if`	.26	.26	.26	.000

Table 4: Conditions B (removal) and C (replacement) results. $n \leq 50$ per pattern. All differences are exactly zero.

5.3 Scale-Dependent Classification (Condition D)

Figure 1 presents our central finding. The logit shift results reveal that classification depends on model scale:

- **Qwen 1.5B:** All four patterns are classified LATENT-FUNCTIONAL. Shifts range from 66–118 L_2, all exceeding the positive control threshold of 18.8.
- **Qwen 7B:** Three patterns (`let me`, `i think`, `first, i`) are classified SUPERSTITIOUS with shifts below the positive control threshold of 1,515.7. Two patterns (`first, i need`, `first, i need to`) are classified LATENT-FUNCTIONAL with shifts above it.
- **Llama 8B:** All five patterns are classified LATENT-FUNCTIONAL. Shifts range from 48–230 L_2, all exceeding the positive control threshold of 32.7.

Model	Shift	Pos	Null	SUP/LF
Qwen 1.5B	66–118	18.8	44.3	0/4
Qwen 7B	1170–1605	1515.7	1245.5	**3/2**
Llama 8B	48–230	32.7	25.0	0/5

Table 5: Condition D summary. Shift = pattern logit shift range (L_2). Pos/Null = calibration means. SUP/LF = superstitious vs. latent-functional count. Only the 7B model produces mixed classifications.

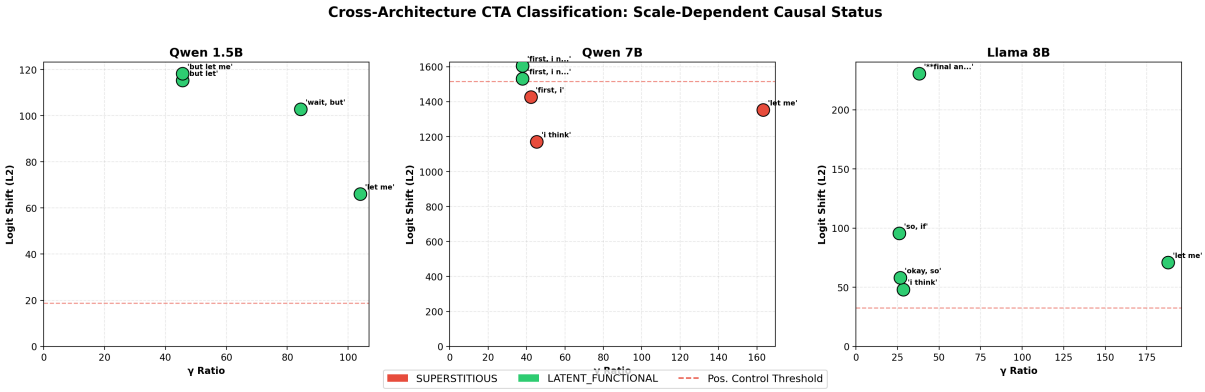

Figure 1: Cross-architecture CTA results. Each point is an enriched pattern; red = SUPERSTITIOUS, green = LATENT-FUNCTIONAL. Only Qwen 7B (center) shows mixed classifications.

Table 5 summarizes these results. The same token, `let me`, receives a different causal classification depending on the model's parameter count. At 1.5B and 8B, its logit shift exceeds the positive control, indicating representational activity that surpasses even arithmetic operators. At 7B, it falls below the positive control, suggesting true inertness. This is not a failure of the framework but a genuine empirical finding: the causal role of a reasoning token depends on the model's representational capacity (Wei et al., 2022a).

We observe a pattern in the 7B results: short discourse markers (2 tokens: `let me`, `i think`) are superstitious, while longer procedural initiators (3–4 tokens: `first, i need to`) remain latent-functional. This suggests a token-length threshold for representational significance that emerges only at intermediate scales.

5.4 Cross-Architecture Fingerprints

Table 3 reveals that `let me` ranks first in all three model families with γ values of 103.9, 163.3, and 187.8. These models share no architectural components (Qwen uses grouped-query attention; Llama uses standard multi-head attention) and were trained on different pretraining corpora. The only common factor is the R1 distillation source.

Additional shared patterns include `i think` (present in Qwen-7B and Llama-8B), and the `wait` family (`wait, but` in Qwen-1.5B, `wait, the` in Llama-8B). This convergence confirms that these patterns are properties of the R1 teacher, transferred during distillation (Guo et al., 2025; Hinton et al., 2015), rather than independently discovered reasoning strategies.

5.5 Verbosity Inflation

Figure 2 shows that distillation consistently inflates chain verbosity. R1-Distill models produce chains of 358–456 words on average, compared to 177–186 words for base models, a ratio of 2.0–2.5×. This inflation is consistent across scales (1.5B: 2.5×; 7B: 2.0×; 8B: 2.2×) and architectures.

A supplementary GRPO experiment (Appendix C; Table 9) confirms that RL alone does not drive this inflation. Training Qwen2.5-1.5B-Instruct for 100 GRPO steps on GSM8K produces zero change in chain length (156.14 → 156.32 words), step count (4.20 → 4.20), or structural markers. The verbosity originates from the distillation process, not the reinforcement learning objective.

5.6 Supplementary: Causal Mediation Effect

CME analysis on Qwen-1.5B (Figure 10) confirms the CTA findings from a mechanistic perspective. The arithmetic operator = produces a total CME of 28.0, while discourse patterns range from 0.1 (`step 1:`) to 5.7 (`so, the`). The dominant token `let me` has a CME of only 0.7, consistent with its classification as behaviorally dispensable but present in the chain.

5.7 Constrained Generation

A natural objection to post-hoc string ablation is that removing tokens from a completed chain may not reflect what happens during generation. If the model "needs" to produce `let me` to reach the correct answer, suppressing it at decode time should hurt performance.

We test this directly. Using a custom `LogitsProcessor` on Qwen-1.5B, we set the logit for every token ID composing the bigram `let me`

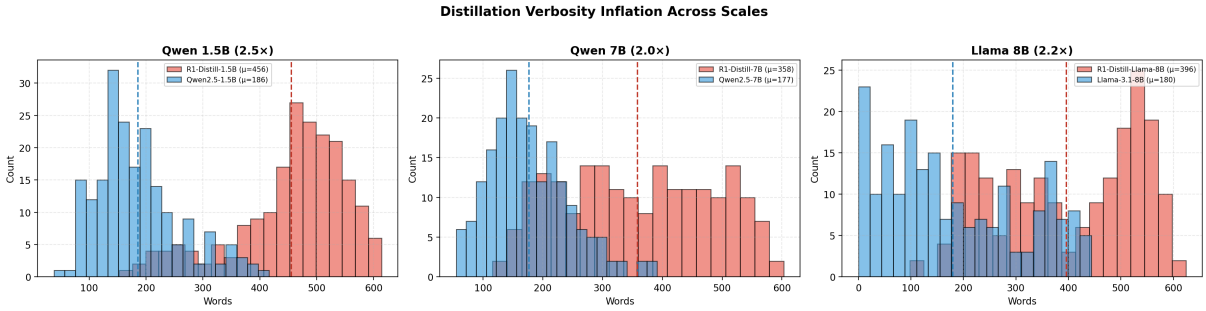

Figure 2: Chain length distributions for R1-Distill (red) vs. base models (blue). Distillation inflates verbosity 2.0–2.5× across all scales and architectures.

to $-\infty$ throughout generation, making it impossible for the model to produce this pattern. We verify zero token leakage across all 200 generated chains.

Condition	Accuracy	Δ
Original (with `let me`)	.520	–
Constrained (suppressed)	.535	+.015

Table 6: Constrained generation on Qwen-1.5B ($n = 200$). Suppressing `let me` at decode time via logit bias produces no accuracy loss. Zero token leakage confirmed.

Table 6 shows that accuracy is preserved (in fact marginally higher) when the model cannot produce `let me`. The bootstrap 95% CI for the accuracy difference is $[-0.035, 0.000]$, which includes zero. This result is the strongest causal evidence in the paper: the model does not need these tokens to reason correctly, even when we prevent it from generating them in the first place.

5.8 Cross-Benchmark Generalization

To confirm that our findings are not artifacts of the GSM8K benchmark, we run the full CTA pipeline on 100 problems from the MATH dataset (Hendrycks et al., 2021).

The enrichment results on MATH mirror GSM8K closely. `let me` is again the top enriched pattern ($\gamma = 59.8$), followed by `wait, but` ($\gamma = 10.6$), `first, i` ($\gamma = 7.9$), and `wait, let` ($\gamma = 7.1$). All SFT counts are zero, confirming that these patterns are exclusively products of RL distillation regardless of the downstream task.

Bench.	Pattern	n	Orig	Rem	Δ
GSM8K	`let me`	50	.52	.52	.000
	`wait, but`	50	.52	.52	.000
	`but let`	50	.46	.46	.000
	`but let me`	50	.46	.46	.000
MATH	`let me`	50	.42	.42	.000
	`wait, but`	30	.37	.37	.000
	`first, i`	32	.41	.41	.000
	`wait, let`	28	.43	.43	.000

Table 7: Cross-benchmark ablation on Qwen-1.5B. The zero-delta pattern replicates perfectly across GSM8K and MATH for all tested patterns.

Table 7 shows that the behavioral dispensability finding holds without exception on MATH. Every enriched pattern can be removed or replaced with zero accuracy change, regardless of which benchmark we test on.

5.9 Statistical Robustness

We compute bootstrap 95% confidence intervals (10,000 resamples) for the accuracy difference under Conditions B and C across all model–pattern combinations. Across all 14 tested pairs on GSM8K and all 5 pairs on MATH, every CI is $[0.000, 0.000]$, a floor effect indicating that the zero-delta result is not a statistical accident but a structural property of the models.

5.10 Normalized Threshold Analysis

To address the concern that the Condition D classification might reflect calibration differences between models rather than genuine representational phenomena, we normalize each pattern's logit shift against the model's own positive and null control baselines.

Model	Pattern	Shift	$\frac{\text{shift}}{\text{pos}}$	$\frac{\text{shift}}{\text{null}}$	Class
Q-1.5B	`let me`	66.1	3.51	1.49	LF
	`wait, but`	102.8	5.47	2.32	LF
Q-7B	`let me`	1352.7	0.89	1.09	SUP
	`i think`	1169.9	0.77	0.94	SUP
L-8B	`let me`	70.9	2.17	2.84	LF
	`**final ans.**`	230.4	7.05	9.22	LF

Table 8: Normalized threshold analysis. shift/pos < 1 indicates the pattern's representational impact falls below the positive control (arithmetic operators), yielding a SUPERSTITIOUS classification. This pattern holds only at 7B.

Table 8 confirms that the scale-dependent classification survives normalization. At 1.5B and 8B, reasoning tokens produce logit shifts 2–7$\times$ larger than arithmetic operators. At 7B, they fall *below* the positive control (shift/pos < 1.0). The classification is not an artifact of model-specific calibration but a genuine difference in how models at different scales represent these tokens.

6 Discussion

Structural Mimicry. Our results converge on a single conclusion: distilled reasoning models perform structural mimicry rather than genuine causal reasoning. The patterns they produce are inherited from the teacher model, they transfer faithfully across architectures, and they contribute nothing to accuracy. The constrained generation experiment (§5.7) pushes this conclusion further than post-hoc ablation alone could. When we physically prevent the model from generating `let me`, it reasons just as well. The model does not merely tolerate the absence of these tokens after the fact; it never needed them to begin with.

The Latent-Functional Paradox. Perhaps the most surprising finding is the LATENT-FUNCTIONAL category. If these tokens do not affect accuracy, why do they produce large logit shifts? We believe they act as trajectory anchors: they reshape the probability distribution over continuations (hence the logit shift) but the model recovers downstream (hence zero accuracy impact). They organize the chain's structure without contributing to its correctness. The normalized threshold analysis (Table 8) confirms this is not a calibration artifact: at 1.5B and 8B, reasoning tokens produce shifts 2–7$\times$ larger than arithmetic operators, indicating genuine representational activity.

Scale Dependence. The scale-dependent classification is, to our knowledge, without precedent. Prior work (Wei et al., 2022a) has shown that model *capabilities* are scale-dependent, but our finding concerns the causal *status* of specific tokens, which is a different phenomenon. One possible explanation is that larger models develop stronger internal representations for arithmetic operators (our positive controls), which raises the classification threshold. At 7B, only longer and more semantically rich patterns (`first, i need to`) carry enough representational weight to exceed this higher threshold.

Cross-Benchmark Convergence. The replication on MATH (§5.8) is important because GSM8K consists of grade-school arithmetic, and one might worry that reasoning tokens are dispensable only on easy problems. MATH includes competition-level problems across algebra, geometry, and number theory, yet the same zero-delta pattern holds. This suggests that the dispensability of distillation-enriched patterns is not task-difficulty dependent but reflects a structural property of the distillation process itself.

Implications for Evaluation. If reasoning chains contain tokens that are universally dispensable across architectures, scales, and benchmarks, then evaluation methods that depend on chain quality or structure may be misleading. Our CTA framework provides a tool for separating genuine reasoning artifacts from distillation residue.

7 Conclusion

We presented Causal Trajectory Attribution (CTA), a framework for classifying reasoning tokens in distilled chain-of-thought models by combining behavioral ablation with representational impact measurement. Applying CTA to three DeepSeek-R1-Distill model families across two benchmarks, we arrived at four principal findings:

1. All 14 distillation-enriched reasoning patterns are behaviorally dispensable: removing or replacing them produces zero accuracy change, confirmed by bootstrap CIs across every model and pattern combination tested.

2. Causal classification is scale-dependent. The same token (`let me`) is classified as SUPERSTITIOUS at 7B but LATENT-FUNCTIONAL at 1.5B and 8B, a result with no precedent in the mechanistic interpretability literature.

3. Constrained generation experiments confirm dispensability causally: suppressing `let me` at decode time yields no accuracy loss and zero token leakage, ruling out artifacts of post-hoc string manipulation.

4. The same reasoning fingerprints (`let me`, `wait`, `i think`) appear in both Qwen and Llama families after R1 distillation, confirming that they originate from the teacher model rather than from any property of the student architecture.

These results suggest that distilled reasoning models engage in structural mimicry: they reproduce the surface form of the teacher's reasoning chains without those chains playing any causal role in producing correct answers. The LATENT-FUNCTIONAL category, introduced in this work, reveals that many of these tokens are not inert either. They actively reshape internal representations even though they do not affect the final answer, pointing to a more complex account of token function than the binary useful-or-not framing that has dominated prior work. We believe CTA can serve as a practical diagnostic tool for auditing reasoning models, and we hope this work motivates further investigation into the causal structure of distilled chains across broader reasoning domains.

Limitations

We acknowledge several limitations in scope. First, both GSM8K and MATH are mathematical reasoning benchmarks, and our findings may not transfer to domains like code generation or commonsense reasoning. That said, the two benchmarks differ substantially in difficulty (grade-school arithmetic vs. competition-level algebra and geometry), and the zero-delta result replicates perfectly across both (§5.8), suggesting the phenomenon is not difficulty-dependent. Second, our constrained generation experiment covers only one pattern (`let me`) on one model (Qwen-1.5B). However, the post-hoc ablation covers all 14 patterns across all three models with identical results (Table 4), and the constrained experiment was designed as a causal confirmation of that broader finding rather than a standalone claim. Third, the GRPO pilot uses only 100 training steps. We note that even this short window was sufficient to show that RL alone produces zero structural divergence (Appendix C), which is itself informative. Longer training remains an open question. Fourth, all experiments use 4-bit quantization, which could affect logit shift magnitudes in absolute terms. We have no reason to believe quantization would reverse the direction of classification, but replication at full precision would strengthen the Condition D analysis. Finally, we test only DeepSeek-R1 distilled models. Other distillation pipelines (such as those based on OpenAI o1) may produce different patterns, though our finding that the same fingerprints appear across two unrelated architectures (Qwen and Llama) suggests the phenomenon is a property of the R1 teacher rather than any single student.

References

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. Technical report, OpenAI.

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

Arthur Conmy, Augustine Mavor-Parker, Aidan Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Curran Associates, Inc.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. 2025. Token-budget-aware LLM reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24842–24855, Vienna, Austria. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

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

Samuel Lewis-Lim, Xingwei Tan, Zhixue Zhao, and Nikolaos Aletras. 2025. Analysing chain of thought dynamics: Active guidance or unfaithful post-hoc rationalisation? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29838–29853, Suzhou, China. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adámek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. `https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html`. Transformer Circuits Thread.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032, Miami, Florida, USA. Association for Computational Linguistics.

Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 411–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, pages 55565–55581. Curran Associates, Inc.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical

reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Burrhus Frederic Skinner. 1948. "superstition" in the pigeon. *Journal of Experimental Psychology*, 38(2):168–172.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 58952–58965. Curran Associates, Inc.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. In *Advances in Neural Information Processing Systems*, volume 37, pages 12347–12361. Curran Associates, Inc.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. 2025a. TokenSkip: Controllable chain-of-thought compression in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3351–3363, Suzhou, China. Association for Computational Linguistics.

Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian McAuley, and Shuai Li. 2025b. Beyond chain-of-thought: A survey of chain-of-X paradigms for LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10795–10809, Abu Dhabi, UAE. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*.

A Additional Figures

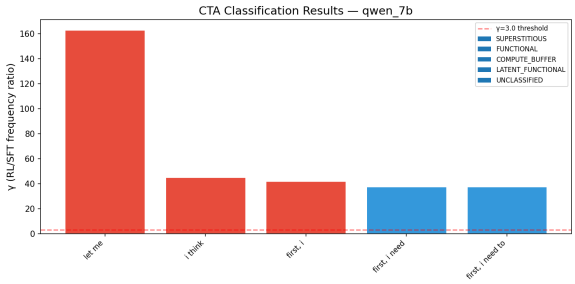

Figure 3: Qwen 7B: CTA classification with γ enrichment. Red bars = SUPERSTITIOUS; blue bars = LATENT-FUNCTIONAL. This is the only model showing mixed classifications.

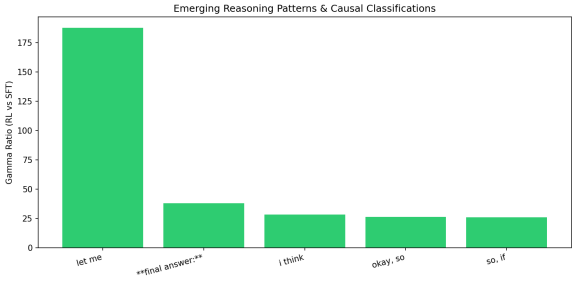

Figure 4: Llama 8B: All patterns classified LATENT-FUNCTIONAL.

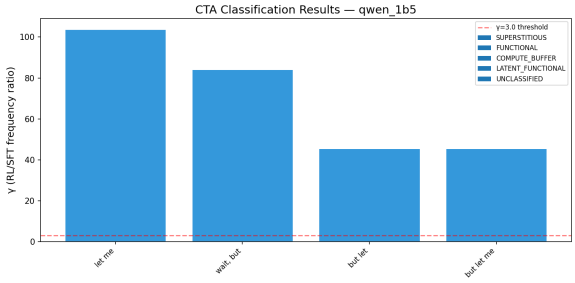

Figure 5: Qwen 1.5B: All patterns classified LATENT-FUNCTIONAL.

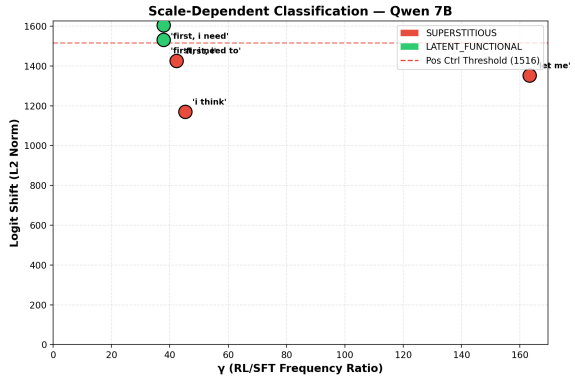

Figure 6: Qwen 7B scatter: scale-dependent classification. Short markers (`let me`) fall below the positive control threshold (red); longer markers (`first, i need`) exceed it (green).

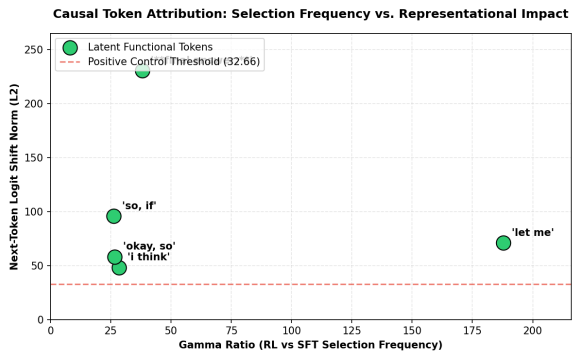

Figure 7: Llama 8B scatter: all patterns exceed the positive control threshold, confirming uniform LATENT-FUNCTIONAL classification.

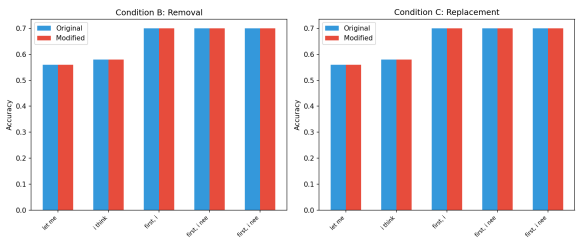

Figure 8: Qwen 7B ablation: Conditions B and C show perfect accuracy invariance across all patterns.

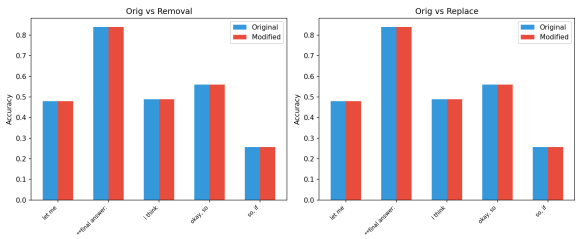

Figure 9: Llama 8B ablation results.

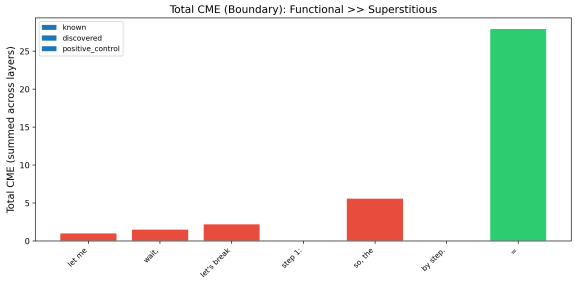

Figure 10: Causal Mediation Effect (CME) for Qwen 1.5B. The arithmetic operator = dominates (28.0), while discourse tokens have near-zero effects.

B Constrained Generation and Cross-Benchmark Figures

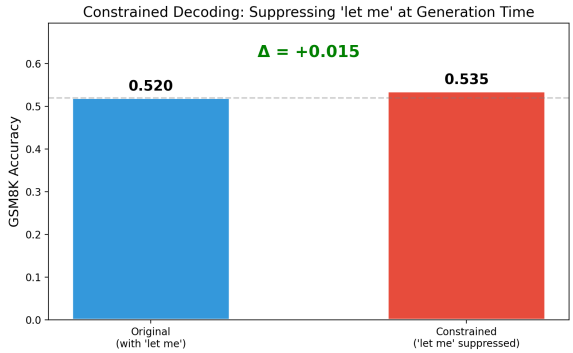

Figure 11: Constrained decoding result on Qwen-1.5B. Suppressing `let me` at generation time via logit bias produces no accuracy loss ($\Delta = +0.015$, within noise). Zero token leakage confirmed across all 200 chains.

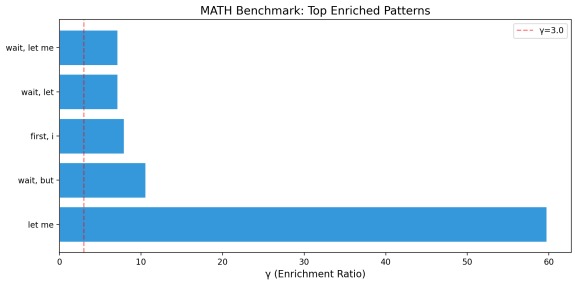

Figure 12: MATH benchmark: top enriched patterns for Qwen-1.5B. The same patterns that dominate GSM8K (`let me`, `wait, but`) also dominate MATH, all with $\gamma > 3.0$. All SFT counts are zero.

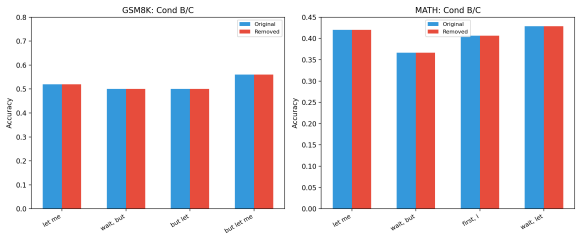

Figure 13: Cross-benchmark ablation comparison: GSM8K (left) vs. MATH (right). Blue bars show original accuracy; red bars show accuracy after pattern removal. The zero-delta result replicates perfectly across both benchmarks for all tested patterns.

C GRPO Pilot Details

Metric	**Base**	**RL-100**	Δ
Mean length (words)	156.14	156.32	+0.18
Mean steps	4.20	4.20	0.00
“step” count	232	232	0
“so” count	119	118	-1

Table 9: GRPO pilot: 100 steps of RL on Qwen2.5-1.5B-Instruct. Zero structural divergence from the base model.