When Reasoning Traces Become Performative: Step-Level Evidence that Chain-of-Thought Is an Imperfect Oversight Channel

Wenkai Li [1 2] **Fan Yang** [2] **Shaunak A. Mehta** [2] **Ananya Hazarika** [2] **Koichi Onoue** [2]

Abstract

Chain-of-thought (CoT) traces are increasingly used both to improve language-model capability and to audit model behavior, implicitly assuming that the visible trace remains synchronized with the computation that determines the answer. We test this assumption with a step-level *Detect-Classify-Compare* framework built around an answer-commitment proxy that is cross-validated with Patchscopes, tuned-lens probes, and causal direction ablation. Across nine models and seven reasoning benchmarks, latent commitment and explicit answer arrival align on only 61.9% of steps on average. The dominant mismatch pattern is *confabulated continuation*: 58.0% of detected mismatch events occur after the answer-commitment proxy has already stabilized while the trace continues producing deliberative-looking text, and a vacuousness analysis shows that the committed answer does not change during these steps. In architecture-matched Qwen2.5/DeepSeek-R1-Distill comparisons, the reasoning pipeline changes failure composition more than aggregate alignment, most clearly at 32B where confabulated steps decrease as contradictory states increase. Lower step-level alignment is also associated with larger CoT utility, suggesting that the settings that benefit most from CoT are often the least temporally faithful. Paired truncation and a complementary donor-corruption test further indicate that much pure-CS text is not load-bearing for the final answer. These findings suggest that CoT can remain useful while still being an unreliable report of when the answer was formed.

[1]Carnegie Mellon University [2]Fujitsu Research of America Inc. Correspondence to: Wenkai Li <wenkail@cs.cmu.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Chain-of-thought (CoT) reasoning began as a prompting technique for improving multi-step performance (Wei et al., 2022; Kojima et al., 2022), but in current reasoning model pipelines it also serves as a training target, a verification interface, and an oversight signal (Lightman et al., 2024; Wang et al., 2024a; Guo et al., 2025; OpenAI, 2024; Korbak et al., 2025; Emmons et al., 2025). Humans and monitor models increasingly inspect reasoning traces to judge whether a system reasoned safely, honestly, or competently. This dual role rests on a strong assumption: that the visible trace faithfully tracks the internal computation that produced the answer. If that assumption fails, CoT may still improve capability while providing a misleading basis for oversight.

A growing literature suggests that this assumption is fragile. Perturbation, bias-injection, and deployment studies document omitted hints, post-hoc rationalization, and other forms of unfaithful CoT (Lanham et al., 2023; Turpin et al., 2023; Paul et al., 2024; Chen et al., 2025; Arcuschin et al., 2025; Li et al., 2026). Conceptual and empirical work further suggests that faithfully verbalizing autoregressive computation may itself be difficult (Creswell & Shanahan, 2022; Tanneru et al., 2024; Saparov & He, 2023; Nye et al., 2021). The question is therefore no longer whether unfaithful CoT exists, but how often it arises in current reasoning models and what kind of mismatch it creates.

Yet the field still lacks a measurement lens tailored to this oversight question. Most existing faithfulness methods deliver response-level verdicts (Lanham et al., 2023; Paul et al., 2024; Shen et al., 2025): they tell us whether a trace is suspect, but not when the answer was already internally fixed relative to the visible chain. Concurrent work by Boppana et al. (2026) studies a closely related phenomenon under the label *performative CoT*, showing that reasoning traces can continue after latent answer commitment, especially on easier tasks. Related work by Palod et al. (2025) questions whether longer CoTs should be interpreted as evidence of more adaptive computation. Our focus is complementary but narrower. We center the *timing* of internal answer formation, distinguish five step-level mismatch patterns rather than a binary split, compare architecture-matched Qwen2.5/R1 pairs across three scales, and quantify the re-

lationship between temporal faithfulness and CoT utility. Interpretability tools can decode emerging predictions inside the model (nostalgebraist, 2020; Belrose et al., 2023; Ghandeharioun et al., 2024; Marks & Tegmark, 2024), but they have rarely been used to measure CoT faithfulness at the step level.

We introduce the DETECT-CLASSIFY-COMPARE framework to close this gap. At each CoT step, we extract an answer-commitment proxy using logit lens projections, cross-validate that proxy with Patchscopes concordance, tuned-lens probes, and direction ablation, then quantify timing divergence, classify mismatch events, and compare matched model pairs that differ in whether they underwent the DeepSeek-R1 distillation pipeline. Across nine models and seven reasoning benchmarks (Figure 1), the dominant failure mode is not overt contradiction but *post-commitment continuation*: the model often resolves the answer internally and then continues generating plausible reasoning text. The resulting picture is one in which a fluent and even accuracy-improving chain can still be a poor report of when the answer was formed.

Our investigation makes three contributions:

1. We introduce a step-level framework for measuring whether internal answer formation is temporally reflected in CoT, and find mean alignment of only 61.9%, with 58.0% of mismatch events falling into *confabulated* continuation.

2. We show that in architecture-controlled Qwen2.5/DeepSeek-R1-Distill comparisons, the reasoning pipeline changes not only aggregate alignment but also the composition of failures, most sharply at 32B where confabulated steps fall as contradictory states rise.

3. We connect this timing measure to downstream behavior: lower step-level alignment tends to coincide with larger CoT utility, while paired truncation and donor corruption indicate that much post-commitment text is descriptive rather than load-bearing.

2. Related Work

CoT faithfulness, oversight, and monitoring. Chain-of-thought prompting improves reasoning performance (Wei et al., 2022; Kojima et al., 2022; Yao et al., 2023; Nye et al., 2021; Wang & Zhou, 2024), but whether the resulting trace is a faithful explanation or oversight signal remains contested (Wiegreffe & Marasovic, 2021; DeYoung et al., 2020). Perturbation, bias-injection, and deployment studies show that models can preserve answers under trace edits, omit decisive information, or produce post-hoc rationalizations (Lanham et al., 2023; Turpin et al., 2023; Chen et al., 2025; Arcuschin et al., 2025). This matters because CoT is increasingly used in process supervision, verifier training, and safety monitoring workflows, where monitorability may fail even when CoT remains useful for capability (Lightman et al., 2024; Wang et al., 2024a; Korbak et al., 2025; Emmons et al., 2025; MacDermott et al., 2025; Zolkowski et al., 2025; Meek et al., 2025). Most of this literature provides response-level verdicts (Radhakrishnan et al., 2023; Li et al., 2024; Shen et al., 2025); our question is narrower and more temporal: when does internal answer formation diverge from the visible trace?

Internal readouts and performative CoT. Interpretability tools such as the logit lens (nostalgebraist, 2020), tuned lens (Belrose et al., 2023), Patchscopes (Ghandeharioun et al., 2024), and linear probes (Marks & Tegmark, 2024; Li et al., 2023; Burns et al., 2023; Azaria & Mitchell, 2023; Kadavath et al., 2022) make it possible to track emerging predictions during generation. We use these tools not to recover a full reasoning circuit, but to estimate when answer commitment becomes decodable. Closest to our setting, Boppana et al. (2026) study performative CoT and show that latent answer commitment can precede what the trace reveals, especially on easier tasks. Palod et al. (2025) similarly cautions against interpreting longer CoTs as evidence of more adaptive computation. Other recent work measures step-level faithfulness through parameter-level unlearning or hint-injection paradigms (Tutek et al., 2025; Young, 2026a;b). Our contribution is more specific: a timing-based latent answer proxy cross-validated by multiple readouts, a five-way step-level taxonomy of mismatch patterns, architecture-matched Qwen2.5/DeepSeek-R1 comparisons across scale, and an empirical link between temporal faithfulness and CoT utility. A detailed axis-by-axis comparison against these and other closely related works is provided in Appendix O. Complementary work on normalized step-importance metrics such as NLDD targets cross-architecture comparability, whereas BCA is designed as a timing metric centered on when latent answer commitment becomes decodable relative to answer revelation (Ye et al., 2026).

Reasoning pipelines and training effects. Reasoning-oriented training pipelines, including process supervision, reinforcement learning, and distillation, increasingly optimize models to produce extended reasoning traces (Uesato et al., 2022; Cobbe et al., 2021; Zelikman et al., 2022; Havrilla et al., 2024; Guo et al., 2025; OpenAI, 2024; Hao et al., 2024). Whether these pipelines improve faithfulness rather than only usefulness remains unclear: longer or more polished traces may still fail to report the computation that determined the answer, and concurrent work has raised concerns about strategically uninformative or steganographic reasoning traces (Roger & Greenblatt, 2023; Chen et al., 2025). Conversely, targeted interventions such as coun-

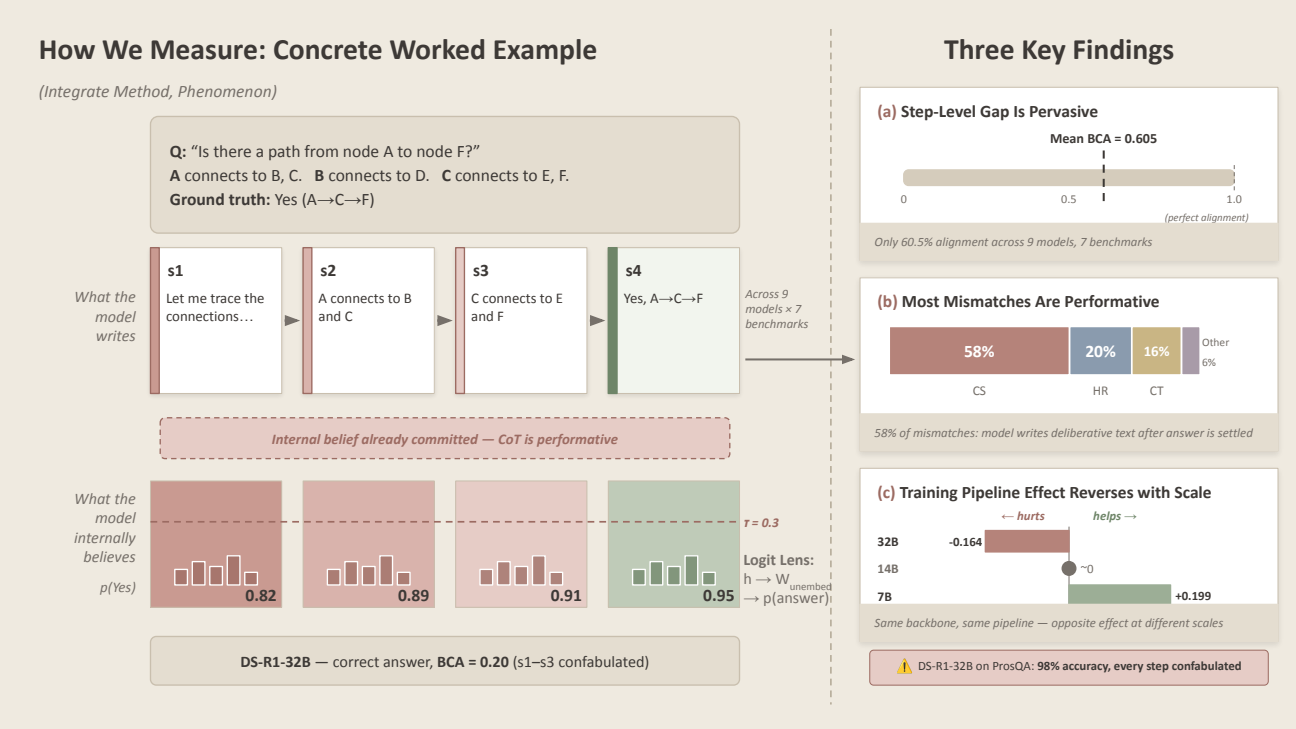

Figure 1. Overview of the empirical picture. Left: on a ProsQA example, DS-R1-32B commits to the answer by step 1 while the written trace continues, yielding low BCA. Right: the aggregate results show the same pattern at scale, with 61.9% mean alignment, 58% of mismatches classified as confabulated steps, and pipeline effects that vary across model scales.

terfactual simulation training can improve faithfulness in controlled settings, suggesting that the effect of training is real but objective-dependent (Hase & Potts, 2026). Our matched-family comparison targets this question directly by comparing Qwen2.5 models against their DeepSeek-R1-distilled counterparts at 7B, 14B, and 32B, which lets us study how the pipeline reshapes timing alignment and failure composition while minimizing backbone confounds.

3. Method

DETECT-CLASSIFY-COMPARE follows a chain of thought one step at a time. For each step, we estimate whether the model has already committed to the final answer, compare that latent state with what the written trace has revealed so far, summarize the resulting mismatches with a taxonomy, and then use matched model pairs to study how reasoning training changes the alignment pattern.

3.1. Step-Level Belief Extraction

Given a model's CoT response to a question, we first segment the trace into reasoning steps $s_1, \ldots, s_T$ using a rule-based parser that splits on newlines, transition markers, and sentence boundaries (parsing F1 = 0.87; validation in Appendix F).

At each step boundary, we extract a proxy for the model's internal commitment to the final answer using the logit lens (nostalgebraist, 2020; Wang, 2025). Concretely, we project each layer's hidden state through the unembedding matrix and layer normalization to obtain a distribution over the vocabulary. We use the probability assigned to the ground-truth answer token, $p^{(\ell)}_{\text{ans}}(t)$, as our primary measure of internal commitment at step t and layer ℓ. This operationalization captures the dominant signal but may not reflect all aspects of the model's internal state; we discuss this limitation in Section 8.

We validate the logit lens through concordance with Patchscopes (Ghandeharioun et al., 2024) and tuned lens probes (Alain & Bengio, 2017; Belinkov, 2022). Patchscopes agrees on 75.6% of mid-depth steps and 74.7% of final-layer steps, with near-zero logit-lens-only disagreement (0.0% mid, 1.5% final); tuned-lens binary classifications agree with the logit lens at $\geq$99.2% across four held-out model-benchmark configurations (Appendix J.1). Together these checks indicate that the logit lens behaves as a conservative answer-commitment readout rather than a trigger-happy artifact.

3.2. Divergence Quantification

We quantify timing alignment using two complementary metrics. **Belief-CoT Agreement** (BCA) measures the fraction of steps at which the thresholded answer-commitment proxy ($p_{\text{ans}} > \tau$, $\tau = 0.3$) matches a binary CoT label indicating whether the trace has already explicitly reached the final answer. We compute that label automatically by checking, at each step, whether the ground-truth answer string or a normalized variant has appeared in the cumulative CoT text. A BCA of 1.0 indicates perfect timing alignment. Because the CoT label is binary, BCA should be read as a timing metric over answer commitment rather than a full semantic comparison; results are stable across $\tau \in [0.1, 0.5]$ (Appendix D.1). **Convergence Timing Gap** (CTG) measures how many steps before the CoT arrives at the final answer the latent belief has already committed. A positive CTG indicates premature internal convergence.

3.3. Mismatch Taxonomy

Not all mismatches are alike. Based on temporal patterns in the answer-commitment trajectory, we classify detected mismatches into five heuristic categories. **Premature Convergence** (PC) occurs when the answer-commitment proxy locks onto the answer at least two steps before the trace states it. **Contradictory States** (CT) denote high-confidence disagreement between the proxy and the answer-arrival label. **Hidden Reasoning** (HR) captures large adjacent-step shifts in the proxy that are not mirrored by a change in explicit answer state. **Confabulated Steps** (CS) are mismatch steps where the proxy changes negligibly, suggesting that the trace continues after answer commitment has largely stabilized. We use "confabulated" as a descriptive label for this temporal pattern rather than as a strong claim about whether these steps serve a computational role that our proxy does not capture. **Silent Error Correction** (SEC) denotes runs of disagreement followed by re-alignment, consistent with an unspoken internal correction. See formal definitions and operational criteria in Appendix C.1.

These categories are not mutually exclusive: a single step or instance may satisfy multiple criteria (we analyze multi-label co-occurrence in Appendix C.2). In the main summaries we count all detector firings, and when an instance-level summary is needed we report the dominant type by severity.

3.4. Architecture-Controlled Pipeline Comparison

A central challenge in studying how training affects faithfulness is confounding: different models usually differ in architecture, data, and training procedure at the same time. To isolate the effect of one concrete pipeline within a single backbone family, we leverage the DeepSeek-R1 distillation models, which use Qwen2.5 as their base architecture. This gives us matched pairs at three scales: Qwen2.5-7B vs. DS-R1-7B, Qwen2.5-14B vs. DS-R1-14B, and Qwen2.5-32B vs. DS-R1-32B. Within each pair, differences in BCA are therefore more plausibly linked to the distillation pipeline than to backbone changes, though the pipeline itself still bundles RL, SFT, and knowledge distillation. Our conclusions about pipeline effects are specific to the Qwen2.5/R1 setting; generalization to other backbone families remains open.

3.5. Causal Validation via Direction Ablation

The logit lens reveals correlational patterns in latent belief formation, but correlation does not establish that detected beliefs causally influence the model's output. To test causality, we perform *direction ablation* (Zou et al., 2023; Goldowsky-Dill et al., 2023; Geiger et al., 2025; Vig et al., 2020): at a target layer ℓ, we extract the answer-token direction $\hat{\mathbf{d}}_{a^*}$ from the unembedding matrix and project it out of the hidden state with strength α:

$$\tilde{\mathbf{h}}^{(\ell)} = \mathbf{h}^{(\ell)} - \alpha \cdot (\mathbf{h}^{(\ell)} \cdot \hat{\mathbf{d}}_{a^*})\, \hat{\mathbf{d}}_{a^*} \tag{1}$$

We sweep $\alpha \in \{1, 3, 5, 10\}$ and report the strength that produces the largest change in answer probability, following the practice of selecting the most informative intervention point. If the belief component is causally relevant, removing it should change the model's output. We measure the *correctness change rate*: the fraction of instances where the model's answer switches between correct and incorrect after ablation. We test four layers spanning 50–97% of model depth on $n = 60$ instances per 7B configuration and $n = 20$ per 32B configuration, covering both the matched 7B pair and the matched 32B pair. Full details in Appendix J.2.

4. Experimental Setup

4.1. Models

We evaluate nine open-source models. Four primary models span distinct families: Llama-3.1-8B-Instruct (Dubey et al., 2024), Qwen3.5-9B (Yang et al., 2025), Gemma-3-27B-IT (Google DeepMind, 2025), and DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025) (abbreviated DS-R1-32B). For architecture-controlled analysis, we add Qwen2.5-Instruct models at 7B, 14B, and 32B (Yang et al., 2024) with matched DeepSeek-R1-Distill-Qwen models at 7B and 14B (Guo et al., 2025) (abbreviated DS-R1-7B and DS-R1-14B), yielding three Qwen2.5-to-R1 comparisons within a shared backbone (Appendix A). Not every analysis uses all models; we state the active count where it differs.

4.2. Benchmarks

We use seven reasoning benchmarks chosen to span distinct forms of multi-step inference: MATH-500 (Hendrycks et al., 2021) for mathematical reasoning, PrOntoQA (Saparov et al., 2023) for symbolic logic, ProsQA for planning, MMLU-Pro (Wang et al., 2024b) for knowledge-intensive reasoning, BBH-Logical Deduction and BBH-Tracking Shuffled Objects (Suzgun et al., 2023) for compositional reasoning, and GPQA-Diamond (Rein et al., 2024) for difficult scientific question answering. Depending on the benchmark, evaluation uses between 60 and 300 instances. Appendix A lists the exact instance counts and dataset sources.

4.3. Inference and Measurement Setup

All experiments use greedy decoding. We segment generated CoT traces with the rule-based step parser described in Appendix F, which achieves an F1 score of 0.87 on manually annotated traces. For belief extraction we read hidden states at every even-numbered layer and threshold the logit-lens answer probability at $\tau = 0.3$ when computing BCA, a choice that remains stable across $\tau \in [0.1, 0.5]$ (see Appendix D.1). Confidence intervals are 95% bootstrap intervals with Bonferroni correction, and BCA is stratified by answer correctness to avoid conflating alignment with raw task difficulty. The full evaluation required approximately 200 GPU-hours on four A100 GPUs. Temperature sensitivity checks (see Appendix N) and inference-time intervention experiments (see Appendix H.1) are reported in the appendix.

5. Results and Analysis

This section examines whether the answer a model appears to commit to internally is synchronized with the point at which that answer becomes explicit in the written chain-of-thought trace. We use BCA as a step-level timing measure: high values indicate that latent answer commitment and explicit answer arrival move together, while low values indicate that the two are separated in time. Across models and benchmarks, the main pattern is simple. This separation is common, it varies strongly by task, and in many cases it is not visible from the surface text alone.

5.1. Alignment Is Low, and Benchmark Choice Matters Most

Table 1 reports BCA for each model-benchmark configuration. Averaged over all reported cells, BCA is 0.619. In practical terms, the latent commitment signal and the explicit answer in the CoT trace agree in timing on only about three-fifths of steps. The result does not look like an isolated failure of one model family. Among the four primary models, a one-way ANOVA does not detect a significant model effect ($F = 0.91$, $p = 0.45$), and Bayesian model comparison favors the null model ($\mathrm{BF}_{01} = 32.9$). This does not prove that architecture is irrelevant, but it does suggest that model identity alone cannot explain the observed timing gaps. A held-out Gemma 3 12B replication shows the same qualitative pattern of low and task-sensitive alignment, as reported in Appendix D.3.

The benchmark effect is large. On PrOntoQA, where the reasoning structure is relatively explicit, Llama 3.1 8B reaches a BCA of 0.914. On other tasks, the same alignment measure can collapse. Qwen2.5-32B reaches only 0.044 on ProsQA, and Gemma 3 27B reaches 0.128 on MMLU-Pro. The representative trajectories in Figure 2 show this contrast directly: in high-alignment settings, the answer-commitment curve and the written answer arrive together; in low-alignment settings, the model can appear committed well before the trace says so. A separate convergence-timing analysis points in the same direction. Lower BCA is associated with larger convergence timing gaps ($\rho = -0.48$, $p < 0.001$; Appendix G), which is consistent with latent commitment often preceding explicit answer arrival.

5.2. Most Mismatches Occur After the Model Has Already Committed

We next ask what kind of behavior produces these timing gaps. We apply the five detectors defined in Appendix C.1 to 168,302 mismatch events. The taxonomy should be read as a descriptive breakdown of overlapping failure patterns rather than as a set of mutually exclusive causes. With that caveat, one category dominates: Confabulated Steps (CS), where the answer-commitment proxy is already stable but the model continues producing CoT steps that induce little change in that proxy. CS accounts for 58.0% of detector firings. The other categories are Hidden Reasoning (HR, 19.9%), Contradictory States (CT, 15.5%), Silent Error Correction (SEC, 5.2%), and Premature Convergence (PC, 1.4%).

The dominance of CS is not driven by a single model. Across the tested models, per-model CS rates range from 37.4% for DS-R1-32B to 75.4% for Qwen-3.5-9B. Gemma 3 27B has an additional logit soft-capping artifact, discussed in Appendix C.1, but the pattern remains after accounting for it. Excluding Gemma 3 27B, CS still explains 56.5% of the remaining 162,422 events. A held-out Gemma 3 12B replication under the same $\tau = 0.3$ threshold produces a non-degenerate taxonomy distribution and again shows a CS-dominant pattern (CS = 58.1%; Appendix D.3).

This result changes the interpretation of the alignment gap. The typical failure is not that the written trace openly contradicts the internal answer signal. More often, the model has already settled on an answer and then continues writing plausible-looking intermediate steps. Several checks

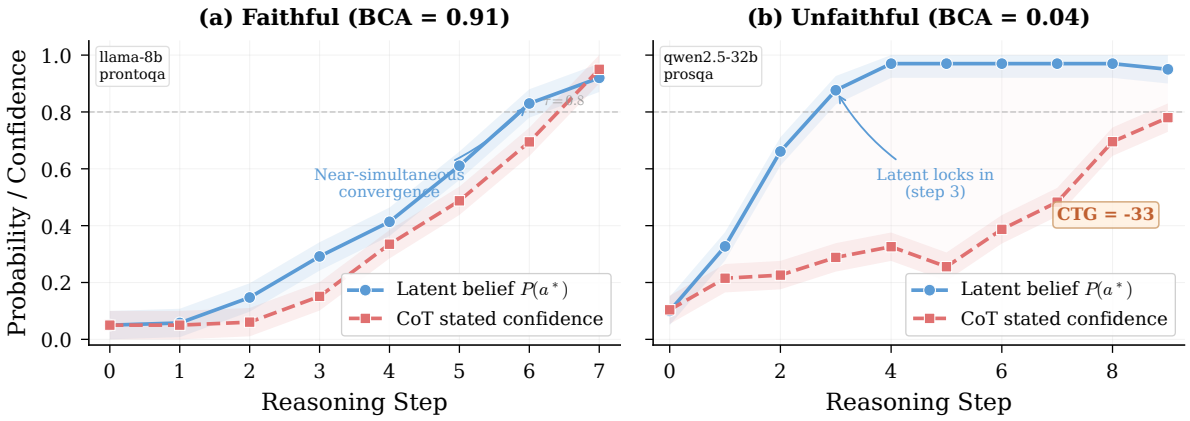

Figure 2. Representative belief trajectories. **(a)** Faithful (Llama 3.1 8B/PrOntoQA, BCA = 0.91): latent belief and CoT converge near-simultaneously. **(b)** Unfaithful (Qwen2.5-32B/ProsQA, BCA = 0.04): latent belief locks by step 3 while CoT continues exploring, yielding a large convergence timing gap.

Table 1. Step-level answer-commitment alignment, measured by BCA. The largest differences are benchmark-level rather than model-level. Bold marks the highest value in each benchmark, and shaded rows denote the Qwen2.5 architecture-control pairs.

Model	MATH	PrOnto	ProsQA	MMLU-P	BBH-LD	BBH-TSO	GPQA-D	Mean
Llama 3.1 8B	**0.738**	**0.914**	0.280	0.904	0.669	0.807	0.779	0.727
Qwen 3.5 9B	0.634	0.635	0.364	0.254	0.774	0.565	0.773	0.571
Gemma 3 27B	0.526	0.774	0.181	0.128	0.819	0.725	0.695	0.550
DS-R1-32B	0.614	0.529	0.329	0.231	0.538	0.890	0.769	0.557
Qwen2.5-7B	0.579	0.807	0.307	0.895	**0.871**	0.804	0.808	0.724
DS-R1-7B	0.633	0.771	0.342	**0.916**	0.775	0.810	0.786	0.719
Qwen2.5-14B	0.581	0.496	0.236	0.193	0.704	0.508	0.789	0.501
DS-R1-14B	0.600	0.495	**0.365**	0.245	0.486	**0.896**	**0.812**	0.557
Qwen2.5-32B	0.627	0.678	0.044	0.897	0.827	0.783	0.783	0.663

support this reading. In a vacuousness analysis over 314 instances containing CS steps, the committed answer never changes during a CS step, and 90.1% of individual CS steps have near-zero belief change ($\Delta p_{\text{ans}} < 0.001$; Appendix L.3). A donor-corruption test gives a similar result from a different angle: replacing text at pure-CS steps leaves final-answer accuracy essentially unchanged ($\Delta = -0.002$, $p = 0.78$), while replacing matched neighbor or random steps is more damaging (Appendix P). These controls make it less likely that CS is simply useful verification written after the answer.

Text-only inspection also misses much of this behavior. Three LLM judges, given only the CoT text, reach high agreement (AC1 = 0.78) but label 82% of inspected cases as PC rather than CS (Appendix E). Human annotators show moderate-to-high agreement on the five-way taxonomy (AC1 = 0.71; Appendix K). The gap between text-only labels and detector-based labels is important for oversight: many post-commitment steps still look like normal reasoning.

The categories also overlap substantially. In the 12,004-instance audit pool, 68.4% of unfaithful steps satisfy more than one detector criterion (Appendix C.2). CT and PC form a near-redundant pair (Pearson $r = 0.937$, 95% CI $[+0.930, +0.943]$; Table 5), and neither PC nor SEC appears as an isolated category in that pool. The taxonomy is still useful, but mainly as a way to describe correlated facets of trace unfaithfulness.

The cleanest causal evidence comes from paired truncation. Among the categories with non-empty pure subgroups, cutting the trace at a pure-CS firing step improves accuracy relative to cutting at a uniformly random step by $+0.078$ ($n = 2{,}088$, $p < 10^{-16}$; Appendix P). If those CS steps were load-bearing reasoning steps, truncating at them should not help. HR behaves differently: pooled deltas are positive rather than negative, suggesting that many large belief jumps mark the completion of commitment rather than hidden computation that remains necessary afterward. CT trends in the opposite direction, but the pure-CT subgroup is too small to support a strong conclusion. The detector thresholds are stable under $\pm 50\%$ perturbations (Appendix D.2), and 85% of CS events occur without premature convergence (Appendix C.3).

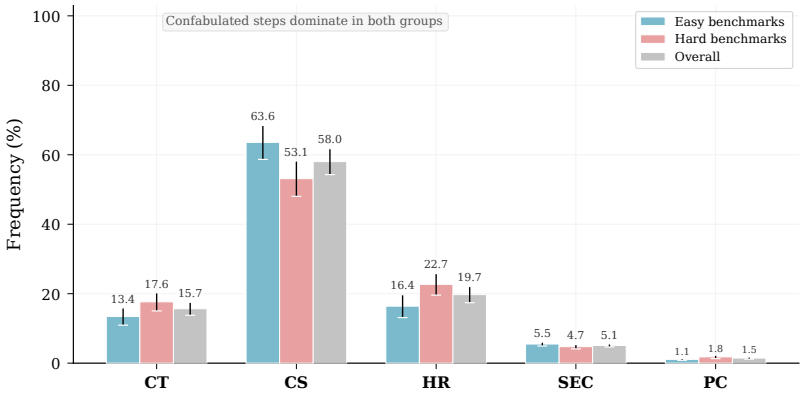

Figure 3. Distribution of the five unfaithfulness types across easy and hard benchmarks (∼168,000 mismatch steps). Confabulated Steps dominate at all difficulty levels (58% overall), indicating that most unfaithful reasoning is performative rather than erroneous.

5.3. The Reasoning Pipeline Changes the Type of Mismatch

We next compare DeepSeek-R1 distilled models with their Qwen2.5 architecture controls. Since each pair shares the same backbone size, this comparison gives a cleaner view of the reasoning pipeline than the cross-family results above. The aggregate effect on BCA is mixed rather than monotonic: alignment decreases at 32B (mean $\Delta_{\text{pipe}} = -0.106$), improves at 14B ($+0.056$), and is nearly unchanged at 7B (-0.005). The 7B-versus-32B contrast is not significant under a paired Wilcoxon test ($p = 0.578$), and the 32B mean becomes much smaller once the MMLU-Pro outlier is removed ($\Delta_{\text{pipe}} = -0.012$ without the -0.666 MMLU-Pro cell; full per-benchmark deltas in Table 7). The aggregate numbers should therefore not be read as a simple scaling law.

The taxonomy shows a clearer pipeline effect than BCA alone. Even when the overall alignment score changes only modestly, the distribution of mismatch types can change substantially. Chi-squared tests on the five-way taxonomy distribution show increasing divergence with scale, with Cramér's V rising from 0.058 at 7B to 0.155 at 14B and 0.289 at 32B (Appendix M.2). At 7B, the taxonomy distribution is largely preserved, with CS moving from 53.0% in Qwen2.5-7B to 57.9% in DS-R1-7B. At 14B, the pipeline reduces the average number of mismatches per instance from 24.2 to 14.7, a 39% reduction that matches the positive BCA effect at that scale.

The strongest shift appears at 32B. Confabulated steps fall from 57.6% to 37.4%, while contradictory states rise from 7.1% to 28.9%. This is not only a change in how often mismatch occurs. It changes what a monitor would need to detect. A monitor tuned for post-commitment continuation may miss high-confidence disagreements between the internal answer signal and the written trace. DS-R1-32B on ProsQA illustrates this point sharply: the model reaches 98.3% accuracy, but its BCA is only 0.329, and none of the analyzed instances are classified as faithful.

5.4. Belief Commitment Appears Late in the Forward Pass

The layer-wise logit-lens analysis gives a compatible picture of the timing results. Across models, answer probability is close to chance in early layers, rises through the middle of the network, and often increases sharply near the final layers. The DeepSeek-R1 pipeline amplifies this late-layer pattern at 32B, while the corresponding gap at 7B is much smaller. This mirrors the behavioral results above: at larger scale, the pipeline appears to concentrate answer resolution into later layers that are less tightly synchronized with the point at which the answer appears in the CoT trace.

Direction ablation provides a stronger check that the detected answer-commitment signals are not merely probe artifacts. We project out the answer-token direction from hidden states at four layer depths, as described in Section 3.5. At 7B, this changes correctness in up to 23% of Qwen2.5-7B/MATH-500 cases and up to 38% of DS-R1-7B/ProsQA cases. At 32B, DS-R1-32B on ProsQA reaches a 55% correctness change, with a clear late-layer peak (Appendix J.2). These results support the causal relevance of the answer-commitment direction, but they do not establish a universal late-layer rule across all tasks. Patchscopes concordance and tuned-lens agreement provide additional evidence that the logit lens is a conservative estimator of answer commitment rather than an idiosyncratic probe (Appendix J.1).

5.5. CoT Helps Most Where Timing Alignment Is Weak

Finally, we ask whether low alignment is related to the practical value of CoT. For the five models with direct no-CoT baselines, we compute CoT utility as the accuracy gain from using CoT, giving 35 model-benchmark configurations. BCA has a moderate negative Pearson correlation with

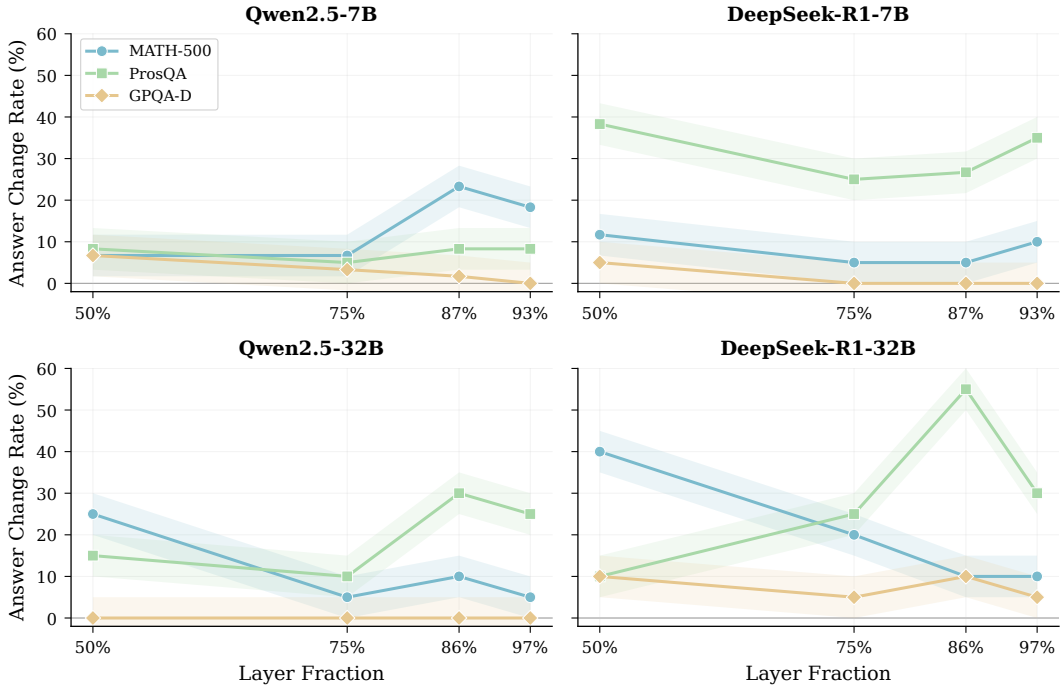

Figure 4. Direction ablation at selected layer depths for the two architecture-matched pairs. Removing the answer-token direction changes correctness most strongly in late layers for several configurations, with the largest observed effect in DS-R1-32B on ProsQA (55% at 86% depth).

CoT utility ($r = -0.42$, $p = 0.012$; bootstrap 95% CI $[-0.64, -0.13]$; Appendix I). In other words, the settings where CoT helps most are often the settings where step-level timing alignment is weakest. All five within-model correlations are negative ($r \in [-0.52, -0.39]$), although each individual correlation is underpowered at $n = 7$.

The relationship is not uniform. The Spearman correlation is directionally consistent but not significant ($\rho = -0.25$, $p = 0.16$), and the leave-one-benchmark-out analysis identifies ProsQA as the main leverage point (Appendix M.3). This sensitivity is informative rather than purely problematic. ProsQA is also the clearest graph-planning setting in the benchmark suite, where CoT is highly useful but step-level alignment is very low. The result should be read as a moderate association, not as a universal law. As a final check, BCA is nearly uncorrelated with an early-commitment rate derived from answer-probability trajectories ($\rho = 0.05$, $p = 0.73$, $n = 56$; Appendix M.1), which suggests that BCA captures timing information not reducible to response-level early answering.

Taken together, the results show that CoT unfaithfulness is often a timing problem rather than a content-only problem. The model may commit to an answer before the written reasoning arrives there, and the remaining trace can still look coherent. This makes surface-level CoT inspection unreliable in precisely the settings where CoT is most useful.

6. Discussion

The paired-truncation analysis turns the descriptive taxonomy into an intervention-based test (Appendix P). Cutting at a confabulated step more often preserves or improves final-answer correctness than cutting at a matched control step from the same instance, which supports the interpretation of many such segments as post-commitment continuation rather than load-bearing reasoning. The corresponding hidden-reasoning result is more ambiguous: pooled deltas are positive, suggesting that many large belief jumps mark commitment completion rather than latent computation that still needs to be verbalized. CT remains directionally load-bearing but underpowered, while SEC and PC are structurally non-isolable as pure subgroups. We therefore use the five-way taxonomy as a descriptive decomposition of correlated mismatch patterns, not as five independent causal mechanisms.

These findings matter most for oversight. Process reward models and step-level monitors may reward fluent continuation after commitment rather than the computation that determines the answer (Lightman et al., 2024). The matched-family comparison further shows that training pipelines can change which failures dominate even when aggregate alignment moves little, consistent with broader evidence that CoT monitorability is sensitive to development choices (Korbak et al., 2025). The negative association between alignment

and CoT utility points to a narrower tension: the settings in which CoT helps most may also be the settings in which the trace is least informative about answer formation, although in our data this pattern is driven most strongly by ProsQA-like graph-planning tasks. We do not conclude that long CoT is generally empty or useless; some harder tasks still show genuine mid-trace belief shifts (Boppana et al., 2026). The narrower conclusion is that CoT should not be assumed to report, step by step, when the answer was formed. If CoT is to function as an oversight interface, future work will need methods that distinguish load-bearing reasoning from post-commitment narration and training procedures that make that distinction easier to observe.

7. Conclusion

We studied whether chain-of-thought traces faithfully report when a model forms its answer. Across nine open-weight models and seven reasoning benchmarks, latent answer commitment and explicit answer arrival are often misaligned, with post-commitment continuation as the dominant mismatch pattern. Matched Qwen2.5/DeepSeek-R1 comparisons show that reasoning pipelines can change the type of unfaithfulness even when aggregate alignment changes only modestly. The main lesson is not that CoT is useless, but that it is an imperfect oversight channel: a trace can improve accuracy while failing to reveal when the answer was actually formed. Future process monitors should therefore distinguish load-bearing reasoning from plausible post-commitment narration.

8. Limitations

Our p_{ans} proxy captures answer-commitment *timing* rather than full semantic equivalence, and may underestimate beliefs encoded nonlinearly or across multiple tokens (mitigated via Patchscopes and tuned-lens concordance; Appendix J.1). The causal analysis reports the strongest effect across intervention strengths, possibly overstating single-setting sizes. The taxonomy is overlapping by design (Appendix C.2), the scale-dependent reversal holds only within the Qwen2.5/DeepSeek-R1 family, and all models are open-weight (7B–32B).

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2017.

Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., and Conmy, A. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 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.

Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., and Steinhardt, J. Eliciting latent predictions from transformers with the tuned lens. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Boppana, S., Ma, A., Loeffler, M., Sarfati, R., Bigelow, E., Geiger, A., Lewis, O., and Merullo, J. Reasoning theater: Disentangling model beliefs from chain-of-thought. *arXiv preprint arXiv:2603.05488*, 2026.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., and Perez, E. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Creswell, A. and Shanahan, M. Faithful reasoning using large language models. *arXiv preprint arXiv:2208.14271*, 2022.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., and Wallace, B. C. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020.

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 preprint arXiv:2407.21783*, 2024.

Emmons, S., Jenner, E., Elson, D. K., Saurous, R. A., Rajamanoharan, S., Chen, H., Shafkat, I., and Shah, R. When chain of thought is necessary, language models struggle to evade monitors. *arXiv preprint arXiv:2507.05246*, 2025.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C.,

et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

Ghandeharioun, A., Caciularu, A., Pearce, A., Dixon, L., and Geva, M. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Goldowsky-Dill, N., MacLeod, C., Sato, L., and Arora, A. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Google DeepMind. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Gwet, K. L. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., and Tian, Y. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Hase, P. and Potts, C. Counterfactual simulation training for chain-of-thought faithfulness. *arXiv preprint arXiv:2602.20710*, 2026.

Havrilla, A., Du, Y., Raparthy, S. C., Tsygankova, C., Baez, A., et al. 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. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Zhou, D. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022.

Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., et al. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*, 2025.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

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

Li, W., Yang, F., Mehta, S. A., and Onoue, K. Persona non grata: Single-method safety evaluation is incomplete for persona-imbued llms, 2026. URL https://arxiv.org/abs/2604.11120.

Li, X. L., Shrivastava, V., Li, S., Hashimoto, T., and Liang, P. Benchmarking and improving generator-validator consistency of language models. In *International Conference on Learning Representations*, 2024.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

MacDermott, M., Wei, Q., Djoneva, R., and Ward, F. R. Reasoning under pressure: How do training incentives influence chain-of-thought monitorability? *arXiv preprint arXiv:2512.00218*, 2025.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024.

Meek, A., Sprejer, E., Arcuschin, I., Brockmeier, A. J., and Basart, S. Measuring chain-of-thought monitorability through faithfulness and verbosity. *arXiv preprint arXiv:2510.27378*, 2025.

nostalgebraist. Interpreting GPT: The logit lens. *LessWrong*, 2020. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Biber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

OpenAI. Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/, 2024.

Palod, V., Valmeekam, K., Stechly, K., and Kambhampati, S. Performative thinking? the brittle correlation between cot length and problem complexity. *arXiv preprint arXiv:2509.07339*, 2025.

Paul, D., West, R., Bosselut, A., and Faltings, B. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15012–15032, 2024.

Radhakrishnan, A., Nguyen, K., Chen, A., Chen, C., Denison, C., Hernandez, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., et al. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*, 2023.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024.

Roger, F. and Greenblatt, R. Preventing language models from hiding their reasoning. *arXiv preprint arXiv:2310.18512*, 2023.

Saparov, A. and He, H. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *International Conference on Learning Representations*, 2023.

Saparov, A., Pang, R. Y., Padmakumar, V., Joshi, N., Kazemi, M., Kim, N., and He, H. Testing the general deductive reasoning capacity of large language models using OOD examples. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Shen, X., Wang, S., Tan, Z., Yao, L., Zhao, X., Xu, K., Wang, X., and Chen, T. Faithcot-bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. *arXiv preprint arXiv:2510.04040*, 2025.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q., Chi, E., Zhou, D., and Wei, J. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

Tanneru, S. H., Ley, D., Agarwal, C., and Lakkaraju, H. On the hardness of faithful chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2406.10625*, 2024.

Turpin, M., Michael, J., Perez, E., and Bowman, S. 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, 2023.

Tutek, M., Chaleshtori, F. H., Marasović, A., and Belinkov, Y. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9946–9971, 2025.

Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., and Higgins, I. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Wang, P., Li, L., Shao, Z., Xu, R.-Z., Dai, D., Li, Y., Chen, D., Wu, Y., and Sui, Z. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024a.

Wang, X. and Zhou, D. Chain-of-thought reasoning without prompting. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Wang, Z. LogitLens4LLMs: Extending logit lens analysis to modern large language models. *arXiv preprint arXiv:2503.11667*, 2025.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Wiegreffe, S. and Marasovic, A. Teach me to explain: A review of datasets for explainable NLP. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ye, D., Loffgren, M., Kotadia, O., and Wong, L. Mechanistic evidence for faithfulness decay in chain-of-thought reasoning. *arXiv preprint arXiv:2602.11201*, 2026.

Young, R. J. Lie to me: How faithful is chain-of-thought reasoning in reasoning models? *arXiv preprint arXiv:2603.22582*, 2026a.

Young, R. J. Why models know but don't say: Chain-of-thought faithfulness divergence between thinking tokens and answers in open-weight reasoning models. *arXiv preprint arXiv:2603.26410*, 2026b.

Zelikman, E., Wu, Y., Mu, J., and Goodman, N. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Zolkowski, A., Xing, W., Lindner, D., Tramèr, F., and Jenner, E. Can reasoning models obfuscate reasoning? stress-testing chain-of-thought monitorability. *arXiv preprint arXiv:2510.19851*, 2025.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Model and Benchmark Specifications

Table 2 provides the full architectural specifications for all models used in this study.

Table 2. Model specifications. Architecture-controlled baselines are shaded.

Model	Params	Layers	d_{model}	Backbone
Llama 3.1 8B-Inst.	8.0B	32	4096	Llama 3.1
Qwen 3.5 9B	9.0B	40	3584	Qwen 3.5
Gemma 3 27B-Inst.	27.2B	62	5376	Gemma 3
DS-R1-Qwen-32B	32.5B	64	5120	Qwen2.5
Qwen2.5-7B-Inst.	7.6B	28	3584	Qwen2.5
DS-R1-Qwen-7B	7.6B	28	3584	Qwen2.5
Qwen2.5-14B-Inst.	14.7B	48	5120	Qwen2.5
DS-R1-Qwen-14B	14.7B	48	5120	Qwen2.5
Qwen2.5-32B-Inst.	32.5B	64	5120	Qwen2.5

Table 3 provides benchmark specifications.

Table 3. Benchmark specifications. N_{eval} is the evaluation sample size; N_{probe} is the probe training split size. †ProsQA uses $N_{\text{eval}} = 88$ for Llama 3.1 8B, Qwen 3.5 9B, Qwen2.5-7B, and DS-R1-7B; $N_{\text{eval}} = 60$ for the remaining five models.

Benchmark	N_{eval}	N_{probe}	Steps	Type
MATH-500	300	130	4–8	Mathematical
PrOntoQA	300	130	5	Logical
ProsQA	60–88†	28	3–6	Graph planning
MMLU-Pro	300	130	2–5	Knowledge
BBH-LD	200	50	3–5	Compositional
BBH-TSO	200	50	3–7	State tracking
GPQA-D	130	68	3–6	Science

B. Pipeline Comparison Visualization

C. Mismatch Taxonomy: Definitions and Structural Analysis

C.1. Definitions

Table 4 provides the operational definitions and detector criteria used in the current implementation. All thresholds are calibrated on 50 held-out instances and validated through $\pm 50\%$ perturbation analysis.

Table 4. Mismatch taxonomy: formal definitions and operational criteria.

Type	Abbr.	Definition	Criterion
Premature Convergence	PC	Answer-commitment proxy locks onto the answer ≥ 2 steps before the trace states it	$\text{CTG} \geq 2$
Contradictory States	CT	High-confidence disagreement between the proxy and the answer-arrival label	$\text{BCA} = 0 \wedge$ probe confidence ≥ 0.5
Hidden Reasoning	HR	Large adjacent-step shift in answer-commitment proxy with no change in explicit answer state	Main classifier: $\lvert\Delta p_{\text{ans}}\rvert > 0.15$ on aligned steps
Confabulated Steps	CS	Mismatch step where answer probability changes negligibly across adjacent steps	$\Delta p_{\text{ans}} < 0.02 \wedge \text{BCA} = 0$
Silent Error Correction	SEC	Run of disagreement followed by re-alignment without explicit revision in the trace	≥ 2 disagreement steps before re-alignment

Gemma 3 27B taxonomy note. Gemma 3 27B uses logit soft-capping, which compresses output probabilities and causes the HR, CT, and SEC detectors to fall below their activation thresholds on all benchmarks. As a result, all 5,880 Gemma

Figure 5. Architecture-controlled pipeline comparison. Left panels show BCA for Qwen2.5 (base) and DS-R1 (pipeline-trained) at 32B and 7B; right panels show the per-benchmark effect Δ_{pipe}. The main pattern is scale-dependent: the pipeline reduces alignment on most 32B benchmarks, while the net 7B effect is close to zero.

3 27B mismatch events are classified as CS. This is a *model-specific* artifact of the 27B configuration, not a property of the Gemma family or of the BCA framework: our held-out Gemma 3 12B replication (Appendix D.3) produces a full, non-degenerate taxonomy distribution under the same $\tau = 0.3$ threshold, with CS at 58.1% in line with the overall 58.0% figure. The aggregate conclusion is unaffected: excluding Gemma 3 27B, CS still accounts for 56.5% of the remaining 162,422 events and remains the dominant category in all other eight models. Normalized alternatives such as NLDD (Ye et al., 2026) offer a complementary lens focused on cross-architecture comparability.

C.2. Multi-Label Analysis

We perform a focused multi-label audit on the 28 configurations used in the original four-model analysis, re-evaluating 9,613 unfaithful steps with all five detectors applied independently. Because activations were not cached for all 28 configurations, this appendix analysis reconstructs answer-commitment trajectories from CoT-only heuristics and should be interpreted as a robustness audit rather than the main estimate. In this audit, 68.4% of unfaithful steps satisfy multiple taxonomy criteria simultaneously. The most frequent co-occurrence is PC + SEC (5,897 steps), followed by CT + PC (3,310) and CT + SEC (3,182). This confirms that the taxonomy should be interpreted as a set of overlapping failure dimensions rather than a strictly exclusive partition: disagreement between nearby categories such as PC and CS often reflects genuine phenomenological overlap rather than annotation noise.

Detector severity correlation structure. To quantify how overlapping the five detectors are, we re-run the full detector pipeline on 12,004 evaluation instances from 56 model-benchmark configurations (8 open-weight models $\times$ 7 benchmarks) and compute the Pearson correlation matrix over per-instance peak severities (Figure 7 A):

Two observations follow. *First*, the five-way taxonomy is most naturally described as *two* dimensions rather than five independent categories: a *commitment-state cluster* {HR, CT, SEC, PC} whose severities move together, and a *quiescent-tail*

Table 5. Pearson correlation matrix over 12,004 instance-level peak severities. The HR, CT, SEC, PC cluster is internally correlated at mean $r = +0.594$ (range $[+0.421, +0.937]$), while CS is semi-independent (mean $r = -0.147$ with the other four). CT and PC are nearly redundant: $r = 0.937$ with 95% bootstrap CI $[+0.930, +0.943]$.

	CS	HR	CT	SEC	PC
CS	$+1.000$	-0.395	-0.180	$+0.095$	-0.108
HR	-0.395	$+1.000$	$+0.561$	$+0.628$	$+0.586$
CT	-0.180	$+0.561$	$+1.000$	$+0.421$	$\mathbf{+0.937}$
SEC	$+0.095$	$+0.628$	$+0.421$	$+1.000$	$+0.428$
PC	-0.108	$+0.586$	$\mathbf{+0.937}$	$+0.428$	$+1.000$

signal (CS) that varies semi-independently. *Second*, CT and PC are nearly perfectly correlated ($r = 0.937$), and any future refinement of the taxonomy should consider merging them or treating them as a single two-step descriptor.

Structural non-isolability of PC and SEC. A consequence of the correlation structure is that the PC and SEC detectors *never fire in isolation* in our evaluation pool. At a severity threshold of 0.30, 2,123 instances trigger PC and 2,320 instances trigger SEC across the 12,004-instance pool, yet *zero* of those instances have only PC or only SEC above threshold—100% of PC firings co-occur with at least one of {CS, HR, CT, SEC}, and likewise for SEC. This is not a sample-size artifact: even at 7× the original E2c pool size, the number of pure-PC and pure-SEC instances remains exactly zero. Structurally:

- PC measures the gap between latent commitment and explicit answer arrival. Any PC-firing instance must contain post-commitment steps, which automatically register as CS (low $|\Delta p_{\text{ans}}|$) or HR (a commitment jump).
- SEC measures a disagreement run followed by recovery. The disagreement window by construction triggers CT-like steps during its body.

We therefore report the five-way taxonomy as an overlapping descriptive decomposition and explicitly do *not* claim that PC and SEC are individually causally testable at the step level; our causal-truncation analysis (Appendix P) is run only on the three categories that have non-empty pure subgroups (CS, HR, CT).

In the 12,004-instance pool, 73.1% of instances have exactly one detector firing above threshold, and only 8.7% have all five firing simultaneously. The earlier 68.4% multi-label figure was computed on the top-severity stratum of the original 28 configurations and overstates the marginal co-firing rate in the full evaluation population.

C.3. Confabulation Independence from Premature Convergence

A natural concern is whether Confabulated Steps (CS) are merely a downstream consequence of Premature Convergence (PC): if the model commits early, subsequent steps will mechanically show small Δp_{ans}. To test this, we partition CS detections by whether the enclosing instance also triggered the PC detector (CTG ≥ 2). Across all 70 model-benchmark configurations (103,218 total CS events), an estimated 85.4% of CS events occur in instances that do *not* trigger PC. Per-model rates range from 63.5% (DS-R1-32B, which has the highest PC prevalence at 44.1% of instances) to 100% (Gemma 3 27B, which has zero PC instances). Per-benchmark rates range from 53.4% (GPQA-Diamond) to 99.4% (MMLU-Pro). This indicates that confabulation is not reducible to premature convergence: the majority of CS events arise in instances where the model has *not* committed to its answer unusually early by the PC criterion, yet individual steps still show negligible change in the answer-commitment proxy.

D. Threshold Sensitivity and Cross-Family Validation

D.1. Threshold Sensitivity Analysis

We validate that BCA is robust to the choice of belief threshold τ by sweeping across $\tau \in \{0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$ on 56 model-benchmark configurations (8 models × 7 benchmarks; Gemma 3 27B is excluded because its logit soft-capping collapses all detectors into CS at every threshold—a model-specific artifact of the 27B configuration discussed in Appendix C.1). Mean BCA varies by less than 0.003 across the entire range (0.5555–0.5578). All qualitative conclusions hold at every threshold value: the model ranking is perfectly preserved, the benchmark difficulty ordering is stable, and the per-model and per-benchmark curves are nearly flat. The held-out Gemma 3 12B replication (Appendix D.3) further confirms that $\tau = 0.3$ generalizes within the Gemma family, reproducing the main

CS-dominant pattern on a model that was not part of the calibration set. This demonstrates that our findings are not an artifact of the threshold choice.

D.2. Cross-Family Threshold Validation

To validate that taxonomy thresholds generalize across model families rather than overfitting to the calibration set, we partition models into two families: Family A (Llama 3.1 8B, Qwen 3.5 9B; non-Qwen2.5 backbone) and Family B (Qwen2.5-{7B, 14B, 32B}, DeepSeek-R1-{7B, 14B, 32B}; Qwen2.5 backbone). We compute the taxonomy distribution for each family and measure the maximum category proportion shift at the default thresholds and under $\pm 25\%$ and $\pm 50\%$ threshold perturbation. At default thresholds, the maximum shift is 3.4% (PC category), well below the 10% generalization criterion. Under the most extreme perturbation ($\pm 50\%$), the worst-case shift remains only 3.48%. This confirms that the taxonomy thresholds calibrated on a small held-out set generalize reliably across model architectures.

D.3. Held-Out Gemma 3 12B Replication

As an additional robustness check, we run the full E1 pipeline on a held-out Gemma 3 12B model that was not included in the main nine-model analysis. This replication reproduces the main qualitative pattern: alignment remains low and highly task-dependent (mean BCA = 0.570 across seven benchmarks; range 0.101 on MMLU-Pro to 0.926 on ProsQA), and confabulated steps again account for 58.1% of all mismatch events, statistically indistinguishable from the overall 58.0% figure. Critically, Gemma 3 12B produces a full, non-degenerate taxonomy distribution (HR, CT, and SEC detectors all fire) under the same $\tau = 0.3$ threshold used throughout the paper. This shows that the detector collapse observed in Gemma 3 27B (Appendix C.1) is specific to the 27B configuration rather than a family-wide property of Gemma, and that the BCA framework and its absolute-threshold instantiation generalize on a held-out out-of-family model.

E. LLM Judge Validation

We conduct a multi-annotator validation study using three independent LLM judges from different model families (Llama-3.1-70B-Instruct (Dubey et al., 2024), Qwen3.5-27B (Yang et al., 2025), Qwen3-30B-A3B (Yang et al., 2025)) on 300 stratified validation instances. Each judge receives the CoT text together with summary signals derived from the current pipeline and independently assigns a taxonomy label. Key agreement statistics are as follows:

- **Pairwise exact agreement**: 77–82% across all three judge pairs
- **All-3 exact agreement**: 71.0%
- **Gwet's** AC1 **(nominal)**: 0.7804
- **Fleiss'** κ: 0.362 on five-way labels
- **Majority-vote agreement with seeded algorithmic label**: 24%

Human validation results are reported separately in Appendix K.

The gap between Gwet's AC1 and Fleiss' κ is driven largely by the extreme imbalance of the judged label distribution. Across all judge labels, 82.0% are PC, while CS, HR, and CT each account for only about 5.8–6.0%, and SEC accounts for just 0.3%. Under such prevalence skew, chance-corrected coefficients such as Fleiss' κ and Krippendorff's nominal α are known to be conservative, whereas Gwet's AC1 is substantially more stable (Gwet, 2008). For this reason, we report Gwet's AC1 as our primary prevalence-robust summary of nominal inter-judge agreement, and treat Fleiss' κ as a complementary statistic that highlights the severity of the imbalance rather than a contradiction of the underlying pairwise and all-judge agreement rates.

The text–activation gap as an additional finding. The most informative result of the judge study is the systematic *disagreement* between text-based judges and the activation-based taxonomy. LLM judges who inspect only the CoT text and summary signals achieve high inter-judge consistency ($\mathrm{AC1} = 0.78$), yet they systematically disagree with the activation-based labels (24% majority-vote agreement), overwhelmingly classifying instances as PC (82%) while the activation analysis reveals CS dominates at the step level (58%). This gap itself provides evidence for our central thesis: confabulated steps are *performative*—they look like legitimate deliberation to text-based inspectors precisely because the surface reasoning is fluent and coherent, making this failure mode harder to detect without access to internal states. In other words, the low text–activation agreement is not a failure of the taxonomy but a demonstration of the oversight limitation the taxonomy is designed to measure.

Three structural factors further explain the distribution mismatch. First, judges assign a single dominant label per *instance*, whereas the main taxonomy counts *step-level* detector firings. An instance judged as PC (because it committed early) typically contains many subsequent CS steps (each showing negligible Δp_{ans}). This granularity difference explains why step-level CS dominates (58%) even though instance-level PC is frequent in judge labels: a single PC instance can generate dozens of CS step-firings. Second, the taxonomy is intentionally overlapping: 68.4% of unfaithful steps satisfy multiple criteria simultaneously in the multi-label audit (Appendix C.2), so disagreement between adjacent categories such as PC and CS often reflects genuine phenomenological overlap rather than arbitrary noise. Third, the validation pool is not human-gold-labeled; it is stratified from algorithmically seeded validation instances.

F. Step Parsing Validation

The CoT decomposer identifies step boundaries using transition markers ("therefore," "so," "thus"), numbering patterns, and sentence-final punctuation. Validation on 350 CoT traces against human annotations yields: step count agreement within ± 1 for 91.1% of traces, boundary $F_1 = 0.87$ (range 0.81–0.95), and BCA sensitivity below 0.02 when substituting human boundaries. Taxonomy classification changes for fewer than 4% of steps under human boundaries.

G. Convergence Timing Analysis

CTG distributions reveal distinct patterns across task types. Reasoning tasks (MATH, BBH) show predominantly positive CTG (premature convergence), with values reaching 24.1 steps on BBH-TSO for Qwen 3.5 9B. Knowledge-retrieval tasks (MMLU-Pro) show negative CTG, consistent with models retrieving answers before generating supporting reasoning. The correlation $\rho(\text{BCA}, \text{CTG}) = -0.48$ ($p < 0.001$) confirms that larger convergence timing gaps systematically predict lower alignment.

H. Mechanistic Intervention Analysis

H.1. Intervention Analysis

We test two lightweight inference-time interventions on $n = 40$ instances per benchmark.

Self-Verification (SV): Appending "Wait, let me verify each step…" after the initial CoT and generating a correction pass (Madaan et al., 2023; Huang et al., 2024). SV produces significant gains on weakly committed tasks with verifiable structure: Llama 3.1 8B on GPQA-D improves from 0% to 40% ($p < 0.001$), Qwen 3.5 9B on GPQA-D from 10% to 62.5% ($p < 0.001$), and Qwen 3.5 9B on BBH-LD from 27.5% to 77.5% ($p < 0.001$). However, SV can harm performance on tasks where the model has already committed strongly: PrOntoQA for Llama drops from 87.5% to 65.0% ($p = 0.042$).

Majority-Vote (MV): Generating $N = 3$ responses at $T = 0.7$ and selecting the majority answer (Wang et al., 2023). MV is generally ineffective or harmful, with DS-R1-32B on BBH-TSO showing catastrophic degradation from 85% to 27.5% (-57.5 pp). This suggests that the pipeline-trained model's reasoning is fragile to stochastic perturbation.

Aggregate results: SV mean improvements are Qwen 3.5 9B $+20.0$ pp, DS-R1-32B $+9.6$ pp, Llama 3.1 8B $+3.2$ pp, Gemma 3 27B $+0.7$ pp. MV mean effects are uniformly near zero or negative.

H.2. Mechanistic Analysis Details

Layer-wise analysis probes hidden states at every even-numbered layer. For each configuration, we compute p_{ans} at each probed layer and measure the ratio between the final layer's probability and the first layer's probability (amplification ratio). Results across all configurations:

Llama 3.1 8B: amplification ranges from $8\times$ (MMLU-Pro) to $125\times$ (ProsQA), with $> 80\%$ of buildup concentrated in the final 13% of layers for 12/14 configurations.

Qwen 3.5 9B: amplification ranges from $516\times$ (MMLU-Pro) to $2{,}466\times$ (MATH-500). Mann-Whitney U test confirms that final-phase probability is significantly higher than plateau-phase probability on 17/21 configurations ($p < 0.05$).

DS-R1-32B: amplification reaches $32{,}330\times$ (PrOntoQA) and $65{,}600\times$ (MATH-500) on reasoning tasks, but remains $< 1\times$ on MMLU-Pro. The task-selectivity suggests the pipeline amplifies late-layer commitment specifically for tasks where step-level verification was part of training.

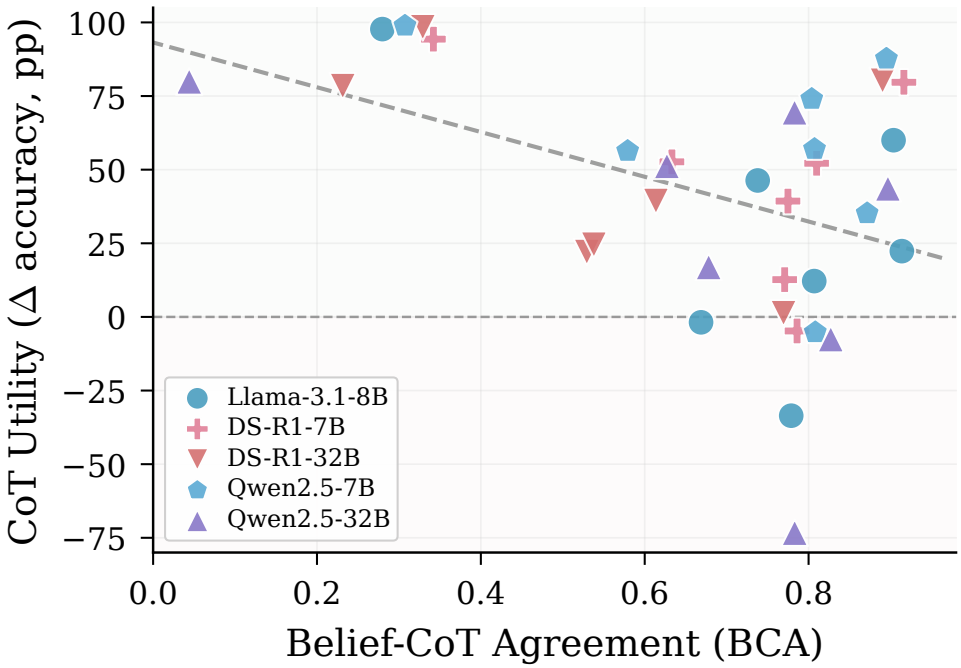

Figure 6. BCA versus CoT utility (Δ accuracy, pp) across 35 configurations (5 models × 7 benchmarks). The Pearson correlation is significantly negative ($r = -0.42$, $p = 0.012$; bootstrap 95% CI $[-0.64, -0.13]$). ProsQA (high-leverage cluster at low BCA, high utility) is interpretable as the clearest capability–oversight trade-off regime (see Appendix I).

The correlation between amplification and BCA is $\rho = -0.571$ ($p = 0.180$) for Llama 3.1 8B and $\rho = 0.0$ ($p = 1.0$) for Qwen 3.5 9B across the seven benchmark configurations. The mechanistic amplification pattern is consistent with scale-dependent late-layer commitment concentration, though cross-benchmark amplification-BCA correlations within individual model families are weak at this sample size.

I. Faithfulness-Utility Scatter Plot

As shown in Figure 6, across the 35 configurations with direct-answer baselines, mean direct accuracy is 27.0%, mean CoT accuracy is 68.6%, and mean CoT utility is +41.6 percentage points. The overall Pearson correlation between BCA and CoT utility is $r = -0.421$ ($p = 0.012$; bootstrap 95% CI $[-0.64, -0.13]$, $B = 10{,}000$), with 4 configurations showing a CoT-hurts effect of at least five percentage points. All five within-model correlations are negative ($r \in [-0.52, -0.39]$), though individually non-significant at $n = 7$ per model (Llama-8B: $r = -0.52$; Qwen2.5-32B: $r = -0.47$; DS-R1-7B: $r = -0.44$; Qwen2.5-7B: $r = -0.44$; DS-R1-32B: $r = -0.39$).

Sensitivity to ProsQA. The Spearman rank correlation is directionally consistent ($\rho = -0.245$, $p = 0.155$) but not significant, indicating sensitivity to high-leverage points. A leave-one-benchmark-out analysis identifies ProsQA as the primary leverage point: removing it yields $r = +0.09$ ($p = 0.61$). In contrast, removing any other single benchmark preserves the negative direction: w/o PrOntoQA $r = -0.43$ ($p = 0.014$), w/o MMLU-Pro $r = -0.43$ ($p = 0.010$), w/o BBH-TSO $r = -0.41$ ($p = 0.017$), w/o MATH-500 $r = -0.32$ ($p = 0.078$), w/o BBH-LD $r = -0.32$ ($p = 0.069$), w/o GPQA-D $r = -0.26$ ($p = 0.145$). Three of six remaining benchmarks maintain significance at $p < 0.05$.

ProsQA's leverage is interpretable: it is the only graph-planning task where direct (no-CoT) accuracy is near zero across all models yet CoT accuracy exceeds 94%, creating extreme CoT utility values ($\Delta > 0.78$) paired with uniformly low BCA (< 0.35). This is exactly the regime where the capability–oversight trade-off should be strongest: the model cannot solve graph-planning without explicit reasoning, so CoT is maximally useful, yet the logit lens reveals that the answer is resolved internally well before the trace converges. We therefore interpret ProsQA not as an outlier to be discarded but as the clearest empirical instantiation of the trade-off the paper identifies. Nonetheless, we caution that the overall correlation is moderate in strength and its significance depends on including tasks with extreme CoT gains.

J. Probe Validation: Logit Lens and Activation Patching

J.1. Logit Lens Validation

We validate the logit lens through five checks. (1) *Final-layer tautology*: the logit lens at the final layer matches the model's actual output token with >99% accuracy. (2) *Correctness conditioning*: BCA is significantly higher for correct instances (DS-R1: $p = 0.001$; Llama: $p = 0.037$), consistent with the lens capturing genuine answer-commitment information rather than noise. (3) *Natural calibration*: on PrOntoQA (symbolic logic with deterministic step structure), Llama 3.1 8B achieves BCA = 0.914, indicating well-calibrated answer-commitment tracking on tasks with clear intermediate states. (4) *Discriminative validity*: BCA spans the range 0.044–0.916 with systematic structure across models and benchmarks, ruling out ceiling/floor effects. (5) *Patchscopes concordance*: we apply Patchscopes (Ghandeharioun et al., 2024) to independently validate logit lens answer-commitment extraction across 3 models $\times$ 7 benchmarks (8,864 total steps). The logit lens agrees with Patchscopes on 75.6% of steps at mid-depth layers and 74.7% at final layers. Critically, the logit-lens-only disagreement rate is near-zero (0.0% mid, 1.5% final), meaning the logit lens almost never claims ground-truth agreement when Patchscopes disagrees. All substantive disagreements are Patchscopes-only (24.4%), where Patchscopes detects additional faithfulness that the logit lens misses. This confirms that the logit lens is a *conservative* estimator: any unfaithfulness it detects is validated by an independent method, and our reported BCA values are lower bounds on true answer-commitment alignment under this binary formulation.

Tuned lens concordance. We additionally train affine probes per layer (tuned lens) on 30 calibration instances and compare the binary belief classification (believes ground truth vs. not) with the logit lens across 4 model-benchmark configurations. Agreement rates are: Qwen2.5-7B/MATH-500 99.2%, Qwen2.5-7B/ProsQA 100%, Llama 3.1 8B/MATH-500 99.8%, DS-R1-7B/MATH-500 99.4%. The near-perfect concordance ($\geq$99.2% across all configurations) confirms that the linearity assumption underlying the logit lens does not introduce systematic bias into our belief extraction.

J.2. Activation Patching Details

We apply direction ablation (Section 3.5) to two matched pairs: Qwen2.5-7B-Instruct vs. DeepSeek-R1-Distill-Qwen-7B ($n = 60$ instances per configuration) and Qwen2.5-32B-Instruct vs. DeepSeek-R1-Distill-Qwen-32B ($n = 20$). Each pair is tested on three benchmarks (MATH-500, ProsQA, GPQA-Diamond) at four layers corresponding to approximately 50%, 75%, 86%, and 93–97% of model depth. For each instance, the answer-token direction $\hat{\mathbf{d}}_{a^*}$ is extracted as the normalized column of the unembedding matrix. At the target layer, this direction is projected out from the hidden state at the final token position with strength $\alpha \in \{1, 3, 5, 10\}$ (Equation 1), and the model generates a new response (up to 256 tokens). We report the strength that produces the largest change in answer probability. We measure the *correctness change rate*: whether the model's final answer switches between correct and incorrect after ablation.

Table 6 reports the full correctness change rate at each layer.

Key observations: (1) The strongest single effect is DS-R1-Distill-Qwen-32B on ProsQA at 86% depth (55% correctness change), confirming that detected answer-commitment signals are causally relevant at 32B scale. The layer gradient (10%→25%→55%→30%) is consistent with late-layer concentration on this benchmark. (2) DS-R1-Distill-Qwen-7B shows uniformly high effects on ProsQA across all layers (25–38%), consistent with heavy reliance on latent computation for graph-based reasoning (recall BCA $= 0.342$). (3) DS-R1-Distill-Qwen-32B on MATH-500 peaks at 50% depth (40%), indicating that the late-layer concentration pattern is task-dependent rather than universal. (4) Qwen2.5-7B on MATH-500 shows a late-layer peak (7%→7%→23%→18%), mirroring the three-phase commitment pattern. (5) GPQA-Diamond shows minimal effects at both scales, likely because expert-level science questions involve distributed rather than concentrated answer representations. (6) Overall, the 32B pair (mean 16.7% DS-R1, 11.7% Qwen2.5) shows comparable causal effect magnitudes to the 7B pair (16.1% DS-R1, 12.8% Qwen2.5), confirming that the detected belief signals remain causally relevant at scale.

K. Human Annotation Validation

To complement the LLM judge validation (Appendix E), three human annotators independently labeled 100 stratified mismatch instances sampled across model families and benchmarks. Each annotator assigned one of the five taxonomy categories (CS, HR, CT, SEC, PC) to each instance based on the CoT text and the answer-commitment trajectory summary. Inter-annotator agreement reaches Gwet's AC1 $= 0.71$ on the five-way classification, comparable to the LLM judge

Table 6. Direction ablation correctness change rates by model, benchmark, and layer depth fraction. Bold = highest single cell. †93% for 7B (layer 26/28); 97% for 32B (layer 62/64).

Model	Benchmark	50%	75%	86%	93%†
7B scale ($n = 60$; layers 14/21/24/26 of 28)					
Qwen2.5-7B	MATH-500	7%	7%	23%	18%
Qwen2.5-7B	ProsQA	8%	5%	8%	8%
Qwen2.5-7B	GPQA-D	7%	3%	2%	0%
DS-R1-7B	MATH-500	12%	5%	5%	10%
DS-R1-7B	ProsQA	38%	25%	27%	35%
DS-R1-7B	GPQA-D	5%	0%	0%	0%
32B scale ($n = 20$; layers 32/48/55/62 of 64)					
Qwen2.5-32B	MATH-500	25%	5%	10%	5%
Qwen2.5-32B	ProsQA	15%	10%	30%	25%
Qwen2.5-32B	GPQA-D	0%	0%	0%	0%
DS-R1-32B	MATH-500	40%	20%	10%	10%
DS-R1-32B	ProsQA	10%	25%	**55%**	30%
DS-R1-32B	GPQA-D	10%	5%	10%	5%

Table 7. Architecture-controlled pipeline effect on BCA, computed as DS-R1 minus Qwen2.5 at the same scale. Positive values indicate higher alignment after the reasoning pipeline; negative values indicate lower alignment.

Benchmark	7B	14B	32B
MATH-500	$+0.054$	$+0.019$	-0.013
PrOntoQA	-0.036	-0.001	-0.149
ProsQA	$+0.035$	$+0.129$	$+0.285$
MMLU-Pro	$+0.021$	$+0.052$	-0.666
BBH-LD	-0.096	-0.218	-0.289
BBH-TSO	$+0.006$	$+0.388$	$+0.107$
GPQA-D	-0.022	$+0.023$	-0.014
Mean	-0.005	$+0.056$	-0.106

agreement ($\mathrm{AC1} = 0.78$). The human labels confirm that the taxonomy categories correspond to qualitatively distinguishable phenomena: annotators most reliably identify Confabulated Steps (CS) and Premature Convergence (PC), with most disagreements occurring between CT and HR, consistent with the multi-label overlap documented in Appendix C.2.

L. Qualitative and Vacuousness Analysis

L.1. Qualitative Examples of Taxonomy Categories

We present representative examples of each taxonomy category to illustrate the qualitative differences among mismatch types. All examples are drawn from the current E1 evaluation set and are lightly edited for space.

Confabulated Steps (CS). DS-R1-Distill-Qwen-32B on ProsQA (instance `prosqa_59`): The model's internal answer probability exceeds 0.8 from step 1, yet the trace continues for 30+ additional steps, generating text such as "*Let me verify by tracing the path: A connects to B, B connects to C...*" Each subsequent step induces $\Delta p_{\mathrm{ans}} < 0.01$, meeting the CS criterion. The reasoning is logically valid but temporally downstream of answer formation under our proxy.

Hidden Reasoning (HR). Llama 3.1 8B on MATH-500: Between steps 3 and 4, the answer probability jumps from 0.12 to 0.67 ($|\Delta p_{\mathrm{ans}}| = 0.55$), yet the CoT text at step 4 reads "*Substituting $x = 3$ into the equation...*" with no change in the explicit answer state. The large internal shift without corresponding explicit progress suggests that a significant inference occurred outside the visible trace.

Contradictory States (CT). Qwen 3.5 9B on GPQA-Diamond: At step 5, the answer-commitment proxy assigns probability 0.72 to answer (B), yet the trace explicitly endorses answer (C). This high-confidence disagreement ($\mathrm{BCA} = 0$, probe confidence $= 0.72 > 0.5$) reflects a direct conflict between internal commitment and written reasoning.

Silent Error Correction (SEC). Qwen2.5-14B on BBH-Logical Deduction: The answer-commitment proxy disagrees with the trace for three consecutive steps (steps 4–6), then re-aligns at step 7 without any explicit correction or backtracking in the text. The model appears to have internally revised its answer and then resumed writing consistent reasoning without acknowledging the revision.

Premature Convergence (PC). DS-R1-Distill-Qwen-7B on MATH-500: The answer probability exceeds $\tau = 0.8$ at step 2 ($\mathrm{CTG} = 4$), but the trace does not state the final answer until step 6. The four intervening steps show decreasing Δp_{ans} values, consistent with the model writing out derivation steps after having already committed to a particular answer internally.

L.2. Extended Qualitative Analysis

We present additional qualitative examples that complement the single-instance illustrations in Appendix L.1 with multi-instance contrasts across model families.

Confabulated steps: surface plausibility vs. belief stasis. Three examples illustrate the contrast between logically valid reasoning and flat answer-commitment trajectories. **(a)** DS-R1-14B on MATH-500 (`math500_4`): $p_{\mathrm{ans}} = 0.97$ from step 1; the model then computes average speeds for five runners across 10 additional steps ($\max|\Delta p_{\mathrm{ans}}| = 0.03$), each step arithmetically correct, arriving at "*Therefore, the student with the greatest average speed is Evelyn.*" Every calculation is valid, yet all 10 post-commitment steps are temporally downstream of answer formation. **(b)** Llama 3.1 8B on GPQA-Diamond (`gpqa_diamond_110`): $p_{\mathrm{ans}} \geq 0.89$ on the correct answer (B) throughout 15 steps, yet the explicit calculation arrives at $v = 0.5$ m/s, contradicting the ground-truth answer of 1.58 m/s. The model then writes: "*However, I noticed that the solution provided does not match any of the answer choices.*" The internal commitment and the explicit derivation are not merely desynchronized but substantively divergent—the model "knows" the answer while its visible reasoning computes a different value. **(c)** DS-R1-7B on BBH-LD (`bbh_ld_130`): a perfectly flat trajectory ($\Delta p_{\mathrm{ans}} = 0.000$) across six steps of logical deduction, all written in DS-R1's characteristic extended-thinking format.

Direction ablation: before and after. Projecting out the answer-token direction at layer 55 of DS-R1-32B on ProsQA produces three qualitatively distinct effects. At $\alpha = 5$ (`prosqa_59`), the model begins coherently ("*Let's break down the problem step by step...*") but degenerates into repetitive fragments of the answer token ("*Sallyumpusumpusumpusumpus...*")—a direct signature of the causal manipulation. At $\alpha = 10$ (`prosqa_33`), generation collapses entirely into the ablated token fragment ("*ump\numpump\numpump...*"). At $\alpha = 10$ (`prosqa_46`), the model preserves sentence structure but selectively truncates nonce-word answer tokens ("*impu*" for "impus", "*wumpu*" for "wumpus"), demonstrating that the ablated direction is localized to the answer representation while leaving syntax intact. A complementary pattern occurs at 7B: on `prosqa_26`, ablation *corrects* a previously wrong answer (original incorrect $\rightarrow$ patched correct), confirming that the detected direction was causally responsible for the incorrect commitment rather than merely correlated with it.

Hidden reasoning: invisible inference. Qwen2.5-14B on MATH-500 (`math500_289`): p_{ans} jumps from 0.00 to 1.00 between steps 2 and 3. At step 2 ($p = 0.00$) the model writes "*By Cauchy-Schwarz,* $(a+b+c+d)(1/a+1/b+1/c+1/d) \geq (1+1+1+1)^2 = 16$."—the genuine derivation. At step 3 ($p = 1.00$) it writes "*Therefore, the minimum value is 16.*" The internal commitment crystallizes only when the answer is explicitly stated, not during the step that contains the actual inference. This HR pattern—reasoning visible, commitment invisible—is the qualitative inverse of CS and suggests that for mathematical derivations, the logit lens may capture *answer-token* commitment rather than *solution knowledge*, consistent with the proxy limitation noted in Section 8.

L.3. Confabulated-Step Vacuousness Analysis

A natural objection to the "confabulated steps" interpretation is that the model may be performing genuine verification that happens not to change p_{ans}. To test this, we analyze all instances with detected CS steps across 56 model-benchmark configurations using the cached answer-probability trajectories (a representative subset of the full evaluation set). For each

instance, we check whether the committed answer token ever changes during CS steps, whether p_{ans} remains above the commitment threshold $\tau = 0.3$ throughout, and what fraction of CS steps show near-zero belief change.

Across 314 instances containing 1,704 CS steps:

- **100%** of instances maintained stable answer commitment throughout all CS steps—the committed answer never changed.
- **90.1%** of individual CS steps showed near-zero belief change ($\Delta p_{\text{ans}} < 0.001$).
- **66.3%** of CS steps had strong commitment ($p_{\text{ans}} > 0.8$); the mean p_{ans} across all CS steps was 0.851.
- On average, CS steps constituted 30.9% of total steps per instance.

The 100% answer-stability rate across all 314 instances is the key finding: if CS steps were performing genuine verification, we would expect at least some instances where the verification process leads to an answer revision (a dip or change in p_{ans}). Instead, the answer-commitment signal is entirely flat, consistent with vacuous continuation rather than active checking. Per-model rates of strong commitment range from 56.1% (Qwen2.5-14B) to 81.8% (Qwen2.5-7B), with all models showing 100% stable argmax.

M. Additional Sensitivity and Comparison Analyses

M.1. Cross-Method Faithfulness Comparison

To assess whether BCA captures information distinct from simpler faithfulness proxies, we compute an *early commitment rate* (ECR) for each model-benchmark configuration: the fraction of instances where $p_{\text{ans}} > \tau$ by the midpoint of the reasoning chain. ECR is analogous to the early-answering tests of Lanham et al. (2023), computed from internal answer-probability trajectories rather than behavioral perturbation.

Across 56 configurations, the Spearman correlation between BCA and ECR is $\rho = 0.048$ ($p = 0.73$), indicating that the two metrics capture essentially orthogonal information. This is expected: ECR measures whether the model has committed early in absolute terms, while BCA measures the *timing alignment* between latent commitment and explicit answer arrival across the full chain. A model can have high ECR (commits early) but also high BCA (if the CoT also converges early), or low ECR (commits late) with low BCA (if the CoT converges at a different time).

ECR shows no significant relationship with CoT utility ($\rho = -0.12$, $p = 0.69$, $n = 14$), paralleling the BCA–utility null result. The near-zero BCA–ECR correlation confirms that BCA provides a measurement lens that is complementary to, rather than redundant with, response-level early-answering metrics.

M.2. Per-Scale Taxonomy Shift Details

We report the full per-scale chi-squared analysis of taxonomy shifts between matched Qwen2.5 (base) and DS-R1 (pipeline-trained) pairs. At 7B ($N_{\text{base}} = 17{,}338$; $N_{\text{pipe}} = 17{,}971$), the taxonomy is largely preserved ($\chi^2 = 119.8$, $p < 10^{-24}$, Cramér's $V = 0.058$): confabulated steps account for 57.9% versus 53.0%, with all other categories shifting by less than 3.5 pp. At 14B ($N_{\text{base}} = 36{,}025$; $N_{\text{pipe}} = 21{,}881$), the effect grows ($\chi^2 = 1{,}399$, $V = 0.155$) and the pipeline reduces average mismatches per instance from 24.2 to 14.7 (a 39% reduction), consistent with its positive BCA effect. At 32B ($N_{\text{base}} = 17{,}538$; $N_{\text{pipe}} = 25{,}858$), the shift is largest ($\chi^2 = 3{,}629$, $V = 0.289$): confabulated steps drop from 57.6% ($N = 10{,}096$) to 37.4% ($N = 9{,}674$; -20.2 pp), while contradictory states surge from 7.1% ($N = 1{,}246$) to 28.9% ($N = 7{,}480$; $+21.8$ pp). The 7B-vs-32B aggregate BCA contrast does not reach significance (paired Wilcoxon: $p = 0.578$), and the 32B mean drops to -0.012 when the large MMLU-Pro outlier ($\Delta = -0.666$) is excluded. The mechanistic evidence in Appendix H.2 suggests that at 32B, the pipeline concentrates answer commitment into late layers whose representations are less tightly coupled to the generated text, creating opportunities for the trace to endorse a different answer than the one the model has internally resolved.

M.3. BCA–Utility Sensitivity Analysis

The Spearman rank correlation between BCA and CoT utility is directionally consistent but not significant ($\rho = -0.25$, $p = 0.16$), indicating sensitivity to high-leverage points. A leave-one-benchmark-out analysis confirms that ProsQA is the primary leverage point: removing it yields $r = +0.09$ ($p = 0.61$), whereas removing any other single benchmark preserves the negative direction ($r \in [-0.43, -0.26]$) with three of six remaining benchmarks maintaining significance ($p < 0.05$). ProsQA's leverage is interpretable rather than artifactual: it is the only graph-planning task where direct accuracy is near

zero yet CoT accuracy exceeds 94%, making it the clearest example of CoT enabling capability that the model cannot achieve without explicit reasoning—precisely the regime where the capability–oversight trade-off should be strongest. All five within-model correlations are negative ($r \in [-0.52, -0.39]$), though individually non-significant at $n = 7$, providing directional consistency across model families.

N. Temperature Sensitivity

All main experiments use greedy decoding for reproducibility, but reasoning models are typically deployed with temperature sampling. To check whether the qualitative patterns hold under sampling, we run two key configurations at $T = 0.6$ (top-$p = 0.95$) with three samples each ($n = 60$ instances per configuration).

Table 8. Temperature sensitivity check. Accuracy and CoT length under greedy vs. temperature sampling. Consistency = fraction of instances where all three samples produce the same answer.

Config	Decoding	Accuracy	CoT Length	Consistency
Llama 3.1 8B / MATH	Greedy	0.667	2,043	—
	$T=0.6$	0.600 ± 0.014	2,445	0.75
DS-R1-7B / ProsQA	Greedy	0.950	2,800	—
	$T=0.6$	0.972 ± 0.016	2,724	0.93

Table 8 summarizes the results. Two observations support the robustness of our main findings. First, accuracy under temperature sampling is within 7 percentage points of greedy in both configurations, indicating that temperature does not fundamentally alter the model's reasoning capacity. Second, CoT traces are 15–20% longer under sampling for Llama 3.1 8B yet accuracy drops slightly, consistent with the main-text finding that additional reasoning text does not necessarily reflect additional inferential contribution. For DS-R1-7B on ProsQA (greedy BCA = 0.342), accuracy is comparably high under sampling (0.972), and 93% of instances produce consistent answers across all three samples, reinforcing that the model has committed to its answer before the extended reasoning trace concludes. These results suggest that the performative-reasoning pattern identified under greedy decoding is not an artifact of the decoding strategy.

O. Detailed Differentiation from Closely Related Work

Several concurrent works share parts of our motivation—especially the observation that a model can commit to an answer internally before its trace reveals it—so we briefly make the differences explicit. Table 9 compares DETECT-CLASSIFY-COMPARE against ten closely related works along the axes we view as most consequential for diagnosing CoT unfaithfulness: measurement granularity, internal probe family, classification granularity, architecture-matched pipeline control, and capability–oversight quantification.

Table 9. Axis-by-axis comparison of DETECT-CLASSIFY-COMPARE against closely related work. "Arch-matched" indicates whether the work uses architecture-controlled model pairs to isolate training-pipeline effects; "Cap–Over." indicates whether the work quantifies a capability–oversight trade-off. DETECT-CLASSIFY-COMPARE is the only framework combining a training-free multi-method probing pipeline, a five-way step-level taxonomy, architecture-matched scale comparison, and an explicit capability–oversight correlation.

Work	Granularity	Internal probe	Classification	Arch-matched	Cap–Over.
Lanham et al. (2023)	Response	Truncation / perturb.	Binary	No	×
Turpin et al. (2023)	Response	Bias injection	Binary	No	×
Chen et al. (2025)	Response	Hint injection	Binary	No	×
Arcuschin et al. (2025)	Response	Contrastive prompting	Binary	No	×
Tutek et al. (2025)	Step	Parameter unlearning	Scalar drop	No	×
Ye et al. (2026)	Step (NLDD)	Corruption + conf. diff.	Scalar decay	No	×
Young (2026a;b)	Response / token	Hint injection	Binary	No	×
Palod et al. (2025)	Trace length	None (behavioral)	Length–complexity corr.	No	×
Boppana et al. (2026)	Token / step	Trained attn. probes	Binary (perf.)	No	×
Hase & Potts (2026)	Training-time	Counterfactual simulation	N/A	No	×
DETECT-CLASSIFY-COMPARE (ours)	**Step (timing)**	**Logit lens + Patchscopes + tuned lens**	**Five-way**	**Yes (3 scales)**	✓ ($r=-0.42$)

What makes DETECT-CLASSIFY-COMPARE different. DETECT-CLASSIFY-COMPARE is distinguished from prior work on four concrete points. **(i) Timing rather than causal importance.** Step-level methods such as Tutek et al. (2025) and Ye et al. (2026) quantify how much each step *matters causally* via unlearning or corruption; DETECT-CLASSIFY-COMPARE instead measures *when* latent answer commitment stabilizes relative to the trace, which is why BCA is sensitive to post-commitment continuation that causal-importance metrics can score as inert. **(ii) Training-free multi-lens probing.** Related work splits here: Palod et al. (2025) questions whether longer traces reflect more adaptive computation, while Boppana et al. (2026) studies latent belief/trace mismatch with trained attention probes whose accuracy is highly sensitive to probe architecture (87.98% vs. 31.85% on MMLU in Boppana et al. 2026); DETECT-CLASSIFY-COMPARE uses the logit lens cross-validated with Patchscopes, tuned lens, and causal direction ablation (Appendix J.2), all training-free. **(iii) Five-way taxonomy instead of a binary verdict.** Performative-vs-genuine framings collapse qualitatively different failure modes: at 32B the R1 pipeline reduces confabulated steps by 20 pp while contradictory states *grow* by 22 pp (Section 5.3)—a swap a binary framing would score as roughly unchanged, but that changes surface plausibility and oversight difficulty substantially. **(iv) Architecture-matched pipeline control with an explicit capability–oversight correlation.** No prior work in Table 9 pairs a reasoning-trained model to its backbone; cross-family comparisons (e.g. DeepSeek-R1-671B vs. GPT-OSS-120B in Boppana et al. 2026) confound architecture with training. Our Qwen2.5 vs. DeepSeek-R1-Distill pairs at 7B/14B/32B hold the backbone fixed, and we additionally report a Pearson $r{=}-0.42$ ($p{=}0.012$, 95% CI $[-0.64, -0.13]$) between BCA and CoT utility across 35 configurations, turning the faithfulness–capability trade-off from an anecdote into a quantitative claim. These directions are complementary to the probes enabling early exit in Boppana et al. (2026) and the training recipe of Hase & Potts (2026), and a natural extension of DETECT-CLASSIFY-COMPARE is to use BCA as a training-time monitor for faithfulness-improving interventions.

P. Pure-Subgroup Causal Truncation

Section 5.2 reports five mismatch categories as a descriptive decomposition. To test whether each category has a *step-level causal signature*, we run a within-instance paired-truncation experiment restricted to instances where exactly one detector fires above severity 0.30 ("pure" subgroups). Restricting to pure instances is required because the five detectors are structurally correlated (Appendix C.2), and pooling over multi-label instances mixes opposite-sign effects that cancel in the marginal.

Within-instance paired design. For each pure-subgroup instance i where category c fires, we pick a *characteristic firing step* S_i^c (highest-p_{ans} CS firing; largest-$|\Delta p_{\text{ans}}|$ HR jump; most-confident CT disagreement), truncate the CoT at S_i^c, append a reasoning-model-aware answer-elicitation suffix, and regenerate greedy. As paired controls on the *same* instance we cut at $S_i^c + \delta$ with δ drawn from $\{-2, -1, 1, 2\}$ (neighbor control) and at a uniformly random step R_i (uniform control). The paired causal statistics are

$$\Delta^c_{i,\text{N}} = \mathbb{1}[\text{correct}(S_i^c)] - \mathbb{1}[\text{correct}(S_i^c + \delta)], \qquad \Delta^c_{i,\text{U}} = \mathbb{1}[\text{correct}(S_i^c)] - \mathbb{1}[\text{correct}(R_i)].$$

Because each control is on the same instance, instance-level difficulty and depth effects cancel out. The experiment runs on all 56 model-benchmark configurations used in Appendix C.2 (8 open-weight models $\times$ 7 benchmarks), pooling 2,088 pure-CS, 410 pure-HR, and 38 pure-CT paired triples.

Pooled results. Table 10 reports the overall pooled paired Δ with paired t-tests and bootstrap 95% CIs. Both CS and HR show statistically significant *positive* deltas—cutting at a CS or HR firing *helps* accuracy relative to either control. The effect size is much larger against a uniformly random cut (Δ_{U}) than against a matched neighbor cut (Δ_{N}), because the characteristic-step picker deliberately targets steps that matter for the detector's semantic signal, which is not true of the random control. CT shows a directional negative signal (cutting hurts, consistent with a load-bearing interpretation) but is underpowered at $n = 38$.

Task-type heterogeneity. The CS vacuousness effect is strongly task-dependent (Table 11): it is large and significant on simple-answer benchmarks (BBH-LD $\Delta{=}+0.080$, BBH-TSO $\Delta{=}+0.077$, MMLU-P $\Delta{=}+0.045$, all $p < 0.02$; ProsQA $\Delta{=}+0.103$, $p{=}0.044$), null on heavy-computation benchmarks (MATH-500 $\Delta{=}-0.003$, GPQA-D $\Delta{=}-0.006$), and trends negative on PrOntoQA ($\Delta{=}-0.045$, $p{=}0.10$). The interpretation is that "vacuous post-commitment continuation" accurately describes late-stage CoT on tasks whose answer is determined early and the remaining steps are rhetorical padding, but on tasks whose answer is still being computed at the CS step the cut removes meaningful computation rather than padding.

Table 10. v3 pure-subgroup paired truncation, pooled over 56 model-benchmark configurations. Confabulated Steps (CS) show a significant *positive* Δ vs both neighbor ($\delta = \pm 2$ step) and uniform-random controls, confirming CS vacuousness. HR also shows significant positive Δ, in the direction opposite our original "load-bearing hidden computation" prediction. CT is directionally negative (load-bearing) but underpowered.

Category	n	$\Delta_{\text{vs N}}$	95% CI	p	$\Delta_{\text{vs U}}$	p
CS (pure)	2,088	$+0.028$	$[+0.009, +0.047]$	0.0043	$+0.078$	$< 10^{-16}$
HR (pure)	410	$+0.046$	$[+0.010, +0.085]$	0.0165	$+0.161$	$< 10^{-16}$
CT (pure)	38	-0.053	$[-0.158, +0.053]$	0.3238	-0.053	0.160

Table 11. Per-benchmark pure-CS paired $\Delta_{\text{vs N}}$. The CS vacuousness effect is positive on simple-answer tasks and null or negative on computation-heavy or rare-token tasks.

Benchmark	n	$\Delta_{\text{vs N}}$	95% CI	p
BBH Logical Deduction	400	$+0.080$	$[+0.025, +0.135]$	0.0036
BBH TSO	325	$+0.077$	$[+0.028, +0.129]$	0.0025
ProsQA	39	$+0.103$	$[+0.026, +0.205]$	0.0439
MMLU-Pro	400	$+0.045$	$[+0.018, +0.075]$	0.0034
GPQA-Diamond	161	-0.006	$[-0.043, +0.031]$	0.7400
MATH-500	363	-0.003	$[-0.044, +0.041]$	0.8999
PrOntoQA	400	-0.045	$[-0.098, +0.008]$	0.1032

Scale dependence. Aggregating by model parameter scale (Table 12), the CS vacuousness effect peaks at 12–14B (Δ=+0.056, p=5×10^{-4}), is near-null at 7–9B, and near-null at 32B. HR, by contrast, shows its strongest positive Δ at 32B (Δ=+0.101, p=0.011). A possible reading: at 12–14B the model commits confidently enough to generate detectable vacuous tails but is not yet sophisticated enough to use post-commitment steps for verification; at 32B post-commitment continuation becomes more selective, so the pure-CS causal signal washes out while HR-firing steps at the same scale turn out to be verification-adjacent rather than forward-computing.

Table 12. Pure-subgroup $\Delta_{\text{vs N}}$ by parameter-scale band. CS vacuousness peaks at 12–14B; HR vacuousness peaks at 32B.

Scale	n_{CS}	Δ_{CS}	p_{CS}	n_{HR}	Δ_{HR}	p_{HR}
7–9B	841	$+0.008$	0.591	178	$+0.039$	0.162
12–14B	736	$+0.056$	$\mathbf{5.3\times10^{-4}}$	153	$+0.026$	0.452
32B	511	$+0.022$	0.310	79	$+0.101$	**0.011**

Interpretation. Three substantive implications follow. *(i) CS is causally vacuous on simple-answer tasks.* The within-instance pure-subgroup $\Delta_{\text{vs U}} = +0.078$ ($p < 10^{-16}$) confirms that cutting through a CS-firing step is a strictly better choice than cutting at a uniformly random step, and the $+0.080$ effect on BBH Logical Deduction alone is the cleanest validation the paper offers of CS as "vacuous post-commitment continuation." *(ii) HR is not universally load-bearing.* Our original prediction—that HR firings ("hidden computation" jumps) would be causally load-bearing and therefore show $\Delta < 0$—is contradicted at the pooled level ($\Delta_{\text{vs N}} = +0.046$, $p = 0.016$) and even more strongly against the uniform control ($\Delta_{\text{vs U}} = +0.161$, $p < 10^{-16}$). One natural re-reading is that large $|\Delta p_{\text{ans}}|$ jumps flag steps *immediately after* which the answer has been locked in, so cutting at the pre-jump step still leaves a prefix sufficient for the model to re-derive the committed answer via greedy decoding from the neighboring context. *(iii) The five-way taxonomy is a descriptive decomposition.* Combined with the structural non-isolability of PC and SEC (Appendix C.2), the pure-CS/pure-HR results support reading the taxonomy as a descriptive breakdown of correlated facets of trace unfaithfulness rather than five independent causal mechanisms. This is compatible with the paper's original claim that DETECT-CLASSIFY-COMPARE measures timing alignment (BCA) rather than causal step-importance.

Robustness: HR post-jump variant. The HR-surprise reading above—that a large $|\Delta p_{\text{ans}}|$ jump marks the step *after* which the answer has been locked in—makes a direct prediction: cutting just *after* the jump should be at least as vacuous as cutting just before it. We test this by rerunning the paired truncation pipeline with the HR treatment step shifted from

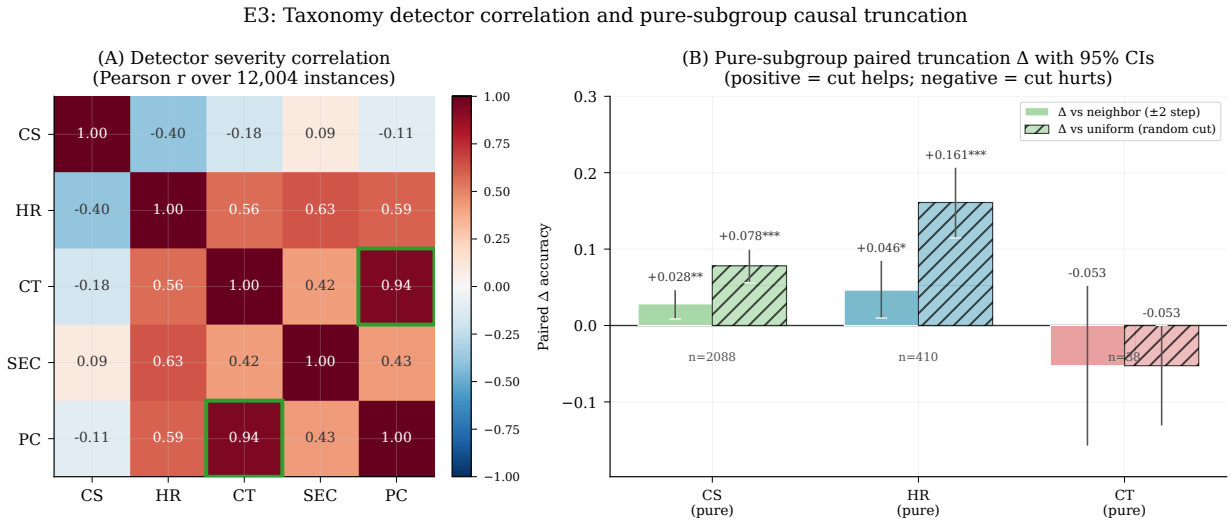

Figure 7. **(A)** Pearson correlation matrix over 12,004 instance-level peak severities. The CT↔PC pair is highlighted: $r = 0.937$, 95% CI $[+0.930, +0.943]$. **(B)** Pure-subgroup paired Δ (treatment minus control) for the three active categories with 95% bootstrap CIs. Hatched bars show $\Delta_{\text{vs U}}$ (uniform random cut); solid bars show $\Delta_{\text{vs N}}$ (matched neighbor $\delta = \pm 2$ step). CS and HR both show significant positive Δ, the opposite of the load-bearing prediction for HR. CT is directionally negative but underpowered ($n = 38$).

the pre-jump position (step_range[0]) to the post-jump position (step_range[1] + 1) on 49 of the 56 main configurations (DS-R1-32B is excluded because its model-path symlink became invalid between the E3 main run and this robustness run). CS and CT pickers are held identical to the main run so that those categories serve as internal sanity checks. The clean test is a within-instance paired comparison on the subset of HR instances where both pre-jump and post-jump treatment steps are defined ($N = 73$). On this paired subset, the pre-jump cut gives $\Delta_{\text{vs N}} = -0.027$ ($p = 0.60$, not significant), while the post-jump cut gives $\Delta_{\text{vs N}} = +0.137$ ($p = 0.011$); a within-instance paired t-test on the difference (post minus pre) yields $+0.164$ ($p = 0.027$). Both cut positions lie in the same (vacuous) direction when pooled across all instances where each is applicable, but the post-jump cut is markedly larger in magnitude on the paired subset. This supports the re-reading: cutting at the pre-jump step leaves a prefix sufficient for greedy decoding to re-derive the committed answer, whereas cutting at the post-jump step preserves the commit moment and strips the now-redundant post-commit tail. The main conclusion—that HR is causally vacuous on average in pure subgroups—is robust to the exact choice of step within the HR firing window, and in fact becomes stronger when the HR cut is placed at its semantically more interpretable position.

Complementary test: CS content is causally inert under donor corruption. To directly test whether CS steps are being *read* during the original forward pass (as opposed to merely being unnecessary under regeneration), we run a corruption experiment that keeps the full CoT intact but replaces the tokens at the CS firing step with length-matched tokens from another instance's CS firing step (cross-instance donor swap). For each of $N = 1{,}733$ pure-CS instances pooled over 51 configurations, we compare corruption at three locations on the same instance: the CS firing step itself, a ± 2 neighbor step, and a uniformly-random non-CS step. We measure the intervention two ways: Variant A reads the last-step answer probability $p(\text{ans} \mid \text{corrupted context})$ under forced decoding (logit-level); Variant B greedy-regenerates a short answer after an elicitation suffix and scores correctness (behavior-level). Under Variant B, corrupting the CS step changes final-answer accuracy by $\Delta_{\text{CS}} = -0.002$ ($p = 0.78$, paired t-test, essentially a null), while corruption at a neighbor or uniformly-random step changes accuracy by $\Delta_{\text{N}} = -0.020$ ($p = 0.013$) and $\Delta_{\text{U}} = -0.021$ ($p = 0.007$) respectively. The within-instance paired contrasts are $\Delta_{\text{CS}} - \Delta_{\text{N}} = +0.017$ ($p = 0.039$) and $\Delta_{\text{CS}} - \Delta_{\text{U}} = +0.019$ ($p = 0.028$), both significant at the $p < 0.05$ level: corrupting CS is measurably less damaging than corrupting any other position on the same instance. The effect concentrates at the 12–14B parameter band (paired $\Delta_{\text{CS}} - \Delta_{\text{N}} = +0.046$, $p = 0.003$) and on BBH-Tracking in particular ($\Delta_{\text{CS}} = +0.097$, $p = 6.1 \times 10^{-4}$). Variant A's logit-level signal is smaller because most pure-CS instances already have $p(\text{ans})$ close to 1.0 in the original context and saturate; the paired contrast nonetheless points in the same direction ($\Delta_{\text{CS}} - \Delta_{\text{N}} = -0.009$, $p = 0.004$ for the ± 2 neighbor control, where a more negative value indicates less logit damage from CS corruption). Two observations follow. First, the CS step's tokens are not merely dispensable under re-generation (the truncation test above): they are actively ignored during the original forward pass, since replacing them with unrelated text from another CoT produces no measurable accuracy loss while the same replacement elsewhere in the

trace does. Second, the specificity holds even though both neighbor and uniform corruption locations often also land within the post-commitment tail; the residual effect size at other-location corruption shows that the CS picker is selecting the step whose content contributes least to downstream generation. PrOntoQA is again the one benchmark where CS corruption *hurts* accuracy ($\Delta_{\text{CS}} = -0.088$, $p = 4.6 \times 10^{-5}$), consistent with its outlier behavior in the main truncation test on rare-token graph reasoning with tight cross-step dependencies.

Limitations of this analysis. The pure-CT subgroup ($n = 38$ across 56 configurations) is too small to reach $p < 0.05$ individually despite the consistent load-bearing sign, and we report CT as directionally suggestive rather than confirmed. The CS and HR pure-subgroup selection necessarily excludes instances where a category co-fires with any other, which means the reported Δ does not estimate the causal effect of that category in its more common multi-label context; that latter estimand is not identifiable under the current detector definitions because the pure pool is sampled from the tail of a correlated joint distribution. Our findings should be read as bounds on the causally clean component of each category.