Beyond the Commitment Boundary: Probing Epiphenomenal Chain-of-Thought in Large Reasoning Models

Anonymous ACL submission

Abstract

Chain-of-thought (CoT) reasoning is the dominant paradigm for inference-time scaling in language models, yet the causal influence of individual steps on the final answer poorly understood. We estimate each step's causal importance via early exit and use this measure to study how answers form across the reasoning traces of several model families. Across diverse tasks, we find that reasoning typically crosses a *commitment boundary* — a sharp transition from transient intermediate guesses to a stable, high-confidence answer. This transition often happens in a single step, well before the model's reasoning block ends, and is followed by *epiphenomenal* CoT steps that leave the final answer probability unaltered. Using attention probes, we show that answer-formation stages can be linearly decoded from intermediate reasoning steps with high accuracy and generalize robustly to unseen reasoning tasks. We exploit this signal to early-exit reasoning blocks at the commitment boundary, reducing the length of CoTs up to 55% on average with negligible impact on model performance. [1]

1 Introduction

Large reasoning models produce Chain-of-Thought (CoT) traces before committing to a final answer (Wei et al., 2022; OpenAI, 2024; Snell et al., 2025; Guo et al., 2025). While CoT reasoning may seem well-suited for safety monitoring thanks to its legible outputs (Korbak et al., 2025), a growing body of work highlighted how CoT outputs do not provide faithful accounts of the model's internal computations (Turpin et al., 2023; Agarwal et al., 2024; Madsen et al., 2024; Tutek et al., 2025). Notable examples of unfaithful CoT include deliberative language with expressions of uncertainty not being reflected in model predictions (Venhoff et al., 2025) and *performative* steps in which model internals encode the final answer well before its explicit verbalization (Boppana et al., 2026).

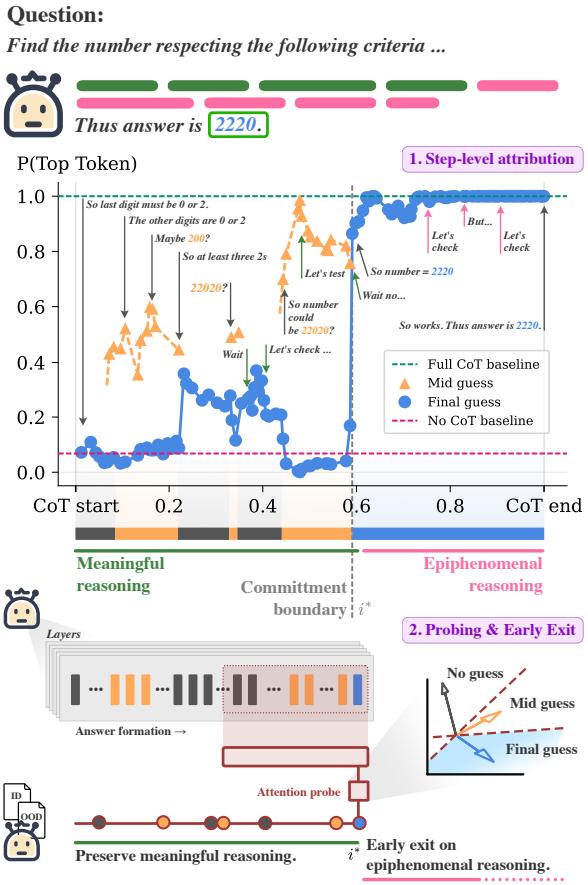


Figure 1: Overview of our approach. **Top:** We use early exit to measure the causal contribution of CoT steps to the model's **final answer** and **mid-guesses** probabilities. We frequently encounter a *commitment boundary* i^*, marking a sharp transition from meaningful reasoning with mid-guesses to a final answer at full-CoT confidence. **Bottom**: We train lightweight attention probes to predict answer-formation stages from model activations, and use them to early-exit reasoning at i^* to skip performative continuations at little performance cost.

In this work, we ask: can we track how models converge to their final answer across CoT steps and precisely locate where they stabilize? If so, can we predict when that transition occurs?

[1] Code and data will be released upon acceptance.

We address this via a step-level causal framework that uses CoT early exits to measure shifts in the probability of the model's *final* answer and *mid-guesses* being temporarily entertained at each reasoning step (Figure 1, top). This yields a causally grounded, fine-grained signal to monitor answer formation throughout reasoning. Across three model families and four popular reasoning tasks, we find that final-answer commitment typically occurs after a single, pivotal reasoning step, producing a large shift in final-answer probabilities. We dub this transition the **commitment boundary** at step i^*, and show that, on average, it falls around the *midpoint* of the CoT. Beyond i^*, models often engage in **epiphenomenal reasoning**: despite frequent hedging and re-verification, the final answer probability remains essentially unaltered. After annotating CoT steps with their respective answer-formation stages (no answer, mid-guess, or final answer), we show that these can be reliably predicted using lightweight attention probes trained on model activations, achieving high precision across in- and out-of-distribution reasoning tasks. Finally, we use these probes to identify the commitment boundary and perform CoT early-exiting at inference time, avoiding superfluous continuations while incurring minimal performance degradation. We summarize our contributions as follows:

- We introduce a causal framework for step-level analysis of CoT reasoning traces, tracking step-level final commitment and intermediate hypothesis changes in any closed or open-ended verifiable task (§3).
- We identify and characterize a commitment boundary i^* across three models and four reasoning tasks (§4), highlighting the causal effect of pre-boundary reasoning and redundancy of post-boundary tokens (§4.1, §4.2).
- We demonstrate that i^* can be reliably located **in- and out-of-distribution** using lightweight attention probes, suggesting a general structure for answer commitment information across reasoning tasks (§5).
- We demonstrate that probes enable CoT early-exiting, saving up to 55% of the reasoning trace with minimal answer accuracy loss (§6).

2 Related Work

A growing body of work has investigated whether CoT traces faithfully reflect the model's internal reasoning process, and more recently, whether individual steps can be attributed a causal role in producing the final answer.

Faithfulness in CoT reasoning. A recurring finding is that CoT explanations provide a false sense of transparency: models asked to explain their reasoning are often inconsistent and unreliable (Madsen et al., 2024), and the generated explanations do not necessarily align with the underlying reasoning process (Agarwal et al., 2024). At a finer grain, Venhoff et al. (2025) identifies specific linguistic patterns in CoTs that superficially resemble deliberate reasoning but whose causal status remains unclear. Other efforts to improve CoT faithfulness through in-context learning, fine-tuning and activation editing have met limited success, suggesting fundamental limitations (Tanneru et al., 2024; Paul et al., 2024; Shen et al., 2026). The ambiguity between genuine and epiphenomenal reasoning is a central concern our work addresses directly.

Causal attribution in reasoning traces A complementary line of work moves beyond behavioral analysis and attempts to assign causal credit to individual CoT steps. Zhao et al. (2026) identify unfaithful steps by perturbing reasoning steps and observing the downstream effect on the final answer. Lanham et al. (2023) evaluate perturbation strategies for measuring CoT faithfulness, finding significant variation across approaches, while Tutek et al. (2025) use erasure techniques to measure CoT faithfulness for individual reasoning steps. Such perturbation-based methods are inherently noisy: results depend on how the model interprets the intervened CoT, and their causal interpretation is confounded by the model's ability to reason implicitly without explicit CoT (Deng et al., 2023; Yee et al., 2024). We sidestep these issues by *truncating* rather than perturbing the trace, and causally estimate the model's confidence in its answers as a function of CoT prefixes of increasing length across reasoning steps.

Interpretability of reasoning Recent work has begun decoding the model's internal state to shed light on CoT faithfulness beyond the generated text. Wang et al. (2025) exploit mid-layer confidence signals to show that model confidence can be predicted throughout reasoning. Bogdan et al. (2026) suppress attention between reasoning steps to show that a small subset of sentences has outsized influence on the final answer distribution.

Concurrent to our work, Boppana et al. (2026)

define performative reasoning as the accuracy gap between a CoT monitor and a probe,[2] showing that the final answer is decodable from reasoning step activations before it is explicitly mentioned in the trace. Our notion of *epiphenomenal reasoning* is strictly broader, including also steps that — while appearing to challenge or revise the model answer through hedging, rechecking and expression of uncertainty — do not impact the decoded answer elicited after truncation. While Boppana et al. (2026) train probes to predict a limited set of answers from multiple-choice options, we instead focus on predicting whether an answer is temporary or final to identify genuine hypothesis revision across CoT traces. Finally, while their work finds performative reasoning to be difficulty-dependent, we find the model family to have a stronger effect on epiphenomenal reasoning.

CoT Efficiency. A shared implication of these findings is that not all CoT steps carry equal informational value, motivating work on reducing the computational cost of long reasoning traces via confidence signals (Scalena et al., 2026; Fu et al., 2026), compression (Li et al., 2026; Xia et al., 2025) or internal geometry (Ge et al., 2024; Liu et al., 2025). Rather than treating token reduction as the primary goal, we use early exit to confirm the redundancy of post-boundary tokens, thereby grounding inference-time savings in a better understanding of answer-commitment mechanisms.

3 Measuring CoT Answer Formation

To verify the presence of epiphenomenal reasoning in CoT traces, we introduce a truncation-based causal framework measuring the marginal contribution of each CoT step to the model's final answer.

Step-Level Attribution. Let P be a set of prompt tokens and C be the resulting CoT.[3] To isolate the effect of the CoT from the answer-generation phase, we construct the answer-forcing sequence $X_{\text{full}} = P + \texttt{[BOT]} + C + \texttt{[EOT]} + S$ where `[BOT]` and `[EOT]` are model-dependent beginning- and end-of-thinking markers, $+$ denotes sequence concatenation, and S is a task-specific suffix that elicits a direct answer.[4]

We segment each decoded CoT C into n sentence-level spans by splitting on terminal punctuation and construct $n+1$ prefix sequences as:

$$X_i = P + \texttt{[BOT]} + C_i + \texttt{[EOT]} + S \qquad (1)$$

where C_i is C truncated at the i-th sentence-level span, X_0 is the no-CoT baseline and $X_n = X_{\text{full}}$ is the full-CoT condition. Importantly, we restrict our analysis to traces in which intermediate reasoning improves the output, dropping those in which the model answers correctly when prompted with X_0.

For each prefix X_i, we greedily decode the model's answer $\hat{A}_i$ and compare whether it is semantically equivalent to the full-CoT answer $\hat{A}_n$ using a verifier $\equiv$ (see Appendix A). Our goal is not to verify the model's correctness, but rather to track the step at which its answer stabilizes to $\hat{A}_n$. This decoupling ensures that ground-truth correctness is not a confounding factor in our assessment.

We define the *step confidence* at CoT step i as:

$$p_i = \arg\max P(\cdot \mid X_i), \quad 0 \le i \le n, \qquad (2)$$

as a scalar proxy for the model's commitment to any answer after reading the first i CoT sentences. To prevent spurious high confidence caused by token collisions when different answers $\hat{A}_i \not\equiv \hat{A}_n$ share the same leading token, we discard traces containing this collision for at least one prefix X_i.[5]

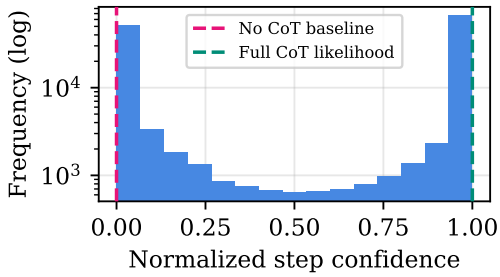


Figure 2: **Answer confidence in reasoning is bimodal.** Normalised step confidences $\tilde{p}_i$ across all CoT steps on `gpt-oss-20b` MATH-500 traces. Probability mass concentrates near 0 (no-CoT baseline) and 1 (full-CoT).

Answer formation stages We categorize each step C_i as follows, using a pre-defined threshold τ:

- **No-guess** ($p_i \le \tau$): the model does not commit to any answer.
- **Mid-guess** ($p_i > \tau$ and $\hat{A}_i \not\equiv \hat{A}_n$): the model is confident on a mid-guess, but has not settled on its final answer.
- **Final-guess** ($p_i > \tau$ and $\hat{A}_i \equiv \hat{A}_n$): the model's answer matches the full-CoT answer.

[2] Or an answer from a truncation procedure similar to ours.

[3] C includes model-specific markers such as `<think>`.

[4] For mathematical and multiple-choice tasks we use the suffix `"Therefore, the final answer is \boxed{"`

[5] Collision rates with final guesses are, on average, 9%, but they vary significantly across examples and tasks. See Appendix B for details.

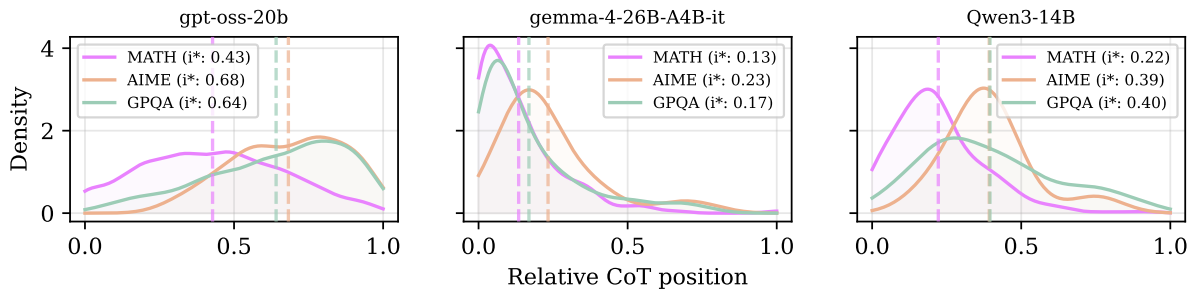


Figure 3: Confidence improvement over CoT tokens, across models and datasets. The relative CoT position of the commitment boundary (i^*/n) varies greatly across model families, but generally occurs well before the CoT end.

For our main experiments, we set the threshold as $\tau = p_0 + \frac{1}{2}(p_n - p_0)$, i.e., the midpoint between the no-CoT and full-CoT confidences, to reflect the CoT confidence gain in individual traces. Appendix C shows that our conclusions remain robust regardless of the exact value of τ.

4 Locating the Commitment Boundary

We evaluate `gpt-oss-20b` (OpenAI, 2025), `Qwen3-14B` (Yang et al., 2025) and `gemma-4-26B-A4B-it` (Google, 2026) across four reasoning datasets: **MATH-500** (Lightman et al., 2023), comprising 500 competition-style mathematics problems, **AIME 2025** (Zhang and Math-AI, 2025), with harder, more recent mathematical problems, **ZebraLogic** (Lin et al., 2025), with open-ended logical reasoning problems requiring multi-step deductive inference, and **GPQA-Diamond** (Rein et al., 2024), with multiple-choice graduate-level questions across biology, chemistry, and physics.

We begin by observing that `gpt-oss-20b` normalized step confidences $\tilde{p}_i = (p_i - p_0)/(p_n - p_0)$ across a subset of MATH-500 traces is strongly bimodal (Figure 2): mass concentrates for the most part near $\tilde{p}_i = 0$ (no-CoT level) and $\tilde{p}_i = 1$ (full-CoT level). This suggests that the final answer emerges abruptly as the preferred choice, rather than gradually. To quantify this phenomenon, we compute the confidence improvement $\Delta_i = p_i - p_{i-1}$ at each step i for the same model, showing that $\max_i \Delta_i$ is on average 4.6 times larger than the second-largest step improvement (median; 95% CI: $[4.3, 4.9]$; $n = 3{,}119$ traces). These results indicate that in the majority of CoT traces, a single sentence C_{i^*} drives the model from the no-guess regime to its final answer. We term this sentence *commitment boundary*, found at position:

$$i^* = \arg\max_{1 \le i \le n} \Delta_i, \tag{3}$$

We find that i^* is on average closer to the midpoint of the CoT than to its end, meaning a substantial fraction of CoT tokens are generated after the answer has already stabilized (Figure 3). In particular, we show that *the position of i^* is mainly influenced by model family*, with `gemma-4-26B-A4B-it` often converging to the final answer after only $13-23\%$ of CoT tokens have been produced, while `gpt-oss-20b` traces converge to the final solution much later ($43-68\%$). This said, across all models and datasets, we find that *a significant portion of the CoT is produced after the model reaches its final solution*. Finally, we also observe that post-boundary confidence for steps $C_{>i^*}$ remains above τ across all CoT traces in our evaluation datasets, suggesting that later tokens do not significantly impact the model's final answer.

4.1 Stress-testing the Commitment Boundary

The truncation result above shows that the prefix ending at i^* is sufficient for reproducing the full-CoT answer. As a complementary stress test, we ask whether the two sides of the boundary behave differently under targeted perturbations of the reasoning text. We focus on AIME2025 traces, where CoTs contain many numeric literals. Following Zhao et al. (2026), we randomly replace a fixed percentage of eligible numbers by adding a small integer offset sampled from $\{-5, \ldots, -1, +1, \ldots, +5\}$. We then append the usual answer-forcing suffix and greedily decode a new answer, counting a run as preserved if the new boxed answer is mathematically equivalent to the original full-CoT answer $\hat{A}_n$.

We compare two interventions, focusing on `gpt-oss-20b` for this analysis. In the POST condition, we perturb only numbers occurring in sentences after the commitment boundary, $C_{>i^*}$, while keeping the full CoT. In the PRE condition, we keep only the prefix X_i^* up to the boundary, and perturb

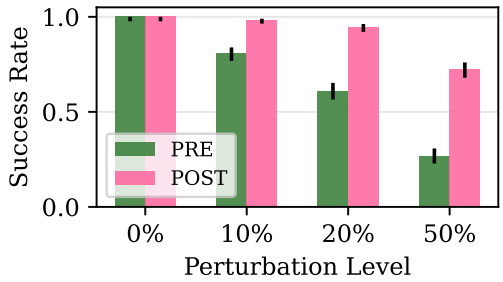


Figure 4: **Perturbations to $C_{<i^*}$ are most damaging.** Fraction of `gpt-oss-20b` AIME2025 traces whose elicited answer stays $\equiv \hat{A}_n$ under numeric corruption of the pre- (PRE) and post-boundary (POST) tokens ($n = 158$, three samples per setting).

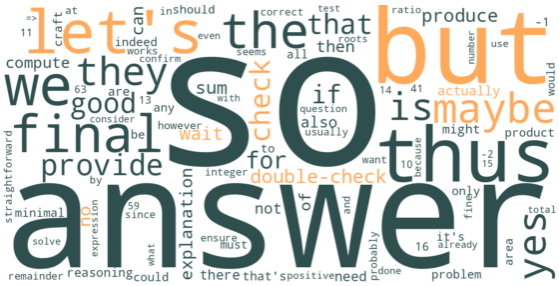


Figure 5: **Hedging language is frequent in post-commitment steps.** Word cloud of content words appearing at the beginning of post-commitment sentences $C_{>i^*}$ across all `gpt-oss-20b` MATH-500 traces. Words associated with self-verification (e.g., *"but"*, *"let's check"*) are disproportionately frequent.

only numbers inside it. The unperturbed PRE baseline reproduces the full-CoT answer on all retained traces, confirming that i^* is indeed a sufficient decision point before any perturbation is applied.

Figure 4 shows that corrupting the prefix up to i^* is much more destructive than corrupting the post-boundary tail. At 20% corruption, PRE preserves the answer in only 61% of runs, compared to 95% for POST; at 50%, the gap remains large (27% vs. 72%). These results provide strong evidence for a phenomenon we term **epiphenomenal reasoning**: after the model forms its final answer, the remaining tail of reasoning steps has little to no impact on its final prediction. We note this significantly extends the *performative reasoning* phenomenon highlighted by Boppana et al. (2026), where intermediate reasoning steps were shown to internally encode the final answer before its explicit mention. In particular, we find that the epiphenomenal tail of reasoning traces often includes steps where the model explicitly challenges or rechecks its own answer, language that superficially resembles genuine deliberation, corroborating previous results by

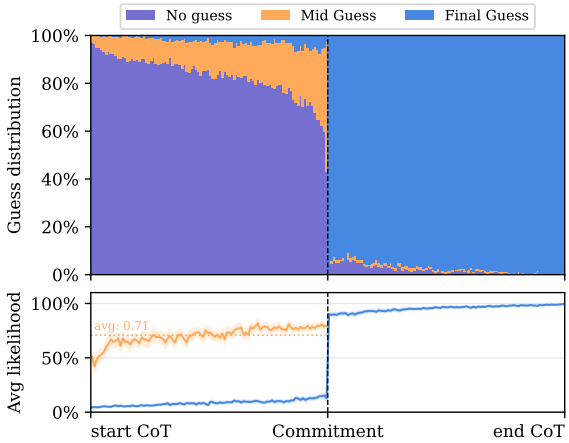


Figure 6: **Pre-boundary steps show genuine answer uncertainty.** Distribution of normalised step confidences $\tilde{p}_i$ for `gpt-oss-20b` MATH-500 traces across the three answer formation stages. Pre-commitment steps show an growing proportion of mid-guesses that disappear after the commitment boundary is reached.

Venhoff et al. (2025) (Figure 5).[6] However, *our truncation experiments confirm these steps have little to no influence on final outputs*.

4.2 All Meaningful Reasoning Precedes the Commitment Boundary

Having established that the commitment boundary i^* marks the start of epiphenomenal reasoning, we now ask whether the pre-boundary CoT is itself informative. Figure 6 (top) shows the progression of guesses across `gpt-oss-20b` MATH-500 traces. Almost all traces begin without guess (purple) above our threshold τ, and we observe a large fraction of these transitions directly to the model final guess when reaching the commitment boundary. However, in a good proportion of traces the model produces one or more mid-guesses (orange), each corresponding to a distinct hypothesis the model entertains with high confidence (Avg. likelihood of 0.69, Figure 6, bottom), and subsequently revises. This provides evidence that the reasoning performed in the pre-commitment region of the CoT is meaningful. On the contrary, the guess distribution after the commitment boundary is almost entirely skewed towards the final guess, providing additional evidence that post-boundary steps seem to have no causal influence on the model's final answer. We also note that traces containing mid-guesses are generally *more than twice as long* as those transitioning directly to the final guess (688 $\pm$57 vs. 306.15 $\pm$24 tokens), with median trace

[6] See Figure 10 and Appendix D for more details.

length increasing monotonically with the number of mid-guesses on MATH-500, GPQA-Diamond and ZebraLogic (Appendix E). These results expand upon previous findings by Boppana et al. (2026): while more complex tasks may yield less performative reasoning, models employ additional inference-time computation to choose among multiple options rather than reallocating epiphenomenal steps to meaningful considerations.

Different traces of the same problem have consistent mid-guesses. For each question with at least two valid traces, we measure two forms of guess consistency across different traces sampled for the same question: the Jaccard similarity between sets of mid-guesses (set overlap), and the standard deviation of the relative position within the CoT at which each mid-guess appears. On MATH-500, `gpt-oss-20b` achieves a median Jaccard of 0.71 across trace pairs, indicating that the majority of mid-guess values are shared between independently sampled traces of the same problem. The median standard deviation of relative CoT position is 0.005, roughly half the empirical baseline of 0.011 obtained by randomly permuting mid-guess positions across problems. This indicates that mid-guesses tend to recur at similar points across different reasoning chains (Appendix E).

5 Probing the Commitment Boundary from Model Activations

While useful for post-hoc analysis of reasoning dynamics post-hoc, computing commitment boundary positions i^* for new traces is prohibitively expensive in practical settings, requiring up to $n+1$ forward passes to sample answer completions across steps using our truncation method. However, we hypothesize that the existence of a commitment boundary may reflect a stable internal mechanism rather than a purely behavioral artifact, and that its occurrence could be efficiently decoded from model activations in a single forward pass.

To test this hypothesis, we begin by extracting activations $(\mathbf{h}_1, \ldots, \mathbf{h}_n) \in \mathbb{R}^d$ at a designated layer ℓ for the last token of each CoT step $C_1, \ldots, C_n$ across all n traces. Each activation $\mathbf{h}_i$ is associated to a label $y_i \in \{\text{NO-GUESS}, \text{MID-GUESS}, \text{FINAL-GUESS}\}$ corresponding to the answer formation stage of the corresponding step C_i, obtained from our previous truncation experiments. Following Boppana et al. (2026), we train a small causal attention probe to perform three-way classification given a fixed context window of w previous CoT steps, enabling its predictive usage at inference time.[7]

We train our probes on a stratified, question-level split of MATH-500 traces (50% train / 10% validation / 10% test, with 30% held out for Section 6), and use the remaining datasets as test sets that span various degrees of similarity to the training data distribution. We optimize a class-weighted cross-entropy loss using Adam and select the final checkpoint via early stopping on validation macro-F_1. At inference time, we predict $\hat{y}_i = \arg\max f(\mathbf{h}_{i-w}, \ldots, \mathbf{h}_i)$.[8]

Boundary localization diagnostics. Rather than reporting sentence-level classification accuracy, which does not reflect the ability of the probe to precisely detect the location of the commitment boundary, we report various trace-level metrics that summarize the probe's ability to predict FINAL-GUESS for usage in monitoring and early-exiting:

- **Detection rate (Det.):** Fraction of traces on which the probe fires at or briefly after i^* (using a $k = 2$-sized tolerance window).
- **Early-Fire Rate:** Fraction of traces on which the probe fires *before* i^*.
- **Miss Rate:** Fraction of traces for which commitment is never detected.
- **Exact-boundary rate:** Fraction of traces in which the probe fires *exactly at* i^*.
- **Detection delay:** Median fraction of traces between the true i^* and the probe nearest post-boundary firing.
- **Saved tokens:** Median fraction of CoT tokens that would be skipped by early-exiting at the probe's predicted location.

The tolerance window k is defined as a run of k consecutive FINAL-GUESS predictions:

$$i^{(k)}_{\text{exit}} = \min\{\, i : \hat{y}_{i-k+1:i} = \text{FINAL-GUESS}^k \,\}, \tag{4}$$

where $\hat{y}_{a:b}$ denotes subsequence $(\hat{y}_a, \ldots, \hat{y}_b)$.

We find that a tolerance window of $k = 2$ provides a good balance between efficiency and precision, providing enough context to elicit a FINAL-GUESS prediction without accumulating unnecessary post-boundary tokens.[9] Table 1 reports our

[7] Full probe architecture described in Appendix F.

[8] Full hyperparameters, architectural choices (ℓ, w), and label distributions are reported in Appendix G.

[9] See full sweep over k in Appendix G.

Model	Eval	Det.↑	Early-Fire↓	Miss↓	Exact↑	Delay↓	Saved↑
gpt-oss-20b	ID	92%	6%	0%	27%	9%	35%
	OOD	69 – 97%	3 – 20%	0 – 0%	20 – 23%	3 – 23%	6 – 35%
gemma-4-26B-A4B-it	ID	62%	16%	25%	15%	30%	55%
	OOD	98 – 99%	0 – 0%	1 – 2%	0 – 7%	25 – 36%	40 – 44%
Qwen3-14B	ID	98%	2%	0%	4%	19%	48%
	OOD	78 – 98%	0 – 21%	0 – 2%	4 – 25%	5 – 47%	13 – 55%

Table 1: **Boundary localization diagnostics** for causal attention probes firing at FINAL-GUESS positions. **Det.**: % of positive traces with firing within $k = 2$ steps after i^*. **Early-Fire**: % of traces with with firing before i^*. **Miss**: % of positive traces with no firing. **Exact**: % of traces with firing exactly at i^* (k=1). **Delay**: Median # of steps between i^* and probe firing on non-early detections. **Saved**: Median % of CoT tokens saved across positive traces (misses counted as 0). ID = MATH-500, OOD *min–max* ranges span AIME 2025, ZebraLogic, and GPQA-Diamond.

results across tested models, showing MATH-500 performance as ID and min.-max. performance across the other three datasets as OOD.

Across all models, detection rates exceed 90% on in-distribution data and *remain equally high in OOD settings*, while early-fire rates remain low, indicating that the probe commits at the right place or slightly after in the vast majority of traces. The generally low mean detection delay implies a small but non-zero residual tail of post-boundary tokens even when the probe succeeds. The consistent pattern confirms that *the commitment boundary can be reliably detected across model families*.

6 Early Exiting at the Commitment Boundary

Section 5 probes predict a per-step answer class $\hat{y}_i$ that can be computed during generation. If the model is genuinely aware of its commitment boundary, then halting generation as soon as $\hat{y}_i =$ FINAL-GUESS should preserve accuracy while skipping the post-boundary tokens that Section 4 identified as output-redundant.

To test the early-exiting performances of trained probes, we adopt the same $k = 2$ tolerance window used in Section 5. If no exit index exists, we fall back to the full CoT. At inference time, we evaluate the effectiveness of our approach by truncating the trace at the end of sentence $i^{(k)}_{\text{exit}}$, appending the model-specific end-of-thinking marker and the task-dependent answer-forcing suffix S, and decoding the answer greedily as in our truncation experiments of Section 4 (see Appendix A).

We compare against three black-box baselines for early-exiting:

- **Full CoT**: Use the entire generated trace as an upper bound on accuracy, with no token saving.
- **No CoT**: Baseline performance without access to reasoning tokens.
- **Fixed Exit**: Using the insight that commitment boundary is far from the true end of the CoT (Figure 3), we truncate each trace at a fixed fraction $\rho \in \{50, 70, 80, 90, 95\}\%$ of its total CoT length, using the same answer-forcing protocol.

The Fixed Exit baseline acts as a natural non-adaptive control for our evaluation, showing whether probe's per-trace exit decisions are doing anything beyond exploiting the population-level statistics of i^*/n reported in Section 4.

For each variant we report (i) accuracy relative to the ground-truth answer a^*, (ii) the average fraction of CoT tokens saved $1 - n_{\text{kept}}/n_{\text{total}}$, and (iii) the boundary localisation error $\mathbb{E}[(i^{(k)}_{\text{exit}} - i^*)/n]$, signed so that negative values indicate exiting before the true commitment boundary.

Figure 7 summarises the trade-off curve across our tested datasets.[10] The probe consistently outperforms fixed-percentage truncation at every operating point, confirming that it exploits per-trace structure rather than population-level length statistics. On the in-distribution MATH-500 dataset, the two strategies diverge sharply at matched average truncation. The fixed-percentage baseline loses 23 percentage points of full-CoT accuracy (from 80% to 57%), whereas the probe retains most of the performance across all values of k. At k=1, it recovers 89% of full-CoT accuracy (71% vs. 80%) while saving 39% of CoT tokens on average, with a median boundary error of just one sentence past i^*. Increasing k trades token savings for accuracy monotonically: k=5 reaches 98% of full-CoT accuracy (78% vs. 80%) at 26% token savings.

[10] See Figure 11 for results on other models.

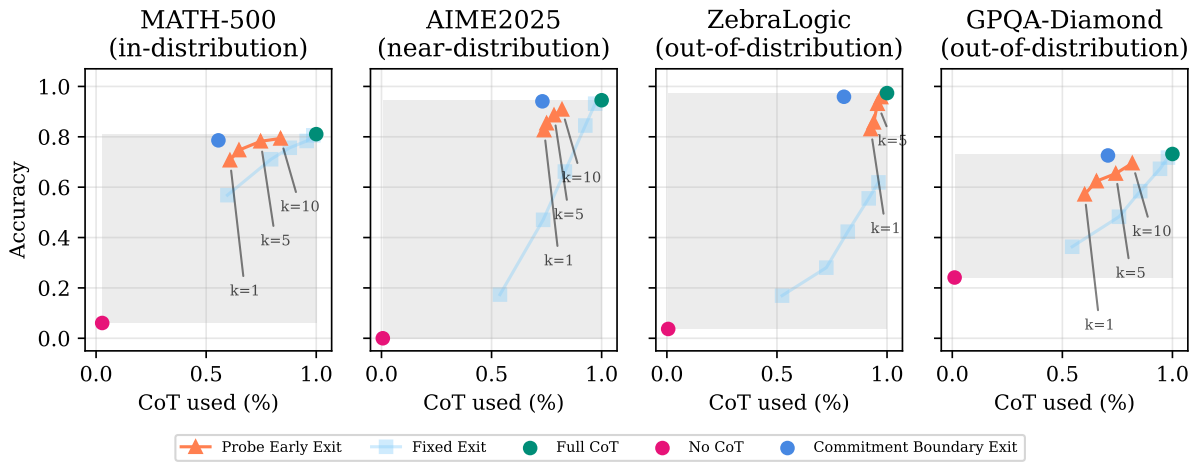


Figure 7: **Probe-mediated early exit dominates fixed-percentage truncation at every operating point,** with results consistent across in- and out-of-distribution datasets suggesting robust detection capabilities.

When applied without modification to AIME 2025, ZebraLogic, and GPQA-Diamond, the probe continues to consistently outperform fixed baselines with small accuracy loss compared to full-CoT (at most 11% on ZebraLogic, 97% $\rightarrow$ 86%) while saving 25%, 6%, and 35% of CoT tokens respectively. The Fixed Exit controls lose more accuracy on every OOD dataset, with the largest gap for AIME 2025, where traces are longest and adaptive exit is therefore most beneficial. Together with probe results of Section 5, these findings suggest that *the commitment boundary is a stable, model-internal property that can be reliably acted on at inference time*.

7 Conclusion

We introduced a step-level causal attribution framework based on CoT early exits to study how large reasoning models converge to their final answer. Across three model families and four reasoning benchmarks, we identified the *commitment boundary*: a sharp, often single-step transition after which the model's final answer probability is essentially fixed. Truncation and targeted numeric-perturbation experiments confirmed an asymmetry between the two sides of this boundary: pre-boundary steps causally shape the answer and frequently entertain genuine mid-guesses that are later revised, whereas post-boundary steps form an *epiphenomenal* tail of hedging and re-verification that leaves the elicited answer essentially unchanged. Its relative position is primarily a function of the model family rather than of task difficulty, with up to 87% of generated reasoning tokens occurring after the answer has stabilized in some configurations.

Beyond post-hoc analysis, we showed that answer-formation stages are linearly decodable from intermediate activations using lightweight causal attention probes that generalize across in- and out-of-distribution benchmarks. Used as online exit signals, these probes consistently outperform fixed-percentage truncation across all operating points, saving up to 35% of CoT tokens with minimal degradation in accuracy. Together, these results suggest that the commitment boundary is not a behavioral artifact but a stable mechanism for answer commitment in reasoning models.

Our findings have implications beyond inference-time efficiency. The prevalence of epiphenomenal reasoning, in which the verbalized trace continues to deliberate, hedge, and challenge an answer the model has already internally fixed, complicates the use of CoT as a faithful window into model computation and reinforces the need for monitoring tools that combine surface-level inspection with representation-level signals. We hope our framework provides a useful lens for future work disentangling genuine deliberation from performative continuation, and for designing training objectives that align the externalized trace more tightly with the underlying answer-formation process.

Limitations

Greedy decoding and answer elicitation. Our step-level attribution relies on greedy decoding from each truncated prefix X_i with a task-specific answer-forcing suffix S. This choice yields a deterministic, reproducible signal but ignores the full predictive distribution: a step could meaningfully

shift probability mass between answers without changing the arg-max, and would be invisible to our framework. Relatedly, the suffix S is itself a soft intervention that pushes the model to commit to an answer that it might have continued to deliberate over. Our conclusions, therefore, concern the answer the model would commit to if forced at step i, which is a useful but not exhaustive notion of commitment. Sampling-based or distribution-level extensions of the framework are a natural direction for future work.

Trace and token filtering. We restrict our analysis to traces where the no-CoT baseline is incorrect and discard traces with first-token collisions between intermediate and final answers (Appendix B). While necessary to avoid confounds, these filters bias our sample toward problems where CoT is genuinely informative and where answers are tokenization-distinguishable.

Sentence-level granularity. We segment CoTs at terminal punctuation, treating sentences as the atomic unit of reasoning. This is a coarse approximation: commitment may occur mid-sentence, and our reported boundary positions i^*/n should be read as upper bounds on localization precision. Finer-grained attribution (clause- or token-level) would likely shift i^* earlier in some traces, but was too computationally prohibitive for our approach.

Model and task coverage. Our experiments span three open-weight model families (`gpt-oss-20b`, `Qwen3-14B`, `gemma-4-26B-A4B-it`) and four reasoning benchmarks dominated by mathematics, logic, and multiple-choice science. We do not evaluate on agentic, tool-using, or long-horizon generation tasks, where the notion of a single final answer is less well-defined and the commitment boundary may not exist or may take a qualitatively different form. We also do not study closed-weight frontier models, where activation-based probing is not possible.

Probe supervision and labels. Our probes are trained on labels derived from the same truncation procedure they are meant to replace, making them prone to inherit the biases of our methodology. While they seem to generalize to OOD benchmarks, performance on substantially different domains (code, open-ended QA, multilingual reasoning) remains a promising avenue for future work.

Early-exit safety. Probe-mediated early exit trades a small accuracy loss for substantial token savings, but the loss is not zero, and the early-fire rate is non-negligible on some OOD settings (up to 21% on `Qwen3-14B`). In deployment contexts where individual answers carry high stakes (medical, legal, safety-critical reasoning), even a small fraction of prematurely truncated traces may be unacceptable. Calibrating exit thresholds per task and per acceptable-risk profile is left to future work.

Interpretation of epiphenomenal reasoning. Finally, we caution against over-interpreting epiphenomenal reasoning as evidence that post-boundary tokens are fully functionally useless to the model. Our framework measures their effect on the *elicited final answer* under answer-forcing; it does not rule out roles such as calibrating internal confidence, supporting follow-up turns, or shaping behavior under different decoding strategies. The post-boundary tail may also serve training-time objectives (e.g., RL reward shaping) that are not visible in a single inference pass. Characterizing what epiphenomenal reasoning is for remains a promising open question.

References

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *Preprint*, arXiv:2402.04614.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. 2026. Thought anchors: Which LLM reasoning steps matter?

Siddharth Boppana, Annabel Ma, Max Loeffler, Raphael Sarfati, Eric Bigelow, Atticus Geiger, Owen Lewis, and Jack Merullo. 2026. Reasoning theater: Disentangling model beliefs from chain-of-thought. *Preprint*, arXiv:2603.05488.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. 2023. Implicit chain of thought reasoning via knowledge distillation. *Preprint*, arXiv:2311.01460.

Jaden Fiotto-Kaufman, Alexander Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Nikhil Prakash, Carla Brodley, Arjun Guha, Jonathan Bell, Byron Wallace, and David Bau. 2025. Nnsight and ndif: Democratizing access to open-weight foundation model internals. In *International Conference on Learning Representations*, volume 2025, pages 92337–92370.

Yichao Fu, Xuewei Wang, Hao Zhang, Yuandong Tian, and Jiawei Zhao. 2026. Deep think with confidence. In *The Fourteenth International Conference on Learning Representations*.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. 2024. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *The Twelfth International Conference on Learning Representations*.

Google. 2026. Gemma 4: Byte for byte, the most capable open models.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, and 22 others. 2025. Chain of thought monitorability: A new and fragile opportunity for ai safety. *Preprint*, arXiv:2507.11473.

Hynek Kydlíček. 2024. Math-verify: Math verification library. `https://github.com/huggingface/math-verify`. Version 0.6.1.

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

Zeju Li, Jianyuan Zhong, Ziyang Zheng, Xiangyu Wen, Zhijian Xu, Yingying Cheng, Fan Zhang, and Qiang Xu. 2026. Making slow thinking faster: Compressing LLM chain-of-thought via step entropy. In *The Fourteenth International Conference on Learning Representations*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. 2025. Zebralogic: On the scaling limits of LLMs for logical reasoning. In *Forty-second International Conference on Machine Learning*.

Yanyu Liu, Jingying Fu, Sixiang Liu, Yitian Zou, You Fu, Jiehan Zhou, and Shouhua Zhang. 2025. Kv cache compression for inference efficiency in llms: A review. *Preprint*, arXiv:2508.06297.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

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

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Daniel Scalena, Leonidas Zotos, Elisabetta Fersini, Malvina Nissim, and Ahmet Üstün. 2026. Eager: Entropy-aware generation for adaptive inference-time scaling. In *Proceedings of the 43rd International Conference on Machine Learning (ICML 2026)*.

Xu Shen, Song Wang, Zhen Tan, Laura Yao, Xinyu Zhao, Kaidi Xu, Xin Wang, and Tianlong Chen. 2026. Faithcot-bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. In *The Fourteenth International Conference on Learning Representations*.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.

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

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasovic, and Yonatan Belinkov. 2025. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9935–9960, Suzhou, China. Association for Computational Linguistics.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. 2025. Understanding reasoning in thinking language models via steering vectors. In *Workshop on Reasoning and Planning for Large Language Models*.

Zezhong Wang, Xingshan Zeng, Weiwen Liu, Yufei Wang, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2025. Chain-of-probe: Examining the necessity and accuracy of CoT step-by-step. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2586–2606, Albuquerque, New Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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

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

Evelyn Yee, Alice Li, Chenyu Tang, Yeon Ho Jung, Ramamohan Paturi, and Leon Bergen. 2024. Faithful and unfaithful error recovery in chain of thought. In *First Conference on Language Modeling*.

Yifan Zhang and Team Math-AI. 2025. American invitational mathematics examination (AIME) 2025.

Jiachen Zhao, Yiyou Sun, Weiyan Shi, and Dawn Song. 2026. Can aha moments be fake? identifying true and decorative thinking steps in chain-of-thought. *Preprint*, arXiv:2510.24941.

A Candidate solutions verifier

For each prefix X_i, we greedily decode an intermediate answer $\hat{A}_i$ (continuation capped at 64 tokens) and compare it to the full-chain answer $\hat{A}_n$. This comparison requires a semantic equivalence relation $\equiv$ robust to surface-form variation, therefore avoiding to introduce spurious mismatches into the attribution signal.

For consistency across mathematical and multiple-choice benchmarks, all non-code datasets use the same boxed-answer suffix, prompting the model to emit its final answer inside a `\boxed{}` environment. For multiple-choice tasks, the verifier subsequently extracts and normalizes the boxed option letter.

Answer strings are however normalized by stripping surrounding whitespace and math-mode delimiters, removing enclosing `\text{...}` wrappers, and collapsing internal whitespace (e.g. `p-q` vs.`p - q`, MATH-500, Q1). If the normalized reference and candidate strings match exactly, they are immediately declared equivalent.

Because our evaluation benchmarks span heterogeneous answer formats, the verifier dispatches automatically to domain-specific comparison branches to spot non-trivial equivalent guesses.

For mathematical benchmarks we use the `math-verify` library (Kydlíček, 2024) which extracts mathematical expressions, parses them symbolically via `latex2sympy2_extended`, and verifies equivalence using SymPy. To reduce formatting-related parsing failures, we additionally normalize `\tfrac` and `\dfrac` to `\frac` and strip `\left` and `\right` delimiters. For each answer we evaluate multiple surface variants (with and without math-mode delimiters, before and after normalization), and declare two answers equivalent if any variant pair is successfully verified. This procedure is robust to common algebraic simplifications (`\frac{3}{2}` vs. 1.5, e.g. MATH-500 q20), symbolic rewritings, and minor LaTeXformatting differences (`\tfrac{1}{2}` vs. `\frac{1}{2}`, e.g. MATH-500 q0).

For multiple-choice benchmarks (e.g. GPQA-Diamond), we extract the predicted option letter using a lightweight regular-expression matcher after normalization. This canonicalizes variants such as `A`, `(A)`, and `A.`. If option extraction fails, comparison falls back to normalized string matching. Two answers are considered equivalent if and only if their canonicalized forms match exactly.

If no answer can be extracted from a generated continuation, the prefix is assigned the NO GUESS label and treated as non-equivalent. The verifier always compares $\hat{A}_i$ against the model's own final answer $\hat{A}_n$, rather than the benchmark ground-truth label, therefore detecting final answer confidence convergence, not task correctness.

B First token collisions between semantically different guesses

A first-token collision occurs at prefix X_i when the forced target token a_0 (the first token of $\hat{A}_n$) is also the first token of a semantically different answer (i.e. $a_0 = a_n$ but $\hat{A}_i \not\equiv \hat{A}_n$), see example in Table 2. In this scenario, that we will be referring to as cross-class collisions, $p_i = P(a_0 \mid X_i)$ overestimates the model's commitment to $\hat{A}_n$: the model may be confidently generating a distinct answer whose surface form happens to share a leading token with $\hat{A}_n$.

For any prefix X_i at which a cross-class collision is detected, the step confidence p_i is unreliable as a proxy for commitment to $\hat{A}_n$, and the prefix is effectively excluded from the semantic attribution signal. Table 3 reports collision rates across benchmarks.

Full-CoT answer $\hat{A}_n$	`49`
First token a_0	`4`
Candidates at prefix X_i	
wrong-guess	`420`, `406`
no-guess	`40`, `44`
final-equivalent	`49`

Table 2: **Example of a cross-class collision** on a trace (in detail: q16 of AIME2025, `Qwen3-14B`). The token `4` is shared by answers of different semantic classes, making p_i an unreliable commitment signal.

It is important to note that the collision rate is task-dependent. For instance, MATH-500, which involves a significant amount of LaTeX-heavy responses, has a much higher collision rate compared to AIME in most of the cases, where all responses are numerical, or even rarer, GPQA-Diamond, which is a multiple-choice format and thus responses are predetermined from the outset.

The anomalously high collision rate observed for `gemma-4-26B-A4B-it` on AIME-2025 (56.67%) is not a measurement artifact, but reflects a genuine property of the model's reasoning style on competition mathematics: `gemma-4-26B-A4B-it` fre-

Benchmark	gpt-oss-20b		gemma-4-26B-A4B-it		Qwen3-14B	
	Traces	Collision rate	Traces	Collision rate	Traces	Collision rate
MATH-500	7883	10.85%	2855	11.91%	2852	11.68%
AIME-2025	367	0.82%	90	56.67%	124	12.10%
ZebraLogic	335	2.99%	273	0.00%	300	4.00%
GPQA-Diamond	1432	0.28%	1004	0.00%	1210	1.24%
W. Average	–	8.71%	–	9.26%	–	8.36%

Table 3: First-token cross-collision rates across benchmarks.

quently produces intermediate numerical guesses (e.g., `21`) that share a leading token with the final answer (expressed as `21+56-7=70`), causing multiple semantically distinct candidates to collapse onto the same first token despite being semantically distinct quantities.

C Clue sensitivity (τ) ablation

To assess sensitivity to the threshold τ (as defined in Section 3), we study the mid-guesses emergence by varying $\alpha \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$:

$$\tau = p_0 + \alpha \cdot (p_n - p_0)$$

and keeping all other hyper-parameters fixed.

In order to discriminate when an early-exit prediction is an mid-guess or *autoregressive noise*, we use τ as a sensitivity parameter to distinguish between NO-GUESS and MID-GUESS. A lower τ is more permissive: sentences carrying modest confidence gains are still credited as clues, increasing recall of intermediate guesses at the cost of potentially including low-confidence outputs. On the opposite, a higher τ is more conservative: only sentences that have already accumulated a large fraction of the full-CoT confidence gain are labelled as clues, trading recall for precision.

Figure 8 reports the fraction of mid-guess and no-guess spans across models and benchmarks for each value of τ, thus excluding the final guesses from the total. As expected, increasing τ monotonically reduces the mid-guess fraction in all settings, confirming that the parameter controls a smooth recall–precision trade-off. The effect is also dataset-dependent: on ZebraLogic, mid-guesses dominate even at $\tau = 0.7$, reflecting the high density of confident wrong intermediate answers in logical reasoning tasks. On AIME2025, `gpt-oss-20b` shows very few mid-guesses across all τ values, suggesting it tends to commit directly to a final answer with little intermediate exploration; `gemma-4-26B-A4B-it` and `Qwen3-14B` exhibit instead a richer mid-guess structure on the same benchmark.

While `gpt-oss-20b` concentrates almost all wrong-guess masses near zero, specifically on mathematical benchmarks, `gemma-4-26B-A4B-it` and `Qwen3-14B` exhibit broader, higher-confidence distributions. This implies that `gpt-oss-20b` candidate alternative solutions are spotted only for the transitory phase in which the model is seriously committed to the wrong answer, as stated in Section 4.2 and shown in Figure 6. For `gemma-4-26B-A4B-it` and `Qwen3-14B` the pattern is more noisy, as shown in Figure 9.

D Post commitment boundary inflection points

Reasoning traces frequently contain linguistic markers associated with deliberation, namely words and phrases such as *wait*, *but*, *check*, or *let's*, that superficially signal reconsideration or uncertainty. We ask whether these *verbal uncertainty signals* are systematically distributed relative to the commitment boundary i^*.

For each trace, we compute the fraction of sentences whose first three words contain at least one such marker, separately for the pre- and post-boundary portions. We restrict our analysis to the first three words of each sentence, rather than scanning the full text, to target these markers in their role as deliberation signals: words like *wait* or *actually* are indicative of reconsideration when they open a new reasoning step, but lose this interpretation when they appear embedded in mathematical or syntactic contexts mid-sentence. Figure 10 reports these fractions across three models on MATH-500, AIME 2025, and GPQA-Diamond.

The distributions are largely similar on both sides of the boundary, with no marker showing a consistent directional shift across all three models. *wait* and *but* remain among the most frequent markers both before and after i^*; *let's* shows a modest

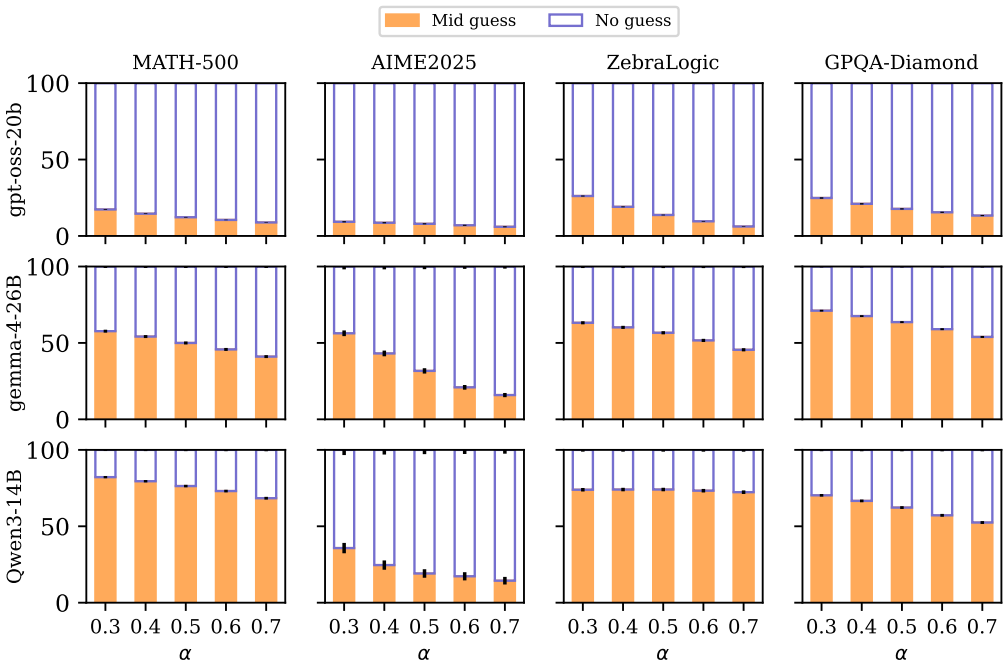


Figure 8: **Mid-guess fraction as a function of τ across models and benchmarks.** Each panel shows the fraction of sentence spans classified as mid-guess (orange) vs. no-guess (grey) for a given model (row) and benchmark (column), as τ varies in $\{0.3, 0.4, 0.5, 0.6, 0.7\}$. Final guesses are excluded from the denominator. `gpt-oss-20b` shows markedly fewer mid-guesses on AIME2025 compared to the other models, suggesting a more direct commitment pattern on hard mathematical reasoning.

pre-boundary preference in `gemma-4-26B-A4B-it` and `gpt-oss-20b`, but not in `Qwen3-14B`. Overall, the presence of these markers alone is not a reliable indicator of whether the model is still actively updating its answer. The same surface signal can occur on either side of the commitment boundary.

E Trace length is midguesses-dependent

Traces exhibiting at least one mid-guess before i^* tend to be longer than those that proceed directly to the final answer. To test whether this relationship holds monotonically with the number of mid-guesses, we apply the Mann-Whitney U-test to compare the distribution of trace lengths between groups of traces binned by their number of mid-guesses n. The test is non-parametric and does not assume normality, making it appropriate for the skewed token-length distributions observed in practice. We report results for both raw and τ-gated mid-guesses separately. Table 4 reports detailed results for `gpt-oss-20b` across all benchmarks.

On MATH-500, all consecutive comparisons are highly significant, with median trace length increasing monotonically from 252 tokens (zero mid-guesses) to over 1000 tokens (more than six mid-guesses). The same monotone pattern holds on GPQA-Diamond and ZebraLogic. On AIME-2025, where problems are substantially harder and the number of mid-guesses is consistently higher, the relationship does not reach significance. We attribute this to high intrinsic variability in trace length across problems, where problem-specific structure dominates computation cost and obscures the mid-guess signal.

F Causal linear probe architecture

The probe operates on sentence-level hidden states: at each sentence boundary C_i, it receives the hidden state $\mathbf{h}_i$ of the last token in that sentence and must predict the commitment label $y_i \in \{\text{NO-GUESS}, \text{MID-GUESS}, \text{FINAL-GUESS}\}$ using only information from sentences $1, \ldots, i$ – a causality constraint that makes the probe directly applicable during generation, without requiring the full trace to be completed first. Rather than pooling over the full trace prefix (which would conflate early and late reasoning steps and thus introduce a position bias), we restrict the attention to a local window of the w most recent sentences. Formally, let $\mathbf{H} = [\mathbf{h}_1; \ldots; \mathbf{h}_n] \in \mathbb{R}^{n \times d}$, and let $\tilde{\mathbf{H}} = \mathrm{LN}(\mathbf{H})$ be its layer-normalised version. For each query position i, the probe pools causally over a window of the w most recent sentences:

$$\alpha_{ij} \propto \exp(\mathbf{w}_q^\top \tilde{\mathbf{h}}_j) \cdot \mathbb{1}[i - w < j \leq i], \quad (5)$$

$$\mathbf{c}_i = \textstyle\sum_j \alpha_{ij} W_v \tilde{\mathbf{h}}_j, \quad (6)$$

$$\mathbf{z}_i = \mathrm{ReLU}\big([\mathbf{c}_i \,\|\, W_\ell \tilde{\mathbf{h}}_i]\big), \quad (7)$$

$$f(\mathbf{h}_{\leq i}) = \mathrm{softmax}(W_o \, \mathbf{z}_i), \quad (8)$$

where $W_q, W_v, W_\ell \in \mathbb{R}^{d \times h}$ and $W_o \in \mathbb{R}^{2h \times 3}$ are the only learned parameters.

The local term $W_\ell \tilde{\mathbf{h}}$ preserves the current sentence representation, while the causal pool $\mathbf{c}_i$ supplies prefix context constrained to the most recent w sentences.

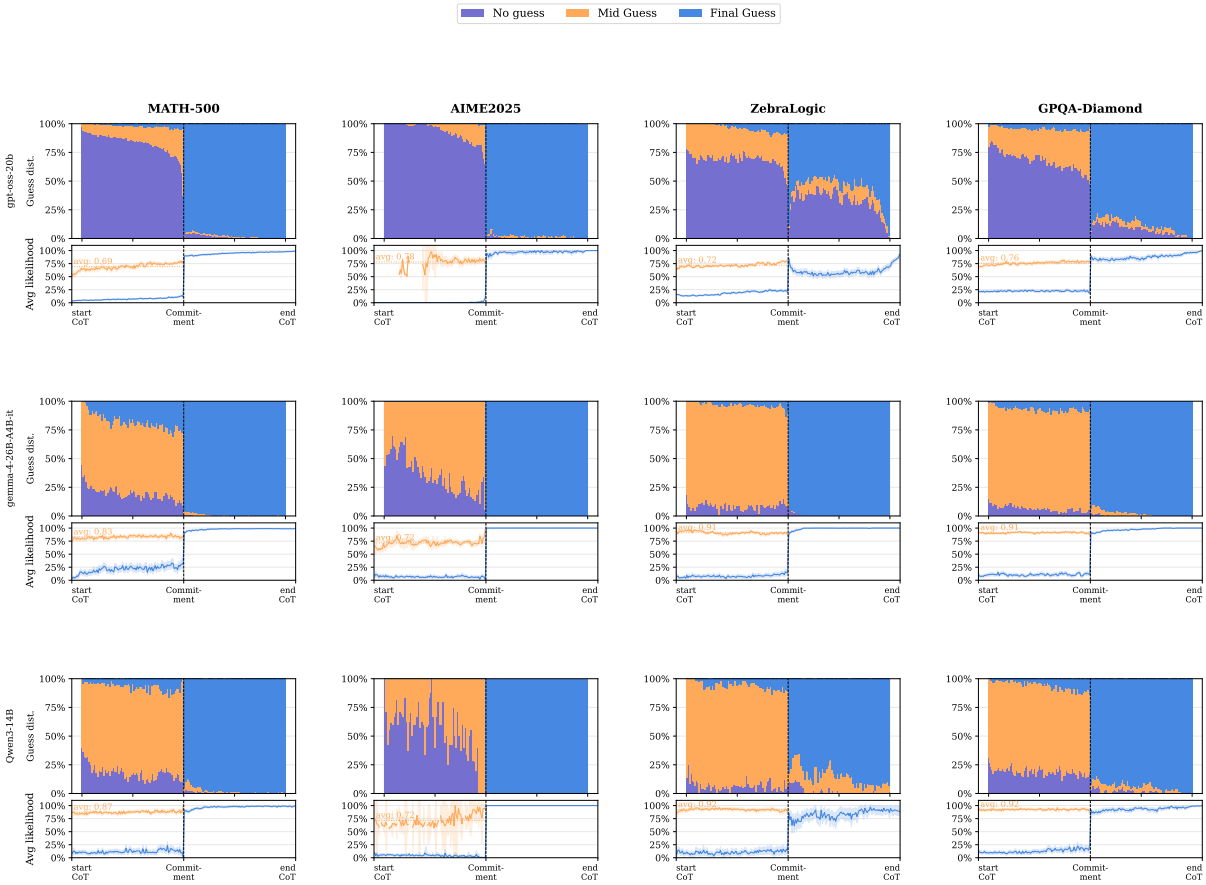


Figure 9: Guess distribution (top) and average token likelihood (bottom) across CoT positions, relative to the commitment boundary, for all three models and four benchmarks. The x-axis is normalised so that first, second and third ticks correspond to start, commitment boundary i^* and end of the CoT respectively. **Top panels:** stacked proportion of sentence spans labelled as NO-GUESS MID-GUESS and FINAL-GUESS at each relative position. **Bottom panels:** token likelihood progresison of MID-GUESS and FINAL-GUESS, with 95% confidence intervals shaded. Across all models, FINAL-GUESS likelihood rises sharply at i^* and stabilises post-commitment, confirming that the boundary captures a genuine, abrupt shift in the model's output distribution. MID-GUESS before i^* remains consistently high for `gemma-4-26B-A4B-it` and `Qwen3-14B`, whereas for `gpt-oss-20b` it is more concentrated in the commitment proximity, consistent with a more decisive reasoning style.

Figure 10: **Uncertainty-signalling language is equally frequent before and after the commitment boundary.** We show the top-20 sentence-initial words across all models (e.g. roughly 12% of Qwen3-14B sentences begin with "*but*", at similar rates on both sides of i^*). Manually highlighted words can signal re-verification behaviour – yet their frequency does not meaningfully change after the commitment boundary, confirming that such language is epiphenomenal: the model produces the same hedging and rechecking vocabulary even after our truncation experiments have causally established that no subsequent reasoning can alter its final answer. Results aggregated across MATH-500, AIME 2025, and GPQA-Diamond.

G Commitment Boundary Probe: Sweep Results

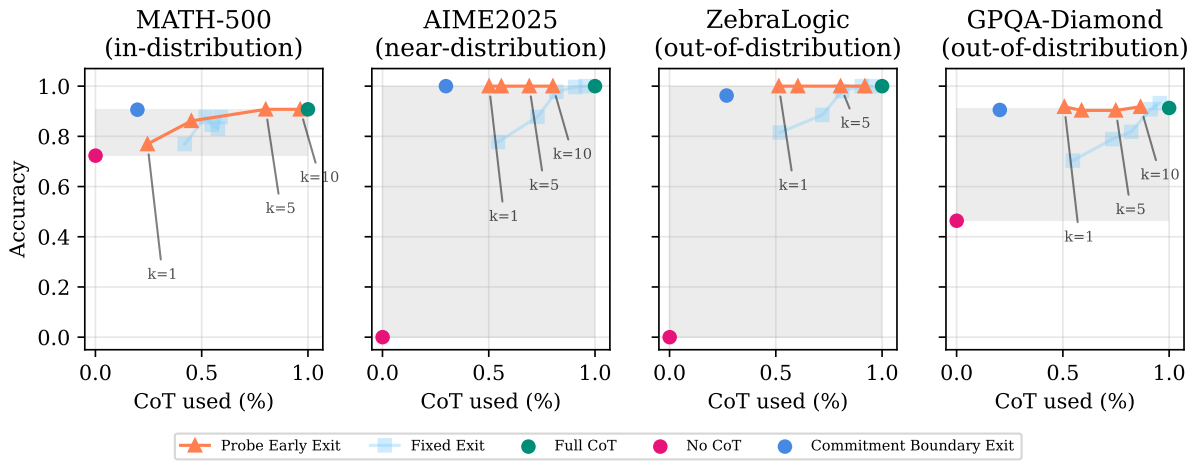
We describe the architecture and hyperparameter selection for the causal attention logistic probe introduced in Section 5. The probe is trained to predict whether a given CoT step i has a guess at all and if so, whether it is MID- or FINAL- from the hidden state $\mathbf{h}_i$ at a fixed transformer layer. We evaluate three models: `gpt-oss-20b`, `gemma-4-26B-A4B-it` and `Qwen3-14B`.

Setup We sweep over three axes: (i) the transformer layer from which $\mathbf{h}_i$ is extracted ($\{50\%, 80\%, 90\%, 100\%\}$ of total depth (exact layer correspondence in different model architectures is detailed in Table 5); (ii) the aggregation strategy over the hidden states of the sentence window preceding C_i: `last` uses only the hidden state of the final token of the window, while `avg` averages the hidden states of all tokens in the window; and (iii) the size of the context window fed to the probe ($\{8, 16, 32, 64, 128, 256\}$ tokens). All probes are trained on a held-out subset of MATH-500 traces and evaluated on a disjoint test set.

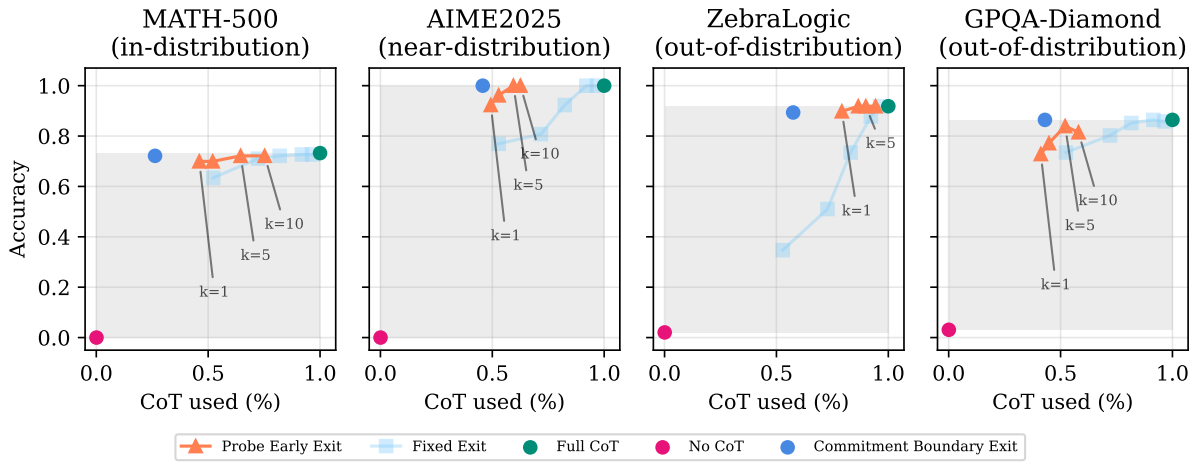
Model	50%	80%	90%	100%
`gpt-oss-20b`	12	19	22	23
`gemma-4-26B-A4B-it`	15	23	27	29
`Qwen3-14B`	19	31	35	39

Table 5: Layer index correspondence for each model architecture at the four depth percentiles used in the probe sweep.

We evaluate each configuration along four boundary-detection metrics at operating point $k = \{1, 2, 5, 10\}$ (i.e., the probe must fire on k consecutive post-boundary steps before triggering): **early-fire rate** (fraction of traces where the probe fires before i^*, lower is better), **detect rate** (fraction of traces where the probe correctly identifies the post-boundary region, higher is better), **miss rate** (fraction of traces where the probe never fires, lower is better), and **saved fraction** (mean fraction of CoT tokens that would be skipped under early exit). We also report **class-2 precision**, i.e. the precision of the probe on the post-boundary class.



(a) GEMMA-4-26B-A4B-IT



(b) QWEN3-14B

Figure 11: Causal optimal early-exit accuracy versus CoT fraction across models for GEMMA-4-26B-A4B-IT and QWEN3-14B.

Training We train on a stratified split of MATH-500 traces (50% train / 10% validation / 10% test, the remaining 30% held out for the early-exit experiments of Section 6. The split is at the question level, ensuring that no sentence from a held-out question appears at train time. We optimise a class-weighted cross-entropy loss with weights inversely proportional to class frequency (clipped to a maximum ratio of 5), Adam with learning rate 10^{-4} and weight decay 10^{-4}, batch size 64, gradient clipping at 1.0, and dropout 0.1. Early stopping based on the validation macro-F_1 (patience 6) selects the final checkpoint after 10 epochs. At inference we predict $\hat{y}_i = \arg\max f(\mathbf{h}_{\leq i})$.

Sensitivity of the τ threshold has already been discussed into Appendix C.

Results Tables 6, 7 and 8 report results for the top configurations of each tested model.

For `gpt-oss-20b` (Table 6), the best overall configuration is `layer100_last256` (layer 23, `last` aggregation, 256-token window), which achieves a k=2 early-fire rate of 6.87%, a detect rate of 93.13%, zero misses, and a class-2 precision of 91.59%, while saving 34.50% of post-boundary tokens on average. Two clear trends emerge from the sweep. First, later layers are more informative: layer 23 consistently outperforms layers 12 and 19 on early-fire and detect rate, confirming that commitment boundary information is encoded in deeper representations. Second, larger context windows improve probe quality up to 256 tokens, suggesting that the probe benefits from attending to a longer span of recent reasoning before mak-

Dataset	Standard mid-guesses			τ-gated mid-guesses (Appendix C))		
	Comparison (n)	p-value	medians	Comparison (n_1/n_2)	p-value	medians
MATH-500	0 vs 1 (493/781)	$\mathbf{9.7 \times 10^{-31}}$	252 / 366	0 vs 1 (2577/1125)	$\mathbf{5.8 \times 10^{-37}}$	428 / 609
	1 vs $[2,3]$ (781/1570)	$\mathbf{1.2 \times 10^{-25}}$	366 / 488	1 vs ≥ 2 (1125/664)	$\mathbf{2.4 \times 10^{-16}}$	609 / 904
	$[2,3]$ vs ≥ 4 (1570/1522)	$\mathbf{1.8 \times 10^{-84}}$	488 / 893			
ZebraLogic	$[1,2]$ vs 3 (59/80)	$\mathbf{6.2 \times 10^{-13}}$	1573 / 3553	1 vs 2 (21/63)	$\mathbf{2.9 \times 10^{-5}}$	1126 / 2236
	3 vs > 3 (80/52)	$\mathbf{1.9 \times 10^{-3}}$	3553 / 5910	2 vs 3 (63/76)	$\mathbf{5.5 \times 10^{-8}}$	2236 / 4016
				3 vs 4 (76/27)	$\mathbf{4.1 \times 10^{-4}}$	4016 / 7034
GPQA-Diamond	0 vs 1 (21/250)	$\mathbf{1.8 \times 10^{-4}}$	671 / 1165	0 vs 1 (217/278)	$\mathbf{1.3 \times 10^{-2}}$	1092 / 1395
	1 vs 2 (250/257)	$\mathbf{3.8 \times 10^{-4}}$	1165 / 1485	1 vs ≥ 2 (278/131)	$\mathbf{1.7 \times 10^{-4}}$	1395 / 1800
	2 vs ≥ 3 (257/98)	1.2×10^{-1}	1485 / 1654			

Table 4: Mann–Whitney U-test results for `gpt-oss-20b` across all benchmarks. Standard and τ-gated mid-guesses (Appendix C) use dataset-specific bins chosen to balance statistical power and interpretability; gated bins are coarser where the τ threshold concentrates traces at low n. Significant p-values ($p < 0.05$) are in **bold**. Median trace lengths are in tokens.

ing a prediction. `avg` aggregation was uniformly worse than `last`, indicating that the hidden state of the most recent token is the primary carrier of boundary information.

For `gemma-4-26B-A4B-it` (Table 7), the best overall configuration is `layer100_last64` (layer 29, `last` aggregation, 64-token window), which achieves an early-fire rate of 5%, a detect rate of 95%, zero misses, and a class-2 precision of 99%. The same depth trend observed for `gpt-oss-20b` holds: layer 29 dominates on early-fire and detect rate, while shallower layers remain competitive on class-2 F1. Notably, the context window exhibits an inverted pattern relative to `gpt-oss-20b`: smaller windows (64 tokens) yield lower early-fire rates, suggesting that for this model a tighter local context is sufficient and larger windows introduce noise.

For `Qwen3-14B` (Table 8), the best overall configuration is `layer90_last64` (layer 35, `last` aggregation, 64-token window), which achieves an early-fire rate of 5%, a detect rate of 95%, zero misses, and a class-2 precision of 97%. The optimal depth shifts to 90% rather than 100%, probably due to the higher number of layers compared to the other tested architectures (40 in place of the 24 of `gpt-oss-20b` and 30 of `gemma-4-26B-A4B-it`). Window size again favours smaller contexts on early-fire, consistent with the pattern seen in `gemma-4-26B-A4B-it`. Across all three models, almost-zero miss rates are achieved consistently, confirming that the probe rarely suppresses a true boundary detection regardless of configuration. For all the tested models and configurations, `avg` aggregation is worse than `last`.

Selected configuration The best probe according to the sweep procedure is adopted as our default probe throughout the paper for early-exit experiments detailed into Section 6. `gemma-4-26B-A4B-it` and `Qwen3-14B` early-exit performance plots can be found in Figure 11.

These results confirm the claim in Section 5 that the commitment boundary is linearly decodable from the model's hidden states with high reliability.

H Model generation arguments and reproducibility details

All reasoning traces were generated using vLLM with `max_model_len=16 384` tokens and we used tokens id instead of plain text to preserve end-of-thinking markers, an information used at attribution time to assess whether or not the trace has been truncated and thus the model has not reached its natural reasoning ending. Prompts were formatted via each model's chat template using `transformers`'s `AutoTokenizer.apply_chat_template`. Reasoning effort was kept to medium where available. For each question, 16 traces were sampled for `gpt-oss-20b` and 8 traces for all other models; for the latter ones, attribution was computed on up to 4 valid traces per question (i.e. traces where the full CoT strictly outperforms the no-CoT baseline and do not cause Out Of Memory errors).

As per generation parameters we used `temperature=0.7`, `top_p=0.9` across all our model suite. We additionally employ `NNsight` (Fiotto-Kaufman et al., 2025).

The distinct pair of start/end markers to delimit its chain-of-thought reasoning section is model-dependent, and we show them as follows:

Model	[BOT] (start) and [EOT] (end) tokens
gpt-oss-20b	<|channel|>analysis<|message|> / <|channel|>final<|message|>
Qwen3-14B	<think> / </think>
gemma-4-26B	<|channel>thought / <channel|>

I Use of AI

We used AI-based coding agents to assist with implementing experimental infrastructure and analysis scripts. All generated code was manually reviewed, validated, and, where necessary, modified by the authors, who take full responsibility for the final codebase and all reported results. Experimental designs, evaluation protocols, and scientific conclusions were conceived and verified by the authors; no experimental results or findings were autonomously generated by AI systems. AI-assisted writing tools were also used during manuscript preparation to improve clarity and presentation.

				Boundary				Class-2		
Config	**Layer**	**Agg**	**Window**	**Early ↓**	**Detect ↑**	**Miss ↓**	**Saved ↑**	**P ↑**	**R ↑**	**F1 ↑**
layer100_last512	23	last	512	**0.06**	**0.93**	**0.00**	0.34	0.91	**0.83**	**0.87**
layer100_last256	23	last	256	**0.06**	**0.93**	**0.00**	0.35	**0.92**	**0.83**	**0.87**
layer100_last128	23	last	128	0.12	0.89	**0.00**	0.36	0.87	**0.83**	0.85
layer100_last64	23	last	64	0.07	0.93	**0.00**	0.34	0.89	**0.83**	0.86
layer100_last32	23	last	32	0.08	0.92	**0.00**	0.35	0.89	0.81	0.85
layer100_last16	23	last	16	0.10	0.90	**0.00**	0.36	0.88	0.79	0.83
layer100_last8	23	last	8	0.15	0.85	**0.00**	**0.38**	0.86	0.77	0.82
layer100_avg64	23	avg	64	0.08	0.91	0.01	0.32	0.87	0.78	0.83
layer100_avg16	23	avg	16	0.10	0.87	0.03	0.33	0.85	0.73	0.79
layer90_last256	22	last	256	0.11	0.89	**0.00**	0.35	0.90	0.82	0.86
layer90_last128	22	last	128	0.11	0.89	**0.00**	0.35	0.86	0.82	0.84
layer80_last256	19	last	256	0.10	0.90	**0.00**	0.35	**0.92**	**0.83**	**0.87**
layer80_last128	19	last	128	0.09	0.91	**0.00**	0.35	0.89	**0.83**	0.86
layer80_last64	19	last	64	0.08	0.92	**0.00**	0.34	0.89	**0.83**	0.86
layer50_last16	12	last	16	0.11	0.89	**0.00**	0.37	0.83	**0.83**	0.84

Table 6: **Probe sweep results on MATH-500 (gpt-oss-20b).** All boundary metrics are reported at operating point k=2. Early-fire and miss rates are lower-is-better; all others are higher-is-better. The best configuration per column is **bold**.

				Boundary				Class-2		
Config	**Layer**	**Agg**	**Window**	**Early ↓**	**Detect ↑**	**Miss ↓**	**Saved ↑**	**P ↑**	**R ↑**	**F1 ↑**
layer100_last256	29	last	256	0.23	**0.77**	**0.00**	0.77	0.97	**0.93**	**0.95**
layer100_last128	29	last	128	0.08	0.92	**0.00**	0.67	**0.99**	0.84	0.91
layer100_last64	29	last	64	**0.05**	**0.95**	**0.00**	0.66	**0.99**	0.85	0.91
layer100_last32	29	last	32	0.07	0.93	**0.00**	0.66	**0.99**	0.82	0.89
layer100_last16	29	last	16	0.09	0.90	0.01	0.66	**0.99**	0.84	0.91
layer100_last8	29	last	8	0.14	0.86	**0.00**	0.70	0.98	0.81	0.89
layer100_avg64	29	avg	64	0.16	0.81	0.04	0.66	0.97	0.83	0.90
layer100_avg16	29	avg	16	0.19	0.79	0.07	0.68	0.95	0.79	0.87
layer90_last256	27	last	256	0.10	0.90	**0.00**	0.70	0.98	0.89	0.93
layer90_last128	27	last	128	0.06	0.94	**0.00**	0.68	**0.99**	0.88	0.93
layer90_last64	27	last	64	0.08	0.92	**0.00**	0.67	**0.99**	0.84	0.91
layer80_last256	23	last	256	0.12	0.88	**0.00**	0.72	0.98	0.87	0.92
layer80_last128	23	last	128	0.10	0.90	**0.00**	0.71	**0.99**	0.90	0.94
layer80_last64	23	last	64	0.08	0.92	**0.00**	0.70	**0.99**	0.91	0.94
layer50_last256	15	last	256	0.17	0.83	**0.00**	0.73	0.98	0.90	0.94
layer50_last128	15	last	128	0.11	0.89	**0.00**	0.73	0.98	0.88	0.93
layer50_last64	15	last	64	0.09	0.91	**0.00**	0.72	**0.99**	0.92	**0.95**

Table 7: **Probe sweep results on MATH-500 (gemma-4-26B-A4B-it).** All boundary metrics are reported at operating point k=2. Early-fire and miss rates are lower-is-better; all others are higher-is-better. The best configuration per column is **bold**.

				Boundary				Class-2		
Config	**Layer**	**Agg**	**Window**	**Early ↓**	**Detect ↑**	**Miss ↓**	**Saved ↑**	**P ↑**	**R ↑**	**F1 ↑**
layer100_last256	39	last	256	0.28	0.72	**0.00**	0.68	0.92	0.93	0.93
layer100_last64	39	last	64	0.19	0.81	**0.00**	0.68	0.95	0.92	0.93
layer100_last16	39	last	16	0.24	0.76	**0.00**	**0.69**	0.93	0.89	0.91
layer100_avg64	39	avg	64	0.19	0.81	**0.00**	0.64	0.95	0.85	0.90
layer100_avg16	39	avg	16	0.15	0.85	**0.00**	0.65	0.94	0.87	0.90
layer90_last256	35	last	256	0.22	0.78	**0.00**	0.68	0.94	0.92	0.93
layer90_last128	35	last	128	0.20	0.80	**0.00**	0.68	0.94	0.93	0.94
layer90_last64	35	last	64	**0.05**	**0.95**	**0.00**	0.66	**0.97**	0.91	0.94
layer90_last32	35	last	32	0.16	0.84	**0.00**	0.65	0.93	0.89	0.91
layer90_last16	35	last	16	0.12	0.88	**0.00**	0.67	0.95	0.90	0.93
layer90_last8	35	last	8	0.24	0.76	**0.00**	0.68	0.93	0.87	0.90
layer80_last256	31	last	256	0.18	0.82	**0.00**	0.67	0.95	0.91	0.93
layer80_last128	31	last	128	0.18	0.82	**0.00**	0.67	0.94	0.92	0.93
layer80_last64	31	last	64	0.08	0.92	**0.00**	0.65	0.96	0.91	0.93
layer50_last256	19	last	256	0.11	0.89	**0.00**	0.66	0.96	0.92	0.94
layer50_last128	19	last	128	0.09	0.91	**0.00**	0.66	**0.97**	0.93	**0.95**
layer50_last64	19	last	64	0.19	0.81	**0.00**	0.67	0.94	**0.94**	0.94

Table 8: **Probe sweep results on MATH-500 (Qwen3-14B).** All boundary metrics are reported at operating point k=2. Early-fire and miss rates are lower-is-better; all others are higher-is-better. The best configuration per column is **bold**.