Temporal Predictors of Outcome in Reasoning Language Models

github.com/joeydavid/latent-correctness-probe

Joey David
Independent Researcher
joeydhp@protonmail.com

Abstract

The chain-of-thought (CoT) paradigm uses the elicitation of step-by-step rationales as a proxy for reasoning, gradually refining the model's latent representation of a solution. However, it remains unclear just how early a Large Language Model (LLM) internally commits to an eventual outcome. We probe this by training linear classifiers on hidden states after the first *t* reasoning tokens, showing that eventual correctness is highly predictable after only a few tokens, even when longer outputs are needed to reach a definite answer. We show that, for harder questions, a drop in predictive accuracy highlights a selection artifact: hard items are disproportionately represented in long CoTs. Overall, our results imply that for reasoning models, internal self-assessment of success tends to emerge after only a few tokens, with implications for interpretability and for inference-time control.

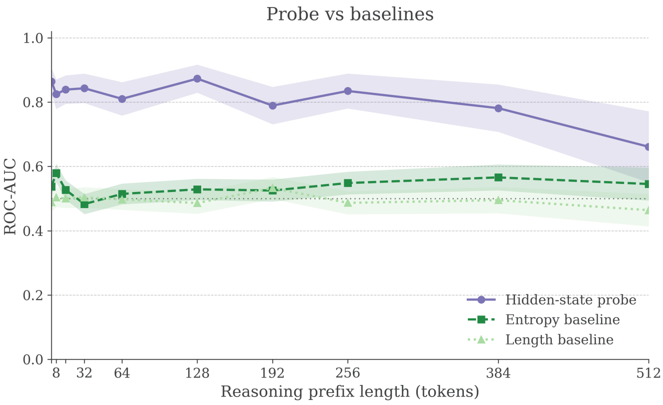

Figure 1: Probe performance vs. entropy and length baselines, depending on CoT length.

1 Introduction

Late 2024 and early 2025 saw the first large-scale implementations of CoT finetuning and prompting, notably in OpenAI's *o1* [OpenAI et al.] and Deepseek's *R1* [DeepSeek-AI et al.] models. This led to LLMs achieving unheard-of performance in cognitively demanding tasks, especially in mathematical problem-solving and commonsense reasoning. The mechanism behind these improvements consists in pushing the model to generate a step-by-step path to a solution, openly formulating calculations or deductions, possibly reflecting on possible mistakes that could be made along the way before settling on a final answer. CoT generations are not mechanistically different from *regular* token generation; they are simply specialized to problem-solving. Thus, deciding whether a model's latent representation has already converged onto a definite answer, or whether all of the CoT remains necessary to reach it, is an important but non-trivial problem. A model able to assess when it is honing in on the right solution may enable early error detection, better calibration, or dynamic halting of reasoning to save computation [Mao et al.].

In this work, we try to assess whether a model is on a trajectory to give the correct solution *throughout* a CoT rollout, before the final answer is produced. Prior work on "early correctness" probes has mostly targeted (i) question-only states or (ii) short-form QA, where surface correlations and entropy cues are strong. Here we consider the harder setting of math-focused CoT with variable-length reasoning, where such cues are weaker. Our experiments reveal a surprising result - a detectable correctness signal generally emerges very early into the reasoning process, even for longer CoT outputs. To ensure robust evaluation, we compare probe performance on the same set of problems across different prefix lengths t. Our analysis confirms that the model's internal state truly encodes, to a significant extent, whether the answer will be correct well before outputting it.

Finally, we highlight how our approach differs from existing methods. Prior works on confidence estimation largely rely on post-hoc signals (the final answer logits or multiple sampled solutions) and added mechanisms at inference time. In contrast, we directly probe the model's own hidden trajectory during reasoning, without requiring any extra decoding passes or fine-tuning of the model. Our results show that a simple linear probe is enough to tap into the model's self-monitoring capabilities. By focusing on challenging long-form math problems, we uncover an early internal self-assessment that was previously elusive (earlier question-only probes struggled on math). In summary, our contributions include:

- **Probing during CoT generation.** We predict answer correctness directly from an LLM's hidden state *while* the chain-of-thought is being generated, not only after initial question or the final answer, using a fixed, auditable recipe.
- **Early, linearly decodable correctness signal.** We show that hidden states from very short prefixes (e.g. $t \approx 4$) already contain linearly decodable information about eventual correctness, reaching AUCs up to **0.8+** on MATH-like data, which shows the model's internal computation anticipates outcome likelihood before verbalization.
- **Difficulty-driven temporal effect.** We highlight that apparent performance "degradation" at later CoT steps is largely a *temporal selection effect*: longer traces correspond to harder items, which both lengthens CoT and makes confidence signals sparser, rather than the hidden-state signal itself vanishing.

2 Related Work

2.1 Probing hidden activations for correctness and factuality.

A growing line of work shows that LLM hidden states encode latent information about truthfulness and correctness that may not surface in its token-space output. [Cencerrado et al.] extract the residual activation right after the question is processed by the model (i.e., before any answer token is produced) and train a linear probe to predict whether the model's eventual answer will be correct. While able to reach above-chance success prediction on trivia-style QA, subsequent results of this work suggest that mathematical reasoning makes it harder to linearly decode internal confidence. Our work expands on this, and targets this gap by (i) focusing on math CoT and (ii) probing hidden states *during* the reasoning trajectory, not only in the pre-answer state.

Similarly, [Gekhman et al.] introduce the *Inside-Out* framework to quantify factual knowledge stored in hidden representations. They show that models frequently encode more correct facts internally than they actually express, with hidden representations yielding markedly higher recall than the generated text. Remarkably, they show that a model can internally embed a fact (with the correct answer being ranked highest in latent space) yet systematically fail to output it. These results strengthen the view that latent activations contain rich information signals, not fully exploited by standard decoding. We build on this observation in the context of CoT reasoning.

2.2 Early stopping in CoT reasoning.

Long CoT traces are expensive, and several methods try to stop decoding once the answer is effectively determined. ES-CoT [Mao et al.] asks the model for its current answer at several points during the procedure; using the length of the resulting *step-answer* compared to other step-answers as a criterion for convergence. While this approach yields substantial token savings with minimal accuracy loss, it relies entirely on output-space heuristics and does not exploit the model's internal signals. This behavior could lag behind the model's internal confidence, and the model might already carry a signal unique to the answer before it begins to repeat it in output. HALT-CoT [Laaouach et al.] instead monitors the entropy of the next-answer distribution and stops when entropy falls below a threshold, cutting 15–30% of tokens on GSM8K while keeping accuracy within about 0.4 points of full CoT. By contrast, our probe uses *internal* signals: it aims to predict correctness from the hidden state *before* the answer has converged. This potentially enables even earlier interventions - e.g., switching to self-reflection or to an alternative decoding strategy as soon as the internal probe flags low confidence; making our method complementary to ES-CoT and HALT-CoT.

2.3 Latent Reasoning and CoT

Our work builds on chain-of-thought prompting [Wei et al.], which showed that providing structured intermediate steps allows LLMs to solve arithmetic and logical problems with much greater consistency than naïve prompting. Prior CoT research has noted that longer reasoning is not always better, models can "overthink" easy questions and introduce errors in late steps [Kojima et al., Zhou et al.]. This observation motivates internal progress monitoring: if the model's activations early in the trajectory already suggest high eventual correctness, we could truncate or adapt the reasoning on the fly. To the best of our knowledge, our results are among the first to show that in complex multi-step math settings, partial CoT hidden states are greatly decodable for eventual correctness.

3 Method

3.1 Difficulty-balanced dataset

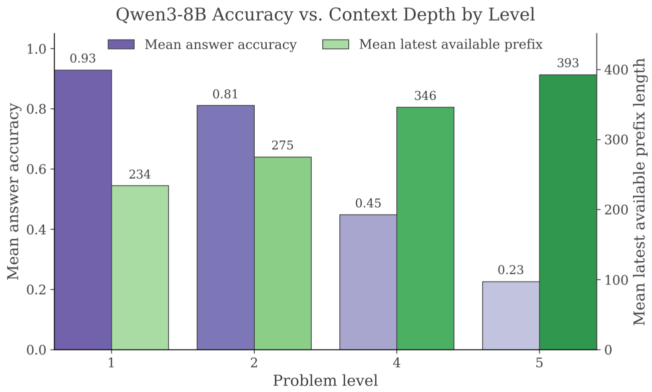

Figure 2: Mean accuracy and CoT length per difficulty level.

We have reasoning LLMs answer questions sampled from Hendrycks MATH [Hendrycks et al.]. Thanks to its explicit labeling of difficulty by question, we separate questions into two main classes: levels 1 and 2 populate the *easy* bucket, and levels 4 and 5 the *hard* bucket. We draw 750 problems per bucket, yielding a total of 1,500 examples. This yields a well-rounded spread in reasoning length and complexity, with Qwen3-8B solving easy instances 85.1% of the time, while hard ones drop to 32.1%, giving a wide accuracy dynamic for probing latent confidence.

3.2 Model Choice and Generation

Because of limited access to computing resources, we restrict our testing to two 8B models:

- **Qwen3-8B** - a model explicitly fine-tuned for reasoning. We enable its structured reasoning mode and initiate generation with the `<think>` token.
- **Llama3.1-8B-Instruct** - an instruction-tuned model used to demonstrate that even models not tailored for reasoning exhibit a confidence signal above chance.

For both models, we choose greedy decoding and set `max_new_tokens`=512.

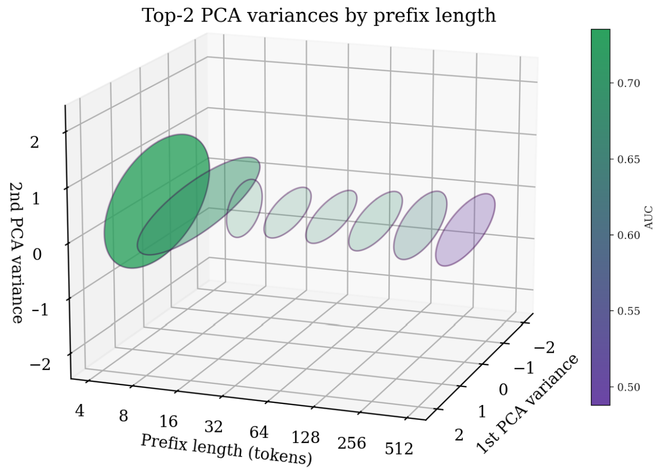

Figure 3: Amplitude of the top-2 components of the PCA for Llama3.1-8B-Instruct.

3.3 PCA and Linear Probe

For every generated solution, we pool the final hidden states of the last four reasoning tokens at fixed prefix lengths $t \in \{4, 8, 16, 32, 64, 128, 192, 256, 384, 512\}$. For each state, we then run PCA down to at most 128 components, then fit an ℓ_2-regularized logistic regression probe to the resulting components. We stratify an 80/20 split and balance class weights to offset label skew. Evaluation reports accuracy and ROC-AUC alongside the class prior in the training fold.

4 Experiments and Results

4.1 Experimental setup

All reported numbers come from the balanced MATH split described in Section 3. Because long reasoning chains are rarer, the effective sample size drops from 1,500 prefixes at t=4 to 566 at t=512.

4.2 Prefix-wise probe accuracy

In a similar fashion to [Cencerrado et al.], we find that the probe is already highly predictive after four reasoning tokens: **Figure 4** shows an average ROC-AUC of **0.84** and an accuracy of **0.76** at t=4. The ROC-AUC and accuracy of the probe remain high - around **0.8** - well into the 100 token mark. Longer traces yield noisier statistics because fewer examples survive; beyond t=256, the label distribution becomes skewed toward incorrect answers as only genuinely hard questions remain.

4.3 Difficulty and CoT-length effects

Difficulty stratification confirms that the probe mirrors how quickly the base model commits to an answer. At t=4, the AUC is of 0.77 on easy items and 0.72 on hard items; by t=32 the gap widens (0.84 vs. 0.76). Hard problems therefore require longer prefixes before their internal state separates correct from incorrect trajectories. When we additionally condition on long chains (*fixed* $\leq$256), the early AUC stays above 0.84, revealing that the information is already present even when we force every example to continue reasoning.

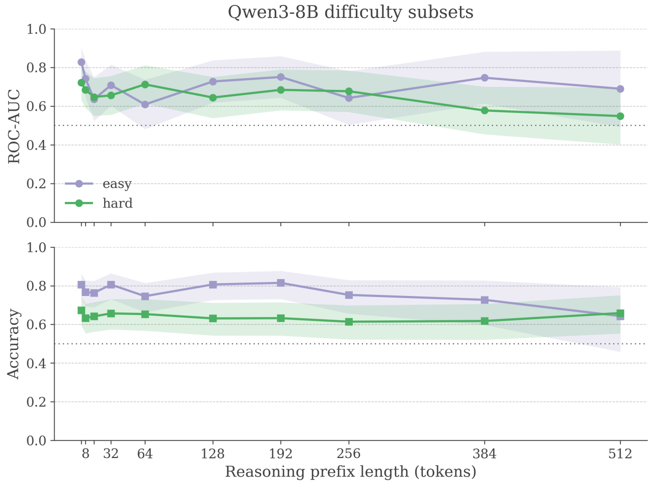

Figure 4: ROC-AUC and Accuracy of the probe for easy and hard questions.

The gradual decline in probe performance seen in **Figure 2** with increasing prefix length does not indicate that the model's confidence deteriorates as reasoning progresses, but rather reflects a shift in the underlying sample composition. As t grows, shorter and easier questions have already completed, leaving only the hardest examples in the remaining pool. Consequently, average probe accuracy drops because these residual items are intrinsically more difficult. This interpretation is confirmed by **Figure 4**, which shows that within each difficulty bucket, the probe's accuracy remains nearly constant across prefixes (higher for easy problems and lower for hard ones), indicating that the hidden state consistently encodes both the model's confidence and the inherent easiness of the question throughout generation.

4.4 Comparison with output-space baselines

Next-token entropy and prefix length provide weak signals relative to the hidden-state probe. Even when combined (*entropy+length*) the best baseline AUC is 0.59 at t=8, more than 0.21 below the hidden-state probe at the same prefix. At the peak around t=32 the margin widens to 0.39. These gaps demonstrate that former uncertainty heuristics cannot explain the strong latent predictability we observe.

5 Discussion and Limitations

Our results show that a linear probe needs only a few reasoning tokens before it predicts the eventual correctness of reasoning language models with high fidelity. The probe also offers a lightweight diagnostic for routing or self-monitoring systems that must decide when to re-run a solution, escalate to human review, or trigger reflective prompting.

The divergence between easy and hard items highlights a second takeaway: long chains mostly arise from inherently difficult questions rather than from meandering solutions. While we observe a modest drop in probe accuracy on the hardest slices, the carry-forward analysis shows that latent evidence for correctness is already present. The missing ingredient appears to be the model's ability to act on that knowledge: for difficult math problems, Qwen3-8B often recognizes its failing path but lacks a corrective policy. This gap aligns with recent evaluations of reflection-heavy reasoning models and points toward combining probes with targeted re-decoding policies.

Despite the strong empirical signal, several limitations are to consider. First, the study focuses on a single dataset and prompting template. This path of exploration would no doubt benefit from being ran on non-mathematical domains, which may exhibit weaker early commitment. Second, our leakage controls remove trivial answer exposures but cannot fully rule out subtle cues, such as partial numeric structure, that a linear classifier could exploit. Third, because of limited compute, long-prefix statistics are based on a few hundred examples, making the tail estimates sensitive to sampling noise and class imbalance. Finally, we only probe the final decoder layer; intermediate layers or non-linear probes might surface richer temporal dynamics.

Future work should replicate the pipeline across model sizes, modalities, and reasoning styles, and pair probe outputs with adaptive decoding strategies (e.g., selective continuation or tool-use triggers). Incorporating causal interventions, such as masking intermediate computations, resampling alternative continuations and pairing with non-greedy decoding could further strengthen the case of the probe capturing genuine self-assessment rather than simple stylistic markers.

6 Conclusion

Linear probes applied to early chain-of-thought prefixes reveal that reasoning models internalize solution quality far earlier than their explicit answers. Through difficulty-balanced sampling, we show that this signal emerges within the first few reasoning tokens and remains stable across prefixes once problem difficulty is controlled. These results indicate that models internally track their own progress and that latent self-assessment precedes verbalized reasoning. Looking forward, such early correctness signals could enable adaptive inference strategies; like halting, rerouting, or self-reflection—based on the model's own hidden confidence, and provide a new diagnostic path into understanding how reasoning unfolds inside large language models.

References

[Hendrycks et al.] Dan Hendrycks, et al. *Measuring Mathematical Problem Solving with the MATH Dataset.* In *NeurIPS*, 2021.

[OpenAI et al.] OpenAI, et al. *OpenAI o1 System Card.* arXiv:2412.16720, 2024.

[DeepSeek-AI et al.] DeepSeek-AI, et al. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.* arXiv:2501.12948, 2025.

[Gekhman et al.] Zorik Gekhman, et al. *Inside-Out: Hidden Factual Knowledge in Language Models.* arXiv:2503.15299, 2025.

[Mao et al.] Minjia Mao, et al. *Early Stopping Chain-of-Thoughts in Large Language Models (ES-CoT).* arXiv:2509.14004, 2025.

[Laaouach et al.] Yassine Laaouach, et al. *HALT-CoT: Model-Agnostic Early Stopping for Chain-of-Thought Reasoning via Answer Entropy.* OpenReview manuscript, 2025.

[Wei et al.] Jason Wei, et al. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.* In *ICLR*, 2022. arXiv:2201.11903.

[Cencerrado et al.] Iván Vicente Moreno Cencerrado, et al. *No Answer Needed: Predicting LLM Answer Accuracy from Question-Only Linear Probes.* arXiv:2509.10625 [cs.CL], 2025. https://doi.org/10.48550/arXiv.2509.10625

[Kojima et al.] Takeshi Kojima, et al. *Large Language Models are Zero-Shot Reasoners.* In *NeurIPS*, 2022. arXiv:2205.11916.

[Zhou et al.] Denny Zhou, et al. *Least-to-Most Prompting Enables Complex Reasoning in Large Language Models.* arXiv:2205.10625, 2023.