Pre-Answer States Encode Correctness but Only Marginally Separate "Cannot Solve" from "Slipped"

A Selection-Corrected Error-Awareness Probe on GSM8K

Ephraiem Sarabamoun

Abstract

When a model answers a problem incorrectly, the failure can be a genuine inability to solve it or a slip in execution on a problem it can solve. We ask whether a model linearly represents which kind of failure is underway before it commits to an answer. On GSM8K with Qwen2.5-3B-Instruct, we label each greedy-wrong problem by self-consistency, drawing K=6 samples and calling the failure a slip if at least one sample is correct (capability present) and a cannot if none are (consistent incapacity), a labeling rooted in sampling consistency [@wang2022selfconsistency] and consistency-based reliability checks [@manakul2023selfcheck]. We then train a linear probe on the pre-answer hidden state, the residual stream at the last token before the answer marker, to separate slip from cannot [@marks2023geometry], following the view that concepts are linear directions [@park2023linearrep]. Two findings, one robust and one marginal. Robustly, a correct-versus-wrong probe on the pre-answer states is well above chance (AUC 0.7600, n=800), so the pre-answer hidden state plainly carries information about whether the answer will be right. The finer slip-versus-cannot separation is only marginal once we correct for layer selection: the best-layer AUC is 0.6547 (n=136), but because we select the best of five layers, the honest significance test is a selection-matched permutation null that re-selects the best layer on each shuffle, against which the observed value gives p=0.040 and sits at roughly the null's 96th percentile, below its 97.5th percentile of 0.659. So the separation does not clear the selection-matched null's upper bound, and a naive single-layer null overstates it. Combined with a small cannot class (37 of 136 wrong items) and a label that conflates the failure mode with problem difficulty and confidence, we report the slip-versus-cannot result as weak, selection-sensitive evidence rather than an established effect.

1. Introduction

Models carry usable internal signals about whether they are right: hidden-state probes detect statement falsehood [@azaria2023internal], latent correctness structure is recoverable without supervision [@burns2022ccs], and models are partially calibrated about their own correctness [@kadavath2022know]. A finer question for error-aware oversight is whether the model distinguishes kinds of failure. Two error-prone situations are very different for a monitor: a slip, where the model could solve the problem but a single rollout went wrong, and an incapacity, where the model cannot solve it at all. If the pre-answer hidden state linearly encodes which is happening, a cheap probe could flag recoverable slips separately from genuine knowledge gaps. We test this directly, using self-consistency sampling to assign the labels [@wang2022selfconsistency], a discrete cousin of semantic-uncertainty estimation [@kuhn2023semantic], and a linear probe on the residual stream to read them, with truthfulness-direction work as precedent that such concepts are linearly available [@li2023iti] and readable as activation directions [@zou2023repe].

2. Method

We evaluate Qwen2.5-3B-Instruct [@qwen2024report] on 800 GSM8K test problems [@cobbe2021gsm8k], prompting for brief step-by-step reasoning [@wei2022cot] ending in a line "#### " and parsing the final integer. For each problem we greedy-decode once (the run we probe) and draw K=6 samples at temperature 0.8 to estimate capability. A greedy-wrong problem is labeled slip if at least one of the six samples is correct and cannot if none are, a discrete form of sampling-based uncertainty [@wang2022selfconsistency; @kuhn2023semantic; @manakul2023selfcheck]. The pre-answer hidden state is the residual stream at the last token before the "####" marker, obtained by re-encoding the prompt and the model's own reasoning truncated before the answer; we take it at five evenly spaced layers. We train a logistic-regression probe on a principal-component reduction of these states with five-fold cross-validation, reporting out-of-fold AUC, and select the best layer by AUC. Three comparisons make the result interpretable: a correct-versus-wrong probe

on the same states (a sanity check that the pre-answer state carries correctness signal, since stated output can hide internal computation [@turpin2023saywhat], and chain-of-thought faithfulness is itself measurable [@lanham2023measuring]), a shuffle-label null at the best layer (100 permutations), and percentile bootstrap 95 percent confidence intervals at seed 42. Calibration of confidence is the closely related scalar quantity our discrete labels coarsen [@guo2017calibration]. A recompute script rebuilds every number from per-example probe scores.

3. Results

The labels split as expected for a capable model: of 800 problems the model answered 664 correctly at greedy decoding (83.0 percent), leaving 136 wrong, of which 99 are slips (some sample correct) and 37 are cannots (no sample correct in six tries). Figure 3 shows the split. The cannot class is the minority and the source of most of the statistical uncertainty.

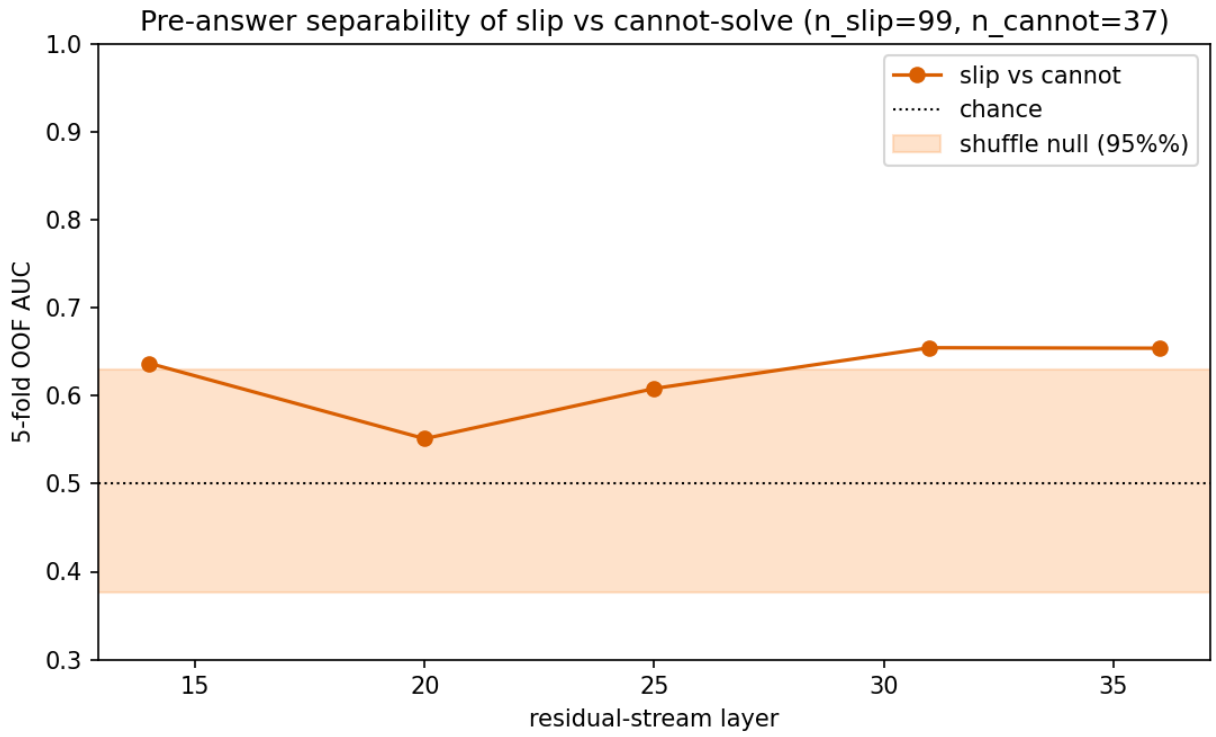

Figure 1: Figure 1. Slip-versus-cannot probe AUC across layers, with the shuffle-null 95 percent band and chance. The probe rises into the mid-to-late layers and peaks at layer 31.

The pre-answer probe separates the two failure modes, but only marginally once layer selection is accounted for. At the best layer (31 of 36) the slip-versus-cannot AUC is 0.6547 (95 percent CI 0.5494 to 0.7547, n=136); the within-layer interval excludes 0.5, and a naive single-layer shuffle null (mean 0.4943, 95 percent interval 0.3764 to 0.6294, Figure 4) is easily cleared. But the best layer was chosen to maximize AUC over five layers, so the naive null understates the chance baseline. The honest test is a selection-matched permutation null that re-selects the best layer on every shuffle: against it the observed 0.6547 gives a permutation p of 0.040 and lands at about the 96th percentile, just below the selection-matched null's 97.5th percentile of 0.659. The effect is therefore real at the 0.05 level but does not clear the selection-matched upper bound, and we treat it as marginal. Figure 1 shows the AUC rising into the mid-to-late layers, consistent with a signal that forms as the model works toward the answer.

The sanity probe behaves as it must for the result to mean anything. Correct-versus-wrong is well above chance on the same pre-answer states, AUC 0.7600 (95 percent CI 0.7128 to 0.8057, n=800), shown across

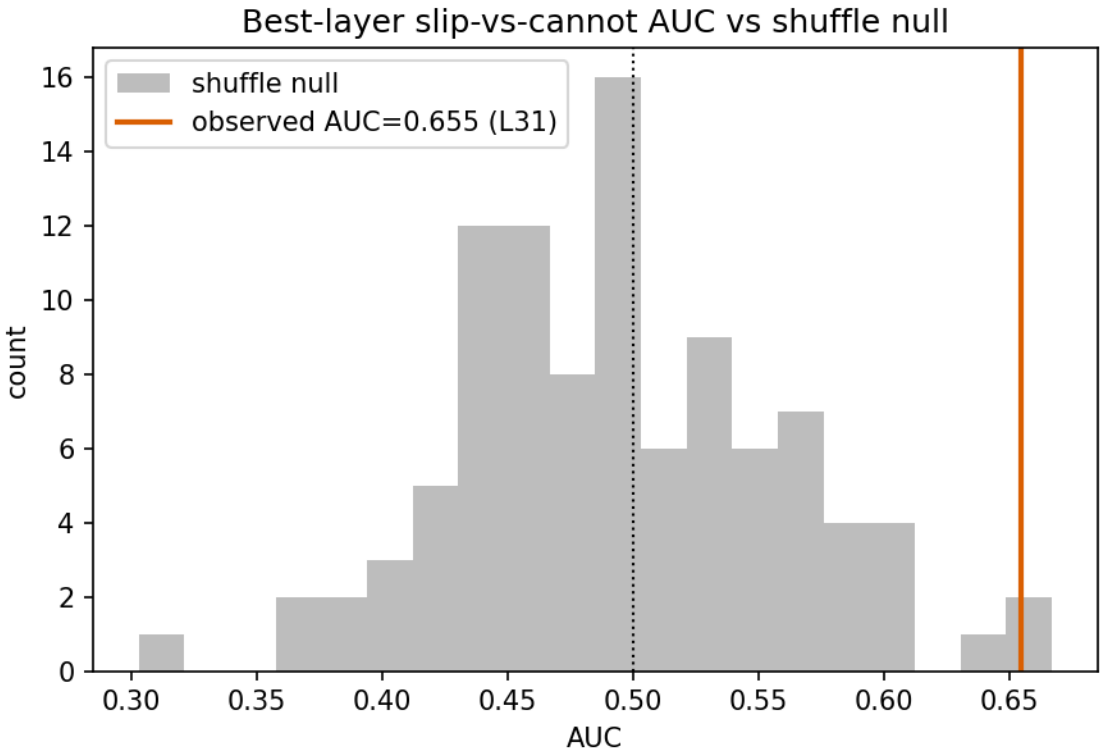

Figure 2: Figure 4. Best-layer slip-versus-cannot AUC against the shuffle-null distribution. The observed value lies above the null's upper tail.

layers in Figure 2. So the pre-answer hidden state plainly carries information about whether the answer will be right, and the more delicate finding is that, within the wrong answers, it carries graded information about whether the model could have been right.

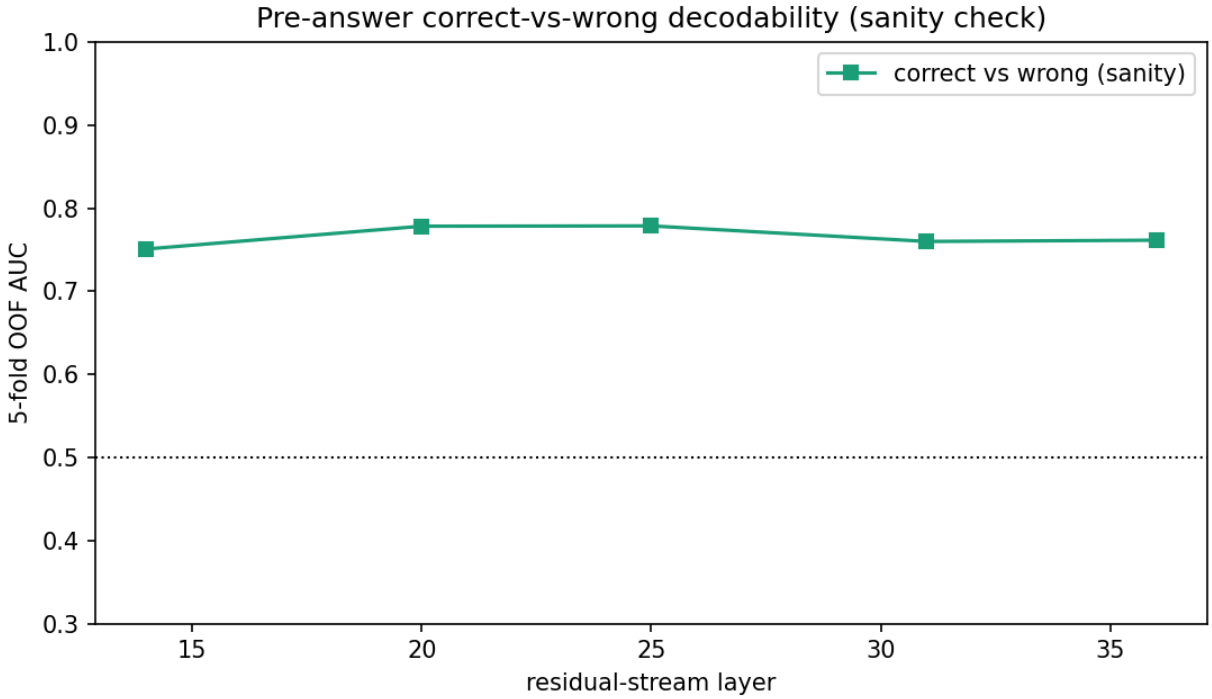

Figure 3: Figure 2. Correct-versus-wrong probe AUC across layers (sanity check). Well above chance everywhere, confirming the pre-answer state encodes correctness.

4. Discussion and limitations

The honest reading separates a robust finding from a marginal one. Robustly, the pre-answer hidden state encodes whether the answer will be correct (correct-versus-wrong AUC 0.76). Marginally, within the wrong answers, a linear probe separates "could have solved this but slipped" from "cannot solve this" at AUC 0.65, which is significant within the selected layer and at a selection-matched permutation p of 0.040, but does not clear the selection-matched null's upper bound. For error-aware oversight the encouraging part is the correctness signal; the failure-mode distinction is at best weakly present here.

The limitations bound how much to make of it. First and most important, the slip-versus-cannot effect is marginal under proper selection correction: the best-of-five-layers AUC clears a naive single-layer null but only reaches the 96th percentile of a selection-matched null (p=0.040), so the separation is not robust, and the cannot class is small (37 items) with a wide interval (0.55 to 0.75). Second, and most important, the self-consistency label conflates the failure mode with problem difficulty and model confidence: a cannot is operationally a problem the model never solved in six tries, which is also what a hard or low-confidence problem looks like, so the probe may be reading a graded capability-or-difficulty signal, closely related to calibrated confidence [@kadavath2022know; @guo2017calibration; @kuhn2023semantic], rather than a categorical slip-versus-cannot representation; we cannot separate these here and we do not claim to. Third, the cannot label is noisy at K=6, since a few more samples could reclassify a borderline cannot as a slip. Fourth, the pre-answer state is located heuristically at the last token before the answer marker, and we study one model at one size on one task. The constructive contribution is the self-consistency-labeled, pre-answer probing design and an honest, bounded positive that a higher-powered study (more cannot items, a difficulty-matched control, and a causal check) should sharpen.

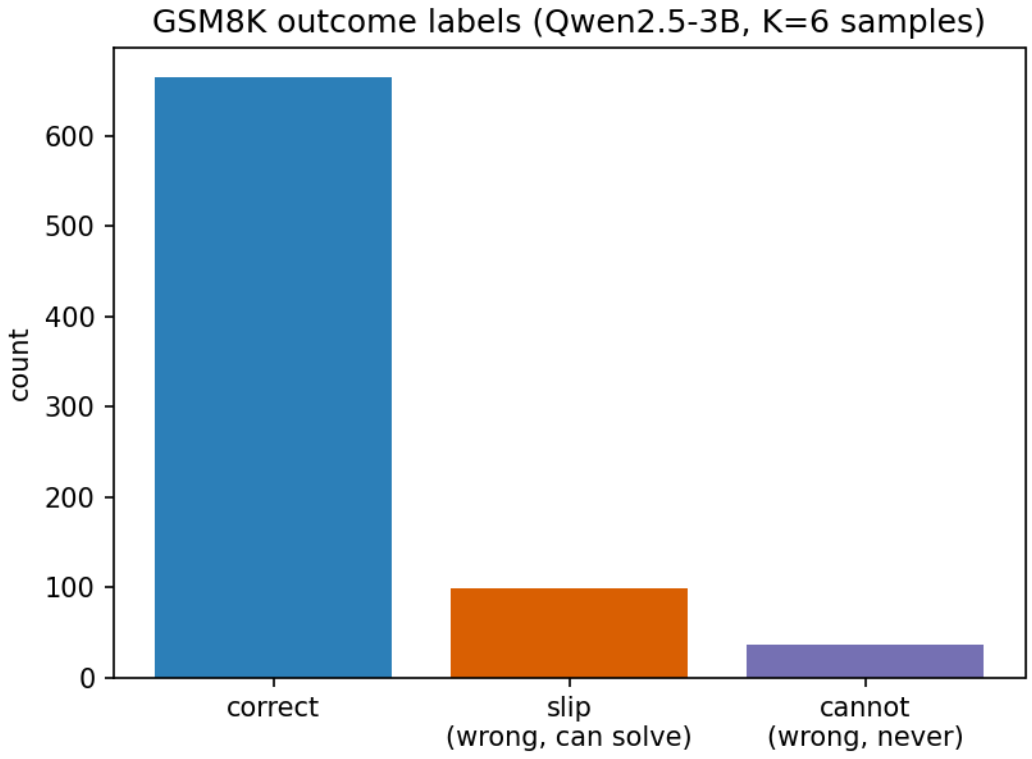

Figure 4: Figure 3. GSM8K outcome label counts: 664 correct, and among 136 wrong, 99 slips versus 37 cannots.

5. Conclusion

We asked whether models linearly distinguish a slip from an incapacity in their pre-answer hidden states. On Qwen2.5-3B-Instruct on GSM8K the answer is: marginally. The pre-answer state robustly encodes correctness (AUC 0.76), and within wrong answers a linear probe separates slip from cannot at AUC 0.65, but that separation only reaches a selection-matched permutation p of 0.040 and does not clear the selection-matched null's upper bound, so it is weak and selection-sensitive. It is also likely entangled with problem difficulty and confidence. We therefore report it as marginal evidence that the pre-answer state encodes a graded capability signal, and we point to a difficulty-matched, higher-powered rerun with a pre-registered layer as the way to tell a genuine failure-mode representation from a confidence gradient [@kadavath2022know; @marks2023geometry].