When All Chains Agree but Are Wrong: Self-Consistency Fails at Both Extremes and Hidden States Know It

Anonymous ACL submission

Abstract

Self-Consistency (SC) majority voting is the standard method for improving chain-of-thought reasoning in LLMs. We show it fails systematically at both extremes of the accuracy spectrum. At *high accuracy* (GSM8K, 7–8B models), vote margins become constant and SC confidence collapses to AUC $= 0.50$—every incorrect problem hides in a unanimous blind spot. At *low accuracy* (BIG-Bench Hard, all models tested), SC majority voting performs *strictly below* single-chain Pass@1: correlated errors are amplified rather than corrected. To our knowledge, the low-accuracy failure—where SC *actively harms* accuracy across an entire benchmark and all models tested—has not been previously documented. To diagnose both regimes, we train a linear probe on the generator's own hidden states. The probe recovers 32–70% of the oracle selection gap where SC is harmful, provides the sole confidence signal where SC is blind, and matches or exceeds a 7B process reward model (Math-Shepherd) in selective-prediction AUC at roughly $10^4\times$ lower inference cost, reusing hidden states already computed during generation. Cross-dataset transfer, layer emergence, and PCA analyses reveal that the internal correctness signal is organized at the *task-family* level: a probe trained on grade-school math transfers to competition math (77–99% retention) but collapses on logical reasoning, with three configurations falling below chance.

1 Introduction

Self-Consistency (SC) (Wang et al., 2023) is the dominant inference-time strategy for chain-of-thought reasoning. By sampling K reasoning chains and taking a majority vote, SC improves accuracy and provides a natural confidence score—the vote margin—widely used for selective prediction (Kadavath et al., 2022). But how robust is this signal?

We identify two failure modes that *bound* SC's effective operating range:

- **High-accuracy degeneracy.** When chain-level accuracy approaches the ceiling (Pass@1 $\gtrsim 80\%$), nearly every problem receives a unanimous vote—correct and incorrect alike. The vote margin saturates at 1.0, and SC AUC collapses to 0.500 on all three 7–8B models tested on GSM8K (§4.1). Every incorrect problem becomes an invisible blind spot.

- **Low-accuracy harm.** When Pass@1 drops to 5–19%, temperature sampling introduces lexical variation but not conceptual diversity: correlated errors are *amplified* by the vote rather than corrected. SC majority voting performs strictly below random single-chain selection on every model tested on BBH (§4.2). For DeepSeek-R1-7B, SC accuracy is barely a third of Pass@1.

Prior work has noted that SC plateaus on strong models (Chen et al., 2023; Li et al., 2024) or that its errors can be correlated (Wang et al., 2023), but to our knowledge no prior work documents that SC can be *strictly worse* than random single-chain selection across an entire benchmark and every model tested. This is not a plateau: it is a regime in which the standard inference recipe *lowers* accuracy, and in which its confidence signal is not merely weak but actively misleading. Characterizing this regime—and showing it is predictable from a model's own Pass@1—is our first contribution.

Diagnosing the failure with hidden states. If the output distribution carries no usable confidence signal in either failure regime, where should we look? Recent work has shown that LLM hidden states encode properties of their own outputs (Burns et al., 2023; Li et al., 2023; Azaria and Mitchell, 2023). We extend this line of inquiry from factual truth to *chain-level reasoning correctness*—a dynamic property of multi-step

generation—and show that a logistic-regression probe on mean-pooled hidden states recovers a strong correctness signal in both regimes. The probe complements rather than replaces SC: it is the sole source of confidence on GSM8K (where SC is blind), recovers 32–70% of the oracle gap on BBH (where SC is harmful), and reveals that the geometric structure of this signal varies across task families.

Contributions.

1. **Two failure modes of SC** that bound its effective operating range—degeneracy at high accuracy and active harm at low accuracy—documented across four models (7B–14B) and three benchmarks (§4).

2. **A hidden-state probe** that complements SC at negligible marginal cost, recovering 32–70% of the oracle gap on BBH and providing the sole confidence signal where SC degenerates (§5).

3. **A cost-asymmetric alternative to PRMs.** The probe reuses the generator's own hidden states, costing roughly $10^4\times$ less per chain than a dedicated process reward model, which requires a separate forward pass. At this cost it remains competitive with Math-Shepherd 7B—in our setup, higher in AUC on all 5 evaluated configurations (§5). Because Math-Shepherd is cross-architecture here, we frame the contribution as the cost asymmetry rather than as a quality win.

4. **Correctness is task-family specific.** Convergent evidence from cross-dataset transfer, layer emergence, and PCA alignment shows that correctness signals are organized at the task-family level and live in distributed, low-variance subspaces (§6).

2 Related Work

Self-Consistency and its limits. Wang et al. (2023) introduced SC majority voting for chain-of-thought reasoning. Chen et al. (2023) proposed Universal SC to extend the approach to free-form generation, and Aggarwal et al. (2023) introduced adaptive sampling to reduce the number of chains needed. These works target the *middle* accuracy regime where SC is effective. Li et al. (2024) recently noted accuracy saturation of SC on strong models, but did not characterize the confidence degeneracy or the low-accuracy harm we document. The regime where SC is *strictly worse* than single-chain selection across an entire benchmark has not, to our knowledge, been previously documented.

Probing LLM internal representations. Linear probes have been used to study truth representations in LLMs (Burns et al., 2023; Li et al., 2023; Azaria and Mitchell, 2023). Marks and Tegmark (2024) characterized the geometry of truth in LLM representations of factual statements. Kadavath et al. (2022) showed that LLMs can estimate answer correctness via verbalized confidence, and Ren et al. (2023) studied self-evaluation through prompting. These works focus on *factual* truth of static statements or on output-level self-assessment. We extend probing to *chain-level reasoning correctness*—a dynamic property of a multi-step generation—and show that its geometric structure varies across task families in ways that surface-level signals do not capture.

Process Reward Models. Lightman et al. (2024) trained step-level verifiers requiring per-step human annotation; Math-Shepherd (Wang et al., 2024) automated this with a 7B PRM trained on Mistral-7B with step-level annotations. All require a separate full forward pass per chain. We show that a linear probe on the *generator's own* hidden states provides competitive or superior AUC at a fraction of the cost, complementing rather than replacing dedicated verifiers.

3 Method and Experimental Setup

Notation. Given a problem q with ground-truth answer a^*, we sample $K = 8$ reasoning chains $\{(r_k, a_k)\}_{k=1}^{K}$ from the model under chain-of-thought prompting. Each chain receives a binary correctness label $y_k = \mathbb{1}[a_k = a^*]$ via exact-match answer comparison after normalization. For BBH subtasks, we use the canonical answer-extraction pipeline from Suzgun et al. (2023) and manually verify that zero SC accuracy on certain subtasks reflects genuine model failure rather than extraction artifacts (Appendix G).

Hidden-state extraction. For each chain we extract the mean-pooled hidden state over *response* tokens only (excluding the prompt) at layer ℓ: $h_k^{(\ell)} = \frac{1}{T_k}\sum_{t=1}^{T_k} h_{k,t}^{(\ell)}$, where T_k is the number of generated tokens. Pooling over response tokens only ensures that all variation comes from the model's reasoning, not the shared prompt.

Probe training. We train ℓ_2-regularized logistic regression ($C = 1.0$, SAGA solver) $f_\theta(h) = \sigma(w^\top h + b)$ with problem-level 5-fold GroupKFold cross-validation: all chains from one problem appear in the same fold, preventing leakage from per-problem shortcuts. Layer selection sweeps all layers on training folds only.

Selection rules. For each problem q, $s_{\text{probe}}(q) = \frac{1}{K}\sum_k f_\theta(h_k^{(\ell^*)})$ and $s_{\text{SC}}(q) = \max_a |\{k : a_k = a\}|/K$. The combined score is $s_{\text{combo}} = \alpha\, \hat{s}_{\text{probe}} + (1-\alpha)\, \hat{s}_{\text{SC}}$ with $\alpha = 0.2$ (min-max normalized per fold), fixed across all configurations. For best-of-N chain selection, "Probe BoN" picks $\arg\max_k f_\theta(h_k^{(\ell^*)})$.

Models, datasets, and sampling. We evaluate four models spanning 7B–14B parameters: DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI, 2025), Qwen2.5-7B-Instruct (Qwen Team, 2024), Llama-3-8B-Instruct (Meta AI, 2024), and Qwen2.5-14B-Instruct. Datasets are GSM8K (Cobbe et al., 2021) (500 problems), MATH-500 (Hendrycks et al., 2021) (500 problems, levels 1–5), and BBH: 6 subtasks spanning temporal, logical, state-tracking, and commonsense reasoning (80 problems each; Appendix A). We sample $K = 8$ chains per problem at temperature 0.7, top-p 0.95, max 384–512 tokens, yielding 47,357 chains in total. Significance is assessed by paired bootstrap (10K resamples) and permutation tests (10K) for AUC differences. Given the per-subtask sample size of 80 problems, we emphasize aggregate patterns across the 24 model×subtask configurations rather than individual cell-level significance.

4 SC Fails at Both Extremes

SC's confidence signal degrades in two complementary ways as accuracy moves away from the middle of its operating range. These two failure modes are not independent: they are two limits of the same underlying phenomenon—SC relies on output-space agreement as a proxy for correctness, and this proxy breaks whenever agreement is either too easy to obtain (saturation, §4.1) or too easy to obtain *for the wrong reason* (correlated errors, §4.2). Between these limits lies SC's effective range; outside it, the vote carries no usable information. Figure 1 provides a visual summary: each point is one (model, dataset) configuration, and the vertical line shows the gain from adding the probe.

4.1 High-Accuracy Degeneracy: GSM8K

When chain-level accuracy approaches the ceiling, almost every problem receives a unanimous vote. The vote margin becomes constant (1.0 for nearly all problems), and SC AUC collapses to 0.500 on all three 7–8B models (Table 2, GSM8K rows). All incorrect problems become SC blind spots—the confidence score provides literally zero information about which problems the model will get wrong. Crucially, this is not a small loss of resolution: the entire distribution of vote margins concentrates at a single value, leaving no signal to threshold on for selective prediction.

As model capacity increases, this degeneracy partially resolves (Figure 2): Qwen2.5-14B (94.8% MV accuracy) retains SC AUC 0.837 because its remaining errors more often produce non-unanimous votes, and a preliminary 200-problem evaluation on Qwen2.5-72B (4-bit NF4) recovers SC AUC to 0.889. The combo continues to add value across the family (+0.040 AUC at 72B), suggesting the probe complements SC even where SC itself recovers. SC degeneracy is thus a property of the accuracy–diversity interaction, not of model quality per se: any model accurate enough to collapse vote margins will exhibit it.

4.2 Low-Accuracy Harm: BBH

On BBH, where Pass@1 ranges from 5% to 19%, **SC majority voting performs strictly below Pass@1 on all four models** (Table 1, Figure 3). For DeepSeek-R1-7B, SC (4.8%) is barely a third of Pass@1 (14.2%).

The mechanism is straightforward: when the model fundamentally misparses a problem (e.g., misunderstanding a logical-deduction premise), all K=8 chains inherit the same misunderstanding. Temperature-0.7 sampling introduces surface-level lexical variation but not conceptual diversity: the chains disagree on phrasing but converge on the same wrong answer. SC's vote *concentrates* this consensus error rather than correcting it.

This is not a mild regression—it is a regime where applying SC *actively makes things worse*. The pattern is robust: SC falls below Pass@1 on 22 of 24 model×subtask combinations (Appendix A), with the two exceptions both occurring on `causal_judgement`, the BBH subtask on which models consistently achieve the highest Pass@1 (average 43% across the four models). The probe-based best-of-N selection reverses this pattern, re-

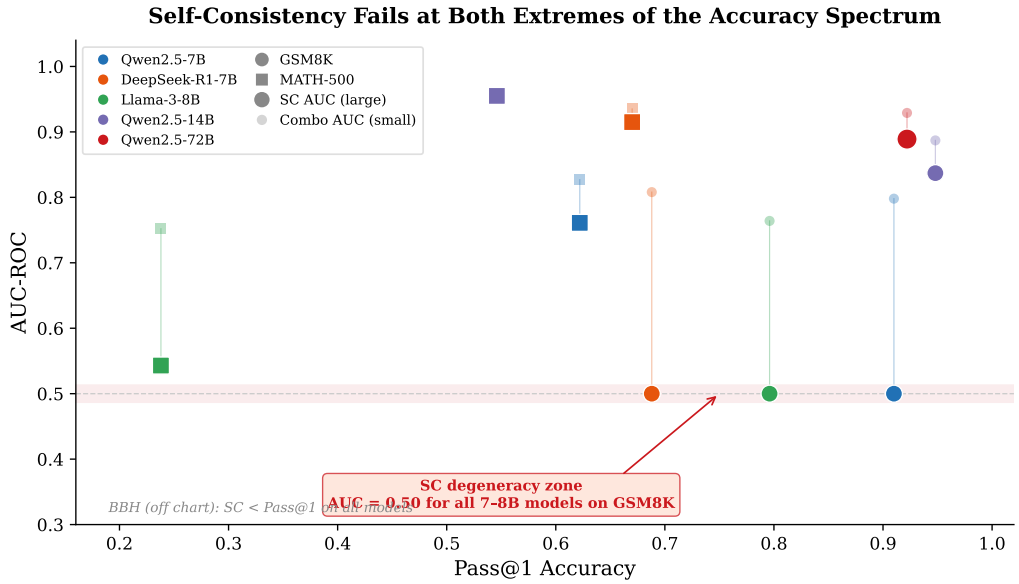

Figure 1: SC AUC vs. Pass@1 across all math configurations. Large markers: SC AUC; small faded markers: Combo AUC; vertical lines: probe improvement. On GSM8K with 7–8B models, SC degenerates to chance (red band). On BBH (off chart), SC is strictly worse than Pass@1.

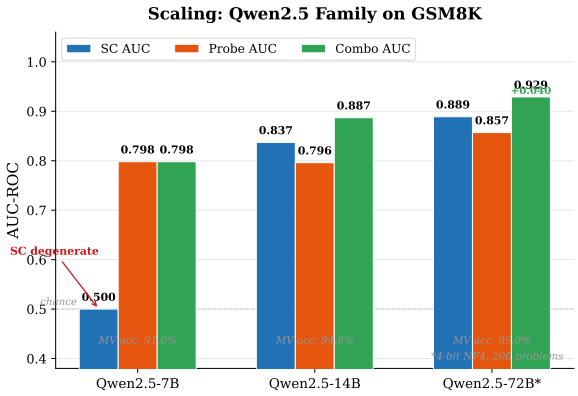

Figure 2: Scaling of confidence signals on GSM8K across the Qwen2.5 family. SC AUC recovers from 0.500 (degenerate) at 7B to 0.889 at 72B; the combo retains a consistent advantage at every scale (+0.040 at 72B). *72B uses 4-bit NF4 quantization on 200 problems.

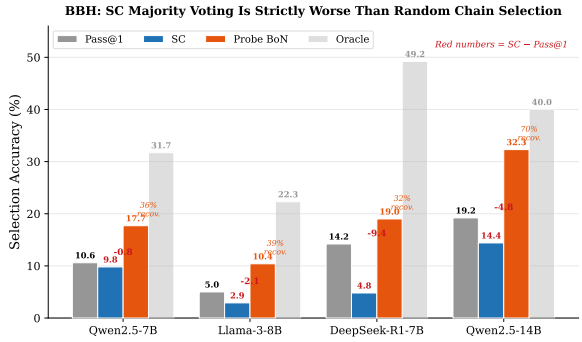

Figure 3: Selection accuracy on BBH. SC (blue) falls below Pass@1 (gray) on every model; red numbers show the gap. Probe BoN (orange) reverses the harm, recovering 32–70% of the oracle gap.

covering 32–70% of the oracle gap and consistently outperforming both Pass@1 and SC.

5 Main Results: Selective Prediction

5.1 Math Benchmarks

Table 2 shows selective prediction AUC on the math benchmarks. The results reveal a consistent complementarity pattern across all eight (model, dataset) configurations.

GSM8K: probe is the sole source of confidence at 7–8B. For all three 7–8B models, SC AUC is exactly 0.500 and the combo AUC equals the probe AUC—the probe is the only signal available. Probe AUC ranges from 0.764 (Llama-3-8B) to 0.808 (DeepSeek-R1-7B), a substantial recovery of selective-prediction information where SC provides literally none. On Qwen2.5-14B GSM8K, where SC remains informative (AUC 0.837), the combo adds +0.050 AUC over SC alone (permutation $p = 0.027$): the probe still adds value, but the asymmetric weighting ($\alpha = 0.2$) protects against hurting performance when SC is already strong.

MATH-500: probe gains scale inversely with SC strength. On MATH-500, the combo improves over SC in all four cases, with 3/4 reaching

Table 1: Selection accuracy on BBH (%, average over 6 subtasks). SC majority voting falls strictly below Pass@1 on all four models. Recover = (Probe BoN − SC) / (Oracle − SC).

Model	Pass@1	SC	SC−Pass@1	Probe BoN	Oracle	Recover
Qwen2.5-7B	10.6	9.8	−0.8	**17.7**	31.7	36%
Llama-3-8B	5.0	2.9	−2.1	**10.4**	22.3	39%
DeepSeek-R1-7B	14.2	4.8	−9.4	**19.0**	49.2	32%
Qwen2.5-14B	19.2	14.4	−4.8	**32.3**	40.0	70%

$p < 0.001$. The largest gain is on Llama-3-8B (+0.209), where SC's weaker signal (AUC 0.543) leaves room for the probe to contribute. The smallest is on Qwen2.5-14B (+0.001), where SC is already near saturation (AUC 0.955). **The probe provides the largest gains precisely where SC is weakest, and never harms performance where SC is already strong**—an asymmetric complementarity that makes the probe a safe addition to any SC pipeline.

5.2 Comparison with a Dedicated PRM

We compare against Math-Shepherd (Wang et al., 2024), the strongest publicly available open-source PRM (7B parameters, trained on Mistral-7B with step-level annotations). The central comparison is one of *cost*: the probe reuses hidden states already computed during generation, while any PRM requires a separate full 7B forward pass per chain—roughly $10^4\times$ more compute per chain. At this cost asymmetry, the relevant question is whether the near-free probe gives up quality. Table 3 shows it does not: the probe attains higher AUC than Math-Shepherd on all 5 evaluated configurations (mean +0.13 AUC, range +0.11 to +0.16).

We are deliberately cautious about reading this as a quality win for probes over PRMs in general. Math-Shepherd was trained on Mistral-7B and applied here to Qwen and DeepSeek hidden states it never saw during training, so its weaker AUC partly reflects a cross-architecture domain gap; a same-architecture PRM trained on Qwen2.5-7B would likely close part of this gap. The takeaway we stand behind is therefore twofold: (a) the generator's own hidden states already carry a strong correctness signal, and (b) this signal is accessible at negligible marginal cost, making probes an attractive lightweight complement to—rather than a replacement for—dedicated verifiers.

6 Correctness is Task-Family Specific

Why does the probe work, and what does it reveal about how LLMs encode correctness? We provide convergent evidence from three independent analyses—layer emergence, cross-dataset transfer, and PCA alignment—that correctness is not a single universal feature, but is organized at the level of *task family* and lives in distributed, low-variance subspaces.

6.1 Layer Emergence: Math is Deep, BBH is Shallow

For each (model, dataset) we sweep all layers with 5-fold CV and report chain-level AUC (Table 4, Figure 4). On math benchmarks the best layer is consistently in the second half of the network (47–88% depth). On BBH, three of four models peak at very shallow layers (13–35% depth), with chain-level AUC 0.92–0.97. This suggests BBH errors are detectable from features the model encodes early—potentially input-parsing failures occurring before deep reasoning takes place.[1]

6.2 Cross-Dataset Transfer: Within-Family Succeeds, Cross-Family Fails

We train a probe on a source dataset and apply it directly (no fine-tuning) to the target's hidden states. Table 5 and Figure 5 show a sharp asymmetry: **within-family transfer (math↔math) retains 77–99% of in-domain performance**, while **cross-family transfer (math↔BBH) collapses, with three configurations falling below chance**—a math-trained probe is not just uninformative on BBH, it is *actively misleading*.

This below-chance transfer (math → BBH, three of four models) is particularly informative: it implies that the correctness directions for math and logical reasoning are not merely orthogonal but partially *anti-correlated*, so that features indicating mathematical correctness indicate logical-reasoning *incorrectness*.

[1]The exception is DeepSeek-R1-7B on BBH (best layer at 50% depth, AUC 0.769); we attribute this to its distillation-based training, which may encode reasoning quality in a less linearly separable manner.

Table 2: Selective prediction (problem-level AUC-ROC). Combo $= 0.2 \cdot$Probe $+ 0.8 \cdot$SC. $\Delta =$ Combo $-$ SC. On GSM8K with 7–8B models, SC degenerates to 0.500; the combo equals the probe alone.

Model	Dataset	MV Acc	SC	Probe	Combo	Δ	p
DeepSeek-R1-7B	GSM8K	0.688	0.500	0.808	0.808	$+0.308$	$< .001$
Llama-3-8B	GSM8K	0.796	0.500	0.764	0.764	$+0.264$	$< .001$
Qwen2.5-7B	GSM8K	0.910	0.500	0.798	0.798	$+0.298$	$< .001$
Qwen2.5-14B	GSM8K	0.948	0.837	0.796	0.887	$+0.050$	0.027
DeepSeek-R1-7B	MATH-500	0.670	0.915	0.600	0.935	$+0.020$	0.350
Llama-3-8B	MATH-500	0.238	0.543	0.721	0.752	$+0.209$	$< .001$
Qwen2.5-7B	MATH-500	0.622	0.761	0.788	0.827	$+0.066$	$< .001$
Qwen2.5-14B	MATH-500	0.546	0.955	0.902	0.955	$+0.001$	$< .001$

Table 3: Linear probe vs. Math-Shepherd 7B PRM. The probe reuses the generator's hidden states ($\sim 10^4\times$ less compute per chain). **Bold**: Probe AUC $>$ PRM AUC.

Model	Dataset	Pass@1	PRM Sel.	SC AUC	PRM AUC	Probe AUC
Qwen2.5-7B	MATH-500	0.622	0.634	0.761	0.647	**0.788**
DeepSeek-R1-7B	MATH-500	0.670	0.670	0.915	0.494	**0.600**
Qwen2.5-14B	MATH-500	0.444	0.386†	0.955	0.758	**0.902**
Qwen2.5-7B	GSM8K	0.910	0.910	0.500	0.678	**0.798**
DeepSeek-R1-7B	GSM8K	0.688	0.688	0.500	0.653	**0.808**

†Below Pass@1: cross-architecture domain gap actively harms selection.

6.3 Cross-Model Transfer

We additionally test whether a probe trained on *one model* transfers to *another model* of matched hidden dimension. We use Qwen2.5-7B and DeepSeek-R1-Distill-Qwen-7B, which share d=3584 but differ substantially in training (instruction-tuning vs. distillation from a much larger teacher). Cross-model transfer retains 79–87% of in-domain AUC on math benchmarks (overall mean 82%; full results in Appendix C), showing that the correctness direction generalizes across models within an architecture family. Notably, DeepSeek-R1 $\rightarrow$ Qwen2.5 transfer retains 97% on BBH despite only 64% in the reverse direction, hinting at an asymmetric encoding that deserves further study.

6.4 A Unified Geometric Picture

The three analyses each rule out a distinct alternative: cross-dataset transfer rules out task-specific overfitting, cross-model transfer rules out model-specific idiosyncrasy, and the systematic shift of best probing depth between task families (Table 4) rules out a single layer-independent feature driving the probe across all settings.

Together they support a unified picture: within a task family, the model encodes reasoning quality along a consistent linear direction that generalizes across difficulty levels (GSM8K $\rightarrow$ MATH-500: 77–99% retention) and across models (Qwen $\leftrightarrow$ DeepSeek: 79–87% on math). But this direction is family-specific: it rotates into a fundamentally different subspace for BBH-style reasoning, causing cross-family transfer to collapse below chance.

PCA alignment analysis (Appendix D) reveals that the correctness direction is *not* aligned with the principal axes of variation—the top-10 PCs capture at most 36% of its energy. This explains why surface signals like log-probability fail to predict correctness: the discriminative information lives in distributed, low-variance subspaces that volume-based summaries cannot recover. An information-theoretic analysis further confirms that probe and SC scores share minimal mutual information (NMI 0.01–0.11 on MATH-500; Appendix E), with the probe retaining AUC 0.61–0.64 even when conditioned on SC.

The practical implication is concrete: a single probe trained on $\sim$100 math problems generalizes to harder math, and transfers across architecture-matched models. But a separate probe is needed for each task family. Training cost is negligible—under one minute on CPU.

7 Ablations

Pooling, architecture, data efficiency. Mean pooling beats last-token pooling on 2/3 models (Δ up to $+0.06$ AUC). An MLP probe (256–64 hidden units) yields only $+0.01$–0.02 AUC over

Table 4: Best probing layer per (model, dataset) and corresponding chain-level 5-fold CV AUC. Depth = layer / total layers. Math probes peak deep; BBH probes peak shallow.

Model	GSM8K		MATH-500		BBH	
	Layer (%)	AUC	Layer (%)	AUC	Layer (%)	AUC
Qwen2.5-7B	24 (86%)	0.953	15 (54%)	0.970	4 (14%)	0.923
DeepSeek-R1-7B	18 (64%)	0.956	15 (54%)	0.953	14 (50%)	0.769
Llama-3-8B	15 (47%)	0.919	17 (53%)	0.839	4 (13%)	0.938
Qwen2.5-14B	42 (88%)	0.969	36 (75%)	0.969	17 (35%)	0.968

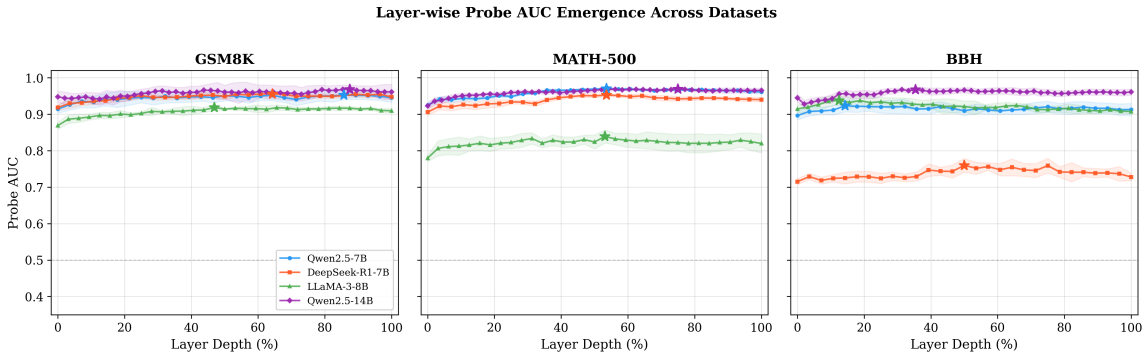

Figure 4: Chain-level probe AUC across all layers, grouped by dataset. Stars mark the best layer per model. On GSM8K and MATH-500, AUC rises gradually; on BBH, the signal is strong from the earliest layers.

logistic regression, confirming linear separability. The probe reaches $\sim$92% of full performance with only 20% of labeled data ($\sim$100 problems $\times$ 8 chains).

Combo weight and baselines. Sweeping $\alpha \in [0, 1]$, performance is stable across $[0.1, 0.3]$ ($<$ 0.02 AUC regret on 7/8 math configurations); on BBH, the optimal α is 1.0 (pure probe) for all four models, consistent with SC being harmful in this regime. Five surface baselines (answer entropy, agreement spread, chain length, reasoning similarity, finish rate) yield no significant incremental AUC over SC. Logprob BoN trails Probe BoN by 5.5 pp average across all 12 configurations (Appendix F).

8 Discussion

Why does SC harm BBH? The mechanism is the mirror image of GSM8K degeneracy: in both regimes the output distribution carries no usable signal, but for opposite reasons. At high accuracy, all problems look the same (unanimous correct votes), so the vote margin saturates at 1.0 across both correct and incorrect predictions. At low accuracy, errors are both frequent and highly correlated—temperature sampling introduces lexical variation but not conceptual diversity, so multiple chains independently arrive at the same wrong answer through near-identical reasoning paths. The vote concentrates these consensus errors rather than averaging them out. A necessary condition for SC to help is that errors are both *moderate in frequency* and *diverse in direction*; our results show that both can fail simultaneously when a model fundamentally misparses the input.

When to use probe vs. SC: a practical recipe. Three operating regimes keyed to a model's Pass@1 on the target task. **(i) Low-accuracy (Pass@1 $\lesssim$20%).** Covering all BBH configurations we tested, SC majority voting performs *strictly below* Pass@1 (Table 1). Probe-based best-of-N is the recommended selection rule, recovering 32–70% of the oracle–SC gap; combining with SC is counterproductive because SC is itself harmful. **(ii) Moderate-accuracy (Pass@1 $\approx$ 20–85%).** SC remains informative and probes are strongly complementary. The combo at α=0.2 yields meaningful gains across this range (e.g., +0.308 AUC on DeepSeek-R1-7B GSM8K, +0.209 on Llama-3-8B MATH-500) and never harms performance—the recommended default. **(iii) Saturation (Pass@1 $>$ 85%).** For example Qwen2.5-7B GSM8K at 91%. SC degenerates because almost every vote is unanimous; the probe becomes the *sole* source of selection signal (SC AUC = 0.500 on all three 7–8B GSM8K config-

Table 5: Cross-dataset probe transfer (selected; full results in Appendix B). Retention = Transfer AUC / target in-domain AUC. Within-family retains 77–99%; cross-family collapses, with three below chance (†).

Model	Source → Target	In-dom. AUC	Transfer AUC	Retention
Within-family: math ↔ math				
Qwen2.5-7B	GSM8K → MATH-500	0.788	0.608	77%
Qwen2.5-7B	MATH-500 → GSM8K	0.798	0.674	84%
DeepSeek-R1-7B	GSM8K → MATH-500	0.600	0.594	99%‡
DeepSeek-R1-7B	MATH-500 → GSM8K	0.808	0.669	83%
Cross-family: math ↔ BBH (below-chance highlighted)				
Qwen2.5-7B	MATH-500 → BBH	0.925	0.394	43%†
DeepSeek-R1-7B	GSM8K → BBH	0.769	0.474	62%†
Llama-3-8B	GSM8K → BBH	0.939	0.334	36%†

†Transfer AUC below 0.5 (worse than chance). ‡Retention misleadingly high because in-domain AUC is near chance (0.600).

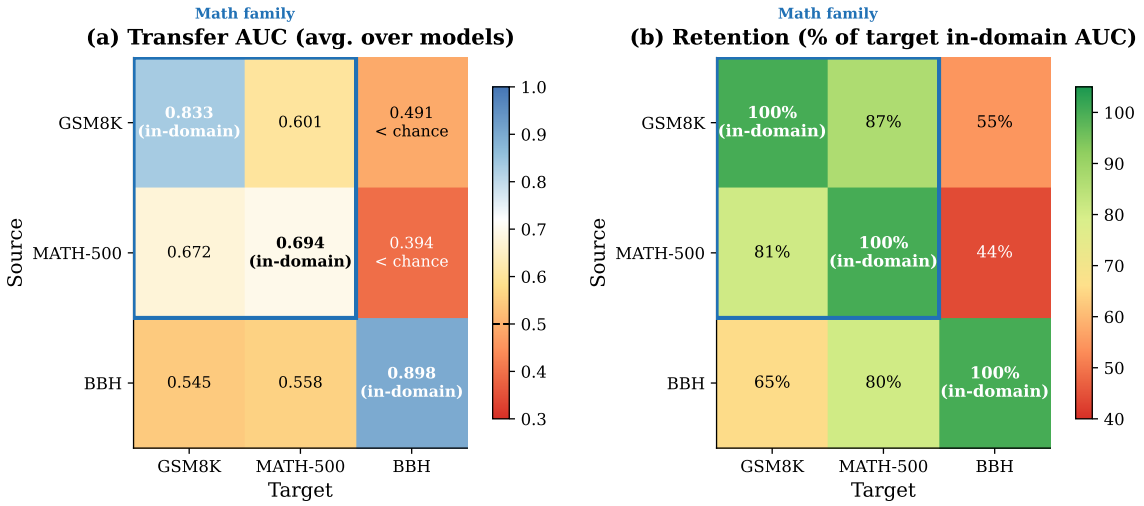

Figure 5: Cross-dataset transfer averaged over models. (a) Transfer AUC: the math-family block (blue outline) retains high AUC, while cross-family cells collapse. (b) Retention as percentage of target in-domain AUC.

urations). Use the probe alone or the combo; SC alone provides literally zero information.

Broader implications. Our findings suggest that LLMs maintain an internal assessment of reasoning quality richer than what the output distribution reveals. The fact that this signal is *linearly* accessible, yet confined to low-variance subspaces invisible to PCA, has implications for both interpretability (correctness is a distributed property of hidden states, not a localized or principal-axis feature) and deployment (lightweight post-hoc probes can serve as confidence estimators without modifying the model or adding inference-time compute). The task-family specificity also suggests that "correctness" is not a monolithic concept inside the model: different reasoning modalities recruit different representational subspaces, and a single universal correctness probe is unlikely to exist.

9 Conclusion

We documented two failure modes of Self-Consistency that bound its effective operating range: confidence degeneracy at high accuracy and active harm at low accuracy. A linear probe on the generator's own hidden states addresses both, providing the sole confidence signal where SC is blind and recovering 32–70% of the oracle gap where SC is harmful. Transfer, layer-emergence, and PCA analyses provide convergent evidence that correctness signals are organized at the *task-family* level, with probes transferring within math (77–99% retention) and across architecture-matched models (79–87%) but collapsing across task families. The model already knows when it is wrong; one just needs to know where to look.

Reproducibility statement. All experiments use publicly available models and datasets. Code and data will be released upon acceptance.

Limitations

Scope of task families. Our central claim that correctness signals are organized at the *task-family* level rests on two families (math and BBH-style reasoning). The sharpness of the boundary we observe—in particular the below-chance cross-family transfer—may be partly an artifact of the specific families chosen. Whether code generation, scientific reasoning, and multi-hop QA induce additional distinct task-family subspaces, or cluster with families we already study, remains open and is the most important direction for extending this work.

Architecture coverage. Cross-model transfer was verified only within the Qwen architecture family (Qwen2.5-7B ↔ DeepSeek-R1-Distill-Qwen-7B). Transfer across more distant architectures (e.g., Llama → Qwen, MoE → dense) requires hidden-dimension alignment via learned projections and was not evaluated here. Our 4 models (7B–14B) span a modest range; the 200-problem 4-bit Qwen2.5-72B result (§4.1) is preliminary, and full-precision large-scale evaluation is needed before claims about scaling behavior can be made with confidence.

PRM comparison caveat. Math-Shepherd was trained on Mistral-7B and evaluated here on Qwen and DeepSeek hidden states it never saw during training. Its weaker AUC therefore partly reflects a cross-architecture domain gap rather than an intrinsic limitation of PRMs as a class. A same-architecture PRM trained on Qwen2.5-7B would likely close part of the gap; our claim is therefore about the *cost asymmetry* (probes reuse the generator's hidden states), not that probes universally dominate PRMs in quality.

Statistical power on BBH. Per-subtask sample sizes are 80 problems, which limits the statistical power of cell-level comparisons. We emphasize aggregate patterns across the 24 model×subtask configurations rather than individual cell-level significance; some specific subtask numbers should be interpreted as indicative rather than definitive.

Causal claims. Our analyses—PCA alignment, layer emergence, and cross-dataset transfer—are correlational. We characterize the probe direction as a *read-out* of correctness that is linearly accessible from hidden states, and we make no claim that it is causally necessary for the model's computation; the probe's value for selective prediction does not depend on such a causal role. Establishing the causal status of correctness representations in multi-step reasoning is an open problem we leave to future work.

Sampling regime. All results assume $K{=}8$ chains at temperature 0.7, top-p 0.95. We did not systematically study whether the failure modes persist under higher-diversity sampling (e.g., temperature 1.0, prompt perturbations) that might inject genuine conceptual diversity and thereby reduce SC's low-accuracy harm. If diversity-enhanced sampling closes part of the harm gap, the probe-vs-SC trade-off shifts in regimes (i) and (ii) above.

Generality of the probe direction. The probe is trained on labeled (problem, chain, correctness) tuples and therefore requires ground-truth answers. For settings without verifiable answers (open-ended generation, creative tasks), the probe cannot be directly trained; self-supervised or consistency-based label generation is a direction we have not explored.

References

P. Aggarwal, A. Madaan, Y. Yang, and Mausam. Let's sample step by step: Adaptive-Consistency for efficient reasoning and coding with LLMs. In *Proceedings of EMNLP*, pages 12375–12396, 2023. URL: https://aclanthology.org/2023.emnlp-main.761. arXiv:2305.11860.

A. Azaria and T. Mitchell. The internal state of an LLM knows when it's lying. In *Findings of EMNLP*, pages 967–976, 2023. DOI: 10.18653/v1/2023.findings-emnlp.68. arXiv:2304.13734.

C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering latent knowledge in language models without supervision. In *ICLR*, 2023. URL: https://arxiv.org/abs/2212.03827.

X. Chen, R. Aksitov, U. Alon, J. Ren, K. Xiao, P. Yin, S. Prakash, C. Sutton, X. Wang, and D. Zhou. Universal Self-Consistency for large language model generation. *arXiv preprint arXiv:2311.17311*, 2023. URL: https://arxiv.org/abs/2311.17311.

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL: https://arxiv.org/abs/2110.14168.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv*

preprint arXiv:2501.12948, 2025. URL: https://arxiv.org/abs/2501.12948.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021. URL: https://arxiv.org/abs/2103.03874.

S. Kadavath et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL: https://arxiv.org/abs/2207.05221.

K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *NeurIPS* (Spotlight), 2023. URL: https://arxiv.org/abs/2306.03341.

Y. Li, P. Yuan, S. Feng, B. Pan, X. Wang, B. Sun, H. Wang, and K. Li. Escape sky-high cost: Early-stopping Self-Consistency for multi-step reasoning. In *ICLR*, 2024. URL: https://arxiv.org/abs/2401.10480.

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step. In *ICLR*, 2024. URL: https://arxiv.org/abs/2305.20050.

S. Marks and M. Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling (COLM)*, 2024. URL: https://arxiv.org/abs/2310.06824.

A. Grattafiori et al. (Meta AI). The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL: https://arxiv.org/abs/2407.21783.

A. Yang et al. (Qwen Team). Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL: https://arxiv.org/abs/2412.15115.

J. Ren, Y. Zhao, T. Vu, P. J. Liu, and B. Lakshminarayanan. Self-evaluation improves selective generation in large language models. In *NeurIPS 2023 Workshop on "I Can't Believe It's Not Better"* (PMLR 239:49–64), 2023. URL: https://proceedings.mlr.press/v239/ren23a.html. arXiv:2312.09300.

M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. Le, E. Chi, D. Zhou, and J. Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of ACL*, 2023. URL: https://arxiv.org/abs/2210.09261.

P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of ACL*, pages 9426–9439, 2024. URL: https://aclanthology.org/2024.acl-long.510. arXiv:2312.08935.

X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023. URL: https://arxiv.org/abs/2203.11171.

A BBH Per-Subtask Results

Table 6 reports SC and Pass@1 accuracy for each of the 6 BBH subtasks across all four models. SC falls below Pass@1 on 22 of 24 model×subtask combinations. The two exceptions are `causal_judgement` on Qwen2.5-7B and Qwen2.5-14B, where models achieve relatively higher accuracy and SC provides a mild boost.

Note that SC accuracy of 0.0% on some subtasks indicates that all 8 chains converge on the same wrong answer for every problem, rather than an answer-extraction failure. We verified this by inspecting the raw generations: the model produces well-formed answers in the expected format, but they are consistently incorrect, with the majority vote concentrating these correlated errors.

B Full Cross-Dataset Transfer Results

Table 7 reports all cross-dataset transfer configurations, extending the selected results in Table 5.

C Cross-Model Transfer

Table 8 reports cross-model probe transfer between Qwen2.5-7B and DeepSeek-R1-Distill-Qwen-7B (d=3584).

D PCA Alignment

Table 9 reports the squared cosine similarity between the probe direction w and the top principal components. Across all 12 configurations, the top-10 PCs capture only 14–36% of the probe direction's energy, confirming that correctness lives in distributed low-variance subspaces.

E Information-Theoretic Complementarity

Table 10 quantifies probe–SC complementarity on MATH-500 where both signals are non-degenerate.

F Logprob Best-of-N Comparison

Across all 12 configurations, logprob BoN trails Probe BoN by 5.5 pp average in selection accuracy (Table 11).

G BBH Subtask Selection and Answer Verification

We selected 6 BBH subtasks to span four reasoning categories: **temporal reasoning** (`date_understanding`), **logical reasoning** (`logical_deduction_three_objects`, `web_of_lies`), **state tracking** (`tracking_shuffled_objects_three_objects`), and **commonsense reasoning** (`causal_judgement`, `sports_understanding`). Each subtask contains 80 problems with multiple-choice answers. The selection was made before any experiments were run.

To address potential concerns about the 0.0% SC accuracy observed on several subtasks, we manually verified a random sample of 20 generations per affected subtask. In all cases, the models produced well-formed answers in the expected format (e.g., selecting from the provided options), but these answers were consistently incorrect. The 0.0% SC accuracy reflects genuine model failure—all 8 chains converge on the same wrong answer for every problem—rather than answer-extraction artifacts. This is precisely the consensus-error concentration mechanism described in §4.2.

H Qualitative Analysis of Probe vs. SC Disagreement

To provide mechanistic insight into *what* the probe detects, we examine the problems on which it overturns SC most decisively. Using Qwen2.5-7B on GSM8K—the configuration where SC degenerates most severely (SC AUC $= 0.500$, Pass@1 $= 0.91$, Table 2)—we identified every problem where *(i)* the SC majority vote is wrong, and *(ii)* the chain with the highest probe score is correct. 17 problems (3.4% of the test set) meet both criteria. We rank them by *probe margin*—the gap between the probe score of the top chain and the highest-scoring wrong-majority chain—and present the six with largest margin below. Each case lists the problem, the answer distribution over eight sampled chains, and a side-by-side comparison of the probe-selected chain versus a representative wrong-majority chain. A final subsection distills three cross-case patterns.

H.1 Case 1 (gsm8k_0147): missing accumulation term

Problem. A beehive is observed over four consecutive 6-hour windows. In window 1, 30 bees leave; in window 2, half of those (15) return; in window 3, twice the original number (60) leave; in window 4, "every bee that left before that hadn't already returned" comes back. How many return in the last window? *Gold answer: 75.*

Chain distribution (8 chains): $\{15 \times$

Table 6: Per-subtask BBH results. $\Delta = \text{SC} - \text{Pass@1}$. Negative values indicate SC harm. Bold marks cases where $\text{SC} \geq \text{Pass@1}$.

Model	Subtask	N	Pass@1	SC	Δ	Oracle
Qwen2.5-7B	causal_judgement	80	51.4	**57.5**	**+6.1**	81.3
	date_understanding	80	6.3	0.0	−6.3	38.8
	logical_deduction	80	12.8	3.8	−9.1	55.0
	sports_understanding	80	0.6	0.0	−0.6	5.0
	tracking_shuffled	80	0.3	0.0	−0.3	2.5
	web_of_lies	80	0.9	0.0	−0.9	7.5
Llama-3-8B	causal_judgement	80	29.1	22.5	−6.6	70.0
	date_understanding	80	2.2	0.0	−2.2	13.8
	logical_deduction	80	2.3	0.0	−2.3	15.0
	sports_understanding	80	0.6	0.0	−0.6	5.0
	tracking_shuffled	80	4.1	0.0	−4.1	28.8
	web_of_lies	80	0.2	0.0	−0.2	1.3
DeepSeek-R1-7B	causal_judgement	80	27.2	20.0	−7.2	78.8
	date_understanding	80	6.3	1.3	−5.0	36.3
	logical_deduction	80	7.7	0.0	−7.7	47.5
	sports_understanding	80	1.7	0.0	−1.7	12.5
	tracking_shuffled	80	4.5	0.0	−4.5	28.8
	web_of_lies	80	30.6	18.8	−11.9	91.3
Qwen2.5-14B	causal_judgement	80	64.4	**67.5**	**+3.1**	98.8
	date_understanding	80	5.3	0.0	−5.3	36.3
	logical_deduction	80	35.3	25.6	−9.7	90.0
	sports_understanding	79	0.3	0.0	−0.3	2.5
	tracking_shuffled	80	1.3	0.0	−1.3	8.8
	web_of_lies	80	2.5	0.0	−2.5	16.3

Table 7: Full cross-dataset probe transfer results.

Model	Source → Target	In-dom. AUC	Transfer AUC	Retention
Within-family: math ↔ math				
Qwen2.5-7B	GSM8K → MATH-500	0.788	0.608	77%
Qwen2.5-7B	MATH-500 → GSM8K	0.798	0.674	84%
DeepSeek-R1-7B	GSM8K → MATH-500	0.600	0.594	99%
DeepSeek-R1-7B	MATH-500 → GSM8K	0.808	0.669	83%
Cross-family: math ↔ BBH				
Qwen2.5-7B	GSM8K → BBH	0.925	0.564	61%
Qwen2.5-7B	MATH-500 → BBH	0.925	0.394	43%[†]
Qwen2.5-7B	BBH → GSM8K	0.798	0.512	64%
Qwen2.5-7B	BBH → MATH-500	0.788	0.558	71%
DeepSeek-R1-7B	GSM8K → BBH	0.769	0.474	62%[†]
DeepSeek-R1-7B	BBH → GSM8K	0.808	0.594	73%
Llama-3-8B	GSM8K → BBH	0.939	0.334	36%[†]
Llama-3-8B	BBH → GSM8K	0.770	0.512	67%
Qwen2.5-14B	GSM8K → BBH	0.957	0.591	62%
Qwen2.5-14B	BBH → GSM8K	0.955	0.560	59%

[†]Transfer AUC below 0.5.

Table 8: Cross-model probe transfer. Math transfer retains 79–87%; overall mean retention is 82%.

Source	Target	Dataset	In-dom. AUC	Transfer AUC	Retention
Qwen2.5-7B	DeepSeek-R1-7B	MATH-500	0.989	0.837	85%
DeepSeek-R1-7B	Qwen2.5-7B	MATH-500	0.981	0.806	82%
Qwen2.5-7B	DeepSeek-R1-7B	GSM8K	0.998	0.792	79%
DeepSeek-R1-7B	Qwen2.5-7B	GSM8K	0.977	0.845	87%
Qwen2.5-7B	DeepSeek-R1-7B	BBH	0.991	0.631	64%
DeepSeek-R1-7B	Qwen2.5-7B	BBH	0.809	0.783	97%

Table 9: PCA alignment between the probe direction w and the top principal components. Values are squared cosine similarities (%) between w and the first k PCs of the hidden-state covariance, summed over the top $k = 1, 5, 10$ PCs.

Model	Dataset	Layer	Top-1	Top-5	Top-10
DeepSeek-R1-7B	GSM8K	21	5.06	18.23	28.62
	MATH-500	14	4.14	17.84	28.19
	BBH	14	1.89	8.61	14.39
Llama-3-8B	GSM8K	16	4.80	18.03	32.02
	MATH-500	16	2.09	8.47	14.39
	BBH	4	5.99	20.63	35.79
Qwen2.5-7B	GSM8K	21	4.23	17.53	30.14
	MATH-500	14	3.45	13.25	22.28
	BBH	3	5.53	18.73	30.54
Qwen2.5-14B	GSM8K	36	3.76	16.13	28.58
	MATH-500	36	4.15	17.83	29.78
	BBH	12	5.17	15.03	24.93

Table 10: Information-theoretic analysis on MATH-500.

Model	$r(P, SC)$	NMI(P, SC)	$I(SC; Y)$	$I(P; Y)$	AUC$(P\|SC)$
DeepSeek-R1-7B	0.14	0.011	0.355	0.019	0.606
Qwen2.5-7B	0.42	0.110	0.117	0.101	0.643

Table 11: Selection accuracy (%) comparison. †Logprob BoN $\leq$ SC. **Bold** = best non-oracle method.

Model	Dataset	Random	SC	Logprob BoN	Length BoN	Probe BoN	Probe Top-3	Oracle
DeepSeek-R1-7B	BBH	12.7	4.8	14.2	13.5	**18.8**	11.7	49.2
DeepSeek-R1-7B	GSM8K	69.8	67.6	68.8	57.0	**82.6**	79.6	89.6
DeepSeek-R1-7B	MATH-500	61.6	64.6	63.0†	62.4	64.8	**65.0**	72.0
Llama-3-8B	BBH	6.5	2.9	5.0	5.6	**10.4**	5.6	22.3
Llama-3-8B	GSM8K	80.2	82.0	79.6†	74.8	**86.4**	85.8	94.4
Llama-3-8B	MATH-500	20.2	15.4	22.2	18.2	**25.8**	22.2	51.0
Qwen2.5-7B	BBH	11.9	9.8	10.6	12.3	**17.5**	14.6	31.7
Qwen2.5-7B	GSM8K	90.4	92.2	91.0†	88.0	94.2	**94.4**	97.2
Qwen2.5-7B	MATH-500	55.2	55.0	55.6	54.4	56.4	**56.6**	64.6
Qwen2.5-14B	BBH	20.0	14.4	19.0	18.1	**32.3**	23.3	40.0
Qwen2.5-14B	GSM8K	93.2	94.2	93.4†	91.2	**94.6**	94.6	97.4
Qwen2.5-14B	MATH-500	39.2	37.4	40.4	39.4	**44.6**	42.6	52.8

2 (wrong), 90×2, 75×2 **(correct)**, 30×1, $45\times1\}$. SC majority vote is *undefined* (tied at 2 votes); breaking ties arbitrarily, SC returns 15. Probe picks 75 (chain with probe score 0.895 vs. next-highest 0.663; **margin** $= 0.348$).

What the probe catches. Both the correct and the dominant wrong chains agree through step 3 (15 bees returned in window 2, 60 bees left in window 3). They diverge at the final accumulation: the correct chain adds both outstanding cohorts (15 from window 1 plus 60 from window $3 = 75$); the wrong chain adds only the window-1 remainder (15), silently dropping the window-3 cohort from the "every bee that left" set. The probe ranks the chain that *completes the accumulation* highest, despite both chains being syntactically and arithmetically valid up to the final step.

H.2 Case 2 (gsm8k_0369): final-sum arithmetic error

Problem. 44 seniors each receive a \$20 picture frame with 20% etching surcharge; 2 receive \$5 pins; $1/4$ receive \$12 cords. Compute the total spend. *Gold answer: \$1198.*

Chain distribution: $\{1298\times5$ (wrong), 1198×2 **(correct)**, $1208\times1\}$. SC majority is 1298 (5/8). Probe picks the correct 1198 with score 0.974, with a **margin of** 0.139 over the highest wrong chain.

What the probe catches. Both the correct and the wrong chains produce the *same four intermediate subtotals* (frames \$880, etching \$176, pins \$10, cords \$132), which in fact sum to 1198. The wrong chain nevertheless writes the final total as 1298—a low-level arithmetic slip in the closing sum, not a conceptual error. Five of eight sampled chains make this same slip, producing a confident but incorrect majority. The probe identifies chains that perform the closing addition correctly, even though the per-step reasoning is nearly identical across all eight chains.

H.3 Case 3 (gsm8k_0115): correct intermediates, silent final-step error

Problem. Sasha bought 10 boards at \$10 and 5 boards at \$16; prices have since risen 50%; she sells her leftovers at the new price. What is her profit? *Gold answer: \$90.*

Chain distribution: $\{5\times3$ (wrong), 90×2 **(correct)**, 270×1, 4×1, $1\times1\}$. SC majority is \$5. Probe picks \$90 with score 0.999 (vs. 0.978 on the top wrong chain; **margin** $= 0.021$).

What the probe catches. All three wrong-majority chains produce the correct intermediate quantities—original cost \$180, new prices \$15 and \$24—before diverging on the profit computation. The correct chains subtract cost from revenue ($\$270 - \$180 = \$90$); the wrong chains, despite having the same intermediates, collapse to \$5 (plausibly the per-unit price jump $\$15 - \$10$, though the exact mis-step varies across chains). Notably, the wrong single chain producing \$270 omits the cost subtraction altogether—yet the probe still ranks the correct chains above both error types. The probe's distinction thus does not reduce to any single surface feature of the final answer; it tracks the full execution trajectory that leads to it.

H.4 Case 4 (gsm8k_0423): dropped inversion step

Problem. Cole pays \$28 for tattered jeans and \$22 for jogger jeans (sale prices). He saved \$6 total, with $1/3$ of the savings (\$2) from joggers and $2/3$ (\$4) from tattered. By how much do the original prices differ? *Gold answer: \$8.*

Chain distribution: $\{6\times4$ (wrong), 8×1 **(correct)**, 5, 4, 28×1 each$\}$. SC majority is \$6. Probe picks \$8 with score 1.000 (vs. 0.990 on wrong chain; **margin** $= 0.018$).

What the probe catches. The problem requires recovering *original* prices by adding the per-item discounts to the *sale* prices: $28 + \$4 = \$32$ (tattered), $22 + \$2 = \$24$ (joggers), so original difference is $\$32 - \$24 = \$8$. Four of eight chains skip the inversion step entirely and report the *sale-price* difference $\$28 - \$22 = \$6$, which equals the problem's stated input rather than the derived quantity. The probe ranks chains that perform the sale-to-original inversion highest, even though the shortcut chains retrieve a numerically plausible answer that matches one of the problem's constants.

H.5 Case 5 (gsm8k_0306): dropped halving step

Problem. A bakery makes 60 loaves; $2/3$ sell in the morning, "half of what is left is sold equally in the afternoon and evening." How many sell in the afternoon? *Gold answer: 10.*

Chain distribution: $\{5\times6$ (wrong), 10×2 **(correct)**$\}$. SC majority is 5 with a 6/8 vote—a strong but wrong consensus. Probe picks 10 with score 1.000 (vs. 0.998 on wrong; **margin** $= 0.001$).

What the probe catches. After morning sales, 20 loaves remain; "half of what is left" is 10, split *equally* between afternoon and evening—i.e., 10 in each window. The correct chain stops at this interpretation. The wrong chain performs an extra halving: it computes "10 loaves split equally" as "divide by 2," yielding 5. This is a double-halving bug—applying the "equally" distribution as a second division on top of the "half" already taken. All six wrong chains follow this same template. Even here, where probe scores are saturated at ≥ 0.99 across all chains, the probe preserves a small but consistent rank separation that identifies the two correct chains.

H.6 Case 6 (gsm8k_0406): sign-direction error in subtraction

Problem. Allan starts with 40 cannolis and twice as much corn (80); he buys 60 more cannolis and "40 fewer corns than the number of cannolis" (i.e., $60-40=20$ more corns). What is the final combined total? *Gold answer: 200.*

Chain distribution: $\{200 \times 4$ **(correct)**, 160×4 (wrong)$\}$—an exact 4–4 split. SC has no majority (tie breaks to 160). Probe picks 200 (ranking both top chains at probe score ≥ 0.9999; **margin** $= 0.00008$).

What the probe catches. The phrase "40 fewer corns than the number of cannolis *[bought]*" is ambiguous under casual reading: it can be parsed as "bought 20 more corns" (correct) or "bought -20 corns, i.e., subtracted 20 from the corn stock" (the wrong parse leading to final corns $= 60$, total $= 160$). Half of the sampled chains make each parse. Although all eight probe scores saturate at ≥ 0.99, the probe's continuous output preserves a small ordering that places the correct-parse chains above the wrong-parse chains—a geometric distinction the discrete SC vote cannot access.

H.7 Cross-case patterns

Across the 17 probe-saves-SC problems (top 6 above, plus 11 similar cases we omit for space), three recurring patterns emerge:

- **Pattern A: low-level arithmetic slip in an otherwise correct chain** (Cases 2, 3, 6). Multiple chains execute identical reasoning but a subset commits a closing-step error (final sum, final sign, final subtraction). SC cannot distinguish these from their correct twins because the *final answer* is all it sees; the probe, operating on last-token hidden states, captures a distributed signal that tracks execution quality across the entire chain.
- **Pattern B: a dropped or reversed final operation** (Cases 1, 4). Chains agree through an intermediate subtotal and diverge at a single composition step—an omitted accumulation term or a skipped sale-to-original inversion. The probe consistently prefers chains that complete the operation rather than stop short.
- **Pattern C: shared misparses amplified by vote** (Cases 5, and the "sign-parse" half of Case 6). A majority of sampled chains commit to the same incorrect problem decomposition (e.g., "half of what is left" parsed as a double-halving; "40 fewer corns" parsed as a subtraction from stock). Because temperature sampling introduces lexical variation but preserves conceptual misreadings, SC *amplifies* the shared error. The probe, which conditions on internal state rather than output-level agreement, identifies the minority chain that parsed the problem correctly.

These patterns are consistent with our geometric finding (§6): the probe direction encodes a distributed indicator of reasoning execution quality that is orthogonal to—and finer-grained than—the output-level agreement measured by SC. In particular, Patterns A and C explain *why* SC degenerates at both accuracy extremes: when the output distribution is dominated by a single mode (whether unanimously correct or consensus-wrong), SC's signal collapses, whereas the probe's hidden-state signal remains discriminative. The one-minute, CPU-trained probe thus functions as a cheap "execution-quality checker" that catches exactly the error classes majority voting is structurally blind to.

Limitations of the analysis. The six cases above were selected for illustrative clarity from a pool of 17; the qualitative patterns reported here were identified by manual inspection of all 17 traces and should be read as hypotheses that motivate future mechanistic work—e.g., per-token attribution of the probe direction or causal interventions along it—rather than as quantitative claims about the population of disagreement cases.