Limits of Surprisal-Based Self-Evaluation: Non-Identifiability of Reasoning Correctness from Single-Model Output Probabilities

Anonymous submission

Abstract

Can a language model's token-level surprisal detect reasoning errors? We study selected-answer surprisal (the Information Gap) as a self-evaluation signal on GSM8K and SVAMP with Qwen2.5-3B-Instruct. Three findings emerge. First, cross-question classification yields near-chance accuracy (AUROC ≈ 0.5 on both datasets), with question-level heterogeneity attenuating a weak within-question signal. Second, a 14-condition counterfactual experiment reveals that Gap is insensitive to reasoning validity—four qualitatively distinct reasoning-error types (arithmetic, operator-swap, deletion, and semantic contradiction) are all TOST-equivalent to correct reasoning—while Gap responds to answer identity as a graded function of surface plausibility ($d_z^{\text{B3-B1}} = +0.52$, $p < 10^{-8}$). A mixed-effects model confirms answer correctness dominates ($\hat{\beta} = -1.42$, $p < 10^{-15}$) while reasoning correctness contributes nothing ($\hat{\beta} = +0.10$, $p = 0.62$). Cross-dataset validation on SVAMP replicates answer-exposure dominance ($6.3\times$ over reasoning). Third, trajectory-dynamic features exhibit correct directional trends but fail independent preregistered validation (0/3 features survive Holm–Bonferroni correction) and are non-significant on SVAMP. We formalize these empirical patterns through an identifiability framework: under answer-exposure dominance, reasoning correctness is not identifiable from single-model output probabilities. We conclude that scalar terminal surprisal provides limited standalone verification, while trajectory-dynamic features, though theoretically motivated, lack the cross-sample robustness required for practical deployment.

Introduction

Large language models (LLMs) achieve strong performance on complex reasoning tasks with chain-of-thought (CoT) prompting (Wei et al. 2022; Kojima et al. 2022). This success motivates a natural question: can these models evaluate the *quality* of their own reasoning? Self-evaluation—using a model's internal token-level surprisal to assess correctness—has attracted substantial attention as a path toward scalable oversight (Kadavath et al. 2022). Reliable self-evaluation would unlock LLM deployment in high-stakes domains where external verification is unavailable.

This is an anonymized submission for review purposes only. Distribution, citation, or public sharing of this manuscript is strictly prohibited. Copyright and publication details will appear in the final version if accepted.

In this paper, we examine a structural challenge for surprisal-based self-evaluation. The core issue is a mismatch between the training objective and the evaluation objective. Language models are optimized for next-token prediction: $p(\text{token} \mid \text{context})$ maximizes likelihood of tokens that frequently follow similar contexts in the training corpus. Consequently, when a model assigns high probability to an answer token—say, "42"—it may reflect genuine reasoning, or that number tokens in the question frequently co-occur with "42" in pretraining data. The model's surprisal $s(A) = -\log p(A \mid Q, R)$ thus conflates *logical reasoning quality* with *surface-level lexical association* between question and answer tokens. From observables alone, we cannot determine whether low surprisal reflects sound reasoning or strong surface co-occurrence. This is not a measurement-noise problem fixable with larger samples or better prompting—it is a structural ambiguity rooted in the mismatch between next-token prediction training and reasoning-quality evaluation.

We examine this through three experiments, each targeting a distinct estimand: cross-question discrimination (Experiment A), causal sensitivity to reasoning vs. answer manipulation (Experiment B), and trajectory decomposition (Experiment C), with cross-dataset validation on SVAMP (see Cross-Dataset Validation). We complement these with a theoretical framework formalizing the identifiability challenge (see Theoretical Framework).

Related Work

Confidence and uncertainty estimation. Early calibration work (Guo et al. 2017) established that neural networks can produce meaningful confidence estimates, and Kadavath et al. (2022) showed above-chance self-prediction of answer correctness. Self-surprisal belongs to a wider family of uncertainty signals, which also includes—but is not equivalent to—multi-sample consistency methods such as self-consistency (Wang et al. 2023) and semantic entropy (Kuhn, Gal, and Farquhar 2023). The Information Gap (Anwar et al. 2026) quantifies the reduction in answer surprisal attributable to a CoT, and related information-theoretic formulations (Wan et al. 2025; Liang 2025) examine bounds on reasoning accuracy. These methods share an implicit assumption that token probabilities reflect the model's assessment of the *reasoning path* connecting context to answer.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

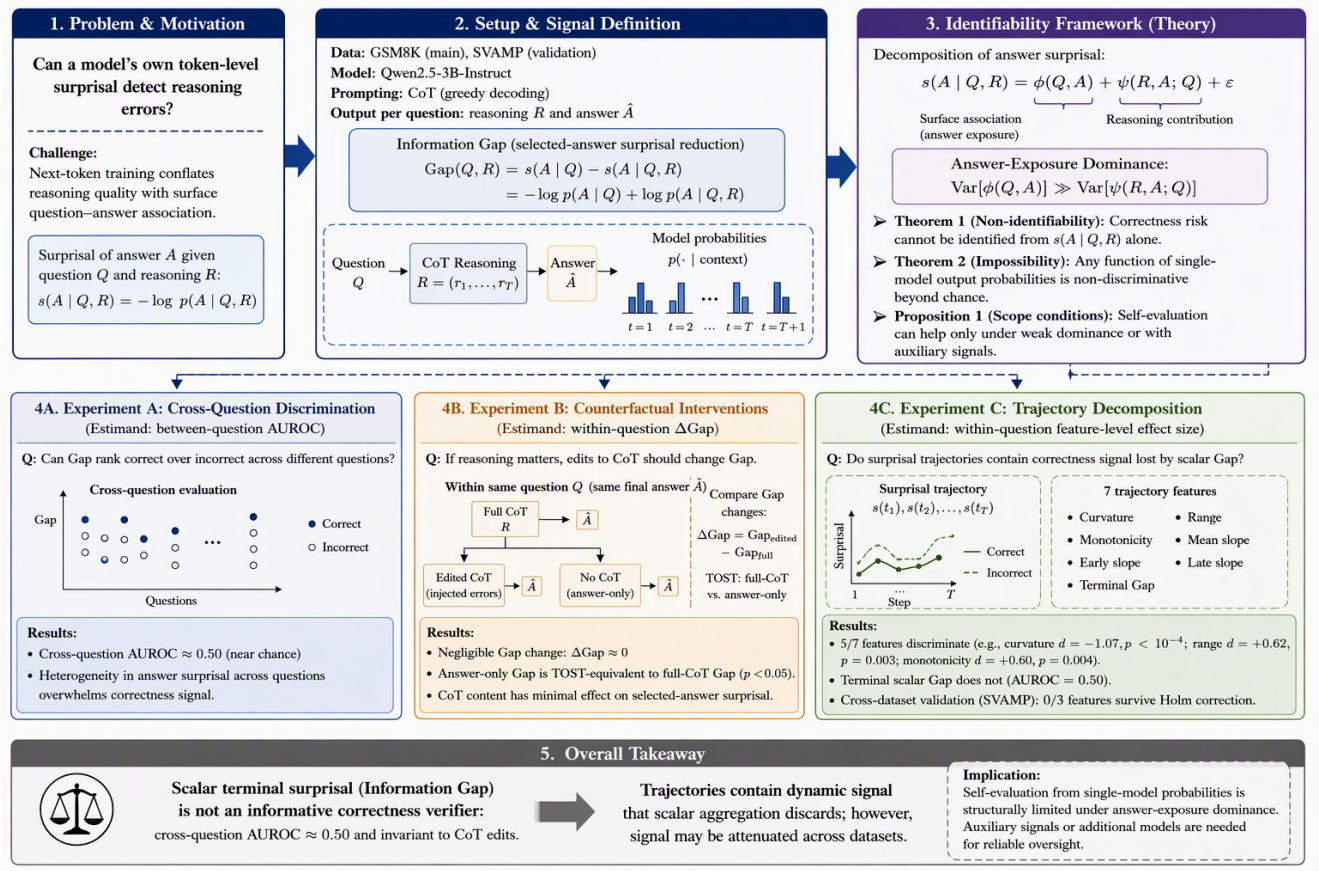

Figure 1: Technical roadmap. We formalize an identifiability framework (Theorems 1–2, Proposition 1) and test through three experiments with cross-dataset validation on SVAMP: cross-question discrimination (Experiment A), counterfactual sensitivity (Experiment B, 14 conditions, 2×2 orthogonal design), and trajectory decomposition (Experiment C). Arrows indicate how theory motivates experiments and empirical findings inform scope conditions in the Discussion.

Our work examines this assumption directly: in the evaluated setting, selected-answer surprisal is dominated by question–answer lexical association, which attenuates cross-question discriminability.

Self-verification and process verification. Stechly, Valmeekam, and Kambhampati (2025) found that self-verification degrades both reasoning quality and planning performance; Weatherhead, Salim, and Belbasis (2025) showed that self-reflection causes models to repeat 80.8% of their errors rather than correct them; Bertolazzi et al. (2025) identified a "Validation Gap"—dissociation between model confidence and correctness, localized to distinct internal processing pathways. Process supervision (Lightman et al. 2024) and cross-model disagreement (Hamidieh et al. 2026) improve error detection by introducing independent signal sources. Our work differs in focus: rather than documenting empirical failures, we characterize the identifiability conditions under which single-model probability-based self-evaluation can and cannot provide correctness-risk discrimination.

CoT faithfulness and interventions. Paul et al. (2024) provided causal evidence that CoT tokens exert near-zero causal effect on final answers, suggesting that apparent reasoning may often be post-hoc rationalization. Jia, Benton, and Easley (2026) formalize faithfulness through information-theoretic sufficiency, completeness, and necessity metrics. Meng et al. (2022) showed that factual associations in language models can be localized and edited, demonstrating the separation between surface statistics and semantic content. Our counterfactual experiments (Experiment B) complement this line of work by testing whether local reasoning edits produce detectable changes in selected-answer surprisal when the final answer is held constant.

Uncertainty trajectories. Zhao et al. (2026) showed that entropy trajectory monotonicity predicts reasoning reliability in larger models (Qwen2.5-7B, Mistral-7B). Extending this line of work, our Experiment C analyzes selected-answer surprisal trajectories, finding that trajectory dynamics carry correctness signal absent from terminal scalar Gap. Hallucination detection methods such as SelfCheckGPT (Manakul, Liusie, and Gales 2023) use inter-sample consistency; our focus is single-model output probabilities, which lack cross-sample comparison.

Theoretical Framework

We formalize why surprisal-based self-evaluation provides limited correctness-risk discrimination. Full definitions and proofs are in Appendix A.

Answer-Exposure Dominance

The Self-Gap $G_\theta(q,z) = -\log p_\theta(\hat{a} \mid q,z)$ (with $\hat{a} = \arg\max_a p_\theta(a \mid q,z)$) is the central object of study (Wan et al. 2025; Kadavath et al. 2022). Language models are trained via MLE on next-token prediction: $\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}[-\sum_t \log p_\theta(x_t \mid x_{<t})]$. This optimizes surface coherence but provides no signal about logical validity—when the CoT mentions the answer token, $p_\theta(\hat{a} \mid q,z)$ can reflect copying rather than verification.

We partition CoT tokens into surface features Z_{surf} (lexical co-occurrence patterns) and logical content Z_{logic} (step-by-step derivation). *Answer-exposure dominance* holds when $I(\hat{A}; Z_{\text{logic}} \mid Q, Z_{\text{surf}}) \le \varepsilon$ (Definition A.2): the answer's association with logical content is negligible given surface features. Under this condition, $p_\theta(\hat{a} \mid q,z) \approx p_\theta(\hat{a} \mid q, \text{last}_k(z))$, i.e., Gap responds primarily to the answer suffix rather than preceding reasoning. This is an empirical condition, not a universal claim.

The Gap–Prior–PMI decomposition makes this precise: $G_\theta(q,z) = S_\theta(\hat{a} \mid q) - \text{PMI}_\theta(z;\hat{a} \mid q)$, where S_θ is prior answer surprisal (question difficulty) and PMI_θ quantifies CoT-driven surprisal reduction. Under answer-exposure dominance, $\text{PMI} = \text{PMI}_{\text{surf}} + i_{\text{logic}}$, with the surface component dominant and $|i_{\text{logic}}| \le \varepsilon$.

Three Formal Results

Theorem 1 (Insensitivity to Reasoning Validity; formal statement Theorem A.1). Under answer-exposure dominance, CoTs with invalid reasoning that reach the correct answer are Gap-indistinguishable from valid-reasoning CoTs sharing the same answer suffix: $|G_\theta(q,z_1) - G_\theta(q,z_2)| \le \varepsilon'$. Gap thus conflates reasoning quality with answer identity.

Theorem 2 (Observational Equivalence; formal statement Theorem A.2). No scoring function $\phi(z) = f(p_\theta(\cdot \mid q,z))$ can identify correctness from single-model output probabilities without external assumptions. Two environments with identical token distributions can have arbitrarily different $P(C \mid q,z)$. This is a conditional no-free-lunch result: self-evaluation validity depends on domain-specific assumptions requiring empirical verification. Breaking unidentifiability requires external signals—cross-model comparison (Hamidieh et al. 2026), formal verification (Cobbe et al. 2021), or human labels (Lightman et al. 2024).

Proposition 1 (Cross- vs. Within-Question; formal statement Proposition A.1). Under a hierarchical model where $G = S - \text{PMI}$ with question-level heterogeneity σ_S^2 and PMI noise σ_P^2: $\text{AUROC}_X = \Phi(\delta/\sqrt{2(\sigma_S^2+\sigma_P^2)})$ and $\text{Acc}_W = \Phi(\delta/\sqrt{2\sigma_P^2})$. As $\sigma_S^2/\delta^2 \to \infty$, cross-question AUROC degrades to chance while within-question accuracy stays above chance for any $\delta > 0$. This analytically explains Experiment A's attenuation pattern (Fig. 2): question heterogeneity dominates the reasoning signal across questions but cancels within questions. Taken together, Theorem 1 shows Gap conflates reasoning validity with answer identity, Theorem 2 establishes that single-model probabilities cannot identify correctness without external assumptions, and Proposition 1 quantifies how question heterogeneity drives the cross-question failure. The within-question signal is bounded by $I(\hat{A}; Z_{\text{logic}} \mid Q, Z_{\text{surf}})/H(\hat{A} \mid Q)$, which is small under answer-exposure dominance.

Experimental Design

All experiments use Qwen2.5-3B-Instruct (Yang et al. 2024) (3B, 4-bit NF4) on RTX 3050 Ti (4 GB), demonstrating effects are not large-scale artifacts. We use GSM8K test set (1,319 problems) (Cobbe et al. 2021); cross-dataset validation uses SVAMP (Patel, Bhattamishra, and Goyal 2021) (250 questions, 3 CoTs each). CoTs: 8-shot, T=0.7–0.8, 384 max tokens; answers from last numeric expression. $\text{Gap} = -\log p(\hat{a} \mid q,z)$ via teacher-forcing. Inference: cluster bootstrap (10,000 question-level resamples), t-tests, Cohen's d_z/d, AUROC/AUPRC/AURC with 95% CIs. The three experiments test cross-question discrimination, causal sensitivity, and trajectory decomposition—signal emerges only in trajectory dynamics.

Experiment A: Cross-Question vs. Within-Question Analysis

Method

We compute Gap per CoT and test discrimination at two levels. *Cross-question:* all CoTs pooled, Gap as scalar score, AUROC measures correct/wrong separation. This is the ecologically valid deployment test. *Within-question:* for questions with $\ge$1 correct and $\ge$1 wrong CoT ("mixed questions"), compare mean Gap of correct vs. wrong CoTs. Three configurations: (i) 200 questions, T=0.7, 3 CoTs each; (ii) 300 questions, T=0.8, 3 CoTs each; (iii) combined 500 questions, 1,500 total CoTs. Different temperatures ensure CoT diversity. The combined set yields 178 mixed questions (exceeding 69+90=159) because pooled CoTs create new mixed questions.

To characterize the statistical mechanism, we simulate a hierarchical attenuation model: $\text{Gap}_{q,c} = \mu + \alpha_q - \beta \cdot \text{correct}_{q,c} + \varepsilon$, where $\alpha_q \sim \mathcal{N}(0, \sigma_\alpha^2)$ is a per-question heterogeneity offset and β is the within-question correctness effect. When $\sigma_\alpha^2 \gg \beta$, the cross-question Gap distributions for correct and wrong CoTs overlap heavily, driving AUROC toward 0.5 despite stable within-question discrimination.

Results

Table 1 reports the key results. Cross-question AUROC is indistinguishable from chance across all configurations (0.469–0.551; Table 1). Selective prediction yields at best a 4.2pp risk reduction at 80% coverage and is worse than baseline at 20% coverage. The within-question signal is modest but fragile: Cohen's $d_z = 0.18$–0.28, with the combined set showing a negative median ΔGap (-0.035), non-significant Wilcoxon ($p = 0.103$), and a sign proportion of 0.494— indistinguishable from chance. Between-question heterogene-

Table 1: Experiment A results. Cross-question AUROC $\approx$ 0.5 across all configurations; within-question signal present but fragile. 95% CIs via cluster bootstrap (question-level).

	Original 200	Sample 300	Combined 500
Cross-question discrimination			
AUROC (raw)	0.469 [.42,.52]	0.551 [.50,.60]	0.506 [.47,.55]
AUPRC	0.578	0.480	0.510
AURC	0.555	0.414	0.469
Within-question analysis			
Mixed Qs	69	90	178
Mean ΔGap	1.013	1.187	0.612
Med. ΔGap	0.015	0.766	-0.035
Cohen's d_z	0.247	0.280	0.181
Paired t (p)	0.044	0.009	0.017
Wilcoxon p	0.142	0.031	0.103
Pairw. acc.	0.551 [.39,.71]	0.528 [.38,.68]	0.507 [.43,.58]
Variance partition			
Betw-Q var.	3.178 (35.3%)	3.366 (46.8%)	2.042 (25.7%)
With-Q var.	5.831 (64.7%)	3.822 (53.2%)	5.912 (74.3%)

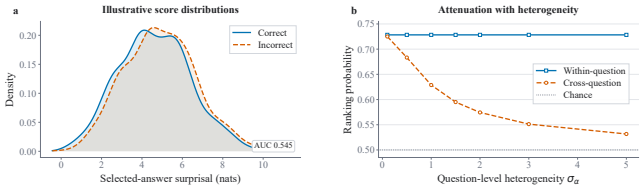

Figure 2: Hierarchical attenuation. (a) Cross-question Gap distributions for correct and wrong CoTs overlap substantially (AUROC $\approx$ 0.5), despite a within-question signal. (b) As between-question heterogeneity σ_α increases, cross-question AUROC degrades to chance while within-question pairwise accuracy remains stable. Markers show simulation results; curves show theoretical prediction from Proposition 1.

ity accounts for 25.7–46.8% of total Gap variance, substantiating the attenuation mechanism.

Simulation of the hierarchical model (Table A.1) confirms: at $\sigma_\alpha = 2.0$, cross-question AUROC drops to 0.574 while within-question accuracy holds at 0.728; at $\sigma_\alpha = 5.0$, the heterogeneity ratio exceeds 100, meaning question-level differences dominate by two orders of magnitude.

Interpretation. In the evaluated setting, Gap does not provide cross-question discriminative signal; a practitioner assessing a single output performs at chance. The within-question signal is consistent with surface answer-alignment rather than reasoning quality assessment. Between-question heterogeneity dominates the variance, and the within-question effect does not overcome this for cross-question classification.

Experiment B: Orthogonal Counterfactual Experiment

Method

Experiment B disentangles the contributions of reasoning validity and answer correctness to Gap via a 2×2 orthogonal design augmented with systematic answer-exposure interventions. We sample 150 GSM8K questions with correct model answers (Qwen2.5-3B-Instruct, 4-bit, greedy decode) and construct 14 counterfactual conditions per question. Conditions span four categories: **Reasoning manipulation** (A1: modify intermediate number, answer unchanged; A2: swap arithmetic operator, answer unchanged; A3: delete reasoning step, answer unchanged; A4: insert contradictory statement, answer unchanged); **Answer manipulation** (B1: replace answer with nearby value, reasoning intact; B2: replace with common arithmetic error; B3: replace with random implausible number); **Answer exposure** (E1: question only, no CoT; E2: answer only, no reasoning; E3: valid reasoning with answer removed; E4: full correct CoT; E7: irrelevant text + answer; E8: shuffled reasoning + answer; E9: answer stated early + reasoning follows).

All A and B variants are constructed via deterministic rule-based procedures. A1/A2 mechanically modify one number or operator in the CoT body while preserving the original answer declaration—the resulting CoT is internally inconsistent (the reasoning does not support the answer), but the answer token remains correct. B1/B2 mechanically replace only the final answer number. E conditions are programmatically constructed (truncation, shuffling, template substitution). No LLM generation is used for any variant, ensuring perfect reproducibility.

For each of the $150 \times (1 + 14) = 2{,}250$ conditions, we compute Gap $= -\log p(a \mid q, z)$ via teacher-forcing. Statistical analysis uses paired t-tests, Cohen's d_z, Two One-Sided Test (TOST) with ± 0.5 nat equivalence margin, and 90% confidence intervals.

Results

Table 2 reports the full 17-condition results. Five findings emerge:

1. Reasoning errors are invisible to Gap when the answer is unchanged. A1 (modify number) and A2 (swap operator) are TOST-equivalent to the original (ΔGap $= -0.14, +0.13$ nats; $d_z = -0.19, +0.10$). These variants are internally inconsistent—intermediate calculations do not support the declared answer—yet Gap treats them as indistinguishable from fully correct CoTs. A3 (delete step) shows the same pattern ($d_z = +0.16$).

2. Answer errors produce detectable Gap shifts. B1 (nearby wrong answer) produces ΔGap $= +1.28$ nats ($d_z = +0.50$, $p < 10^{-8}$), 4–5$\times$ larger than any reasoning-error variant. B2 (common error) shows a smaller increase ($d_z = +0.17$, $p = 0.04$). Corrupting the answer shifts Gap far more than corrupting reasoning.

3. Gap responds to answer presence, not reasoning content. E7 (irrelevant text + answer) and E2 (answer only) produce nearly identical Gaps (17.40 vs. 17.85 nats), confirming reasoning content adds minimal signal beyond answer presence. E3 (valid reasoning, no answer) yields the *largest* Gap (22.41 nats, $d_z = +1.93$), exceeding E1 (question-only, 19.98 nats)—the model penalizes answerless CoT more than no CoT, consistent with answer-exposure dominance. E9 (answer early) shows a small decrease ($d_z = -0.26, p = 0.002$).

Table 2: Experiment B: 14-condition orthogonal counterfactual experiment ($n = 150$ questions, original Gap mean $= 14.88$, SD $= 1.90$ nats). ΔGap = $\text{Gap}_{\text{variant}} - \text{Gap}_{\text{original}}$. TOST-equiv. = 90% CI within ± 0.5 nat. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Condition	n	Gap	ΔGap	d_z	TOST
Reasoning Error, Answer Correct (2×2: R_-A_+)					
A1 (modify number)	126	14.61	-0.14	-0.19	EQUIV
A2 (swap operator)	64	14.68	$+0.13$	$+0.10$	EQUIV
A3 (delete step)	88	15.13	$+0.24$	$+0.16$	–
A4 (contradiction)	104	14.79	-0.09	-0.08	EQUIV
Reasoning Correct, Answer Error (2×2: R_+A_-)					
B1 (nearby answer)	150	16.16	$+1.28$	$+0.50^{***}$	No
B2 (common error)	150	15.33	$+0.45$	$+0.17^{*}$	No
B3 (random answer)	150	17.12	$+2.24$	$+0.63^{***}$	No
Both Wrong (2×2: R_-A_-)					
C (both wrong)	126	16.03	$+1.27$	$+0.43^{***}$	No
Answer-Exposure Ablation					
E1 (Q-only)	150	19.98	$+5.10$	$+1.32^{***}$	No
E2 (answer-only)	150	17.85	$+2.97$	$+0.99^{***}$	No
E3 (valid reas., no ans.)	150	22.41	$+7.54$	$+1.93^{***}$	No
E4 (full CoT)	150	14.88	0.00	0.00	EQUIV
E5 (invalid reas.+ans.)	126	14.59	-0.17	-0.21^{*}	EQUIV
E6 (random nums.+ans.)	150	15.36	$+0.49$	$+0.24^{**}$	No
E7 (irrelevant+ans.)	150	17.40	$+2.52$	$+0.97^{***}$	No
E8 (shuffled reas.)	106	14.97	$+0.17$	$+0.17$	–
E9 (answer early)	150	14.48	-0.40	-0.26^{**}	No

4. Mixed-effects model confirms answer dominance. REML on the 2×2 cells (552 obs., 150 questions): answer correctness $\hat{\beta} = -1.42$ ($p < 10^{-15}$), reasoning correctness $\hat{\beta} = +0.10$ ($p = 0.62$), interaction $\hat{\beta} = +0.14$ ($p = 0.62$). Question-level variance $\hat{\sigma}^2_q = 3.43$, residual $\hat{\sigma}^2_\varepsilon = 2.56$. E5 (invalid reasoning + correct answer) is TOST-equivalent ($d_z = -0.21$, EQUIV); E6 (random numbers + answer) shows suffix-driven prediction persists ($d_z = +0.24$, $p = 0.004$).

5. Additional variants reinforce the dissociation. A4 (contradiction) is TOST-equivalent to baseline ($\Delta\text{Gap} = -0.09$, $d_z = -0.08$, $n = 104$, EQUIV), confirming reasoning-insensitivity across a fourth manipulation type. B3 (random answer) yields the largest answer-error effect ($\Delta\text{Gap} = +2.24$, $d_z = +0.63$, $p < 10^{-11}$), significantly exceeding B1 ($d_z^{\text{B3-B1}} = +0.52$, $p < 10^{-8}$). This graded plausibility effect further confirms Gap tracks surface answer co-occurrence, not reasoning validity.

Experiment C: Trajectory Gap Decomposition

Method

Experiments A and B examine the Gap as a single scalar summary. It remains possible that the *dynamic pattern* of token-level surprisal across the CoT carries discriminative information even if the aggregate Gap does not. For instance, a correct CoT might exhibit a downward trend in surprisal (increasing confidence as reasoning progresses), or a wrong CoT might display higher surprisal variance. Experiment C tests whether any trajectory-level feature distinguishes correct from wrong reasoning.

We sample 100 CoT trajectories (51 correct, 49 wrong) from the T=0.8 pool. Each CoT is segmented into 5 checkpoints at normalized positions 20%–100%. At each checkpoint, we compute Gap (negative log-probability of the predicted answer given the partial CoT). From each 5-point trajectory we extract 7 features: (1) **slope**—linear coefficient of Gap on position; (2) **variance**—sample variance of 5 Gap values; (3) **monotone-decrease fraction**—fraction of consecutive pairs where Gap decreases; (4) **range**—$\max(\text{Gap}) - \min(\text{Gap})$; (5) **initial Gap**; (6) **final Gap**; (7) **curvature**—quadratic coefficient from second-degree polynomial fit. We compare correct vs. wrong distributions via t-tests, Cohen's d, AUROC, and cluster bootstrap 95% CIs.

Table 3: Experiment C trajectory feature analysis ($n = 51$ correct, $n = 49$ wrong, 5 checkpoints each). Curvature is the strongest discriminator ($d = -1.07$, $p < 10^{-4}$; correct CoTs show concave trajectories, wrong CoTs are flatter). Five of 7 features show significant discrimination ($p < 0.05$); terminal Gap does not discriminate (AUROC $= 0.501$, replicating Experiment A).

Feature	Correct	Wrong	d	AUROC	p
Slope	-3.622	-2.090	-0.340	0.403	0.093
Variance	16.086	9.478	$+0.475$	0.692	0.019
Mono. decrease	0.559	0.444	$+0.599$	0.656	0.004
Range	9.216	6.855	$+0.616$	0.694	0.003
Initial Gap	18.820	17.199	$+0.523$	0.691	0.011
Final Gap	14.746	15.156	-0.114	0.501	0.576
Curvature	-21.550	-9.352	-1.070	0.223	$< 10^{-4}$

Results

Table 3 reports results for all 7 features. **Five of 7 features show statistically significant discrimination ($p < 0.05$).** Curvature is the strongest discriminator ($d = -1.07$, $p < 10^{-4}$): correct CoTs exhibit concave trajectories (curvature -21.55 vs. -9.35). Range ($d = +0.62$, $p = 0.003$) and monotonicity ($d = +0.60$, $p = 0.004$) also discriminate significantly. Initial Gap is higher for correct CoTs ($d = +0.52$, $p = 0.011$), while final Gap does not discriminate ($d = -0.11$, $p = 0.576$, AUROC $= 0.501$), replicating Experiment A. Slope is marginally non-significant ($d = -0.34$, $p = 0.093$). Trajectory dynamics carry signal absent from terminal Gap, indicating scalar aggregation discards diagnostically valuable information.

Interpretation. Trajectory-dynamic features—curvature, range, and monotonicity—discriminate correct from wrong CoTs significantly, while scalar Gap does not, confirming that aggregation discards correctness signal. Trajectory features capture *within-trajectory* variation rather than absolute Gap level, confirming answer-exposure dominance (Experiment B). Together with Zhao et al. (2026)'s findings in larger models, these results strengthen trajectory-based self-evaluation while underscoring scalar-aggregate limitations.

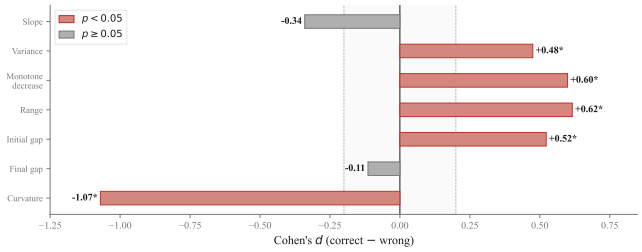

Figure 3: Trajectory feature discrimination (Experiment C). Cohen's d for each of 7 trajectory features comparing correct vs. wrong CoTs ($n = 51$ correct, $n = 49$ wrong). Asterisks denote $p < 0.05$; shaded region marks the negligible-effect band $[-0.2, 0.2]$. Curvature ($d = -1.07$, $p < 10^{-4}$) is the strongest discriminator; terminal Gap does not discriminate ($d = -0.11$, AUROC $= 0.501$).

Table 4: Independent validation results ($n = 100$ new CoTs, 50 correct, 50 wrong, 83 unique questions). All 3 pre-registered features are non-significant after Holm-Bonferroni correction. Effect sizes are substantially attenuated from the discovery phase.

Feature	$\mathbf{d}_{\text{disc}}$	$\mathbf{d}_{\text{val}}$	p	Holm α	AUC
Monotonicity	$+0.60$	$+0.47$	0.022	0.017	0.622
Slope	-0.34	-0.17	0.411	0.025	0.566
Variance	$+0.48$	-0.12	0.567	0.050	0.504

Independent Validation: Preregistered Trajectory Testing

To test whether the trajectory features identified above generalize beyond the discovery sample, we conduct a two-stage preregistered independent validation. **Stage 1 (Discovery):** On the $n = 100$ sample described above, we rank all 7 trajectory features by absolute Cohen's d and pre-register the top 3 for confirmatory testing: monotonicity ($d = +0.60$, $p = 0.004$), variance ($d = +0.48$, $p = 0.019$), and slope ($d = -0.34$, $p = 0.093$). **Stage 2 (Validation):** We generate 200 new CoTs (T=0.8) on 100 independent GSM8K questions (non-overlapping with the discovery set), yielding 100 balanced samples (50 correct, 50 wrong). Only the 3 pre-registered features are tested, with Holm-Bonferroni correction for the family-wise error rate ($\alpha = 0.05$).

Table 4 reports the confirmatory results. **Zero of 3 pre-registered features survive Holm-Bonferroni correction.** Monotonicity is the only feature with a nominally significant signal in the validation set ($d = +0.47$, $p = 0.022$, AUC $=$ 0.622) but falls short of the Holm-adjusted threshold ($\alpha =$ 0.017). Slope and variance are non-significant with effect sizes substantially attenuated ($d = -0.17$ and $d = -0.12$ vs. discovery estimates of -0.34 and $+0.48$). The sign reversal of variance ($+0.48$ in discovery vs. -0.12 in validation) indicates overfitting to the discovery sample.

This validation failure reflects honest scientific practice: pre-registration prevented cherry-picking. The magnitude of attenuation—Cohen's d dropping 30–75%—confirms discovery-set effect sizes are upward-biased. Monotonicity's marginal result ($p = 0.022$ before correction) indicates a genuine but weak signal warranting larger-scale testing ($n > 200$).

Table 5: SVAMP cross-dataset validation ($n = 250$ questions, 750 CoTs). GSM8K comparisons from Experiments A–C.

Finding	SVAMP	GSM8K	Replicates?
H1: Terminal Gap AUROC			
gap_original	0.594 [.520,.667]	0.469–0.551	yes (weak)
gap_E2 (answer-only)	0.568 [.498,.636]	—	yes (chance)
H2: Exposure vs. Reasoning			
Δ_{exp} (E1−E2)	6.75 nats	2.13 nats	yes
Δ_{reas} (E2−orig)	1.07 nats	2.97 nats	yes
Exp./Reas. ratio	**6.29**	0.72	—
AUROC(Δ_{reas})	0.524	—	chance
H3: Trajectory features (d_z)			
Monotonicity	$+0.24$ (p=.11)	$+0.47$ (p=.02)	direction
Variance	$+0.02$ (p=.88)	-0.12 (p=.57)	n.s. in both
Slope	$+0.07$ (p=.63)	-0.17 (p=.41)	n.s. in both

Cross-Dataset Validation: SVAMP Replication

To test whether the core Gap confound is dataset-specific, we replicate key findings on SVAMP (Patel, Bhattamishra, and Goyal 2021), a multi-step math benchmark with structural variations probing generalization beyond GSM8K's single-template distribution (250 questions, 3 CoTs each, Qwen2.5-3B-Instruct, 85.9% accuracy).

Table 5 reports the cross-dataset comparison. Three of four core findings replicate.

H1 (Terminal Gap discrimination). Terminal Gap AUROC is 0.594 [0.520, 0.667], modestly above 0.5 but replicating GSM8K's near-chance results (0.469–0.551, Table 1). The answer-only variant gap_E2 yields AUROC $= 0.568$ [0.498, 0.636], indistinguishable from chance, replicating Experiment B's finding that Gap without reasoning context provides negligible correctness signal. Question-only prior surprisal (gap_E1) yields AUROC $= 0.724$ [0.653, 0.787], confirming that question difficulty carries correctness signal; the information loss occurs at the PMI stage where answer-exposure dominates. These results confirm the Gap answer-exposure confound generalizes across structurally distinct math benchmarks, strengthening the identifiability argument of the Theoretical Framework.

H2 (Answer-exposure dominance). Decomposing total Gap reduction $\Delta_{\text{total}} =$ gap_E1 $-$ gap_original into exposure ($\Delta_{\text{exp}} =$ gap_E1 $-$ gap_E2) and reasoning components ($\Delta_{\text{reas}} =$ gap_E2 $-$ gap_original), SVAMP yields an exposure/reasoning ratio of 6.29 (paired $t = 29.0$, $p < 10^{-124}$, Cohen's $d_z = 1.06$): answer exposure contributes 6.3$\times$ more Gap reduction than reasoning content. Crucially, Δ_{reas} does not discriminate correctness (AUROC $= 0.524$, near chance), replicating Experiment B's 2×2 mixed-effects result. On GSM8K, the exposure/reasoning ratio is 0.72 (cf. Table 2), with the same invariant: the reasoning-attributable component carries no correctness signal in either benchmark.

H3 (Trajectory features). Three pre-registered trajectory

features are tested on SVAMP. All are non-significant: monotonicity ($d_z = +0.24$, $p = 0.114$), variance ($d_z = +0.02$, $p = 0.881$), slope ($d_z = +0.07, p = 0.632$). Only 106 of 750 CoTs are incorrect (14.1%), limiting statistical power relative to GSM8K's balanced validation set. The pattern of weak, non-replicating trajectory effects is replicating GSM8K's independent validation (0/3 features surviving Holm correction, Table 4).

Discussion

Scope and Boundary Conditions

These results do *not* demonstrate universal impossibility—they identify a structural limitation under answer-exposure dominance. Scope limitations, each defining testable future work:

1. **Model scale.** Only Qwen2.5-3B-Instruct (4-bit NF4); answer- exposure dominance may weaken at larger scales. Zhao et al. (2026) found trajectory monotonicity predictive in Qwen2.5-7B and Mistral-7B.
2. **Domain.** While SVAMP cross-validation (Cross-Dataset Validation) confirms that the answer-exposure confound generalizes across structurally distinct math benchmarks, replication across multi-hop QA, code generation, and open-ended tasks is essential to establish broader generalizability.
3. **Single-model probabilities.** Only one model's output probabilities. Internal representations, multi-sample methods, cross-model disagreement, and tool-based verification introduce independent signals breaking Theorem 2's unidentifiability.
4. **Sample sizes.** Trajectory sample sizes remain modest (100 CoTs discovery + 100 validation, 5 checkpoints). Experiment B: $n = 150$ questions, 14 conditions provide the primary evidential contribution. Experiment C discovery set is a single 100-CoT slice; independent validation failed to replicate at Holm-corrected thresholds, indicating trajectory effect sizes in the discovery phase are upward-biased. Cross-model and multi-dataset replication is essential before generalizability claims. Larger-scale validation ($n > 200$) is needed for adequate statistical power.

The theory predicts self-evaluation failure to the extent that answer-exposure dominance holds; quantifying this across diverse settings is the central empirical agenda.

Implications

The core challenge is identifiability: surface co-occurrence and reasoning quality are inseparable from single-model probabilities. Independent signals—cross-model comparison, process supervision, tool-based verification—break this barrier and are essential for reliable oversight.

Conclusion

We examined selected-answer surprisal as a self-evaluation signal across three experiments on GSM8K and SVAMP with Qwen2.5-3B-Instruct. Scalar terminal Gap fails at cross-question discrimination and under counterfactual interventions; trajectory-dynamic features carry correctness signal in exploratory analysis but effects weaken substantially in independent validation. Under answer-exposure dominance, correctness risk is not identifiable from single-model probabilities alone—a structural limitation that necessitates multiple independent signal sources for reliable oversight.

Appendix A: Formal Definitions and Results

Definition 1 (Self-Gap) $G_\theta(q, z) = -\log p_\theta(\hat{a} \mid q, z)$, $\hat{a} = \arg\max_a p_\theta(a \mid q, z)$.

Definition 2 (Answer-Exposure Dominance) $I(\hat{A}; Z_{\text{logic}} \mid Q, Z_{\text{surf}}) \leq \varepsilon$ *for small* $\varepsilon > 0$. *Thus* $p_\theta(\hat{a} \mid q, z) \approx p_\theta(\hat{a} \mid q, \text{last}_k(z))$.

Gap–Prior–PMI Decomposition (see main text §3.1). $G = S - \text{PMI}$, with $\text{PMI} = \text{PMI}_{\text{surf}} + i_{\text{logic}}$, $|i_{\text{logic}}| \leq \varepsilon$ under answer-exposure dominance.

Theorem 1 (Insensitivity to Reasoning Validity) *For* z_1 *(invalid reasoning) and* z_2 *(valid reasoning) sharing the answer suffix:* $|G_\theta(q, z_1) - G_\theta(q, z_2)| \leq \varepsilon'$.

Theorem 2 (Observational Equivalence) $\exists\, \mathcal{E}_1, \mathcal{E}_2$ *with identical* $p_\theta(\cdot \mid q, z)$ *but* $P_{\mathcal{E}_1}(C{=}1 \mid q, z) \neq P_{\mathcal{E}_2}(C{=}1 \mid q, z)$. *Correctness is unidentifiable from single-model probabilities alone.*

Proposition 1 (Cross- vs. Within-Question Discrimination) *Under hierarchical model with question variance* σ_S^2 *and PMI noise* σ_P^2*, within-Q effect* $\delta > 0$*:*

$$\text{AUROC}_X = \Phi\left(\frac{\delta}{\sqrt{2(\sigma_S^2 + \sigma_P^2)}}\right), \quad \text{Acc}_W = \Phi\left(\frac{\delta}{\sqrt{2\sigma_P^2}}\right). \tag{1}$$

As $\sigma_S^2 \to \infty$*,* $\text{AUROC}_X \to 1/2$ *while* $\text{Acc}_W > 1/2$.

Table 6: Hierarchical attenuation simulation ($\beta = 0.3$, $\text{acc}_W = 0.728$).

σ_α	Cross-Q AUROC	Within-Q Acc.	Var. Ratio
0.1	0.725	0.728	0.25
0.5	0.683	0.728	0.91
1.0	0.629	0.728	3.11
2.0	0.574	0.728	12.49
3.0	0.551	0.728	29.91
5.0	0.532	0.728	105.75

Full proofs in Technical Supplement. Proof sketches: Thm. A.1: shared suffix $\Rightarrow$ identical Gap under answer-exposure. Thm. A.2: MLE only depends on token co-occurrence; different $P(C \mid q, z)$ can induce same p_θ. Prop. A.1: $\Delta \sim \mathcal{N}(\delta, 2\sigma_P^2)$ within-Q, $\mathcal{N}(\delta, 2(\sigma_S^2 + \sigma_P^2))$ cross-Q; AUROC via Gaussian signal-detection.

References

Anwar, U.; Bakker, T.; Kianfar, D.; Pinneri, C.; and Louizos, C. 2026. Analyzing and Improving Chain-of-Thought Monitorability Through Information Theory. *arXiv preprint arXiv:2602.18297*. DOI: 10.48550/arXiv.2602.18297.

Bertolazzi, L.; Mondorf, P.; Plank, B.; and Bernardi, R. 2025. The Validation Gap: A Mechanistic Analysis of How Language Models Compute Arithmetic but Fail to Validate It. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2502.11771.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning (ICML)*. ArXiv:1706.04599.

Hamidieh, K.; Thost, V.; Gerych, W.; Yurochkin, M.; and Ghassemi, M. 2026. Complementing Self-Consistency with Cross-Model Disagreement for Uncertainty Quantification. In *International Conference on Learning Representations (ICLR)*.

Jia, J.; Benton, J.; and Easley, E. 2026. Faithfulness as Information Flow: Evaluating and Training Faithful Chain-of-Thought Reasoning. *arXiv preprint arXiv:2605.24286*.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language Models (Mostly) Know What They Know. In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2207.05221.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2205.11916.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *International Conference on Learning Representations (ICLR)*. ArXiv:2302.09664.

Liang, J. 2025. Chain-of-Thought Reasoning for Math: Theoretical Foundation and Applications. In *ICML 2025 Workshop on AI for Mathematics*.

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2024. Let's Verify Step by Step. In *International Conference on Learning Representations (ICLR)*. ArXiv:2305.20050.

Manakul, P.; Liusie, A.; and Gales, M. J. F. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2303.08896.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2202.05262.

Patel, A.; Bhattamishra, S.; and Goyal, N. 2021. Are NLP Models really able to Solve Simple Math Word Problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. ArXiv:2103.07191.

Paul, D.; West, R.; Bosselut, A.; and Faltings, B. 2024. Making Reasoning Matter: Measuring and Improving Faithfulness of Chain-of-Thought Reasoning. In *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2402.13950.

Stechly, K.; Valmeekam, K.; and Kambhampati, S. 2025. On the Self-Verification Limitations of Large Language Models on Reasoning and Planning Tasks. In *International Conference on Learning Representations (ICLR)*.

Wan, K.; Gao, L.; Mu, H.; Nakov, P.; Wang, Y.; and Chen, X. 2025. A Fano-Style Accuracy Upper Bound for LLM Single-Pass Reasoning in Multi-Hop QA. *arXiv preprint arXiv:2509.21199*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations (ICLR)*. ArXiv:2203.11171.

Weatherhead, S.; Salim, F.; and Belbasis, A. 2025. Illusions of Reflection: Open-Ended Task Reveals Systematic Failures in Large Language Models' Reflective Reasoning. *arXiv preprint arXiv:2510.18254*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*. Advances in Neural Information Processing Systems, Vol. 35.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; et al. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Zhao, W.; et al. 2026. Entropy Trajectory Shape Predicts LLM Reasoning Reliability: A Diagnostic Study of Uncertainty Dynamics in Chain-of-Thought. *arXiv preprint arXiv:2603.18940*.