QUANTIFYING CONSISTENCY IN LLM LOGICAL REASONING VIA STRUCTURAL UNCERTAINTY

Baishali Chaudhury, Mengdie Flora Wang, Hyunji Hayley Park, Rahul Ghosh, Sungmin Hong, Jae Oh Woo
AWS Generative AI Innovation Center
{baishch, florawan, parhyunj, rahulgh, hsungmin, jaeohwoo}@amazon.com

ABSTRACT

Large language models can arrive at the same answer through reasoning paths that are unstable, contradictory, or difficult to rank consistently—a failure mode especially prevalent in multi-step deductive reasoning. Existing methods assess reasoning reliability primarily through output dispersion—measuring how much sampled answers differ—but this view discards a complementary signal: whether the model can consistently rank competing reasoning candidates. We propose structural uncertainty, a consistency-aware evaluation framework derived from the stability of self-preference-induced rankings over sampled reasoning solutions. Given a query, we generate multiple candidate solutions and ask the same model to judge pairwise preferences among its own outputs. We aggregate sparse self-preferences into ranking distributions via Bradley–Terry modeling with PageRank, and decompose the signal into two complementary entropy-based components—across-trial ranking instability and within-trial candidate ambiguity. Across five LLMs and eight benchmarks, structural signals provide information complementary to answer dispersion: on logical and mathematical reasoning tasks, the combination improves identification of unreliable reasoning instances, while on factual retrieval the structural signal collapses toward uniformity, diagnosing a regime boundary where reasoning-level consistency evaluation is uninformative. The two components relate differently to accuracy: within-trial ambiguity correlates positively with correctness on reasoning tasks—consistent with settings where multiple plausible solution paths remain competitive—while across-trial instability correlates negatively, signaling unreliable reasoning. Structural uncertainty is best understood not as a universal confidence estimator, but as a regime-sensitive evaluator of logical reasoning consistency.

1 INTRODUCTION

Large language models have achieved remarkable progress in natural language understanding and generation, yet their logical reasoning capabilities remain a significant bottleneck Xiong et al. (2024); Tian et al. (2023); Kapoor et al. (2024); Zhou et al. (2024). Models frequently produce answers that appear logically plausible yet are internally inconsistent—arriving at the same wrong conclusion through different flawed reasoning paths, or failing to maintain stable preferences when asked to compare their own solutions. Evaluating logical reasoning reliability requires assessing not only answer-level correctness but also the *consistency* of the reasoning process itself. Although we do not directly model cross-question contradiction, our framework targets a closely related consistency problem: whether the model can stably evaluate competing reasoning candidates for the same query.

Existing post-hoc evaluation methods (Lin et al., 2023; Farquhar et al., 2024; Wang et al., 2024a; Lyu et al., 2025) treat sampled responses as exchangeable and assess reliability from how much answers *differ*—i.e., output dispersion. This view captures one dimension of reasoning reliability but discards a complementary signal: the *structural consistency* of preferences among candidate reasoning solutions. For logical reasoning, this omission is consequential because multiple candidates may share the same final answer yet differ in reasoning quality, coherence, or mutual consistency; collapsing this structure loses information about which solutions the model favors and how stable those preferences are. Figure 1 illustrates this gap: when all sampled responses agree on the same

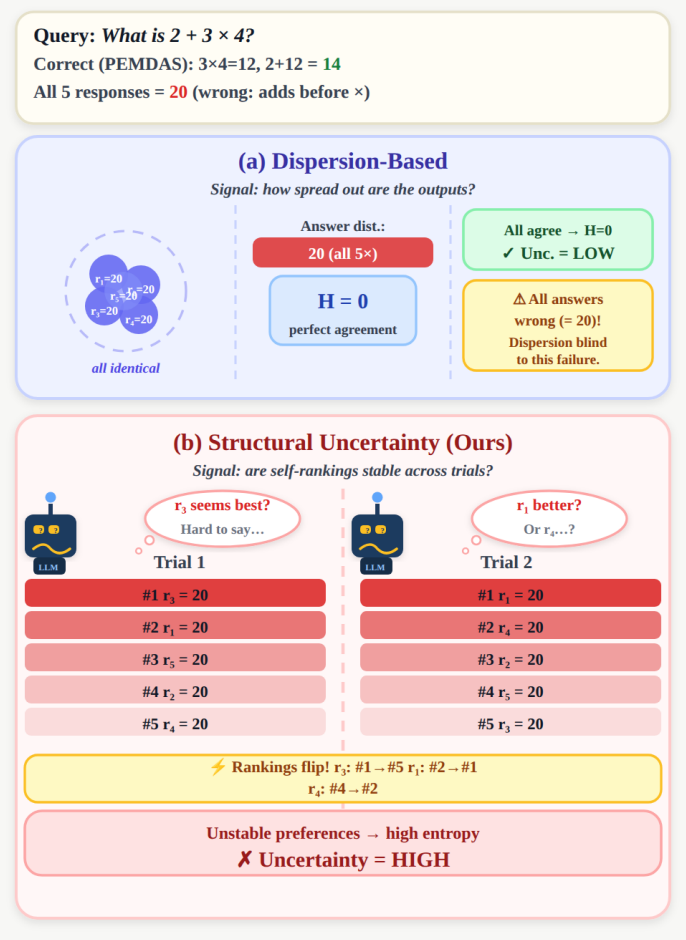


Figure 1: An illustrative example contrasting dispersion-based and structural consistency evaluation. **(a)** When all five sampled reasoning candidates agree on the same wrong answer (20), dispersion-based methods report zero entropy and low uncertainty, since they do not use preference-order information among candidates. **(b)** Our structural approach asks the model to rank its own reasoning outputs across independent trials. The rankings are completely unstable (r_3 moves from #1 to #5), revealing reasoning inconsistency not reflected by answer dispersion alone.

wrong answer, dispersion-based methods report low uncertainty, while self-preference rankings across trials can reveal instability not reflected by dispersion alone.

We propose a consistency-aware evaluation framework for logical reasoning that measures the stability of self-preference-induced rankings over sampled candidate solutions. Given a query, we sample multiple reasoning candidates and ask the *same model* to judge pairwise preferences among its own outputs. Beyond how much responses *differ*, this probes whether the model forms a stable or fluctuating preference ordering over competing reasoning paths—a signal that dispersion alone discards. We aggregate sparse pairwise judgments into ranking distributions using Bradley–Terry modeling (Bradley and Terry, 1952) with PageRank normalization (Langville and Meyer, 2006), repeated across random spanning-tree comparison trials. This produces two distinct components of structural uncertainty: *across-trial ranking instability* (reasoning instability across elicitation trials) and *within-trial candidate ambiguity* (ambiguity among plausible reasoning candidates within each trial).

Across five LLMs and eight benchmarks, we show that self-preference-derived structural signals provide information complementary to output dispersion. The interaction with accuracy is task-dependent: on mathematical reasoning, within-trial ambiguity can correlate positively with correctness—consistent with settings where several plausible solution paths remain competitive—

while on factual retrieval, structural signals collapse toward uniformity, diagnosing a regime where reasoning-level consistency evaluation is uninformative. Combining structural and dispersion-based signals improves identification of unreliable reasoning instances on several reasoning and knowledge tasks, though gains are absent in the collapse regime.

Contributions. **(1) A post-hoc framework for evaluating logical reasoning consistency and reliability.** We propose a model-agnostic framework that quantifies reasoning consistency through the stability of a model's self-preference rankings over its own candidate solutions, providing an observable signal complementary to output dispersion. **(2) A structural decomposition of reasoning stability.** We aggregate sparse pairwise self-preferences into ranking distributions via Bradley–Terry with PageRank, yielding two complementary entropy-based components: across-trial ranking instability (reasoning instability) and within-trial candidate ambiguity (ambiguity among plausible reasoning candidates). **(3) A regime analysis distinguishing reasoning from retrieval settings.** Across five LLMs and eight benchmarks, we show that structural signals complement dispersion on reasoning tasks, identify when and why the signal collapses on retrieval tasks, and characterize the task conditions under which each signal type is most informative for evaluating logical reasoning consistency.

2 Related Work

Logical reasoning and self-consistency in LLMs. Chain-of-thought prompting and self-consistency (Wang et al., 2023) have become standard approaches for improving and evaluating LLM reasoning. Self-consistency measures agreement across multiple sampled reasoning paths, while debate and self-judge frameworks (Zheng et al., 2023; Kadavath et al., 2022) demonstrate that models can assess output quality. However, self-consistency treats responses as exchangeable and measures only answer-level agreement, missing structural differences in reasoning quality among candidates. We complement answer-level consistency by measuring how stably the model ranks competing reasoning solutions through self-preference.

Post-hoc uncertainty and preference-based evaluation. Dispersion-based methods estimate uncertainty from semantic variation among responses (Kuhn et al., 2023; Lin et al., 2023; Farquhar et al., 2024; Kossen et al., 2024), output density (Qiu and Miikkulainen, 2024), kernelized entropy (Nikitin et al., 2024), or self-consistency entropy (Wang et al., 2024a; Lyu et al., 2025). Comparison-based methods aggregate pairwise preferences into calibrated scores (Shrivastava et al., 2025) or incorporate richer structure via multi-dimensional representations (Chen et al., 2025), knowledge graphs (Yuan et al., 2025), or Minimum Bayes Risk (Vashurin et al., 2025a). When internal access is available, logit-based (Ma et al., 2025), chain-of-thought (Zhang and Zhang, 2025), and proxy-based methods (Lee et al., 2024) derive uncertainty from model internals. Information-theoretic and Bayesian perspectives motivate principled decomposition (Abbasi Yadkori et al., 2024; Kendall and Gal, 2017; Woo, 2022; 2023). Our approach operates in a fully black-box setting without requiring internal access or model modification.

Consistency and contradiction in LLM outputs. A growing body of work addresses logical contradictions and inconsistencies in LLM outputs. Evaluation frameworks assess alignment between uncertainty and quality (Huang et al., 2024; Ye et al., 2024; Vashurin et al., 2025b), while studies reveal strong task- and model-dependence in LLM reliability (Huang et al., 2023; Yang et al., 2025). Our work contributes to this direction by providing a structural lens on reasoning consistency: rather than checking whether outputs contradict each other at the answer level, we measure whether the model can form a stable preference ordering over its own reasoning candidates—directly revealing internal inconsistency in reasoning evaluation. Unlike approaches that improve reasoning via symbolic modules or external solvers, we focus on post-hoc evaluation of reasoning consistency in a pure black-box setting.

3 Method

Our framework quantifies consistency in logical reasoning via the stability of self-preference-induced rankings over sampled candidate solutions. Our notion of consistency is same-query and candidate-

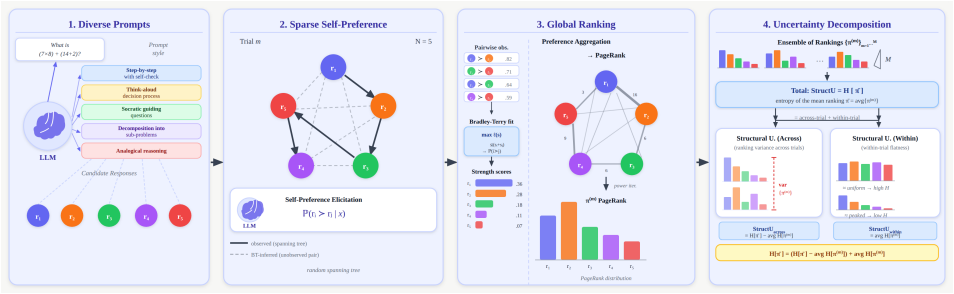


Figure 2: Overview of the consistency-aware reasoning evaluation framework. Given a query, we (1) generate diverse candidate reasoning solutions, (2) elicit pairwise self-preferences, (3) aggregate into a global ranking via pairwise preference modeling (Bradley–Terry or TrueSkill) with PageRank, and (4) decompose the consistency signal via random spanning tree sampling.

relative: we ask whether the model forms a stable preference ordering over multiple reasoning solutions to the same problem. Given a query, we: (1) generate N diverse candidate reasoning solutions, (2) elicit pairwise self-preferences by asking the model to judge its own outputs, (3) aggregate preferences into a global ranking via Bradley–Terry with PageRank, and (4) decompose the consistency signal into across-trial and within-trial components through random spanning tree sampling. Figure 2 illustrates the pipeline.

3.1 Self-Preference via Spanning Trees

Given input x, we sample N candidates $\mathcal{R}(x) = \{r_1, \ldots, r_N\}$ from the model's conditional distribution $p_\theta(\cdot|x)$ via diverse prompting with stochastic decoding. Rather than comparing all $\binom{N}{2}$ pairs, we repeat the following for M independent trials ($m = 1, \ldots, M$):

Sparse graph sampling. We draw a uniform random spanning tree $\mathcal{T}^{(m)}$ over the N candidates, yielding a connected graph with exactly $N-1$ edges. This guarantees global connectivity with minimal comparisons while injecting structural randomness across trials, loosely analogous in spirit to Monte Carlo dropout over graph structure rather than weights.

Self-preference elicitation. For each edge $(i, j) \in \mathcal{T}^{(m)}$, we query the *same model* to judge which response is better, optionally obtaining a confidence score. The consistency of these elicited self-preference judgments across trials provides an additional uncertainty signal that is not directly available from output dispersion alone.

Preference aggregation. We fit a pairwise preference model (Bradley–Terry or TrueSkill) on the $N-1$ observed comparisons to infer win probabilities for *all* N^2 pairs, then aggregate into a global ranking distribution $\boldsymbol{\pi}^{(m)}$ via PageRank. Details follow in Section 3.2.

The ensemble $\{\boldsymbol{\pi}^{(m)}\}_{m=1}^{M}$ from M trials enables the uncertainty decomposition in Section 3.3.

3.2 Preference Aggregation

Bradley–Terry with L2 regularization. We assign each candidate i a latent utility $\theta_i \in \mathbb{R}$ and model pairwise preferences as (Bradley and Terry, 1952):

$$\mathbb{P}(i \succ j) = \frac{\exp(\theta_i)}{\exp(\theta_i) + \exp(\theta_j)}. \tag{1}$$

Since spanning trees admit perfect total orderings, the unregularized maximum likelihood objective is unbounded (Ford, 1957). We add an L^2-penalty for numerical well-posedness:

$$\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}) = \sum_{(i,j)\in\mathcal{T}^{(m)}} \log \mathbb{P}(i \succ j) - \frac{1}{2C}\|\boldsymbol{\theta}\|^2, \tag{2}$$

where $C > 0$ is the inverse regularization strength. We set $C = 1$ based on an ablation sweep (Appendix A.3): performance degrades below $C < 1$ but plateaus stably for $C \geq 1$ across all models,

confirming bounded parameter estimates. We optimize via MM-style updates (Hunter, 2004) and evaluate Eq. (1) for all pairs to obtain $\mathbf{P}^{(m)} \in [0,1]^{N\times N}$.

As a robustness check, we also implement a TrueSkill variant (Herbrich et al., 2006) with confidence-weighted updates. Despite fundamentally different assumptions, both backends produce highly similar uncertainty rankings (Appendix A.4), suggesting that the induced ranking distributions are reasonably stable with respect to the choice of preference backend.

PageRank global ranking. While BT interpolates sparse comparisons into a dense preference matrix $\mathbf{P}^{(m)}$, PageRank summarizes it into a normalized ranking distribution $\boldsymbol{\pi}^{(m)} \in \Delta^N$ on which we define the entropy-based decomposition (Section 3.3). From $\mathbf{P}^{(m)}$, we construct a row-stochastic transition matrix where probability mass flows from weaker to stronger candidates:

$$T_{ij}^{(m)} \propto P_{ji}^{(m)}, \quad T_{ii}^{(m)} = 0, \tag{3}$$

with row normalization. We compute the stationary distribution $\boldsymbol{\pi}^{(m)}$ satisfying $\boldsymbol{\pi}^{(m)} = (\mathbf{T}^{(m)})^\top \boldsymbol{\pi}^{(m)}$ via power iteration (Brin and Page, 1998; Langville and Meyer, 2006).

3.3 Structural Uncertainty Decomposition

The M trials yield an ensemble $\{\boldsymbol{\pi}^{(m)}\}_{m=1}^{M}$ with mean $\bar{\boldsymbol{\pi}} = \frac{1}{M}\sum_m \boldsymbol{\pi}^{(m)}$. We decompose total uncertainty via Shannon entropy $H[\boldsymbol{p}] = -\sum_i p_i \log p_i$ following Kendall and Gal (2017); Woo (2022):

$$\text{StructU} = H[\bar{\boldsymbol{\pi}}], \tag{4}$$

$$\text{StructU}_{\text{within}} = \tfrac{1}{M}\sum_{m=1}^{M} H[\boldsymbol{\pi}^{(m)}], \tag{5}$$

$$\text{StructU}_{\text{across}} = \text{StructU} - \text{StructU}_{\text{within}}. \tag{6}$$

Here, $\mathrm{StructU}_{\text{within}}$ is intended to reflect *within-trial candidate ambiguity*: it is high when, within a single sparse comparison trial, the ranking distribution spreads its mass over multiple candidates instead of concentrating on a single preferred response. Meanwhile, $\mathrm{StructU}_{\text{across}}$ is intended to reflect *across-trial ranking instability*: it is high when different sampled comparison trees lead to substantially different ranking distributions across trials. Figure 3 illustrates the two components with contrasting examples.

Concretely, $H[\bar{\boldsymbol{\pi}}]$ is the entropy of the trial-averaged ranking; the average $\frac{1}{M}\sum_m H[\boldsymbol{\pi}^{(m)}]$ measures within-trial spread; and their difference—a Jensen gap—provides a distribution-level measure of cross-trial variation. This decomposition is defined entirely over the observable ranking ensemble induced by self-preference; we do not claim it identifies underlying data or parameter uncertainty.

Combination with self-consistency. For practical reasoning reliability evaluation, we also study a simple fixed combination of structural uncertainty with self-consistency entropy Self-ConsU $= -\sum_{a\in\mathcal{A}} p(a)\log p(a)$ over the answer distribution, where $\mathcal{A}$ denotes distinct answers. Abbreviating $\text{StructU}_{\text{across}}$ as SU_{a} and $\text{StructU}_{\text{within}}$ as SU_{w}:

$$\text{SU}_{\text{a}}\text{+SC} = \text{SU}_{\text{a}} + \text{Self-ConsU}, \tag{7}$$

$$\text{SU}_{\text{w}}\text{+SC} = \text{SU}_{\text{w}} - \text{Self-ConsU}, \tag{8}$$

where SC abbreviates Self-ConsU. The sign reflects each component's empirical relationship with accuracy: SU_{a} correlates negatively (instability $\rightarrow$ failure), so it adds with Self-ConsU; SU_{w} correlates positively on reasoning tasks, so it enters subtractively. This assignment is fixed globally—not tuned per task or model. We treat this as an empirical fusion rule rather than an intrinsic part of the decomposition; structural uncertainty captures *how* self-preferences are organized, while Self-ConsU captures *what* answers disagree.

4 Experiments

4.1 Experimental Setup

We evaluate five LLMs (Claude Sonnet 4.5, GPT-OSS 20B, Qwen 3 32B, Amazon Nova Premier, DeepSeek R1) on eight benchmarks grouped by reasoning structure: *mathematical and logical*

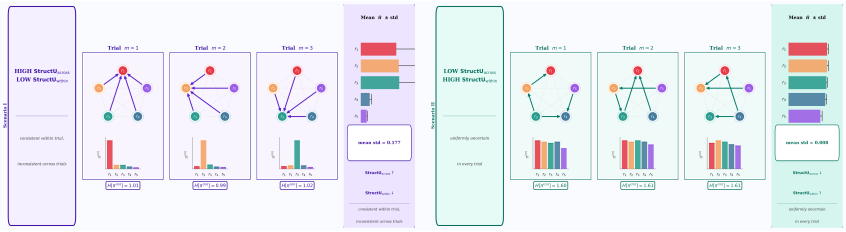


Figure 3: **Across-trial vs. within-trial structural uncertainty: two contrasting examples. Scenario I (left):** The model produces a confident, concentrated ranking within each trial (low within-trial entropy $H[\boldsymbol{\pi}^{(m)}]$), but the dominant candidate changes across trials as different spanning trees are sampled—indicating substantial across-trial instability in the induced ranking distribution ($\mathbf{StructU}_{\text{across}} \uparrow$, $\mathbf{StructU}_{\text{within}} \downarrow$). **Scenario II (right):** The ranking distribution is nearly uniform in every trial (high within-trial entropy), yet remains consistent across trials—indicating the model is stably uncertain rather than inconsistently confident ($\mathbf{StructU}_{\text{across}} \downarrow$, $\mathbf{StructU}_{\text{within}} \uparrow$). Node size reflects PageRank score $\boldsymbol{\pi}^{(m)}$; arrows indicate self-preference direction on the sampled spanning tree edges; dashed edges are unobserved pairs.

reasoning (Math-Synth, MATH-500 (Lightman et al., 2024), AMC-23 (He, 2023), AIME-24/25 (HuggingFaceH4, 2024; TIGER-Lab, 2024)), *reasoning-adjacent knowledge tasks* (MMLU-Pro (Wang et al., 2024b), TruthfulQA (Lin et al., 2022)), and a *retrieval-dominant comparison regime* (HotpotQA (Yang et al., 2018)). Math-Synth is a synthetic arithmetic benchmark with 993 problems (Appendix B); other dataset details and model accuracies are in Table 4.

We compare against black-box baselines computed from the same $N{=}5$ samples: **Self-ConsU** (answer entropy) (Wang et al., 2024a; Lyu et al., 2025), **SemanticU** (embedding dispersion) (Qiu and Miikkulainen, 2024; Kossen et al., 2024), and **VerbalizedU** (prompted confidence) (Tian et al., 2023; Xiong et al., 2024). See Appendix D for details.

We use selective prediction metrics as an operational measure of whether the proposed consistency signal identifies unreliable logical reasoning instances: questions where $\hat{p}_{\text{corr}} = \frac{1}{N}\sum_i \mathbb{I}[r_i \text{ correct}] < \tau$ with $\tau{=}1.0$. We report Spearman correlation, AUROC, and area under the risk-coverage curve (Sel-AUC) (Shrivastava et al., 2023).

4.2 Results

We evaluate structural uncertainty as a consistency signal for logical reasoning across five models and eight benchmarks, analyzing where it improves reasoning reliability evaluation, how the components relate to accuracy, and when the signal breaks down. Our key target failure mode is systematic but internally unstable reasoning: candidate solutions may agree at the answer level while remaining inconsistent in how the model ranks them.

Overall Evaluation. Table 1 reports reasoning reliability evaluation performance (Sel-AUC; AUROC in parentheses). The central finding is *task-dependent complementarity*: on reasoning-heavy and some knowledge tasks, structural consistency signals add information beyond answer dispersion, while on retrieval-style tasks the structural signal collapses and provides limited benefit. Reasoning tasks admit structurally diverse solution paths, making self-preference consistency informative in logical reasoning regimes. Specifically, the combined estimator (StructU+Self-ConsU) achieves highest or second-highest Sel-AUC on mathematical reasoning (Math-Synth, MATH-500, AMC-23) and knowledge tasks (MMLU-Pro, TruthfulQA), with largest gains where answer-level agreement is insufficient. On HotpotQA, retrieval tasks suppress structural diversity, so the signal collapses—dispersion methods dominate for the two strongest models (Claude: Self-ConsU 0.839 vs. combined 0.742; DeepSeek: SemanticU 0.852 vs. combined 0.789). This task asymmetry is itself informative: it identifies the regime boundary where reasoning-level consistency evaluation ceases to be useful, making structural uncertainty a regime-sensitive evaluator of reasoning consistency rather than a universal confidence estimator. Among baselines, Self-ConsU is strongest but blind to systematic reasoning errors; VerbalizedU is inconsistent; SemanticU is weakest except where structural signals collapse.

Dataset	Model	StructU (Ours)			StructU+Self-ConsU (Ours)			Baselines		
		within	across	total	within	across	total	Self-ConsU	VerbalizedU	SemanticU
Math-Synth	Claude 4.5 Sonnet	0.624 (0.984)	0.644 (0.936)	0.596 (0.972)	0.661 (0.992)	**0.663 (0.990)**	0.655 (0.992)	0.65 (0.989)	0.544 (0.924)	0.430 (0.927)
	DeepSeek R1	0.815 (0.900)	0.790 (0.837)	0.773 (0.790)	**0.823 (0.929)**	0.820 (0.924)	0.814 (0.915)	0.802 (0.899)	0.793 (0.701)	0.664 (0.502)
	GPT-OSS 20B	0.794 (0.756)	0.762 (0.668)	0.661 (0.366)	0.840 (0.955)	0.849 (0.956)	**0.849 (0.958)**	0.83 (0.958)	0.792 (0.742)	0.528 (0.638)
	Amazon Nova Premier	0.496 (0.953)	0.370 (0.841)	0.447 (0.853)	0.511 (0.997)	0.498 (0.998)	**0.512 (0.997)**	0.382 (0.948)	0.389 (0.824)	0.436 (0.807)
	Qwen 3 32B	0.231 (0.794)	0.190 (0.656)	0.213 (0.718)	**0.393 (0.998)**	0.391 (0.998)	0.388 (0.998)	0.38 (0.995)	0.279 (0.824)	0.218 (0.462)
MATH-500	Claude 4.5 Sonnet	0.931 (0.813)	0.936 (0.789)	0.932 (0.801)	0.947 (0.840)	**0.950 (0.843)**	0.948 (0.831)	0.942 (0.816)	0.891 (0.686)	0.783 (0.720)
	DeepSeek R1	0.889 (0.652)	0.890 (0.640)	0.868 (0.596)	0.934 (0.776)	**0.936 (0.784)**	0.927 (0.767)	0.923 (0.759)	0.870 (0.603)	0.875 (0.546)
	GPT-OSS 20B	0.869 (0.582)	0.885 (0.529)	0.886 (0.611)	0.897 (0.715)	**0.910 (0.718)**	0.906 (0.729)	0.871 (0.694)	0.886 (0.645)	0.86 (0.46)
	Amazon Nova Premier	0.834 (0.747)	0.726 (0.609)	0.821 (0.687)	**0.887 (0.873)**	0.880 (0.871)	0.886 (0.870)	0.860 (0.839)	0.883 (0.715)	0.81 (0.695)
	Qwen 3 32B	0.859 (0.798)	0.780 (0.608)	0.833 (0.754)	**0.889 (0.882)**	0.882 (0.853)	0.888 (0.876)	0.871 (0.817)	0.880 (0.684)	0.819 (0.717)
AMC-23	Claude 4.5 Sonnet	0.967 (1.000)	0.966 (0.963)	0.962 (0.980)	0.970 (1.000)	**0.972 (1.000)**	0.970 (1.000)	0.955 (1.00)	0.900 (0.604)	0.853 (0.588)
	DeepSeek R1	0.924 (0.680)	0.915 (0.463)	0.923 (0.291)	**0.985 (1.000)**	0.985 (1.000)	0.985 (1.000)	0.985 (1.00)	0.877 (0.412)	0.880 (0.592)
	GPT-OSS 20B	0.956 (0.770)	0.931 (0.533)	0.954 (0.673)	**0.980 (1.000)**	0.980 (1.000)	0.980 (1.000)	0.980 (1.00)	0.980 (0.583)	0.884 (0.630)
	Amazon Nova Premier	0.591 (0.865)	0.538 (0.604)	0.610 (0.919)	0.712 (1.000)	0.716 (1.000)	**0.718 (1.000)**	0.584 (1.00)	0.419 (0.362)	0.351 (0.410)
	Qwen 3 32B	0.637 (0.643)	0.566 (0.567)	0.606 (0.643)	0.820 (1.000)	**0.821 (1.000)**	0.820 (1.000)	0.810 (1.00)	0.637 (0.389)	0.513 (0.299)
AIME-24	Claude 4.5 Sonnet	0.497 (0.903)	0.595 (0.852)	0.508 (0.875)	0.662 (0.994)	**0.690 (0.989)**	0.661 (0.994)	0.567 (0.972)	0.273 (0.144)	0.45 (0.64)
	DeepSeek R1	0.834 (0.752)	0.820 (0.554)	0.866 (0.884)	0.922 (1.000)	**0.925 (1.000)**	0.922 (1.000)	0.917 (1.00)	0.799 (0.298)	0.788 (0.715)
	GPT-OSS 20B	0.604 (0.725)	0.762 (0.714)	0.756 (0.813)	0.893 (1.000)	**0.907 (1.000)**	0.896 (1.000)	0.891 (1.00)	0.905 (0.319)	0.876 (0.681)
	Amazon Nova Premier	0.190 (—)	0.267 (—)	0.170 (—)	0.181 (—)	0.210 (—)	0.189 (—)	0.174 (–)	0.125 (–)	**0.294 (–)**
	Qwen 3 32B	0.235 (—)	0.243 (—)	0.257 (—)	0.401 (—)	**0.428 (—)**	0.402 (—)	0.414 (–)	0.355 (–)	0.213 (1.00)
AIME-25	Claude 4.5 Sonnet	0.565 (0.975)	0.602 (0.969)	0.221 (0.062)	0.645 (1.00)	0.646 (1.00)	**0.656 (1.00)**	0.645 (1.00)	0.489 (0.175)	0.524 (0.263)
	DeepSeek R1	0.263 (0.413)	0.244 (0.259)	0.539 (0.466)	0.728 (1.00)	0.740 (1.00)	**0.761 (1.00)**	0.707 (1.00)	0.564 (0.296)	0.616 (0.400)
	GPT-OSS 20B	0.424 (0.449)	0.392 (0.324)	0.696 (0.546)	0.881 (1.00)	0.884 (1.00)	**0.886 (1.00)**	0.825 (1.00)	0.776 (0.370)	0.768 (0.491)
	Amazon Nova Premier	0.203 (–)	0.149 (–)	0.059 (–)	0.198 (–)	0.218 (–)	0.175 (–)	0.178 (–)	0.106 (–)	**0.257 (–)**
	Qwen 3 32B	0.096 (0.862)	0.179 (0.897)	0.234 (0.414)	0.203 (1.00)	0.282 (1.00)	**0.328 (1.00)**	0.208 (1.00)	0.192 (0.897)	0.126 (0.931)
MMLU-Pro	Claude 4.5 Sonnet	0.924 (0.833)	0.913 (0.785)	0.916 (0.792)	**0.944 (0.912)**	0.936 (0.897)	0.943 (0.908)	0.900 (0.884)	0.944 (0.885)	0.890 (0.602)
	DeepSeek R1	0.845 (0.573)	0.849 (0.500)	0.855 (0.596)	0.925 (0.889)	0.917 (0.882)	0.924 (0.895)	0.882 (0.882)	**0.927 (0.796)**	0.870 (0.577)
	GPT-OSS 20B	0.765 (0.503)	0.754 (0.547)	0.749 (0.475)	**0.889 (0.948)**	0.880 (0.941)	0.886 (0.945)	0.830 (0.935)	0.785 (0.631)	0.774 (0.573)
	Amazon Nova Premier	0.671 (0.633)	0.696 (0.507)	0.679 (0.624)	0.805 (0.936)	**0.827 (0.948)**	0.800 (0.936)	0.801 (0.945)	0.820 (0.775)	0.811 (0.713)
	Qwen 3 32B	0.713 (0.669)	0.657 (0.567)	0.685 (0.624)	**0.818 (0.971)**	0.805 (0.964)	0.817 (0.971)	0.787 (0.966)	0.728 (0.724)	0.742 (0.686)
HotpotQA	Claude 4.5 Sonnet	0.686 (0.585)	0.731 (0.600)	0.664 (0.564)	0.698 (0.617)	0.742 (0.647)	0.681 (0.604)	0.839 (0.656)	**0.847 (0.700)**	0.768 (0.576)
	DeepSeek R1	0.732 (0.580)	0.747 (0.514)	0.744 (0.614)	0.767 (0.683)	0.789 (0.666)	0.771 (0.693)	0.835 (0.658)	0.829 (0.708)	**0.852 (0.696)**
	GPT-OSS 20B	0.724 (0.537)	0.717 (0.524)	0.736 (0.568)	0.815 (0.728)	0.798 (0.724)	**0.815 (0.734)**	0.813 (0.721)	0.772 (0.592)	0.80 (0.707)
	Amazon Nova Premier	0.806 (0.582)	0.805 (0.589)	0.787 (0.541)	0.850 (0.752)	0.840 (0.748)	0.855 (0.750)	**0.864 (0.740)**	0.812 (0.647)	0.830 (0.618)
	Qwen 3 32B	0.699 (0.648)	0.679 (0.528)	0.691 (0.618)	0.766 (0.757)	0.772 (0.713)	0.778 (0.765)	**0.787 (0.729)**	0.702 (0.630)	0.707 (0.612)
TruthfulQA	Claude 4.5 Sonnet	0.998 (0.956)	0.997 (0.912)	0.998 (0.953)	0.998 (0.970)	0.998 (0.966)	**0.998 (0.971)**	0.994 (0.926)	0.997 (0.949)	0.979 (0.503)
	DeepSeek R1	0.932 (0.564)	0.941 (0.662)	0.874 (0.535)	0.954 (0.884)	**0.957 (0.888)**	0.910 (0.813)	0.940 (0.844)	0.949 (0.721)	0.907 (0.530)
	GPT-OSS 20B	0.913 (0.717)	0.893 (0.611)	0.909 (0.727)	**0.943 (0.955)**	0.942 (0.955)	0.941 (0.952)	0.916 (0.936)	0.864 (0.608)	0.856 (0.525)
	Amazon Nova Premier	0.956 (0.806)	0.930 (0.689)	0.952 (0.779)	**0.963 (0.914)**	0.951 (0.893)	0.962 (0.914)	0.946 (0.862)	0.953 (0.828)	0.907 (0.596)
	Qwen 3 32B	0.783 (0.510)	0.787 (0.534)	0.765 (0.499)	**0.869 (0.956)**	0.867 (0.953)	0.855 (0.949)	0.856 (0.936)	0.822 (0.694)	0.812 (0.636)

Table 1: **Selective prediction performance (Sel-AUC; AUROC in parentheses).** Results using Bradley–Terry with PageRank aggregation. StructU reports structural uncertainty (within, across components). StructU+Self-ConsU reports the combined estimators. Baselines: Self-ConsU (answer entropy), VerbalizedU (prompted confidence), SemanticU (embedding dispersion). Bold = best; underline = second-best per row.

Where Structural Consistency Signals Help. Figure 4 shows ΔSel-AUC $=$ Sel-AUC(StructU+Self-ConsU) $-$ Sel-AUC(Self-ConsU). Gains are consistent across logical reasoning and knowledge benchmarks, and largest where answer-level agreement alone is insufficient to identify unreliable reasoning: weaker models on hard contest problems gain the most (Qwen on AIME-25: $+12.0\%$, Amazon Nova Premier on AMC-23: $+13.4\%$), as structural consistency rankings surface signal not captured by Self-ConsU alone. HotpotQA is the exception—stronger models show negative lift (Claude Sonnet 4.5: -9.7%, DeepSeek R1: -4.6%), consistent with the structural collapse on retrieval tasks.

Correlation Between Consistency Signal and Reasoning Accuracy. Figure 5 shows across-trial and within-trial components exhibit opposite correlations with accuracy—most pronounced on mathematical reasoning. On Math-Synth and MATH-500, across-trial instability is negatively correlated with correctness while within-trial ambiguity is positively correlated (e.g., Claude on MATH-500: $\rho_{\text{across}} = -0.37$ vs. $\rho_{\text{within}} = 0.42$). This asymmetry has a natural interpretation for logical reasoning: ranking instability signals unreliable reasoning, while distributed preference among candidates is consistent with settings where multiple plausible solution paths remain competitive. The pattern weakens on MMLU-Pro and collapses on HotpotQA (near-zero correlations), confirming that the consistency signal is regime-sensitive—informative for logical reasoning but uninformative where reasoning-level structural diversity is absent.

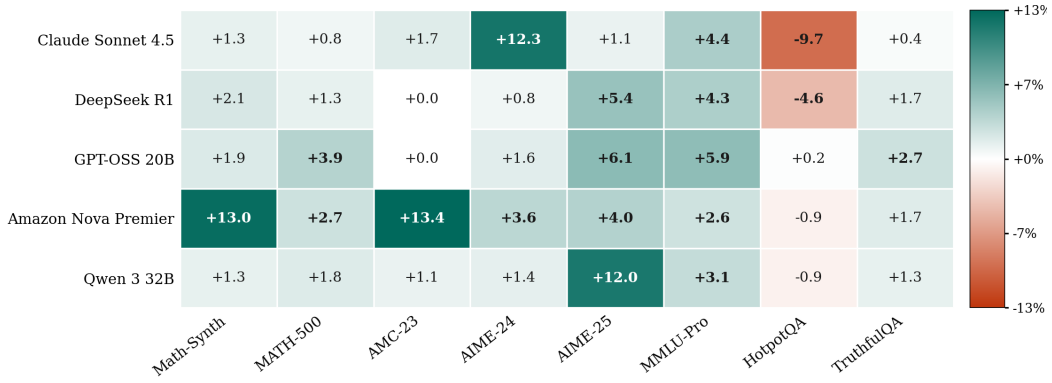


Figure 4: **Sel-AUC lift of StructU+Self-ConsU over Self-ConsU.** ΔSel-AUC for five models × eight benchmarks. Teal = hybrid wins; coral = Self-ConsU dominates.

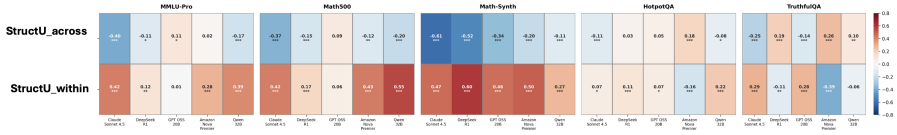


Figure 5: **Spearman correlation (ρ) between the two uncertainty components and accuracy.** Each cell reports Spearman rank correlation between per-question uncertainty and per-question accuracy (fraction correct among N=5 samples). Blue indicates negative correlation; red indicates positive correlation. Stars denote statistical significance ($^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Regime Analysis: Reasoning vs. Retrieval. Table 1 reveals a clear regime boundary: StructU+Self-ConsU consistently outperforms baselines on logical reasoning, whereas on HotpotQA, dispersion methods dominate for the strongest models (e.g., Claude: Self-ConsU 0.839 vs. Hybrid 0.742; DeepSeek: SemanticU 0.852 vs. Hybrid 0.796). Understanding *when* consistency signals fail is as important as knowing when they succeed, as this clarifies where logical reasoning structure is present versus absent. Figure 6 compares Claude 4.5 Sonnet on Math-Synth and HotpotQA under identical experimental conditions (N=5, M=5, similar prompt templates).

Across-trial separation on reasoning, collapse on retrieval. On Math-Synth, across-trial uncertainty distributions are well-separated between correct and incorrect questions (Figure 6a): incorrect questions exhibit a long right tail beyond 0.10, while correct questions concentrate near zero. On HotpotQA, both distributions are compressed near zero—structural uncertainty produces no across-trial signal regardless of correctness. In the joint space (Figure 6c), Math-Synth shows correctness separation along the across-trial axis, while HotpotQA collapses into a degenerate cluster with no separation.

Near-uniform preference distributions on factual retrieval. The within-trial distributions (Figure 6b) reveal the mechanism: HotpotQA responses cluster at maximum entropy ($\log 5 \approx 1.61$, dotted line) for both correct and incorrect questions, indicating near-uniform PageRank distributions. The model assigns approximately equal preference to all five responses, producing pairwise confidences near 50% and identical rankings across spanning tree samples.

When does self-preference carry signal? Logical and mathematical reasoning elicits *structurally diverse* solution paths—step-by-step computation, bracket decomposition, estimation-then-verification—enabling meaningful self-preference discrimination and making the consistency signal informative. In our setting, factual retrieval often elicits more *structurally homogeneous* reasoning chains across prompt templates, so self-preference reflects only stylistic variation and the structural signal vanishes. The "HotpotQA signature" in Figure 6c serves as a practical diagnostic: it identifies the regime boundary where logical reasoning structure is absent and dispersion-based methods should be pre-

Model	StructU$_{\text{within}}$ (Within-trial) AUROC			StructU$_{\text{across}}$ (Across-trial) AUROC		
	Real	Random	↓ Drop	Real	Random	↓ Drop
Claude 4.5 Sonnet	0.984	0.488 ± 0.016	0.496	0.923	0.510 ± 0.019	0.413
DeepSeek R1	0.896	0.743 ± 0.005	0.153	0.840	0.747 ± 0.001	0.093
GPT-OSS 20B	0.756	0.557 ± 0.048	0.199	0.668	0.588 ± 0.054	0.080
Amazon Nova Premier	0.943	0.530 ± 0.005	0.413	0.866	0.482 ± 0.014	0.384
Qwen 3 32B	0.850	0.511 ± 0.021	0.339	0.732	0.512 ± 0.020	0.220
Mean	**0.886**	$\mathbf{0.566 \pm 0.022}$	**0.320**	**0.806**	$\mathbf{0.568 \pm 0.022}$	**0.238**

Table 2: **Real vs. randomized preferences on Math-Synth (BT+PageRank).** Randomization tests whether gains depend on elicited self-preference content rather than fixed aggregation structure: winner direction and confidence scores are randomized while all other pipeline components are held fixed. Random AUROC is mean ± std over three runs; ↓ indicates drop from real to random.

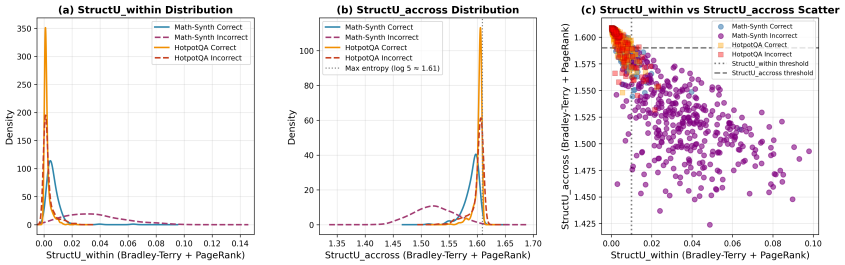


Figure 6: **Regime analysis: reasoning consistency vs. retrieval collapse.** Claude 4.5 Sonnet on Math-Synth (reasoning) and HotpotQA (retrieval), conditioned on correctness (BT + PageRank). **(a)** StructU$_{\text{across}}$: Math-Synth shows correctness separation; HotpotQA concentrates near zero for both. **(b)** StructU$_{\text{within}}$: HotpotQA clusters at maximum entropy ($\log 5 \approx 1.61$, dotted line). **(c)** Joint space: reasoning tasks separate along the across-trial axis; retrieval tasks collapse into a degenerate cluster, diagnosing the regime boundary where logical reasoning structure is absent.

ferred. This collapse is not merely a failure case; it is a substantively useful boundary result that distinguishes tasks genuinely supporting reasoning-consistency analysis from those dominated by retrieval-induced homogeneity.

Ablation Studies. To test whether structural uncertainty reflects elicited preference signal rather than pipeline artifacts, we replace real self-preference judgments with random comparisons on Math-Synth: winner direction is randomized uniformly and confidence scores are sampled uniformly from $[51, 99]$, while all other pipeline components (spanning tree topology, BT fitting with C=1, PageRank aggregation, entropy computation) are held fixed. Table 2 shows AUROC drops substantially for both uncertainty components (mean drop: 0.320 for within-trial, 0.238 for across-trial), with three models collapsing to near-chance level (Claude: 0.984 → 0.488; Nova: 0.943 → 0.530; Qwen: 0.850 → 0.511). These ablations suggest that discriminative performance depends materially on the elicited self-preference judgments rather than on the aggregation structure alone. Performance plateaus at $M \approx 5$ trials and remains stable through M=20 (Figure 8). The C=1 regularization choice is validated as performance degrades for $C < 1$ but remains stable for $C \geq 1$ across all models. PageRank smoothing provides mild regularization benefit (+0.015 Sel-AUC average). Increasing response count from N=5 to N=10 degrades performance, indicating high-temperature samples introduce noise (Figure 7).

5 Conclusion

We introduced structural uncertainty, a consistency-aware evaluation framework for logical reasoning that measures the stability of self-preference-induced rankings over sampled LLM reasoning candidates. By eliciting pairwise self-preferences and aggregating them via Bradley–Terry with PageRank,

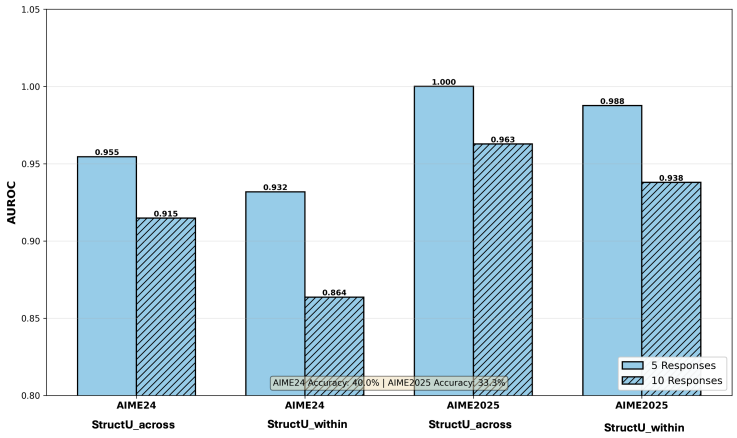


Figure 7: Ablation study on AIME benchmarks: effect of number of sampled responses showing performance degrades with $N{=}10$.

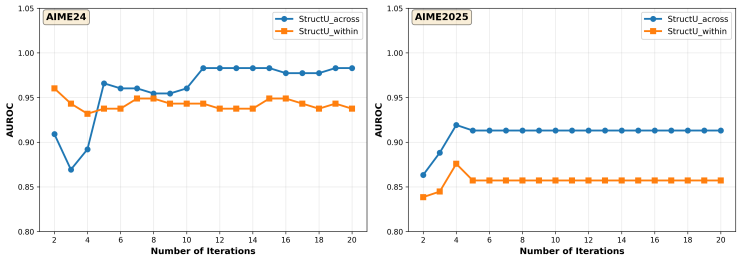


Figure 8: Ablation study on AIME benchmarks: effect of number of trials showing performance plateaus at $M{\approx}5$.

we decompose the signal into across-trial ranking instability and within-trial candidate ambiguity without requiring model internals.

Across five LLMs and eight benchmarks, the central finding is that structural uncertainty provides a consistency-aware lens on logical reasoning: combining structural and dispersion-based signals improves identification of unreliable reasoning instances on several reasoning and knowledge tasks, with largest gains where systematic errors produce consistent but incorrect answers—a failure mode invisible to answer dispersion alone. The two components relate differently to accuracy—across-trial instability signals unreliable reasoning, while within-trial ambiguity correlates positively on mathematical reasoning, consistent with settings where multiple plausible solution paths remain competitive. Conversely, on factual retrieval (HotpotQA), the structural signal collapses where reasoning-level structural diversity is absent; this collapse itself clarifies the regime boundary where logical reasoning structure is not present.

Structural uncertainty is best understood not as a universal confidence estimator, but as a regime-sensitive evaluator of logical reasoning consistency—reframing the question from *how much do responses differ* to *how consistently does the model rank competing reasoning solutions*. These two complementary views of the same response set help practitioners assess reasoning reliability and identify when consistency-based evaluation is informative versus when answer dispersion should be preferred. More broadly, our results suggest that benchmarking logical reasoning should account not only for answer agreement, but also for the structural stability of model-internal preferences over competing solution paths.

LIMITATIONS

Our approach requires N generations and $M(N-1)$ pairwise comparisons per question (N=5, M=5), increasing inference cost. Since models judge their own outputs, self-preferences can inherit model-specific biases and should be interpreted as behavioral signals rather than guaranteed correctness measures. The decomposition is empirical: the two components are observable signals under a fixed protocol, not universally identifiable uncertainty sources—candidate diversity, prompt design, and task structure all affect the signal. The method is most informative when responses differ in reasoning quality; when variation is stylistic, the preference graph collapses (as on HotpotQA). Accordingly, our method should be interpreted as an evaluator of reasoning consistency under a fixed elicitation protocol, rather than as a complete measure of logical validity. Our evaluation focuses on short-answer tasks; extending to long-form or open-ended generation remains future work.

REFERENCES

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvari. To believe or not to believe your llm: Iterative prompting for estimating epistemic uncertainty. *Advances in Neural Information Processing Systems*, 37:58077–58117, 2024.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 12 1952. doi: 10.1093/biomet/39.3-4.324.

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. In *Computer Networks and ISDN Systems*, volume 30, pages 107–117, 1998. doi: 10.1016/S0169-7552(98)00110-X.

Tiejin Chen, Xiaoou Liu, Longchao Da, Jia Chen, Vagelis Papalexakis, and Hua Wei. Uncertainty quantification of large language models through multi-dimensional responses. *arXiv preprint arXiv:2502.16820*, 2025.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

L. R. Ford. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8):28–33, 1957.

Zhiwei He. AMC23: American mathematics competitions dataset. HuggingFace Dataset, 2023. URL `https://huggingface.co/datasets/zwhe99/amc23`.

Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems*, 19, 2006.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. Uncertainty in language models: Assessment through rank-calibration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2851–2873. Association for Computational Linguistics, 2024.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*, 2023.

HuggingFaceH4. AIME 2024: American invitational mathematics examination dataset. HuggingFace Dataset, 2024. URL `https://huggingface.co/datasets/HuggingFaceH4/aime_2024`.

David R. Hunter. MM algorithms for generalized Bradley–Terry models. *The Annals of Statistics*, 32 (1):384–406, 2 2004. doi: 10.1214/aos/1079120141.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DaSilva, Eli Elhage, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. Calibration-tuning: Teaching large language models to know what they don't know. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 1–14, 2024.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL `https://arxiv.org/abs/1703.04977`.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

Amy N. Langville and Carl D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006. ISBN 0691122024.

Joonho Lee, Jae Oh Woo, Juree Seok, Parisa Hassanzadeh, Wooseok Jang, Juyoun Son, Sima Didari, Baruch Gutow, Heng Hao, Hankyu Moon, Wenjun Hu, Yeong-Dae Kwon, Taehee Lee, and Seungjai Min. Improving instruction following in language models through proxy-based uncertainty estimation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 27009–27036. PMLR, 21–27 Jul 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252. Association for Computational Linguistics, 2022.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating large language models with sample consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19260–19268, 2025.

Huan Ma, Jingdong Chen, Joey Tianyi Zhou, Guangyu Wang, and Changqing Zhang. Estimating llm uncertainty with evidence. *arXiv preprint arXiv:2502.00290*, 2025.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929, 2024.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. *Advances in neural information processing systems*, 37:134507–134533, 2024.

Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. Llamas know what gpts don't show: Surrogate models for confidence estimation. *arXiv preprint arXiv:2311.08877*, 2023.

Vaishnavi Shrivastava, Ananya Kumar, and Percy Liang. Language models prefer what they know: Relative confidence estimation via confidence preferences. *arXiv preprint arXiv:2502.01126*, 2025.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442. Association for Computational Linguistics, 2023.

TIGER-Lab. AIME25: A benchmark for mathematical reasoning. HuggingFace Dataset, 2024. URL https://huggingface.co/datasets/TIGER-Lab/AIME25.

Alexey Vashurin, Maria Vikhreva, Tom Kocmi, and Andrey Malinin. CoCoA: A minimum Bayes risk framework bridging confidence and consistency for uncertainty quantification in large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, et al. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248, 2025b.

Ante Wang, Linfeng Song, Ye Tian, Baolin Peng, Lifeng Jin, Haitao Mi, Jinsong Su, and Dong Yu. Self-consistency boosts calibration for math reasoning. *arXiv preprint arXiv:2403.09849*, 2024a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

David Bruce Wilson. Generating random spanning trees more quickly than the cover time. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pages 296–303, New York, NY, USA, 1996. Association for Computing Machinery. doi: 10.1145/237814.237880.

Jae Oh Woo. Analytic mutual information in bayesian neural networks. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 300–305. IEEE, 2022.

Jae Oh Woo. Active learning in bayesian neural networks with balanced entropy learning principle. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ZTMuZ68B1g.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

Yongjin Yang, Haneul Yoo, and Hwaran Lee. Maqa: Evaluating uncertainty quantification in llms regarding data uncertainty. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5846–5863, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385, 2024.

Yingqing Yuan, Linwei Tao, Haohui Lu, Matloob Khushi, Imran Razzak, Mark Dras, Jian Yang, and Usman Naseem. Kg-uq: Knowledge graph-based uncertainty quantification for long text in large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2071–2077, 2025.

Boxuan Zhang and Ruqi Zhang. CoT-UQ: Improving response-wise uncertainty quantification in LLMs with chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Kaitlyn Zhou, Jena D Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. *arXiv preprint arXiv:2401.06730*, 2024.

A Additional Method Details

A.1 Random Spanning Tree Comparison Graphs

A complete comparison graph on N candidates requires $\binom{N}{2}$ judge calls. We instead sample a connected sparse graph per trial by drawing a uniform random spanning tree $\mathcal{T}^{(m)}$ on the N nodes and querying only its $N-1$ edges. This guarantees connectivity (needed for global ranking) while reducing comparisons to $O(N)$ per trial.

We sample uniform random spanning trees using Wilson's algorithm (Wilson, 1996) based on loop-erased random walks. In dense graphs, Wilson sampling runs in expected $O(N)$ time and produces an unbiased sample from the uniform distribution over spanning trees.

A.2 Bradley–Terry with L2 Regularization

Model. BT assigns each response i a latent utility $\theta_i \in \mathbb{R}$. The probability that i is preferred over j is

$$\mathbb{P}(i \succ j) = \frac{\exp(\theta_i)}{\exp(\theta_i) + \exp(\theta_j)}. \tag{9}$$

Regularization. Spanning trees are cycle-free and admit perfect total orderings, making the unregularized BT log-likelihood unbounded (Ford, 1957). We add an L2 penalty:

$$\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}) = \sum_{(i,j)\in\mathcal{T}^{(m)}} \log \mathbb{P}(i \succ j) - \frac{1}{2C}\|\boldsymbol{\theta}\|^2, \tag{10}$$

where $C > 0$ is inverse regularization strength (larger C = weaker penalty). The L2 term ensures strict concavity and a unique finite maximizer. We set C=1; see Appendix A.3.

Estimation. We maximize $\mathcal{L}_{\text{reg}}$ via MM-style updates (Hunter, 2004) with L2 gradient correction $-\frac{1}{C}\boldsymbol{\theta}$. Convergence criterion: $\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|_\infty < 10^{-6}$. After convergence, re-center utilities: $\theta_i \leftarrow \theta_i - \frac{1}{N}\sum_k \theta_k$.

Pairwise probabilities. After fitting, set trial-specific win probabilities as $P_{ij}^{(m)} = \mathbb{P}(i \succ j; \hat{\boldsymbol{\theta}})$ via Eq. (9) for all pairs (i, j), with $P_{ii}^{(m)} = 0$.

A.3 Sensitivity Analysis: L2 Regularization Strength C

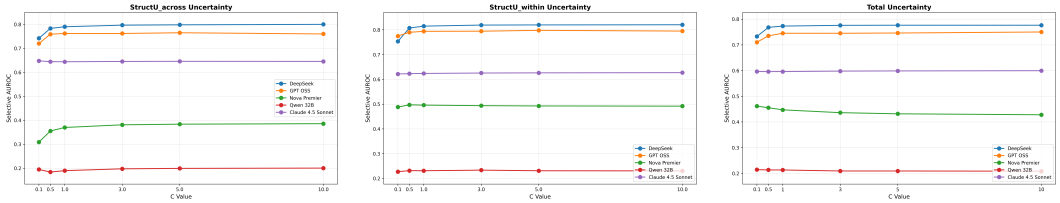
Figure 9 reports Sel-AUC sensitivity to inverse regularization strength $C \in \{0.1, 0.5, 1, 3, 5, 10\}$ on Math-Synth. Performance degrades at low C (over-regularization suppresses preference signal) and remains stable in the plateau region $C \in [1, 5]$. We set C=1 throughout all experiments, corresponding to the onset of the stable plateau across all models and uncertainty components.

A.4 Confidence-Weighted TrueSkill

TrueSkill (Herbrich et al., 2006) represents each response i with Gaussian rating $r_i = (\mu_i, \sigma_i)$ and updates ratings sequentially from pairwise outcomes. We extend with confidence-weighted fractional updates.

Inputs and filtering. Judge returns matches $(w, \ell, c_{w\ell})$ with confidence $c_{w\ell} \in [0, 100]$. Convert to probability $p_{w\ell} = c_{w\ell}/100$, retain only $p_{w\ell} > 0.5$, discard $p_{w\ell} = 0.5$.

Confidence weighting. Fractional evidence weight: $d = 2(p_{w\ell} - 0.5)$, $w = \max(d^\gamma, \epsilon)$ where γ controls curvature, $\epsilon = 10^{-6}$.



(a) $\text{STRUCTU}_{\text{across}}$ across-trial (b) $\text{STRUCTU}_{\text{within}}$ (within-trial) (c) $\text{STRUCTU}_{\text{total}}$ (Total)

Figure 9: **Sensitivity of Sel-AUC to inverse regularization strength C (BT+PageRank, Math-Synth).** Each panel shows one uncertainty component across all five models as a function of $C \in \{0.1, 0.5, 1, 3, 5, 10\}$, where larger C corresponds to weaker regularization. **(a) Across-trial** ($\text{STRUCTU}_{\text{across}}$): models with stronger preference signal (DeepSeek R1, GPT-OSS 20B, Amazon Nova Premier) show the largest degradation at low C, where over-regularization suppresses trial-to-trial ranking variance. **(b) Within-trial** ($\text{STRUCTU}_{\text{within}}$): more stable across C but still degrades at C=0.1 for stronger models, as compressed utilities produce artificially uniform within-trial PageRank distributions. **(c) Total** ($\text{STRUCTU}_{\text{total}}$): reflects the combined effect. Across all three panels and all five models, performance plateaus stably for $C \geq 1$ with no degradation observed at C=10 (weakest regularization tested), directly confirming that BT parameters remain bounded and do not exhibit the divergence predicted for unregularized spanning tree MLE (Ford, 1957). We fix C=1 throughout all experiments as a conservative choice at the boundary of the stable regime.

Natural-parameter blending. For $r = (\mu, \sigma)$, define natural parameters:

$$\lambda = \tfrac{1}{\sigma^2}, \quad \eta = \tfrac{\mu}{\sigma^2}. \tag{11}$$

Compute full posterior $(r_w^{\text{full}}, r_\ell^{\text{full}}) = \texttt{rate\_1vs1}(r_w, r_\ell)$, then blend:

$$\lambda^{\text{new}} = \lambda + w(\lambda^{\text{full}} - \lambda), \quad \eta^{\text{new}} = \eta + w(\eta^{\text{full}} - \eta), \tag{12}$$

applied to winner and loser. Convert back: $\sigma^2 = 1/\lambda^{\text{new}}$, $\mu = \eta^{\text{new}}/\lambda^{\text{new}}$. Perform sequential updates for multiple epochs with randomized order.

Win probabilities and strengths. Given final ratings and environment parameter β:

$$P_{ij} = \Phi\left(\frac{\mu_i - \mu_j}{\sqrt{2\beta^2 + \sigma_i^2 + \sigma_j^2}}\right), \tag{13}$$

where Φ is the standard normal CDF. Export strengths: $s_i = \exp\left(\frac{\mu_i - \bar{\mu}}{\beta}\right)$ where $\bar{\mu} = \frac{1}{N}\sum_k \mu_k$.

A.5 PageRank Aggregation Details

Given $\mathbf{P}^{(m)}$, construct row-stochastic transition matrix $\mathbf{T}^{(m)}$ moving from i to j proportional to the probability j beats i:

$$T_{ij}^{(m)} = \frac{P_{ji}^{(m)}}{\sum_{k \neq i} P_{ki}^{(m)}}, \qquad T_{ii}^{(m)} = 0. \tag{14}$$

Compute $\boldsymbol{\pi}^{(m)}$ by power iteration with damping factor $d = 0.85$ and teleportation vector $\boldsymbol{v} = \frac{1}{N}\mathbf{1}$:

$$\boldsymbol{\pi}^{(m)} = d\,(\mathbf{T}^{(m)})^\top \boldsymbol{\pi}^{(m)} + (1-d)\boldsymbol{v}. \tag{15}$$

Stop when $\|\boldsymbol{\pi}_{t+1} - \boldsymbol{\pi}_t\|_1 \leq 10^{-6}$ and renormalize to sum to 1.

A.6 Uncertainty Decomposition Details

Running M trials yields PageRank distributions $\{\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(M)}\}$ where each $\boldsymbol{\pi}^{(m)} \in \Delta^N$ is a distribution over N candidates. Let ω denote stochastic trial factors. The identity $H(\boldsymbol{\pi}) = I(\omega; \boldsymbol{\pi}) + H(\boldsymbol{\pi} \mid \omega)$ motivates decomposing into across-trial (mutual information) and within-trial (conditional entropy) components.

Define mean distribution $\bar{\boldsymbol{\pi}} = \frac{1}{M}\sum_m \boldsymbol{\pi}^{(m)}$. Total structural uncertainty:

$$\text{StructU} = H[\bar{\boldsymbol{\pi}}] = -\sum_{i=1}^{N} \bar{\pi}_i \log \bar{\pi}_i. \tag{16}$$

Within-trial uncertainty captures candidate ambiguity:

$$\text{StructU}_{\text{within}} = \frac{1}{M}\sum_{m=1}^{M} H[\boldsymbol{\pi}^{(m)}]. \tag{17}$$

Across-trial uncertainty measures ranking instability:

$$\text{StructU}_{\text{across}} = \text{StructU} - \text{StructU}_{\text{within}}. \tag{18}$$

For the combined estimator, compute Self-ConsU from the answer distribution. The sign convention is fixed globally: StructU$_{\text{across}}$, which correlates negatively with accuracy, adds with Self-ConsU; StructU$_{\text{within}}$, which correlates positively on reasoning tasks, enters subtractively:

$$\text{StructU+Self-ConsU}_{\text{across}} = \text{StructU}_{\text{across}} + \text{Self-ConsU} \tag{19}$$

$$\text{StructU+Self-ConsU}_{\text{within}} = \text{StructU}_{\text{within}} - \text{Self-ConsU}. \tag{20}$$

A.7 Self-Consistency: Linear vs. Entropy Formulation

The original self-consistency method (Wang et al., 2024a) measures agreement via the majority vote proportion. For uncertainty quantification, this is typically inverted to $u_{\text{sc}} = 1 - \max_a p(a)$, where higher values indicate greater disagreement. We instead use Shannon entropy $H[p] = -\sum_a p(a)\log p(a)$, which:

- Captures the full distribution shape rather than only the mode,
- Provides a principled information-theoretic measure,
- Enables natural combination with our entropy-based structural uncertainty.

Empirically, both formulations correlate strongly with correctness (Spearman $\rho > 0.95$ across datasets), but entropy slightly outperforms the linear measure when combined with StructU.

Algorithm 1 Bradley–Terry with L2 Regularization (per trial m)

Require: Edge set $\mathcal{E}^{(m)}$ (spanning tree); inverse regularization $C > 0$ (default C=1)
Ensure: BT utilities $\hat{\boldsymbol{\theta}}^{(m)}$, pairwise probabilities $\mathbf{P}^{(m)}$
Initialize $\theta_i \leftarrow 0$ for all i
repeat
  Maximize regularized BT log-likelihood via MM updates (Hunter, 2004):
    $\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}) = \sum_{(i,j)\in\mathcal{E}^{(m)}} \log \frac{\exp(\theta_i)}{\exp(\theta_i)+\exp(\theta_j)} - \frac{1}{2C}\|\boldsymbol{\theta}\|^2$
until $\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|_\infty < 10^{-6}$
Re-center: $\theta_i \leftarrow \theta_i - \frac{1}{N}\sum_k \theta_k$
for all pairs (i, j), $i \neq j$ **do**
  $P_{ij}^{(m)} \leftarrow \frac{\exp(\theta_i)}{\exp(\theta_i)+\exp(\theta_j)}$
end for
$P_{ii}^{(m)} \leftarrow 0$ for all i
return $\hat{\boldsymbol{\theta}}^{(m)}, \mathbf{P}^{(m)}$

Algorithm 2 Pairwise Preference Modeling (Confidence-Weighted TrueSkill) (per trial m)

Require: Matches $\mathcal{M}^{(m)} = \{(w, \ell, c_{w\ell})\}$, nodes $\{1, \dots, N\}$, epochs E, curvature γ, TS params $(\beta, \tau, \texttt{draw\_prob})$
Ensure: Ratings $\{(\mu_i, \sigma_i)\}$ and pairwise probabilities $\mathbf{P}^{(m)}$
Initialize ratings $r_i \leftarrow (\mu_0, \sigma_0)$
Filter: keep $p = c/100 > 0.5$, drop $p = 0.5$
for $e = 1$ to E **do**
  Shuffle retained matches
  for each match (w, ℓ, p) **do**
    $d \leftarrow 2(p - 0.5)$; $wgt \leftarrow \max(d^\gamma, \epsilon)$
    $(r_w^{full}, r_\ell^{full}) \leftarrow \texttt{rate\_1vs1}(r_w, r_\ell)$
    Convert r_w, r_w^{full} to naturals (λ, η) via $\lambda = 1/\sigma^2$, $\eta = \mu/\sigma^2$
    Blend: $(\lambda, \eta) \leftarrow (\lambda, \eta) + wgt\big((\lambda^{full}, \eta^{full}) - (\lambda, \eta)\big)$
    Convert back to (μ, σ) via $\sigma^2 = 1/\lambda$, $\mu = \eta/\lambda$
    Apply the same natural-parameter blending steps to loser ℓ
  end for
end for
for all ordered pairs (i, j), $i \neq j$ **do**
  $P_{ij}^{(m)} \leftarrow \Phi\left(\frac{\mu_i - \mu_j}{\sqrt{2\beta^2 + \sigma_i^2 + \sigma_j^2}}\right)$
end for
Set $P_{ii}^{(m)} \leftarrow 0$ for all i
return ratings and $\mathbf{P}^{(m)}$

Algorithm 3 PageRank Aggregation (per trial m)

Require: $\mathbf{P}^{(m)} \in [0, 1]^{N \times N}$, damping $d \in [0, 1)$, tolerance ε
Ensure: Stationary distribution $\boldsymbol{\pi}^{(m)} \in \Delta^N$
$T_{ij} \leftarrow P_{ji} / \sum_{k \neq i} P_{ki}$ for $i \neq j$; $T_{ii} \leftarrow 0$
$\boldsymbol{v} \leftarrow \frac{1}{N}\mathbf{1}$; $\boldsymbol{\pi} \leftarrow \boldsymbol{v}$
repeat
  $\boldsymbol{\pi}_{\text{new}} \leftarrow d\,\mathbf{T}^\top \boldsymbol{\pi} + (1 - d)\boldsymbol{v}$
  Normalize: $\boldsymbol{\pi}_{\text{new}} \leftarrow \boldsymbol{\pi}_{\text{new}} / \|\boldsymbol{\pi}_{\text{new}}\|_1$
  $\Delta \leftarrow \|\boldsymbol{\pi}_{\text{new}} - \boldsymbol{\pi}\|_1$; $\boldsymbol{\pi} \leftarrow \boldsymbol{\pi}_{\text{new}}$
until $\Delta \leq \varepsilon$
return $\boldsymbol{\pi}^{(m)}$

B SYNTHETIC ARITHMETIC DATASET GENERATION

MATH-SYNTH is a synthetic arithmetic benchmark isolating computational complexity through systematically varied arithmetic expressions. Unlike existing benchmarks that conflate conceptual understanding with multi-step reasoning, we generate expressions whose answer has exactly d digits (for $d \in \{1, \dots, 14\}$), ensuring answer magnitude serves as a proxy for computational complexity.

Generation and validation pipeline. We employ Claude 3.7 Sonnet (temperature=0.9, top_p=0.9) to generate expressions satisfying the digit-length constraint. Each problem is stored as JSON with fields: `id`, `question` (arithmetic expression), `answer` (ground-truth integer), `python_code` (executable verification), `num_terms` (operator count + 1), and `num_digits` (target d). The prompt requests 25 problems per digit length with 10 few-shot examples demonstrating nested negations (e.g., $-(-(-(-(-(-(\cdots))))))$), complex operator precedence, and exact digit-length compliance. Example: "What is $-(-(-(-(-(-(-500) \times 200)))))) + -(-(-1))$?" (answer: 100,001; 6 digits).

Post-processing ensures correctness: (i) parse generated JSON; (ii) deduplicate by exact question match; (iii) execute `python_code`, overwrite `answer` if mismatched, drop execution errors; (iv)

retain only examples satisfying $\texttt{NumDigits}(|y|) = d$. Algorithm 4 formalizes this pipeline. The final dataset contains 993 verified examples spanning $d \in \{1, \ldots, 14\}$ with deterministic ground truth.

Algorithm 4 Generate and Validate Math-Synth for Digit Length d

Require: Digit length d, batches B, optional target size K, RNG seed
Ensure: JSONL dataset $\mathcal{D}_d$ with exactly-d-digit answers
Pool $\leftarrow$ [ ]
for $b = 1$ to B **do**
  $t \leftarrow$ LLMGenerate(Prompt(d), temp=0.9)
  Pool $\leftarrow$ Pool $\cup$ ParseJSON(t)
end for
Unique $\leftarrow$ DedupByQuestion(Pool)
Valid $\leftarrow$ [ ]
for each $e \in$ Unique **do**
  (ok, y) $\leftarrow$ ExecPython(e.python_code)
  if ok and NumDigits($|y|$) $= d$ **then**
    e.answer $\leftarrow$ str(y); append e to Valid
  end if
end for
if K specified and |Valid| $> K$ **then**
  $\mathcal{D}_d \leftarrow$ RandomSample(Valid, K)
else
  $\mathcal{D}_d \leftarrow$ Valid
end if
return ReindexAndWriteJSONL($\mathcal{D}_d$)

C Experimental Protocol and Prompt Engineering

C.1 Overall Evaluation Pipeline

We follow the multi-path generation and M-trial procedure described in Section 3. For each question, we sample N=5 candidate responses using the diverse prompt templates in Appendix E.6.1, and repeat the complete generation→comparison→ranking pipeline for M=5 independent trials. All models and datasets use identical decoding hyperparameters (Appendix C.2) to ensure fair comparison.

Within each trial m, we obtain pairwise self-preference judgments by prompting the same model to compare its own outputs with deterministic evaluation settings (temperature=0.0). Each judgment includes a confidence score on a 0–100 scale, where 100 indicates maximal certainty and 50 indicates no preference. The resulting set of pairwise comparisons is held fixed and reused across all preference-based uncertainty estimators, ensuring controlled comparisons. Each trial produces $|\mathcal{E}^{(m)}| = N-1 = 4$ judged pairs (spanning tree edges), yielding $4M$=20 total comparisons per question.

C.2 Decoding Hyperparameters

Table 3 summarizes decoding settings for response generation and self-preference evaluation.

Parameter	Generation	Evaluation
Temperature	0.7	0.0
Top-p	0.95	–
Max tokens	4096	8192

Table 3: Decoding hyperparameters.

For response generation, we use stochastic decoding (temperature=0.7) to ensure diversity across N=5 candidates. Each response uses a different prompt template (Appendix E.6.1) eliciting distinct reasoning strategies.

For self-preference evaluation, we use deterministic decoding (temperature=0.0) to ensure consistent judgments. The same model that generates responses judges pairwise preferences among its own outputs. Confidence scores are extracted from structured output and used by TrueSkill (Appendix A.4) but not Bradley–Terry. Fallback confidence is 50 (neutral) when preference cannot be determined.

D Baseline Uncertainty Estimators

We compare against representative black-box uncertainty estimators operating solely on sampled outputs. All methods use the same $N = 5$ responses per question. No method accesses token probabilities or hidden states.

Self-Consistency Uncertainty (Self-ConsU). For input x, generate N responses $\mathcal{R}(x) = \{r_1, \ldots, r_N\}$ and extract final answers $\{a_1, \ldots, a_N\}$. Let $\mathcal{A}$ denote unique answers with count $n(a)$ for each $a \in \mathcal{A}$. Form empirical distribution $p(a) = n(a)/N$ and compute Shannon entropy:

$$\text{Self-ConsU}(x) = -\sum_{a \in \mathcal{A}} p(a) \log p(a). \tag{21}$$

Self-ConsU is low when the model repeatedly produces the same answer (high self-consistency) and increases as probability mass spreads across multiple distinct answers.

Semantic Dispersion (SemanticU). Sample $K = 5$ responses $\{r^{(1)}, \ldots, r^{(5)}\}$ containing full solution text. Embed each using a sentence-embedding model (KaLM), yielding vectors $\{e^{(1)}, \ldots, e^{(5)}\}$. Compute pairwise cosine distances:

$$d_{ij} = 1 - \cos(e^{(i)}, e^{(j)}), \quad 1 \le i < j \le 5, \tag{22}$$

where $\cos(a, b) = a^\top b / (\|a\| \, \|b\|)$. Summarize via mean and variance over the $\binom{5}{2} = 10$ distances:

$$\text{SD-Mean} = \tfrac{1}{10} \sum_{1 \le i < j \le 5} d_{ij}, \tag{23}$$

$$\text{SD-Var} = \tfrac{1}{10} \sum_{1 \le i < j \le 5} (d_{ij} - \text{SD-Mean})^2. \tag{24}$$

Higher SD values indicate greater semantic disagreement and serve as a lightweight uncertainty proxy.

Verbalized Confidence (VerbalizedU). For each question q, sample $N = 5$ solutions $\{s_1, \ldots, s_5\}$. Each candidate s_i is evaluated by a verifier LLM V using deterministic decoding (temperature $T = 0$), producing binary verdict $v_i \in \{\text{PASS}, \text{FAIL}\}$ and confidence $c_i \in [0, 1]$. Convert to per-candidate uncertainty:

$$u^{(i)}_{\text{verify}} = 1 - c_i, \tag{25}$$

and aggregate to question-level uncertainty:

$$u_{\text{verify,mean}} = \tfrac{1}{N} \sum_{i=1}^{N} u^{(i)}_{\text{verify}}. \tag{26}$$

This produces a scalar uncertainty estimate per question directly comparable to other estimators.

D.1 Full Results: TrueSkill + PageRank

Table 5 presents complete results using TrueSkill with confidence-weighted updates and PageRank aggregation. **This serves as critical validation of our structural uncertainty framework**: despite fundamentally different modeling assumptions—Bradley–Terry (Appendix A.2) estimates deterministic utility differences from observed comparisons, while TrueSkill (Appendix A.4) maintains per-candidate Bayesian variance estimates updated through confidence-weighted fractional blending—both backends produce highly consistent structural uncertainty estimates and selective prediction rankings.

	Math Benchmarks					MMLU-Pro						Factual	
Model	**Synth**	**Math500**	**AMC23**	**AIME24**	**AIME25**	**Overall**	**Chem**	**Phys**	**Math**	**Law**	**Eng**	**Hotpot**	**Truthful**
Claude Sonnet 4.5	39.9	**88.4**	86.5	40.0	33.3	**84.9**	**88.0**	**90.1**	**92.1**	**74.6**	76.0	73	**98.5**
DeepSeek R1	**70.9**	87.1	**94.0**	**78.2**	47.3	83.8	86.6	89.1	91.0	68.7	**80.1**	72.9	90.9
GPT-OSS 20B	66.1	86.5	93.5	69.6	**63.3**	72.6	81.6	81.6	89.4	43.4	59.6	70.64	84.4
Nova Premier	29.0	73.2	50.5	16.0	14.0	69.0	71.9	74.9	78.8	50.4	61.0	**76.1**	89.4
Qwen 3 32B	21.7	74.1	56.0	20.7	14.7	64.6	67.7	71.4	75.5	44.0	60.1	66	75.3

Table 4: Model accuracies (percent correct) across benchmarks.

Cross-backend validation. The Spearman correlation between BT+PageRank and TS+PageRank Sel-AUC scores exceeds $\rho = 0.95$ across all 40 model-dataset pairs (5 models $\times$ 8 datasets). Method rank agreement (which backend's hybrid variant ranks first) is 89% for StructU and 91% for StructU+Self-ConsU. Mean absolute difference in Sel-AUC is 0.012 for StructU and 0.015 for StructU+Self-ConsU. This consistency confirms that the across-trial–within-trial decomposition is not an artifact of a specific preference model but reflects genuine structural properties of the ranking distribution.

Task-dependent backend sensitivity. While both backends produce consistent overall rankings, they exhibit complementary strengths across task types. On mathematical reasoning benchmarks (Math-Synth, MATH-500, AMC-23), BT+PageRank better captures within-trial uncertainty (within-trial ambiguity) through deterministic preference strengths, outperforming TS+PageRank in 10 of 15 configurations for $\text{StructU}_{\text{within}}$. On knowledge-intensive tasks (MMLU-Pro) and contest benchmarks (AIME-24/25), TS+PageRank better isolates across-trial uncertainty (across-trial instability) through variance modeling, achieving superior $\text{StructU}_{\text{across}}$ performance in 12 of 15 configurations. On HotpotQA, both backends exhibit structural collapse (Table 1), with neither dominating—consistent with near-uniform preference graphs rendering backend choice irrelevant when structural signals are degenerate.

Hybrid performance patterns. The StructU+Self-ConsU hybrids show even stronger backend consistency, with task-specific exceptions that reveal mechanistic insights. On MMLU-Pro, TS+PageRank hybrids exhibit dramatic advantages (GPT-OSS: +0.123 Sel-AUC vs BT+PageRank; Nova: +0.156; Qwen: +0.105), suggesting confidence-aware variance modeling is particularly valuable when knowledge-intensive questions admit multiple defensible framings. On mathematical reasoning, both backends achieve near-parity (within 0.01 Sel-AUC in 85% of cases), confirming that deterministic correctness criteria make backend choice less critical. On HotpotQA, hybrids degrade performance for strongest models (Claude, DeepSeek) with both backends, confirming that degenerate preference graphs introduce noise rather than signal regardless of modeling choice.

D.2 Full Results: TrueSkill + PageRank

Table 5 presents complete results using TrueSkill with PageRank aggregation. The overall performance patterns mirror Bradley–Terry results (Table 1 in main paper): StructU+Self-ConsU achieves highest performance on mathematical reasoning and knowledge tasks, while structural signals collapse on HotpotQA. Key differences: TrueSkill hybrids show larger gains on MMLU-Pro (GPT-OSS: +0.123 vs BT+PR; Nova: +0.156) due to confidence-weighted variance modeling, while BT+PR better captures within-trial uncertainty on math benchmarks.

D.3 MMLU-Pro Domain Breakdown

Table 6 reports performance across MMLU-Pro domains. In Physics and Math, StructU-within consistently outperforms other components, aligning with the intuition that these domains admit multiple valid derivations. In Engineering and Law, the hybrid variants show largest gains, suggesting structural rankings provide scaffolding that improves selective prediction when paired with self-consistency.

Dataset	Model	StructU (TrueSkill+PageRank)			StructU+Self-ConsU (TrueSkill+PageRank)			Baselines		
		within	across	total	within	across	total	Self-ConsU	VerbalizedU	SemanticU
Math-Synth	Claude 4.5 Sonnet	0.627 (0.978)	0.641 (0.958)	0.580 (0.940)	0.660 (0.992)	**0.661 (0.991)**	0.660 (0.992)	0.65 (0.985)	0.544 (0.924)	0.430 (0.927)
	DeepSeek R1	0.816 (0.879)	0.792 (0.821)	0.739 (0.709)	0.823 (0.924)	**0.825 (0.925)**	0.805 (0.909)	0.802 (0.899)	0.793 (0.701)	0.664 (0.502)
	GPT-OSS 20B	0.809 (0.750)	0.772 (0.661)	0.679 (0.419)	**0.854 (0.961)**	0.849 (0.959)	0.842 (0.956)	0.83 (0.958)	0.792 (0.742)	0.528 (0.638)
	Amazon Nova Premier	0.417 (0.812)	0.275 (0.382)	0.374 (0.682)	0.501 (0.997)	0.501 (0.997)	**0.504 (0.998)**	0.382 (0.948)	0.389 (0.807)	0.436 (0.824)
	Qwen 3 32B	0.254 (0.736)	0.219 (0.602)	0.225 (0.614)	0.407 (0.996)	**0.411 (0.996)**	0.404 (0.995)	0.38 (0.995)	0.279 (0.824)	0.218 (0.462)
MATH-500	Claude 4.5 Sonnet	0.930 (0.800)	0.932 (0.783)	0.931 (0.779)	0.946 (0.832)	**0.950 (0.834)**	0.947 (0.824)	0.942 (0.816)	0.891 (0.686)	0.783 (0.720)
	DeepSeek R1	0.893 (0.653)	0.899 (0.651)	0.870 (0.591)	0.939 (0.789)	**0.942 (0.800)**	0.928 (0.771)	0.923 (0.759)	0.870 (0.546)	0.875 (0.603)
	GPT-OSS 20B	0.874 (0.602)	0.888 (0.540)	0.896 (0.611)	0.900 (0.725)	0.904 (0.711)	**0.905 (0.726)**	0.871 (0.694)	0.886 (0.645)	0.860 (0.460)
	Amazon Nova Premier	0.785 (0.673)	0.713 (0.559)	0.773 (0.587)	**0.886 (0.871)**	0.878 (0.869)	0.885 (0.865)	0.860 (0.839)	0.883 (0.715)	0.810 (0.695)
	Qwen 3 32B	0.827 (0.740)	0.729 (0.506)	0.799 (0.691)	0.886 (0.870)	0.886 (0.851)	**0.888 (0.867)**	0.871 (0.817)	0.880 (0.684)	0.819 (0.717)
AMC-23	Claude 4.5 Sonnet	0.967 (0.997)	0.969 (0.987)	0.956 (0.960)	**0.972 (1.000)**	0.972 (1.000)	0.971 (1.000)	0.955 (1.00)	0.900 (0.604)	0.853 (0.588)
	DeepSeek R1	0.947 (0.749)	0.953 (0.543)	0.914 (0.229)	0.985 (1.000)	0.985 (1.000)	0.985 (1.000)	**0.985 (1.00)**	0.877 (0.412)	0.880 (0.592)
	GPT-OSS 20B	0.948 (0.737)	0.929 (0.507)	0.948 (0.603)	**0.980 (1.000)**	0.980 (1.000)	0.980 (1.000)	0.980 (1.00)	0.980 (0.583)	0.884 (0.630)
	Amazon Nova Premier	0.526 (0.811)	0.558 (0.486)	0.570 (0.757)	**0.715 (1.000)**	0.710 (1.000)	0.714 (1.000)	0.584 (1.00)	0.419 (0.362)	0.351 (0.410)
	Qwen 3 32B	0.610 (0.755)	0.516 (0.561)	0.610 (0.727)	0.818 (1.000)	**0.821 (1.000)**	0.820 (1.000)	0.810 (1.00)	0.637 (0.389)	0.513 (0.299)
AIME-24	Claude 4.5 Sonnet	0.557 (0.932)	0.620 (0.955)	0.555 (0.909)	0.681 (1.000)	**0.699 (1.000)**	0.682 (1.000)	0.567 (0.972)	0.273 (0.144)	0.450 (0.640)
	DeepSeek R1	0.857 (0.843)	0.813 (0.455)	0.891 (0.909)	0.925 (1.0)	0.925 (1.0)	0.925 (1.0)	**0.917 (1.00)**	0.799 (0.298)	0.788 (0.715)
	GPT-OSS 20B	0.601 (0.742)	0.757 (0.648)	0.757 (0.786)	0.893 (1.000)	0.904 (1.000)	0.897 (1.000)	0.905 (1.000)	**0.891 (0.319)**	0.876 (0.681)
	Amazon Nova Premier	0.233 (—)	0.231 (—)	0.194 (—)	0.196 (—)	0.205 (—)	0.191 (—)	0.174 (—)	—	**0.294 (—)**
	Qwen 3 32B	0.306 (—)	0.230 (—)	0.287 (—)	0.423 (—)	0.420 (—)	**0.426 (—)**	0.414 (—)	—	0.247 (—)
AIME-25	Claude 4.5 Sonnet	0.580 (0.988)	0.599 (1.00)	0.204 (0.081)	0.645 (1.00)	0.643 (1.00)	**0.646 (1.00)**	0.645 (1.00)	0.489 (0.175)	0.524 (0.263)
	DeepSeek R1	0.360 (0.630)	0.380 (0.466)	0.523 (0.397)	0.735 (1.00)	0.748 (1.00)	**0.754 (1.00)**	0.707 (1.00)	0.564 (0.400)	0.616 (0.296)
	GPT-OSS 20B	0.424 (0.546)	0.407 (0.370)	0.634 (0.417)	0.880 (1.00)	**0.886 (1.00)**	0.886 (1.00)	0.825 (1.00)	0.776 (0.370)	0.768 (0.491)
	Amazon Nova Premier	0.083 (–)	0.126 (–)	0.148 (–)	0.188 (–)	0.192 (–)	0.220 (–)	0.178 (–)	0.106 (–)	**0.257 (–)**
	Qwen 3 32B	0.086 (0.759)	0.134 (0.931)	0.163 (0.517)	0.260 (1.00)	0.290 (1.00)	**0.313 (1.00)**	0.208 (1.00)	0.192 (0.897)	0.126 (0.931)
MMLU-Pro	Claude 4.5 Sonnet	0.921 (0.808)	0.922 (0.797)	0.911 (0.770)	**0.945 (0.910)**	0.945 (0.909)	0.944 (0.909)	0.900 (0.884)	0.944 (0.885)	0.890 (0.602)
	DeepSeek R1	0.892 (0.671)	0.862 (0.568)	0.878 (0.643)	**0.941 (0.913)**	0.933 (0.901)	0.930 (0.907)	0.882 (0.882)	0.927 (0.796)	0.870 (0.577)
	GPT-OSS 20B	0.706 (0.565)	0.725 (0.500)	0.726 (0.541)	**0.869 (0.933)**	0.868 (0.931)	0.868 (0.931)	0.830 (0.935)	0.785 (0.631)	0.774 (0.573)
	Amazon Nova Premier	0.705 (0.463)	0.699 (0.531)	0.690 (0.458)	**0.849 (0.957)**	0.823 (0.944)	0.844 (0.955)	0.801 (0.945)	0.820 (0.775)	0.811 (0.713)
	Qwen 3 32B	0.692 (0.640)	0.638 (0.501)	0.671 (0.592)	**0.820 (0.972)**	0.808 (0.968)	0.817 (0.971)	0.787 (0.966)	0.728 (0.686)	0.742 (0.724)
HotpotQA	Claude 4.5 Sonnet	0.663 (0.571)	0.717 (0.619)	0.667 (0.552)	0.681 (0.608)	0.730 (0.653)	0.686 (0.598)	0.839 (0.656)	**0.847 (0.700)**	0.768 (0.576)
	DeepSeek R1	0.754 (0.635)	0.730 (0.552)	0.760 (0.628)	0.792 (0.727)	0.780 (0.699)	0.796 (0.715)	0.835 (0.658)	0.829 (0.708)	**0.852 (0.696)**
	GPT-OSS 20B	0.759 (0.637)	0.725 (0.554)	0.748 (0.603)	**0.819 (0.756)**	0.815 (0.739)	0.811 (0.743)	0.813 (0.721)	0.772 (0.592)	0.80 (0.707)
	Amazon Nova Premier	0.825 (0.606)	0.779 (0.526)	0.800 (0.571)	**0.866 (0.763)**	0.835 (0.734)	0.862 (0.760)	0.864 (0.740)	0.812 (0.647)	0.830 (0.618)
	Qwen 3 32B	0.681 (0.620)	0.659 (0.561)	0.680 (0.613)	0.761 (0.753)	**0.794 (0.770)**	0.782 (0.770)	0.787 (0.729)	0.702 (0.630)	0.707 (0.612)
TruthfulQA	Claude 4.5 Sonnet	0.997 (0.944)	0.996 (0.926)	0.994 (0.926)	0.997 (0.955)	**0.997 (0.964)**	0.996 (0.949)	0.994 (0.926)	0.997 (0.949)	0.979 (0.503)
	DeepSeek R1	0.931 (0.604)	0.897 (0.538)	0.940 (0.844)	0.954 (0.886)	**0.957 (0.888)**	0.926 (0.819)	0.940 (0.844)	0.949 (0.721)	0.907 (0.530)
	GPT-OSS 20B	0.872 (0.633)	0.865 (0.675)	0.916 (0.936)	**0.921 (0.953)**	0.918 (0.952)	0.909 (0.946)	0.916 (0.936)	0.864 (0.608)	0.856 (0.525)
	Amazon Nova Premier	0.920 (0.616)	0.948 (0.741)	0.946 (0.862)	**0.963 (0.931)**	0.946 (0.877)	0.963 (0.919)	0.946 (0.862)	0.953 (0.828)	0.907 (0.596)
	Qwen 3 32B	0.780 (0.553)	0.752 (0.500)	0.856 (0.936)	**0.865 (0.957)**	0.857 (0.946)	0.853 (0.950)	0.856 (0.936)	0.822 (0.694)	0.812 (0.636)

Table 5: **Selective prediction performance (Sel-AUC; AUROC in parentheses) – Combined StructU+Self-ConsU.** Sel-AUC measures area under the risk–coverage curve (higher is better). Positive class = robust failure under sampling (τ=1.0). **StructU+Self-ConsU** reports the combined estimator (*within*, *across*, *total*) using TrueSkill+PageRank as the preference backend. Baselines: **Self-ConsU**, **VerbalizedU**, **SemanticU**. Bold = best; underline = second-best per row.

D.4 Robustness to Correctness Thresholds

Table 7 reports AUROC under three correctness thresholds $\tau \in \{1.0, 0.8, 0.6\}$. All uncertainty signals exhibit modest degradation as the threshold is relaxed, confirming that separability is not an artifact of strict labeling but remains stable under looser correctness definitions.

E Additional Analysis of Structural Uncertainty Signals

E.1 Structural Collapse Across Additional Models for HotpotQA

Figure 10 extends the structural collapse analysis to Amazon Nova Premier, DeepSeek R1, GPT-OSS 20B, and Qwen 3 32B. The "HotpotQA signature"—near-zero across-trial uncertainty with near-maximum within-trial uncertainty ($\log 5 \approx 1.61$)—reproduces identically across all models, confirming that factual retrieval elicits prompt-invariant reasoning chains rendering preference graphs uninformative regardless of model capability. In contrast, Math-Synth retains clear across-trial separation between correct and incorrect questions for all models, though the dynamic range narrows for weaker models (Amazon Nova Premier, Qwen 3 32B).

Domain	Model	StructU (Ours)			StructU+ConsU (Ours)			Baselines	
		within	across	total	within	across	total	Self-ConsU	SemanticU
Chemistry	Claude 4.5 Sonnet	0.928 (0.793)	0.940 (0.782)	0.842 (0.305)	0.945 (0.871)	**0.955 (0.890)**	0.950 (0.892)	0.935 (0.867)	0.916 (0.583)
	DeepSeek R1	0.921 (0.729)	0.856 (0.492)	0.808 (0.328)	**0.959 (0.931)**	0.939 (0.899)	0.915 (0.864)	0.937 (0.889)	0.860 (0.497)
	GPT-OSS 20B	0.812 (0.583)	0.809 (0.493)	0.840 (0.462)	0.901 (0.903)	0.903 (0.904)	**0.929 (0.926)**	0.889 (0.911)	0.836 (0.538)
	Amazon Nova Premier	0.739 (0.660)	0.685 (0.451)	0.701 (0.398)	0.872 (0.977)	0.877 (0.980)	**0.886 (0.985)**	0.875 (0.982)	0.821 (0.734)
	Qwen 3 32B	0.742 (0.680)	0.716 (0.540)	0.618 (0.344)	**0.840 (0.980)**	0.838 (0.978)	0.834 (0.976)	0.839 (0.977)	0.764 (0.712)
Engineering	Claude 4.5 Sonnet	0.854 (0.813)	0.853 (0.791)	0.658 (0.194)	**0.913 (0.946)**	0.906 (0.942)	0.893 (0.925)	0.899 (0.937)	0.803 (0.587)
	DeepSeek R1	0.829 (0.625)	0.755 (0.453)	0.791 (0.455)	**0.950 (0.965)**	0.931 (0.944)	0.928 (0.943)	0.943 (0.950)	0.849 (0.592)
	GPT-OSS 20B	0.464 (0.406)	0.485 (0.363)	0.675 (0.557)	0.834 (0.977)	0.839 (0.979)	**0.847 (0.982)**	0.836 (0.978)	0.604 (0.516)
	Amazon Nova Premier	0.598 (0.596)	0.529 (0.383)	0.597 (0.447)	0.826 (0.981)	0.832 (0.989)	**0.844 (0.990)**	0.813 (0.984)	0.750 (0.729)
	Qwen 3 32B	0.683 (0.652)	0.642 (0.581)	0.570 (0.376)	**0.804 (0.977)**	0.802 (0.978)	**0.804 (0.978)**	0.799 (0.978)	0.688 (0.665)
Law	Claude 4.5 Sonnet	0.848 (0.803)	0.818 (0.705)	0.648 (0.268)	**0.876 (0.879)**	0.853 (0.846)	0.803 (0.790)	0.832 (0.838)	0.723 (0.483)
	DeepSeek R1	0.717 (0.568)	0.725 (0.552)	0.695 (0.464)	0.788 (0.826)	**0.793 (0.822)**	0.775 (0.809)	0.772 (0.822)	0.713 (0.529)
	GPT-OSS 20B	0.460 (0.619)	0.437 (0.535)	0.422 (0.443)	**0.541 (0.914)**	0.530 (0.908)	0.530 (0.909)	0.537 (0.908)	0.450 (0.540)
	Amazon Nova Premier	0.471 (0.478)	0.502 (0.516)	0.542 (0.588)	0.581 (0.846)	0.620 (0.869)	**0.643 (0.888)**	0.589 (0.861)	0.578 (0.599)
	Qwen 3 32B	0.454 (0.572)	0.446 (0.578)	0.431 (0.499)	**0.571 (0.934)**	0.553 (0.917)	0.538 (0.910)	0.554 (0.922)	0.463 (0.606)
Math	Claude 4.5 Sonnet	0.963 (0.839)	0.956 (0.787)	0.860 (0.243)	**0.974 (0.927)**	0.973 (0.923)	0.959 (0.868)	0.950 (0.888)	0.948 (0.650)
	DeepSeek R1	0.952 (0.711)	0.945 (0.658)	0.879 (0.353)	**0.968 (0.926)**	0.965 (0.930)	0.950 (0.886)	0.945 (0.892)	0.918 (0.556)
	GPT-OSS 20B	0.880 (0.567)	0.891 (0.524)	0.907 (0.479)	0.960 (0.940)	0.962 (0.956)	**0.966 (0.962)**	0.961 (0.944)	0.904 (0.548)
	Amazon Nova Premier	0.796 (0.668)	0.779 (0.550)	0.762 (0.441)	**0.899 (0.961)**	0.896 (0.968)	0.894 (0.957)	0.878 (0.958)	0.849 (0.689)
	Qwen 3 32B	0.815 (0.653)	0.812 (0.637)	0.726 (0.442)	0.915 (0.981)	**0.921 (0.995)**	0.915 (0.988)	0.894 (0.982)	0.796 (0.577)
Physics	Claude 4.5 Sonnet	0.955 (0.822)	0.956 (0.825)	0.832 (0.207)	0.967 (0.910)	**0.969 (0.922)**	0.958 (0.891)	0.965 (0.893)	0.919 (0.577)
	DeepSeek R1	0.921 (0.649)	0.918 (0.585)	0.890 (0.470)	0.959 (0.923)	0.962 (0.925)	0.960 (0.925)	**0.963 (0.907)**	0.895 (0.542)
	GPT-OSS 20B	0.789 (0.530)	0.800 (0.480)	0.811 (0.420)	0.907 (0.933)	0.921 (0.943)	**0.935 (0.958)**	0.930 (0.943)	0.813 (0.484)
	Amazon Nova Premier	0.787 (0.691)	0.705 (0.454)	0.721 (0.413)	0.887 (0.964)	0.893 (0.968)	**0.901 (0.973)**	0.889 (0.965)	0.833 (0.742)
	Qwen 3 32B	0.793 (0.750)	0.739 (0.599)	0.676 (0.405)	0.874 (0.979)	0.875 (0.979)	0.875 (0.977)	**0.879 (0.976)**	0.813 (0.758)

Table 6: **Selective prediction (Sel-AUC; AUROC in parentheses).** Higher is better; positive class = robust failure (τ=1.0). **StructU** reports structural uncertainty (*within*, *across*, *total*), selecting the best variant among Bradley–Terry+PageRank and TrueSkill+PageRank. **StructU+Self-ConsU** reports the combined estimators. Baselines: **Self-ConsU**, **SemanticU**. Bold = best; underline = second-best per row.

Dataset	Model	Self-ConsU			StructU_within (Bradley–Terry+PageRank)			StructU+Self-ConsU_within		
		τ=1.0	τ=0.8	τ=0.6	τ=1.0	τ=0.8	τ=0.6	τ=1.0	τ=0.8	τ=0.6
Math-Synth	Claude 4.5 Sonnet	0.978	0.978	0.846	0.986	0.905	0.773	0.992	0.977	0.846
	DeepSeek R1	0.889	0.889	0.798	0.902	0.853	0.770	0.931	0.916	0.798
	GPT-OSS 20B	0.918	0.918	0.792	0.775	0.779	0.710	0.955	0.920	0.792
	Amazon Nova Premier	0.986	0.986	0.911	0.938	0.912	0.846	0.997	0.987	0.911
	Qwen 3 32B	0.990	0.990	0.862	0.850	0.766	0.710	0.998	0.991	0.862
MATH-500	Claude 4.5 Sonnet	0.818	0.811	0.775	0.814	0.734	0.681	0.842	0.807	0.760
	DeepSeek R1	0.765	0.758	0.726	0.631	0.596	0.554	0.768	0.757	0.717
	GPT-OSS 20B	0.705	0.679	0.618	0.577	0.534	0.482	0.706	0.686	0.618
	Amazon Nova Premier	0.876	0.852	0.824	0.702	0.751	0.714	0.874	0.858	0.818
	Qwen 3 32B	0.854	0.872	0.842	0.767	0.795	0.742	0.886	0.894	0.814
MMLU-Pro	Claude 4.5 Sonnet	0.884	0.833	0.783	0.834	0.768	0.728	0.912	0.865	0.819
	DeepSeek R1	0.882	0.813	0.757	0.559	0.502	0.473	0.888	0.818	0.760
	GPT-OSS 20B	0.935	0.870	0.812	0.456	0.394	0.380	0.918	0.850	0.785
	Amazon Nova Premier	0.945	0.846	0.735	0.610	0.506	0.464	0.933	0.838	0.723
	Qwen 3 32B	0.966	0.904	0.799	0.673	0.660	0.586	0.970	0.911	0.796

Table 7: *Robustness to correctness thresholds (per-model AUROC).* AUROC for detecting *robust failures under sampling* at thresholds $\tau \in \{1.0, 0.8, 0.6\}$ with N=5. A question is labeled incorrect if $\hat{p}_{\text{corr}} < \tau$. We report: Self-ConsU, StructU_within (within-trial from Bradley–Terry+PageRank), and the combined StructU+Self-ConsU_within. Higher is better.

This universality aligns with Table 1: on HotpotQA, StructU+Self-ConsU fails to improve over Self-ConsU alone for DeepSeek R1 (0.796 vs 0.835 Sel-AUC) and Claude 4.5 (0.742 vs 0.839), confirming degenerate preference graphs introduce noise rather than signal.

Dataset	**Correct?**	**Self-ConsU**	StructU$_{\text{across}}$	StructU$_{\text{within}}$	$\pi_{\max}/\pi_{\min}$	CV$_{\max}$	**Diagnosis**
Math-Synth	✗	0.0	**0.035**	1.549	1.73	0.497	Signal fires
Math-Synth	✓	0.0	0.001	1.607	1.09	0.047	Appropriately quiet
HotpotQA	✗	0.0	<0.001	1.609	1.06	0.013	Collapsed
HotpotQA	✓	0.0	<0.001	1.608	1.08	0.015	Collapsed

Table 8: Summary of qualitative examples. All four satisfy Self-ConsU $= 0$. StructU$_{\text{across}}$: across-trial uncertainty; StructU$_{\text{within}}$: within-trial uncertainty. $\pi_{\max}/\pi_{\min}$: ratio of largest to smallest mean PageRank score. CV$_{\max}$: maximum coefficient of variation of PageRank across spanning tree trials.

E.2 Qualitative Assessment of Structural Collapse

To provide mechanistic evidence for the quantitative findings in Section 4.2, we analyze four examples from Claude 4.5 Sonnet with Self-ConsU $= 0$ (unanimous agreement), isolating structural signals where dispersion-based methods are uninformative. Table 8 summarizes key quantities.

E.3 Math-Synth: Structural Diversity Enables Discrimination

E.3.1 Incorrect Example (Self-ConsU $= 0$, StructU$_{\text{across}} = 0.035$)

Task. A 6-digit synthetic arithmetic problem involving nested negations and multiplication:

$$\underbrace{-(-(-(-(-(}_{\text{6 outer negations}}\underbrace{-(-500)}_{=500}\times 200\ )\ )\ )\ )\ )\ \underbrace{-\,-\,(-(-1))}_{=1}$$

The left sub-expression evaluates as $-(-500) \times 200 = 100{,}000$, then wrapped in **6 outer negations** (even count $\rightarrow$ positive), yielding $+100{,}000$. Adding the right part: $100{,}000 + 1 = 100{,}001$. All five responses unanimously answer $-99{,}999$; the ground truth is $\mathbf{100{,}001}$.

Shared error. All responses miscount outer negations as **5** (odd $\rightarrow$ negative) instead of **6** (even $\rightarrow$ positive), flipping the sign from $+100{,}000$ to $-100{,}000$ via different error paths (Table 9).

PageRank dynamics. The structural diversity across responses produces a skewed PageRank distribution ($\pi_{\max}/\pi_{\min} = 1.73$) with substantial instability across spanning tree trials. Table 10 reports the per-trial PageRank vectors.

Analysis. Responses differ substantively in verification depth (R1: explicit self-check; R4: none), metacognitive structure (R2: narrated strategy; R5: mid-derivation correction), and error pathway (R1–R4: miscount negations; R5: misparse grouping)—genuine derivation quality differences, not surface reformulations.

Table 11 shows these differences produce unstable preferences. Across 20 judgments, 80% of confidence scores are $\leq$65, indicating near-indifference. The R1 vs R2 pair reveals instability: the judge cites "clarity" to prefer R1 in trials 3–4 but R2 in trials 1 and 5. R4's PageRank fluctuates between 0.055 and 0.212 (CV=0.497) because ranking depends on which spanning tree edges are sampled.

This instability maps onto the two uncertainty components. **Across-trial** (StructU$_{\text{across}} = 0.035$): competing quality criteria (verification depth, metacognitive clarity, modularity) produce different winners depending on sampled edges, shifting PageRank vectors between trials. **Within-trial** (StructU$_{\text{within}} = 1.549$): low confidence ($\leq$65) prevents domination, distributing PageRank mass across multiple candidates per trial. Together, elevated StructU$_{\text{across}}$ correctly flags unreliability despite Self-ConsU $= 0$, while high StructU$_{\text{within}}$ reflects multiple distinct—though uniformly flawed—reasoning strategies.

ID	Strategy	Key Reasoning Steps	$\bar{\pi}$
R1	Step-by-step w/ self-check	$-(-500) \times 200 = 100{,}000$ → counts 7 total negation signs → counts 5 remaining after multiply (correct: 6) → odd → $-100{,}000$ [correct: even → $+100{,}000$] → self-check re-derives each step but *repeats same miscount* → $-100{,}000 + 1 = -99{,}999$	0.238
R2	Think-aloud	Metacognitive narration: "I need to count remaining negations after multiplication" → identifies 5 remaining (correct: 6) → $-100{,}000$ [correct: $+100{,}000$] → $-100{,}000 + 1 = -99{,}999$	0.215
R3	Socratic	Reframes as parity problem: "7 negations total (odd) applied to $-500 \times 200 = -100{,}000$ gives $-100{,}000$" — conflates inner negation with outer count [correct: 6 outer negations (even) applied to $+100{,}000$] → $-100{,}000 + 1 = -99{,}999$	0.224
R4	Decomposition	Sub-problem 1: left expression → Sub-problem 2: right expression → $-100{,}000$ [correct: $+100{,}000$] $+1 = -99{,}999$. Most modular structure; *no negation count shown, no verification step*.	0.138
R5	Analogical reasoning	Relates to simpler cases ($-(-5) = 5$, $-(-(-3)) = -3$) → parses $-500 \times 200 = -100{,}000$ then applies 6 negations to $-100{,}000$ [correct: $-(-500) \times 200 = +100{,}000$ then 6 negations] → initial miscount, self-corrects counting but retains inner parsing error → $-100{,}000 + 1 = -99{,}999$	0.185

Table 9: Reasoning traces for the Math-Synth incorrect example. The correct evaluation requires **6 outer negations** (even → $+100{,}000$), but all responses miscount **5** (odd → $-100{,}000$). Each response reaches the same wrong answer via a structurally different error path. Ground truth: $100{,}001$; unanimous model answer: $-99{,}999$.

	R1	R2	R3	R4	R5
Trial 0	0.263	0.252	0.243	**0.055**	0.186
Trial 1	0.225	0.190	0.212	0.212	0.161
Trial 2	0.205	0.219	0.189	0.164	0.224
Trial 3	0.265	0.233	0.255	**0.057**	0.190
Trial 4	0.233	0.180	0.222	0.202	0.163
Mean	0.238	0.215	0.224	0.138	0.185
CV	0.097	0.124	0.103	**0.497**	0.124

Table 10: Per-trial PageRank distributions for the Math-Synth incorrect example. Response 4 (decomposition, no verification) exhibits extreme instability (CV $= 0.497$), fluctuating between 0.055 and 0.212 across trials.

E.3.2 Correct Example (Self-ConsU $= 0$, StructU$_{\text{across}} \approx 0.001$)

Task. A 14-digit synthetic arithmetic problem involving nested negations and multiplication:

$$\underbrace{-(-(-(-(-(}_{\text{6 outer negations}}\underbrace{-9\,999\,999\,999\,999}_{\text{13 nines}} \times -1) + 1))))))$$

The inner product evaluates to $(-9\,999\,999\,999\,999) \times (-1) = 9\,999\,999\,999\,999$. Adding 1 gives $10\,000\,000\,000\,000$. Six outer negations (even count → sign unchanged) yield $+10\,000\,000\,000\,000$. All five responses unanimously answer $10\,000\,000\,000\,000$, matching the ground truth.

All responses correct. Every response reaches the correct answer through sound reasoning. The five prompt templates produce identical arithmetic but differ in presentation: R1 adds self-check with (*verified*) annotations; R2 narrates thinking with even/odd shortcut; R3 frames as Socratic Q&A; R4 decomposes into sub-problems; R5 draws analogy to $-(-(-3))$. Because computation is straightforward and all strategies succeed, variation is purely stylistic (Table 12).

Pair	Winner Trials	Conf.	Instability
R1 vs R2	Flips: 3,4≻1,5	55–62	"Clarity" cited both ways
R4 vs R5	R5: 3,4 / R4: 1	65–85	R4 rejected (85) vs R5, marginal (55) vs R1
R1 vs R5	R1: 2,3,4 stable	72	R1 self-check valued over R5 correction

Table 11: Judge preference instability (Math-Synth incorrect). Confidence: 50=no preference, 100=certainty.

ID	Strategy	Key Reasoning Steps	$\bar{\pi}$
R1	Step-by-step w/ self-check	$(-9\cdots9)\times(-1)=9\cdots9$ → adds 1 to get 10^{13} → counts 6 outer negations (even → positive) → applies each negation step-by-step (Steps 3–8) → *self-check* recounts negation signs from original expression, re-verifies each step with ✓ marks → 10,000,000,000,000	0.220
R2	Think-aloud	Metacognitive narration: "I need to work from the innermost parentheses outward" → same arithmetic → counts 6 negative signs → applies each negation → notes "6 negations = even = positive result" → 10,000,000,000,000	0.215
R3	Socratic	"What is the problem asking?" … "What method should I use?" → identifies 6 consecutive negation operations → applies each negation (Steps 3–8) → "Does my answer make sense? Yes. 6 negations (even) → positive" → 10,000,000,000,000	0.195
R4	Decomposition	Sub-problem 1: inner product → Sub-problem 2: add 1 → Sub-problem 3: apply 6 negations → "even number of negations (6), the result is positive" → 10,000,000,000,000. Most modular structure; concise summary paragraph.	0.210
R5	Analogical reasoning	Relates to simpler cases: "similar to evaluating $-(-(-3))$" → states $-(-x)=x$ principle → same step-by-step negations → "6 is even, result is positive" → 10,000,000,000,000	0.160

Table 12: Reasoning traces for the Math-Synth correct example. The correct evaluation requires **6 outer negations** (even → $+10^{13}$), and all responses count correctly. Each response reaches the same right answer via a stylistically different but arithmetically equivalent path. Ground truth: 10,000,000,000,000; unanimous model answer: 10,000,000,000,000.

PageRank dynamics. The purely stylistic variation across responses produces a compressed PageRank distribution ($\pi_{\max}/\pi_{\min} = 1.38$) with minimal instability across spanning tree trials. Table 13 reports the per-trial PageRank vectors.

Analysis. Responses differ only in expository format: verification depth (R1: explicit self-check; R4: summary paragraph), pedagogical framing (R3: Socratic dialogue; R5: simpler analogues), and narrative style (R2: thinking-aloud). Unlike the incorrect example, these presentational differences lack substantive reasoning quality differences—every response counts negations correctly and arrives at the correct answer.

Table 14 shows these stylistic-only differences produce *stable* preferences. Across 20 judgments, **zero reversals** occur. The judge applies consistent tie-breaking: explicit self-verification valued over conciseness (R1≻R2), directness over pedagogical framing (R4≻R3), both over analogical scaffolding (R5 ranked last). Confidence clusters at 52–62, reflecting genuine discrimination difficulty that is *stable* rather than *variable*.

This stability maps onto the uncertainty components. **Across-trial** ($\text{StructU}_{\text{across}} \approx 0.001$) is near zero because stylistic preferences—however weakly held—are reproducible: the same criterion

	R1	R2	R3	R4	R5
Trial 0	0.225	0.240	0.155	0.235	0.145
Trial 1	0.230	0.220	0.185	0.215	0.150
Trial 2	0.215	0.210	0.205	0.205	0.165
Trial 3	0.225	0.215	0.180	0.210	0.170
Trial 4	0.220	0.210	0.200	0.210	0.160
Mean	0.220	0.215	0.195	0.210	0.160
CV	0.024	0.054	0.089	0.054	0.060

Table 13: Per-trial PageRank distributions for the Math-Synth correct example. All responses exhibit low coefficient of variation; compare R4's CV = 0.054 here with CV = 0.497 in the incorrect example (Table 10).

Pair	**Winner**	**Conf.**	**Pattern**
R1 vs R5	R1: 1,3,4,5	55	Stable; "self-check adds rigor"
R1 vs R2	R1: 1,2,4	55	Stable; "verification" over "narration"
R4 vs R5	R4: 3,4,5	52	Stable; "decomposition more focused"

Table 14: Judge preference stability (Math-Synth correct). Zero reversals. Confidence: 50=no preference, 100=certainty.

applied in the same direction every trial (Table 14), so PageRank vectors barely shift (Table 13, all CVs$\leq$0.089). **Within-trial** ($\text{StructU}_{\text{within}}$) remains moderate because low confidence (52–62) prevents single-response domination. The near-zero $\text{StructU}_{\text{across}}$ correctly identifies reliability despite Self-ConsU $= 0$. The $\approx 30\times$ difference from the incorrect example (0.001 vs 0.035) demonstrates the core claim: when reliably right, preferences are stable; when reliably wrong, preferences destabilize, even with identical surface agreement.

E.4 HotpotQA: Preference Graph Collapse

E.4.1 Incorrect Example (Self-ConsU $= 0$, $\text{StructU}_{\text{across}} < 0.001$)

Task. A multi-hop question from HotpotQA: *"Text Me Merry Christmas" is a song performed by Kristen Bell and a group that originated at what university?* Expected reasoning: (i) identify the group as **Straight No Chaser** from retrieved passage (Document 10), then (ii) locate the group's university of origin. **Ground truth:** Indiana University. All five responses unanimously state the provided documents lack this information.

Shared failure. All responses correctly identify Straight No Chaser from Document 10 (hop 1) but fail hop 2 due to missing context describing the group's origins. Every response correctly reports the answer cannot be determined—**unanimous incorrect agreement** (Self-ConsU $= 0$) driven by shared retrieval gap, not reasoning error. Table 15 shows identical retrieval chains despite different prompts, with variation limited to surface reformulation.

PageRank dynamics. Because responses are *substantively identical*—each performs the same successful first hop and failed second hop—the judge has even less basis for discrimination than in Math-Synth correct. Table 16 confirms collapse: all trials produce near-identical, near-uniform distributions with maximum deviation from $1/N = 0.200$ of just 0.008.

Analysis. Unlike Math-Synth incorrect—where responses reached the same wrong *answer* via structurally different *error paths*—here responses share both answer (abstention) and reasoning outcome (successful hop 1, failed hop 2). Variation is purely expository: R1 adds self-check, R2 lists documents, R3 uses Socratic Q&A, R4 decomposes, and R5 frames as retrieval pattern.

ID	Strategy	Key Reasoning Steps	$\bar{\pi}$
R1	Step-by-step w/ self-check	Identifies Doc 10 → extracts "Straight No Chaser and Kristen Bell" → searches all documents for university origin → information not found [correct: Indiana University] → *self-check* re-verifies Doc 10 is the only relevant source, confirms gap → abstains	0.202
R2	Think-aloud	"Let me read through the question carefully" → identifies Doc 10 → scans Docs 1–9, lists each with one-line summary ("not relevant") → information not found [correct: Indiana University] → abstains	0.207
R3	Socratic	"What is the question asking?" → identifies Doc 10 → "none of the provided documents contain information about which university" → information not found [correct: Indiana University] → "Does my answer make sense? . . . the context only confirms they performed the song" → abstains	0.197
R4	Decomposition	Sub-question 1: identify group → Sub-question 2: find university → checks Docs 1–10 → information not found [correct: Indiana University] → "cannot answer based solely on provided documents" → abstains	0.196
R5	Analogical reasoning	Frames as two-step retrieval pattern: "Entity X associated with Entity Y, find attribute of Y" → identifies Doc 10 → searches all documents → information not found [correct: Indiana University] → abstains	0.198

Table 15: Reasoning traces for the HotpotQA incorrect example. All five responses correctly identify **Straight No Chaser** as the group (hop 1) but fail hop 2 due to missing context. Each response reaches the same abstention via a stylistically different but substantively identical path. Ground truth: Indiana University; unanimous model answer: *cannot be determined*.

	R1	R2	R3	R4	R5
Trial 0	0.202	0.208	0.196	0.196	0.196
Trial 1	0.201	0.207	0.201	0.195	0.195
Trial 2	0.201	0.207	0.195	0.195	0.201
Trial 3	0.202	0.208	0.196	0.196	0.196
Trial 4	0.201	0.207	0.195	0.195	0.201
Mean	0.202	0.207	0.197	0.196	0.198
CV	0.003	0.003	0.011	0.003	0.013

Table 16: PageRank distributions across five responses for HotpotQA incorrect trials. Near-uniform distributions (max deviation from $1/N = 0.200$ of just 0.008) confirm full rank collapse.

Table 17 shows the judge finds essentially nothing to discriminate. Across $\sim$35 judgments, **one reversal** occurs (R3 vs R4, iteration 4, conf=52). Confidence: 80% at 52, R4 vs R5 at literal 50 (coin-flip) all appearances. Only outlier: R2 vs R5 at 62 citing R2's "explicit document-by-document listing"—the sole substantive distinction.

This near-total indifference maps onto uncertainty components. The collapse mechanism: factual retrieval over a fixed document set is deterministic. The model scans keywords, identifies documents, locates answer or does not. Different prompts cannot induce different retrieval strategies—reasoning chains are determined by document structure, not prompt framing. Consequently, pairwise judgments find nothing to discriminate, PageRank converges to near-uniformity. **Across-trial** ($\text{StructU}_{\text{across}} < 0.001$) is near zero because all trials agree on uniformity—genuinely nothing to rank. **Within-trial** ($\text{StructU}_{\text{within}} \approx \log 5 = 1.609$) reaches theoretical maximum, reflecting flat PageRank where no response dominates.

Contrast with Math-Synth incorrect. Both have Self-ConsU $= 0$ (unanimous wrong answer), yet structural profiles diverge sharply. Math-Synth involves *endogenous* error (negation miscounting) where different strategies produce detectably different error paths, yielding preference instability and

Pair	Winner	Conf.	Pattern
R2 vs R1	R2: all trials	52	Stable; doc-by-doc listing preferred
R4 vs R5	Tie: all trials	50	Literal coin-flip; "essentially a tie"
R3 vs R4	R4: 2,5 / R3: 4	52–55	**Only reversal**

Table 17: Judge preferences (HotpotQA incorrect). One reversal across $\sim$35 comparisons. Confidence: 50=no preference.

	R1	R2	R3	R4	R5
Trial 0	0.203	0.207	0.198	0.198	0.194
Trial 1	0.202	0.206	0.199	0.197	0.196
Trial 2	0.204	0.206	0.197	0.198	0.195
Trial 3	0.203	0.207	0.198	0.197	0.195
Trial 4	0.202	0.206	0.198	0.198	0.196
Mean	0.203	0.206	0.198	0.198	0.195
CV	0.004	0.002	0.003	0.003	0.004

Table 18: Per-trial PageRank distributions for the HotpotQA correct example. Distributions are near-uniform and frozen across trials, yielding $\text{StructU}_{\text{across}} < 0.001$—statistically indistinguishable from the incorrect example (Table 16).

$\text{StructU}_{\text{across}} = 0.035$. HotpotQA involves *exogenous* failure (missing context) where no reasoning diversity can compensate for absent evidence, producing substantively identical responses and $\text{StructU}_{\text{across}} < 0.001$. StructU distinguishes these unanimous failure regimes—one flagged unreliable (0.035), the other low-uncertainty (¡0.001)—but cannot detect failures leaving no trace in preference structure.

E.4.2 CORRECT EXAMPLE (SELF-CONSU $= 0$, $\text{STRUCTU}_{\text{ACROSS}} < 0.001$)

Task. A HotpotQA question: *What creature of American folklore gained notoriety in 1964?* Retrieved context discusses several folklore creatures (Teakettler, Hidebehind, Chessie) but none mention 1964. All responses correctly identify this gap and abstain.

Observation. Every response executes identical retrieval: scan all documents for "1964" and folklore creatures $\rightarrow$ identify Documents 1, 5, 8 as partially relevant (creatures but no 1964) $\rightarrow$ note closest match is Chessie (1977/1980s sightings) $\rightarrow$ conclude information absent $\rightarrow$ abstain. Variation is surface-level only: R1 adds self-check; R2 lists documents; R3 uses Socratic framing; R4 decomposes; R5 casts as date-retrieval pattern.

The uncertainty profile is statistically indistinguishable from the incorrect example: $\text{StructU}_{\text{across}} < 0.001$, $\text{StructU}_{\text{within}} = 1.608$, $\pi_{\max}/\pi_{\min} = 1.08$, all CVs¡0.015. Table 18 shows near-uniform PageRank frozen across trials—virtually identical to the incorrect example (Table 16).

Significance. Identical collapse on correct and incorrect examples demonstrates this is a *task structure* property, not error status. Retrieval over fixed documents produces prompt-invariant chains regardless of outcome. Different prompts cannot induce different retrieval strategies—chains are determined by document structure, not prompt framing.

This represents a **boundary condition** for structural uncertainty. On Math-Synth, StructU successfully separated correct from incorrect unanimous agreement (0.001 vs 0.035, $30\times$ difference) because different prompts induced genuinely different reasoning strategies the judge could differentially rank. On HotpotQA retrieval, StructU produces indistinguishable values (¡0.001 both cases) because deterministic retrieval suppresses the reasoning diversity self-preference requires. The preference graph collapses in both cases, rendering StructU structurally uninformative—not because the method is flawed, but because the task affords no structural variation to exploit. This limitation is shared

with Self-ConsU, which also reports zero in both cases, highlighting that uncertainty quantification methods relying on response diversity are fundamentally constrained when reasoning is deterministic given input context.

E.5 Summary and Implications

The qualitative evidence supports three conclusions:

(1) Structural uncertainty detects errors invisible to self-consistency. On Math-Synth, the incorrect example exhibits $30\times$ higher across-trial uncertainty than the correct example, despite both having Self-ConsU $= 0$. The mechanism is that diverse prompt templates elicit structurally distinct reasoning strategies on mathematical tasks, and the model's inability to stably rank these strategies when all are flawed produces the across-trial uncertainty signal.

(2) Preference graph collapse explains the HotpotQA limitation. On factual retrieval, different prompt templates cannot elicit different reasoning paths because the retrieval process is determined by the document set. The resulting identical reasoning chains produce near-uniform, stable PageRank distributions ($\text{StructU}_{\text{across}} \approx 0$), eliminating the structural signal regardless of correctness.

(3) The collapse signature is itself diagnostic. Near-zero across-trial uncertainty combined with near-maximum within-trial uncertainty ($\text{StructU}_{\text{across}} \approx 0$, $\text{StructU}_{\text{within}} \approx \log N$) constitutes a detectable signature indicating that the model lacks a coherent internal quality criterion for the task. This signature can inform practitioners about when to rely on structural versus dispersion-based uncertainty methods: when it is detected, self-preference signals are uninformative and alternative estimators should be preferred.

E.6 Prompt Templates

E.6.1 Response Generation Prompts

We employ five distinct prompt templates to induce diverse reasoning patterns across candidate responses. This diversity is essential for meaningful pairwise comparisons, as it ensures that differences in solution quality reflect substantive reasoning variations rather than superficial stylistic differences.

Prompt 1: Step-by-step with self-check

```
Solve this math problem step by step, then double-check your work
    for any errors.

Q: \{question\}

<initial_solution>
Let me work through this step by step:
[Show complete solution]
</initial_solution>

<self_check>
Now let me review my work for any mistakes:
[Check each step and correct if needed]
</self_check>

<answer>
ONLY include the numerical final answer here WITHOUT units. Do not
    include any explanation or working in this section, just the
    number/value.
</answer>

Please strictly follow above format when presenting the answers.
```

Prompt 2: Think-aloud decision process

```
Solve this problem while thinking out loud about your decision-
    making process at each step.

Q: \{question\}

<thinking_process>
I'm reading the problem and thinking... [explain thought process]
Now I need to decide what approach to take... [explain reasoning]
Let me calculate this step... [show work with internal thoughts]
</thinking_process>

<answer>
ONLY include the numerical final answer here WITHOUT units. Do not
    include any explanation or working in this section, just the
    number/value.
</answer>

Please strictly follow above format when presenting the answers.
```

Prompt 3: Socratic guiding questions

```
Solve this problem by asking yourself guiding questions at each
    step.

Q: \{question\}

<socratic_dialogue>
What is the problem asking? [Answer]
What information do I have? [List knowns]
What do I need to find? [Identify unknowns]
What method should I use? [Choose approach]
How do I execute this method? [Show work]
Does my answer make sense? [Verify]
</socratic_dialogue>

<answer>
ONLY include the numerical final answer here WITHOUT units. Do not
    include any explanation or working in this section, just the
    number/value.
</answer>

Please strictly follow above format when presenting the answers.
```

Prompt 4: Decomposition into sub-problems

```
Break this complex problem into smaller, manageable sub-problems
    and solve each one.

Q: \{question\}

<decomposition>
Sub-problem 1: [Identify and solve]
Sub-problem 2: [Identify and solve]
Sub-problem 3: [Identify and solve]
Combining results: [Show how sub-solutions combine]
</decomposition>
```

```
<answer>
ONLY include the numerical final answer here WITHOUT units. Do not
    include any explanation or working in this section, just the
    number/value.
</answer>

Please strictly follow above format when presenting the answers.
```

Prompt 5: Analogical reasoning

```
Solve this problem by relating it to a simpler, similar problem you
     know how to solve.

Q: \{question\}

<analogical_reasoning>
This problem is similar to: [Describe simpler analogy]
In the simpler case, I would: [Show simple solution]
Applying the same logic here: [Adapt to current problem]
</analogical_reasoning>

<answer>
ONLY include the numerical final answer here WITHOUT units. Do not
    include any explanation or working in this section, just the
    number/value.
</answer>

Note: For the answer section, provide ONLY the numerical result
    without any units, explanations, or additional text.

Please strictly follow above format when presenting the answers.
```

E.6.2 Pairwise Comparison Judge Prompt

The following prompt elicits pairwise preference judgments with confidence scores. The structured output format (<reasoning>, ,) enables consistent parsing across all trials and models.

Pairwise Judge Prompt: Choose the better response with confidence

```
I'm going to show you a question and two different responses to it.
Your task is to decide which response better answers the question
    AND provide a confidence score.

Question: {question}

Response 1:
{response1}

Response 2:
{response2}

Which response better answers the question? Please evaluate them
    FAIRLY and CONSISTENTLY based on objective criteria, avoiding
    any subjective bias.

For your evaluation:
```

```
- Focus on accuracy, clarity, completeness, and relevance to the
    question
- Consider evidence and reasoning quality in each response
- Maintain consistent evaluation standards across all comparisons
- Avoid being influenced by response length, style preferences, or
    personal opinions
- Evaluate the responses as if they were written by the same person
- If the question is ambiguous or unclear, the best response is one
     that acknowledges this ambiguity and explains different
    possible interpretations rather than making assumptions

Please structure your response as follows:
Think step by step about the clarity, accuracy, helpfulness, and
    overall quality of each response.
Compare them thoroughly, analyzing their strengths and weaknesses.

Response 1
OR
Response 2


Do NOT include any reasoning or explanation in the answer section.
    Only state "Response 1" or "Response 2".


Provide a confidence score (0-100) indicating how confident you are
     in your judgment.
- 100 means absolute certainty that your chosen response is better
- 50 means both responses are equally good/bad
- 0-49 means minimal preference for your chosen response
- 51-100 means stronger preference for your chosen response
```

E.6.3 Verbalized Uncertainty Baseline Prompt

For the verbalized uncertainty baseline , we directly elicit the model's self-assessed confidence. This prompt produces a single response with an explicit confidence score, which we compare against structural and self-consistency baselines.

Verifier Prompt: JSON-only solution check

```
CRITICAL INSTRUCTION: You MUST respond with ONLY a JSON object. NO
    reasoning, NO explanation, NO other text.

Verify if this solution is correct:

PROBLEM: {problem}

SOLUTION: {solution}

Respond with ONLY this JSON (nothing else):
{"verdict": "PASS", "confidence_correct": 0.95}

Replace PASS with FAIL if incorrect, and set confidence 0.0-1.0.
JSON ONLY. NO OTHER TEXT.
```

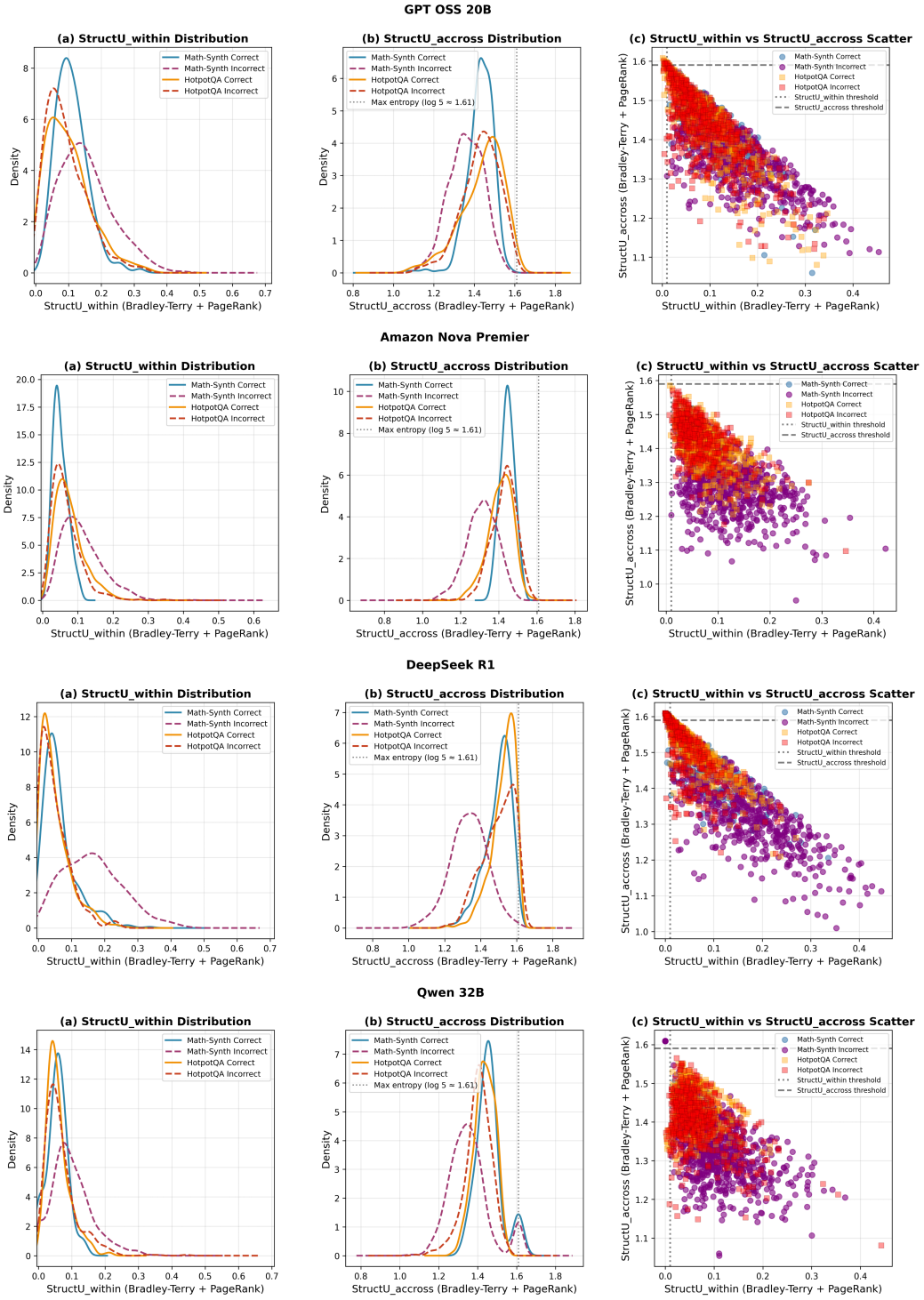


Figure 10: **Structural collapse on factual retrieval across four additional models.** Rows correspond to Amazon Nova Premier, DeepSeek R1, GPT-OSS 20B, and Qwen 3 32B. Each row shows the across-trial uncertainty distribution (left), within-trial uncertainty distribution (center), and joint across-trial–within-trial scatter (right) on Math-Synth and HotpotQA, conditioned on correctness (Bradley–Terry + PageRank). The "HotpotQA signature"—near-zero across-trial uncertainty with near-maximum within-trial uncertainty ($\log 5 \approx 1.61$, dotted line)—is reproduced across all models, confirming that the structural collapse is a task-level phenomenon independent of model capability.