Revisiting the Uniform Information Density Hypothesis in LLM Reasoning

Minju Gwak[1,2] **Guijin Son**[2] **Jaehyung Kim**[1]
[1]Yonsei University [2]OneLine AI
mjgwak@yonsei.ac.kr, jaehyungk@yonsei.ac.kr

Abstract

The Uniform Information Density (UID) hypothesis proposes that effective communication is achieved by maintaining a stable flow of information. In this work, we revisit this principle in the context of Large Language Model (LLM) reasoning, asking whether step-level uniformity reflects reasoning quality. To this end, we introduce a novel framework to quantify uniformity of information flow at both local and global levels, using an entropy-based stepwise density metric. Across experiments on seven reasoning benchmarks, we see a counter-intuitive pattern: while high-quality reasoning exhibit smooth step-by-step transitions (*local uniformity*) and structured, non-uniform information flow at the trajectory level (*global non-uniformity*). The results demonstrate that these uniformities outperform alternative internal signals as predictors of reasoning quality, and such divergence with human communication is not a model deficiency, but a byproduct of distinct objectives between human communication and LLM reasoning. [1]

1 Introduction

Chain-of-Thought (CoT) reasoning has become a central technique for enhancing large language models (LLMs) on complex reasoning tasks (Wei et al., 2023; Kojima et al., 2023; Chae et al., 2023). By generating step-by-step rationales, CoT enables models to decompose problems into simpler subproblems and thereby improve accuracy (Golovneva et al., 2023; Prasad et al., 2023; Yao et al., 2023). Despite these successes, recent studies have highlighted the fragility of this approach (Zhao et al., 2025a). For example, the intermediate rationales are often logically inconsistent or incoherent, and hence models fail to generalize out-of domain tasks even when producing lengthy reasoning traces (Shojaee et al., 2025). This raises a critical question: *how can we determine whether LLMs are reasoning effectively, rather than merely generating superficially coherent text?*

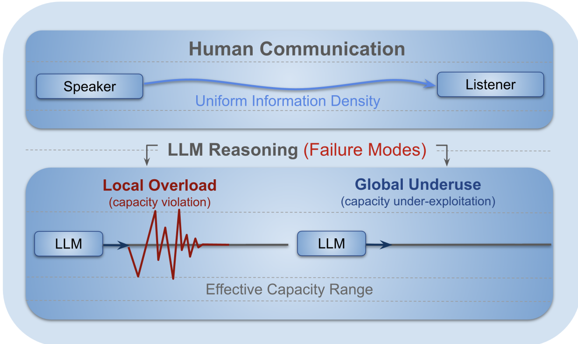


Figure 1: **Reasoning as information flow.** Human communication distributes information smoothly to respect channel capacity, enabling successful understanding. LLM reasoning transmits information across reasoning steps; failures arise from local overload (sharp spikes) and underuse (flat trivial trajectory).

Clues may lie in human communication itself; the psycholinguistic hypothesis of Uniform Information Density (UID) proposes that speakers distribute information as evenly as possible to balance clarity and efficiency (Fenk and Fenk-Oczlon, 1980; Genzel and Charniak, 2002; Clark et al., 2023; Jaeger and Levy, 2006). Namely, a relatively uniform flow of information is necessary for effective communication (Meister et al., 2021; Aylett and Turk, 2004), aligned with the limits of human cognitive processing. When this balance is disrupted by too much or too little information, communication deteriorates. Motivated by this, we ask *whether a similar principle governs reasoning in LLMs.* As human speakers maintain balanced information flow to support comprehension, effective reasoning traces may require comparable uniformity across steps. Recent findings in cognitive science support this view: Bhambri et al. (2025) shows that reasoning paths interpretable to humans

[1]Code is released at: https://github.com/talzoomanzoo/uid-reasoning

Findings of the Association for Computational Linguistics: ACL 2026, pages 31304–31333

are also easier for models to generate and learn, suggesting a shared structure between human cognition and machine reasoning.

To investigate this, we focus on analyzing the information flow of LLM-generated reasoning traces on challenging mathematical benchmarks. Specifically, we begin by defining per-step measurements of information density using entropy of predictive distribution, and examine their relationship to answer correctness. We then introduce two complementary metrics to quantify uniformity at both global and local levels. Our experiments reveal a counter-intuitive pattern: unlike human communication, successful LLM reasoning exhibits high local uniformity but low global uniformity. In our experiments across seven challenging reasoning benchmarks and three LLMs, these uniformity consistently outperform conventional approaches in identifying high-quality trajectories for Best-of-N sampling. We find that this divergence is not a model deficiency, but rather an instrumental byproduct of the distinct objectives between human communication and LLM reasoning.

Overall, our contributions are threefold:

- To our knowledge, we are the first to revisit the Uniform Information Density (UID) hypothesis in the context of LLM reasoning.
- Contrary to our hypothesis, we find that reasoning patterns characterized by global non-uniformity and local uniformity in surprisal correlate with reasoning success on challenging mathematical reasoning tasks.
- Extensive analyses show that deviations from such patterns serve as a trace-level internal signal for predicting failure cases, enabling complementary improvements to response-level aggregation and LLM reasoning evaluation.

2 Exploring the Uniform Information Density Hypothesis in LLM Reasoning

2.1 Background: the UID hypothesis

The UID hypothesis considers language as a signal transmitted through a noisy channel with limited capacity (Meister et al., 2021; Tsipidi et al., 2024). Then, UID posits that speakers aim to convey information efficiently without overwhelming the listener's processing resources. Formally, let an utterance $\mathbf{u} = [u_1, u_2, \ldots, u_N]$ be a sequence of N linguistic units, such as words, subwords, or characters, depending on the granularity of representation. For each unit u_n, *surprisal* is defined as its unexpectedness, given its previous context:

$$s(u_n) = -\log P(u_n \mid \mathbf{u}_{<n}),$$

where $P(u_n|\mathbf{u}_{<n})$ is the probability of seeing unit utterance u_n after the earlier sequence $\mathbf{u}_{<n} = [u_1, \ldots, u_{n-1}]$. Intuitively, the unit with high surprisal is very unexpected and therefore hard to process for the listener. Then, the total processing effort for the utterance $\mathbf{u}$ is expressed as:

$$E_{\texttt{process}}(\mathbf{u}) \propto \sum_{n=1}^{N} s(u_n)^k + c \cdot N.$$

for some constant $c > 0$ and $k > 1$. Here, the exponent k encodes the super-linear nature of processing effort, where rare or unexpected units impose larger effort than predictable ones. Under this formulation, uniform surprisals across units minimize total processing effort, whereas "spiky" linguistic signal with highly uneven surprisal values increase the burden of communication (Meister et al., 2021).

While UID has been validated in human language, its implications for machine reasoning remain underexplored. LLMs, or more specifically, recent reasoning models such as Deepseek-R1 (DeepSeek-AI et al., 2025) and Qwen3 (Yang et al., 2025a) generate step-by-step CoT traces, similar to how human speech unfold over time. If we treat each reasoning step z_i like a unit with surprisal $s(z_i)$, a single reasoning trace $\mathbf{z} = [z_1, z_2, \ldots, z_N]$ can be analyzed in the same way to have the total reasoning effort as below:

$$E_{\texttt{reason}}(\mathbf{z}) \propto \sum_{n=1}^{N} s(z_n)^k + c \cdot N.$$

Then, a natural question arises: *does the UID hypothesis hold for good reasoning patterns in LLMs?* A smooth, uniform surprisal profile may reflect clear and logical reasoning, while sharp spikes may signal confusion or errors. Therefore, in this work, we validate UID hypothesis beyond psycholinguistics to CoT reasoning of LLMs, offering a new lens on why reasoning models succeed or fail.

2.2 Preliminary analyses: Step-wise information density in CoTs of LLMs

We start by defining the step-level information density ID_i for a reasoning trace $\mathbf{z} = [z_1, \ldots, z_N]$, where each reasoning step z_i is composed of M_i tokens, *i.e.*, $z_i = [x_1, \ldots, x_{M_i}]$. We divide the given reasoning trace into multiple reasoning steps using

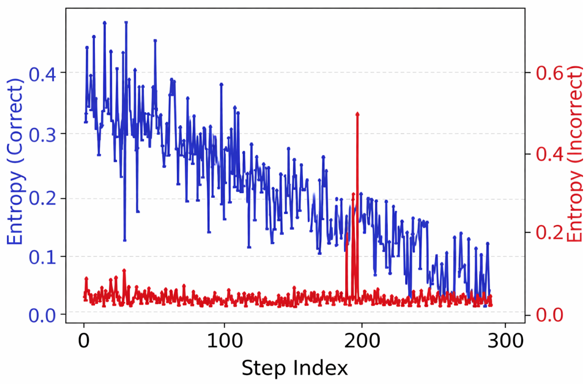


Figure 2: Averaged ID_i scores of LLM reasoning traces on AIME2025. Correct traces show **a downward trend with smooth decay**, while incorrect traces show noisy entropy with **unresolved spikes**.

\n\n, following Lightman et al. (2023).[2] Let p_t be the predictive distribution over the vocabulary $\mathcal{V}$ at the token position t. Then, to characterize ID_i, we consider entropy over tokens in each step:

$$H_t = -\sum_{v \in \mathcal{V}} p_t(v) \log p_t(v),$$

and step-level information density with entropy is:

$$ID_i = \frac{1}{M_i} \sum_{t=1}^{M_i} H_t.$$

Justifications for using entropy as a proxy. We use entropy as a proxy for information density because it reflects both model confidence and variability in reasoning; low entropy indicates confident predictions while higher entropy implies uncertainty between multiple plausible continuations (Shannon, 1948; Kuhn and Johnson, 2013). Also, in an information-theoretic perspective, entropy quantifies the expected number of bits required to encode the predictive distribution, where higher values corresponding to richer informational content (Cover and Thomas, 2006). Therefore, aggregating entropy across tokens offers a compact and interpretable signal of reasoning difficulty. Furthermore, our experiments (see Appendix C.1) suggest that using entropy as information density is more effective, compared to other candidates such as log-probability and confidence-based methods.

[2]While we adopt newline-based segmentation, we demonstrate that our findings are robust to alternative stepwise segmentation strategies in Appendix A.

Figure 2 compares the evolution of ID_i between the averaged reasoning traces for correct and incorrect solutions in AIME2025, respectively. Here, correct traces exhibit a clear global trend: entropy begins with exploratory fluctuations, stabilizes in mid-trace, and then steadily decays toward near zero, reflecting a structured process of resolution and convergence on the final answer. In contrast, incorrect traces instead displays a flat, noisy entropy trajectory with occasional sharp spikes.

Motivation for a structural perspective. As we examine entropy peaks at the level of individual reasoning traces, we find that interpreting them solely through the semantic content of the text is difficult (see Appendix C.2). To be specific, (1) entropy levels in both correct and incorrect traces can be quite varying, (2) number of transition words *(i.e., But, Alternatively, Wait)* may appear more at correct traces contrary to our intuition, and (3) wrong reasoning traces may be concise and have fewer number of steps. More importantly, such an approach does not provide a consistent basis for interpretation across various benchmarks, highlighting the need for a unified structural perspective. Building on such motivation, we introduce a framework for measuring the uniformity of information density in reasoning traces.

2.3 Measuring global and local uniformity of information density in CoTs of LLMs

To quantify the uniformity of information density in a reasoning trace, we first distinguish between two complementary notions of uniformity that have been discussed in prior psycholinguistic work (Meister et al., 2021; Collins, 2014). **Global uniformity** characterizes whether information is distributed evenly across the entire trace, corresponding to a relatively stable surprisal level over long horizons. In contrast, **local uniformity** captures whether information changes smoothly between adjacent steps, reflecting gradual and coherent transitions rather than abrupt jumps. These two notions capture uniformity in different ways and therefore diverge substantially in reasoning traces: a trace may appear globally uniform yet contain sharp local disruptions, or conversely, exhibit smooth local transitions while concentrating information unevenly across steps.

Motivated by this distinction, we introduce two complementary UID-based metrics that separately operationalize global and local uniformity in LLM reasoning traces: (1) *global variance* and (2) *local step-to-step spikes and falls*.

Global uniformity via variance. Global uniformity is measured by the variance of step-level information density across the entire reasoning trace, capturing *whether the information is evenly distributed or concentrated in a small number of steps.* Formally, for a reasoning trace $\mathbf{z} = [z_1, \ldots, z_N]$, let us define the non-negative information density vector $\mathbf{u} = [ID_1, \ldots, ID_N]$. After min–max normalization, where $ID'_i = \frac{ID_i - m}{M - m}$ with $m = \min_{1 \leq i \leq N} ID_i$ and $M = \max_{1 \leq i \leq N} ID_i$, we obtain the normalized vector $\tilde{\mathbf{u}} = [ID'_1, \ldots, ID'_N]$. The variance of the normalized information density values is then defined as:

$$\mathrm{Var}(\tilde{\mathbf{u}}) = \frac{1}{N} \sum_{i=1}^{N} \left(ID'_i - \mu \right)^2, \tag{1}$$

where $\mu = \frac{1}{N} \sum_{i=1}^{N} ID'_i$. High variance indicates global non-uniformity, where information is unevenly concentrated across steps, while lower variance corresponds to globally uniform traces.

Local uniformity via step-to-step smoothness. Local uniformity captures how smoothly information density evolves between adjacent reasoning steps, measuring whether uncertainty is resolved gradually or through abrupt transitions. Given $\tilde{\mathbf{u}}$, we define the step-to-step change as $\Delta_i = ID'_i - ID'_{i-1}$ for $i = 2, \ldots, N$, and compute the mean and standard deviation of the change sequence as $\mu_\Delta = \frac{1}{N-1} \sum_{i=2}^{N} \Delta_i$ and $\sigma_\Delta = \sqrt{\frac{1}{N-1} \sum_{i=2}^{N} (\Delta_i - \mu_\Delta)^2}$. To identify significant local disruptions, we define thresholds $T^+ = \mu_\Delta + \tau\sigma_\Delta$ and $T^- = \mu_\Delta - \tau\sigma_\Delta$, where $\tau \in \{2, 3\}$. Then, an *upward spike* is identified when $\Delta_i > T^+$, and a *downward fall* when $\Delta_i < T^-$. Finally, total local irregularity count is:

$$\begin{aligned} S_{\text{up}}(\tilde{\mathbf{u}}) &= \sum_{i=2}^{N} \mathbb{I}[\Delta_i > T^+], \\ S_{\text{down}}(\tilde{\mathbf{u}}) &= \sum_{i=2}^{N} \mathbb{I}[\Delta_i < T^-], \\ S_{\texttt{local}}(\tilde{\mathbf{u}}) &= S_{\text{up}}(\tilde{\mathbf{u}}) + S_{\text{down}}(\tilde{\mathbf{u}}). \end{aligned} \tag{2}$$

Intuitively, smaller $S_{\texttt{local}}$ indicates higher local uniformity, therefore reflecting smoother stepwise information flow.

These two metrics, $\mathrm{Var}(\tilde{\mathbf{u}})$ and $S_{\texttt{local}}(\tilde{\mathbf{u}})$, provide complementary trace-level characterization of how information is distributed and evolved during reasoning. As reflected in our empirical observations (Figure 2), correct reasoning traces tend to exhibit *low local disruption* (*i.e.*, high local uniformity) alongside *structured global non-uniformity* (*i.e.*, low global uniformity), whereas incorrect traces often exhibit large, abrupt fluctuations in local uniformity or consistently high and flat global uniformity that fails to convey informative structure. This complementary view enables a more precise diagnosis of reasoning quality beyond token-wise confidence or entropy measures.

3 Does the UID Hypothesis Hold in LLM Reasoning Traces?

Setups. For the experiments, we use three widely adopted open-source reasoning models, Deepseek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025), Deepseek-R1-Distill-Llama-8B (DeepSeek-AI et al., 2025), and Qwen3-8B (Yang et al., 2025a). Also, we use four challenging mathematical benchmarks: AIME2025, BRUMO2025, HMMT2025, and MinervaMath (MM). To demonstrate the effectiveness as the selection criteria of Best-of-N, we sample five reasoning traces for each question, with temperature as 0.6, top-p as 0.95, and top-k as 20. Then, we calculate their corresponding UID scores, and select the traces with the highest and lowest scores. We consider the following three baselines that leverage internal signals to find a good logical trail: (1) Self-Certainty (Kang et al., 2025), (2) High confidence, and (3) Low entropy. Although we evaluate both τ = 2, 3, we report only the results of τ = 3 results in the main text. Details are presented in Appendix E.

Main results. Table 1 shows that UID-guided selection outperforms non-UID baselines across all benchmarks, which indicates that, *unlike human communication postulated by the original UID hypothesis, reasoning success in LLMs is best predicted by a combination of local uniformity and global non-uniformity in information distribution.*

For Deepseek-R1-Distill-Qwen-7B under *local uniformity*, the accuracies are improved by +33% on AIME2025, +4% on BRUMO2025, +25% on HMMT2025, and +3% on MinervaMath. These gains are consistent, indicating that smoothing stepwise information flow reliably benefits reasoning regardless of task domain.

On the other hand, *global non-uniformity* also strengthens performance, particularly on harder

Table 1: **Main results.** Performance on four math benchmarks (AIME2025, BRUMO2025, HMMT2025, and MinervaMath (MM)). Results are averaged over seeds 42, 1234, and 2025 with standard deviations. The best and second-best scores are highlighted in **bold** and underline, respectively. Full results are in Appendix F.

	`DS-R1-Distill-Qwen-7B`				`DS-R1-Distill-Llama-8B`				`Qwen3-8B`			
Methods (↓)	AIME25	BRUMO25	HMMT25	MM	AIME25	BRUMO25	HMMT25	MM	AIME25	BRUMO25	HMMT25	MM
Mean Acc.	0.40 ±0.02	0.54 ±0.01	0.24 ±0.00	0.30 ±0.00	0.33 ±0.01	<u>0.40</u> ±0.02	0.20 ±0.01	0.23 ±0.00	0.67 ±0.01	0.68 ±0.02	0.43 ±0.01	<u>0.34</u> ±0.01
Self-Cert.	0.48 ±0.04	0.52 ±0.04	<u>0.28</u> ±0.02	0.30 ±0.00	0.33 ±0.00	0.34 ±0.02	0.19 ±0.02	0.22 ±0.01	0.63 ±0.00	**0.71** ±0.02	**0.50** ±0.03	<u>0.34</u> ±0.01
High Conf.	0.48 ±0.04	0.52 ±0.05	0.27 ±0.00	0.30 ±0.00	0.36 ±0.02	0.36 ±0.03	0.21 ±0.02	**0.23** ±0.01	0.60 ±0.00	0.63 ±0.03	0.42 ±0.06	0.33 ±0.01
Low Ent.	0.48 ±0.04	<u>0.56</u> ±0.05	0.24 ±0.02	0.30 ±0.00	0.34 ±0.04	0.37 ±0.06	0.19 ±0.02	0.22 ±0.01	0.60 ±0.00	0.64 ±0.04	0.43 ±0.05	0.33 ±0.01
UID Metrics (↓)												
Loc. non-uni	0.27 ±0.00	0.39 ±0.07	0.18 ±0.04	0.26 ±0.01	0.36 ±0.04	0.36 ±0.03	0.20 ±0.03	**0.23** ±0.01	0.63 ±0.00	0.63 ±0.07	0.40 ±0.03	0.34 ±0.01
Loc. uni	**0.53** ±0.06	0.56 ±0.02	**0.30** ±0.00	**0.31** ±0.01	**0.39** ±0.04	**0.44** ±0.06	<u>0.24</u> ±0.04	**0.23** ±0.01	<u>0.69</u> ±0.03	<u>0.70</u> ±0.03	<u>0.48</u> ±0.04	<u>0.34</u> ±0.01
Glob. non-uni	<u>0.52</u> ±0.08	**0.64** ±0.05	0.26 ±0.04	<u>0.30</u> ±0.01	<u>0.37</u> ±0.03	0.36 ±0.04	0.18 ±0.04	0.21 ±0.00	**0.70** ±0.00	0.61 ±0.06	0.47 ±0.03	0.33 ±0.01
Glob. uni	0.33 ±0.00	0.44 ±0.02	0.16 ±0.04	0.28 ±0.00	0.27 ±0.12	<u>0.40</u> ±0.06	**0.27** ±0.03	0.23 ±0.01	0.66 ±0.02	0.68 ±0.07	0.41 ±0.02	**0.34** ±0.02

benchmarks. Compared to Self-Certainty selection, global non-uniformity yields an additional +23% improvement on BRUMO2025 and +8% on AIME2025, while maintaining comparable performance on MinervaMath and only marginal degradation on HMMT2025. Overall, this setting achieves the strongest average performance, with relative gains reaching up to 32% over the vanilla accuracy baseline. Together, these results suggest that *local smoothness and global heterogeneity provides the most stable and informative criterion for selecting high-quality reasoning traces.*

Similar trends are observed for Deepseek-R1-Distill-Llama-8B. Under local uniformity, accuracy improves by approximately 15–20% on AIME2025 and BRUMO2025 relative to the overall-accuracy baseline, while yielding more modest gains 5–10% on HMMT2025 and MinervaMath. Enabling global non-uniformity further enhances performance on AIME2025 and HMMT2025, delivering an additional 10–15% improvement over Self-Certainty selection, while remaining competitive on BRUMO2025 and MinervaMath. These results indicate that UID-guided selection remains effective even when the backbone model exhibits weaker absolute reasoning performance, though the gains are naturally bounded by model capacity.

Qwen3-8B follows the same qualitative trend, though with reduced margins. Because its baseline accuracy is already strong, UID-guided selection yields more modest gains (approximately 2–5% on AIME2025 and BRUMO2025), while still consistently outperforming entropy- and confidence-based heuristics. This confirms that UID metrics generalize across models but naturally offer larger benefits when baseline reasoning quality is weaker.

Table 2: **Model size analysis.** Performance of `Qwen3` models (1.7B, 4B, 8B) on AIME2025. Results are averaged over seeds 42, 1234, and 2025. The best and second-best scores are highlighted in **bold** and underline, respectively. Full results are in Appendix F.

Methods (↓)	`Qwen3-1.7B`	`Qwen3-4B`	`Qwen3-8B`
Mean Acc.	0.35	0.65	0.67
Self-Cert.	**0.45**	**0.73**	0.63
High Conf.	0.37	0.62	0.60
Low Ent.	0.37	0.63	0.60
UID Metrics (↓)			
Loc. non-uni	0.24	0.54	0.63
Loc. uni	<u>0.41</u>	<u>0.69</u>	<u>0.69</u>
Glob. non-uni	0.37	0.66	**0.70**
Glob. uni	0.33	0.67	0.66

Takeaway for UID-Guided Selection

Reasoning traces exhibiting *local uniformity* and *global non-uniformity* achieve the highest accuracy, consistently outperforming confidence- and entropy-based selection across benchmarks.

4 Further Analyses

4.1 Trends across different model sizes

Table 2 presents the results with `Qwen3` models of different sizes (1.7B, 4B, 8B) and reveals clear scaling trends. As expected, larger models achieve stronger baselines: mean accuracy increases by approximately 86% from 1.7B to 4B, with a more modest 3% relative gain from 4B to 8B. Self-Certainty shows a similar pattern, rising by 62% from 1.7B to 4B , followed by a 14% relative decrease at 8B, suggesting diminishing returns and

Table 3: **Reasoning performance outside the math domain.** The best and second-best scores are highlighted in **bold** and underline, respectively.

	`DS-R1-Distill-Llama-8B`			`DS-R1-Distill-Qwen-7B`		
Methods (↓)	GPQA-D	LSAT-AR	LSAT-LR	GPQA-D	LSAT-AR	LSAT-LR
Mean Acc.	0.48	0.43	0.52	0.48	0.55	0.50
Self-Cert.	0.51	0.44	0.52	**0.51**	0.51	0.49
High Conf.	**0.53**	0.43	0.51	0.46	0.51	0.51
Low Ent.	0.49	0.42	0.52	0.48	0.51	0.52
UID Metrics (↓)						
Loc. non-uni	0.44	0.39	0.49	0.43	0.48	0.46
Loc. uni	0.52	0.46	0.53	**0.51**	**0.62**	0.49
Glob. non-uni	0.47	**0.49**	**0.54**	0.50	**0.62**	**0.53**
Glob. uni	0.44	0.40	0.50	0.46	0.46	0.46

possible overconfidence effects at larger scales.

Despite reduced headroom from stronger baselines, UID-based methods continue to provide meaningful relative improvements. Local uniformity yields consistent gains across all scales, improving performance by 17% for 1.7B, 6% for 4B, and 3% for 8B, indicating that enforcing local smoothness stabilizes reasoning regardless of model size. In contrast, global non-uniformity exhibits a pronounced scaling effect: while the 1.7B model improves by only 6%, the 4B and 8B models gain 2% and 4%, respectively, with the 8B model achieving the strongest overall performance in the table. *These results suggest a size-dependent effect: smaller models benefit more from local smoothing, whereas larger models increasingly exploit global non-uniformity in surprisal distribution.*

Takeaway 4.1 for Scaling with Model Size

UID-guided reasoning selection scales with model capacity.

4.2 Generalization beyond math reasoning

To evaluate generalization beyond mathematical reasoning, we examine reasoning performance on non-math benchmarks (GPQA-Diamond (GPQA-D), LSAT-AR, and LSAT-LR).

Table 4 reports results under different sampling scales. We observe a consistent tendency for both DS-R1-Distill-Llama-8B and DS-R1-Distill-Qwen-7B; UID-based strategies consistently improve over the mean-accuracy baseline. For DS-R1-Distill-Llama-8B on GPQA-D, locally uniform UID increases accuracy from 0.48 to 0.52, corresponding to an absolute gain of +4%p compared to mean accuracy. Similar trends hold on LSAT tasks: on LSAT-AR, locally uniform UID improves performance by about +7% relative compared to mean accuracy, while on LSAT-LR, it achieves the best overall score with a +1%p absolute gain.

Table 4: **Sample size analysis**. Sample by 3, 5, 10 on AIME using `Qwen3-8B`. Results are averaged over seeds 42, 1234, and 2025 with standard deviations. The best and second-best scores are highlighted in **bold** and underline, respectively. Full results are in Appendix F.

Methods (↓)	Sample by 3	Sample by 5	Sample by 10
Mean Acc.	0.67	0.67	0.68
Self-Cert.	0.70	0.63	0.62
High Conf.	0.63	0.60	0.57
Low Ent.	0.63	0.60	0.56
UID Metrics (↓)			
Loc. non-uni	0.63	0.63	0.53
Loc. uni	**0.73**	0.69	0.72
Glob. non-uni	0.70	**0.70**	**0.70**
Glob. uni	0.70	0.66	0.63

For DS-R1-Distill-Qwen-7B, the effect is even more pronounced. On LSAT-AR, both locally uniform and globally non-uniform UID reach 0.62 accuracy, improving over the mean baseline (55%) by +7%p, or roughly +12.7% relative improvement. On GPQA-D, locally uniform UID matches the strongest non-UID heuristic (self-certainty), while maintaining a +3–4% relative gain over mean accuracy. On LSAT-LR, globally non-uniform UID achieves the best score, yielding a +6% relative improvement compared to the baseline.

Such results show that UID metrics capture meaningful structural signals even in non-mathematical reasoning. Rather than merely matching strong confidence-based heuristics, UID-based selection often yields *consistent percentage-level gains* across diverse reasoning benchmarks.

Takeaway 4.2 for Domain Generalization

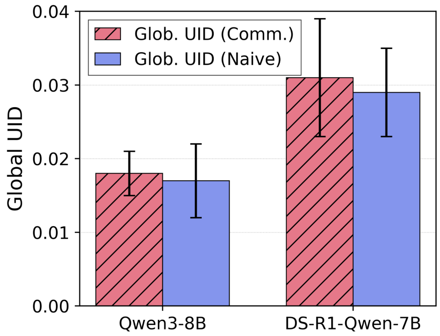
UID-guided selection yields consistent relative accuracy gains (5–13%) across non-math reasoning tasks, demonstrating robust generalization beyond math domains.

4.3 Trends across different sample sizes

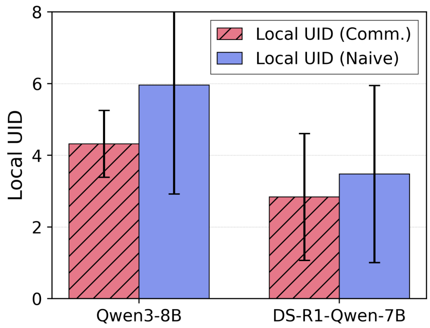
Table 3 reports results under different sampling scales. Across all regimes, we observe consistently higher-performing reasoning traces exhibit *local uniformity* while remaining *globally non-uniform*.

This pattern already emerges under the smallest setting, Sample by 3. Locally uniform traces achieve 0.73 accuracy, outperforming locally non-uniform variants by a clear margin. At the global level, non-uniform traces reach 0.70, while globally uniform traces do not provide additional gains.

When sampled by 5, locally uniform traces



(a) Global UID of comm vs. naive prompting



(b) Local UID of comm vs. naive prompting

Figure 3: Communication prompts induce higher global variance while yielding smoother local transitions, indicating more explanation-oriented pacing.

achieve 0.69 accuracy, again outperforming locally non-uniform variants. At the global level, non-uniform traces reach 0.70, exceeding globally uniform ones by approximately 6% relative.

This trend becomes more pronounced as we scale up to Sample by 10. Locally uniform traces further improve to 0.72 accuracy, while globally uniform traces degrade to 0.63, corresponding to a $\sim$12% relative decrease. Across all sampling settings, locally uniform traces consistently outperform locally non-uniform ones by 3–6 percentage points, while global non-uniformity remains a key factor for maintaining high accuracy.

Takeaway 4.3 for Scaling with Sample Size

Effective reasoning benefits from local uniformity but global non-uniformity: stable step-level information flow with trajectory-level variation yields the highest accuracy, even under small sampling budgets.

5 Why Does UID Differ between Human Communication and LLM Reasoning?

Our results show that high-quality LLM reasoning traces typically exhibit smooth local transitions in information density together with structured global non-uniformity. We conjecture that such deviation from the original UID hypothesis in psycholinguistics is due to different objectives between human communication and LLM reasoning rather than a deficiency in model behavior.

- **Human Communication: Normative, Listener-Optimized UID.** In human language, UID is a *normative* principle of communication. Speakers assume the presence of a listener and aim to distribute information evenly over time to reduce processing difficulty under cognitive constraints. Here, *global uniformity is desirable*, as sharp information spikes can overload the listener.
- **LLM Reasoning: Instrumental, Computation-Driven UID.** LLM reasoning traces, in contrast, are typically generated in a *listener-free* setting. CoT reasoning reflects an internal computational process rather than an act of communication. Consequently, UID in LLMs is better understood as *instrumental*: it characterizes how information is explored and evolved during problem solving, rather than how information is packaged for an external recipient. Here, *global non-uniformity* is not problematic but expected. High-quality reasoning traces exhibit initial exploration with higher entropy, followed by consolidation and a final commitment phase where uncertainty collapses. These transitions naturally produce uneven information distribution across steps, which is captured by higher global variance. In contrast, incorrect traces tend to show noisy fluctuations without coherent phase structure.

To verify the role of communicative objectives, we compare a conventional listener-free reasoning prompt with *a new listener-aware (communication) prompt* (see Appendix H) that instruct the model to explain its reasoning to a listener. Under the same model, communication prompting consistently increases global UID, by approximately 6% for Qwen3-8B and 11% for DS-R1-Qwen-7B (Figure 3, top), indicating more explanation-oriented pacing. Simultaneously, it reduces local UID, with about a 29% decrease for Qwen3-8B and an 18% decrease for DS-R1-Qwen-7B (Figure 3, bottom), reflecting smoother step-to-step transitions with fewer abrupt spikes and falls.

These findings clarify the role of UID as an evaluation signal. High-quality internal reasoning need not resemble human explanations, and penalizing global non-uniformity may harm reasoning assess-

Model	Comm.	Naive
DS-R1-Distill-Qwen-7B	0.37	**0.41**
Qwen3-8B	0.63	**0.67**

Table 5: Performance under comm. and naive settings.

ment even if it improves readability. At the same time, local smoothness remains a shared indicator of coherence. Consistent with this interpretation, Table 5 shows that adopting a listener-aware, communicative style of reasoning leads to a performance decrease, highlighting that communicative optimality and internal reasoning effectiveness optimize for fundamentally different objectives. Overall, our suggested UID-based measures are best interpreted as internal, trace-level diagnostics of reasoning dynamics, rather than as proxies for human-likeness or communicative optimality.

Takeaway for Human vs. LLM Reasoning

Human communication is listener-optimized and therefore favors global uniformity; LLM reasoning is computation-driven and naturally exhibits structured global non-uniformity from phase-based problem solving. Only local smoothness reliably indicates reasoning coherence.

6 Related Work

Fragility of CoT and the role of individual reasoning steps. CoT prompting improves reasoning but remains fragile (Wei et al., 2023; Zhao et al., 2025a; Kojima et al., 2023; Chae et al., 2023; Chen et al., 2023). Small, seemingly irrelevant perturbations in the reasoning chain can sharply reduce accuracy (Mirzadeh et al., 2025; Tang et al., 2023), suggesting that models often produce the appearance of reasoning rather than logically sound traces (Shojaee et al., 2025). Moreover, longer reasoning steps do not necessarily reflect the true difficulty of the problem, and many intermediate steps can be altered or even removed without changing the final answer (Lanham et al., 2023). This raises doubts about the necessity and faithfulness of these step-by-step explanations. Another line of research takes a different perspective: rather than viewing all steps as equally important, it suggests that a small subset of pivotal steps within CoT traces disproportionately drives predictions (Bogdan et al., 2025). Attribution methods and their frameworks identify and highlight these critical steps, emphasizing the need to understand how individual steps shape outcomes (Golovneva et al., 2023; Wu et al., 2023; Bigelow et al., 2024). Despite these advances, prior works have no clear interpretations of what constitutes as a truly good reasoning pattern.

Intrinsic signals in LLM reasoning. Research on LLM reasoning has increasingly turned to internal model signals to gain insight into how reasoning unfolds (Zhao et al., 2025b; Zhang et al., 2025). Many approaches use these signals to improve performance (Zuo et al., 2025), such as self-certainty (Kang et al., 2025; Zhao et al., 2025b), or confidence (Jang et al., 2025) to refine outputs, or using entropy-based measures to encourage diverse reasoning paths (Zhang et al., 2025; Agarwal et al., 2025; Gao et al., 2025; Lee et al., 2025; Li et al., 2025; Zhou et al., 2023). However, these methods largely treat internal signals as heuristics for guiding or controlling reasoning, without providing a principled account of why certain reasoning traces are more coherent than others. In contrast, we ground our analysis in the long-standing psycholinguistic theory of Uniform Information Density (UID) hypothesis, which offers a theoretical lens for understanding how information is introduced, transformed, and propagated through reasoning. Our perspective on information flow in reasoning traces moves beyond performance heuristics, emphasizing structural properties that characterize coherence and clarify how reasoning differs from human communication.

7 Conclusion

This paper revisits the long-standing Uniform Information Density (UID) hypothesis in the context of large language model reasoning. By shifting the focus from output-level correctness to step-level information flow, we demonstrate that entropy-uniformity serves as a meaningful indicator of reasoning quality. Our analysis reveals that coherent reasoning traces tend to distribute information more evenly across steps, while disfluent traces exhibit sharp entropy fluctuations. On the other hand, non-uniform traces at the global level with higher variance leads to a more coherent reasoning trace. These findings bridge psycholinguistic theory with computational analysis, providing a new lens for interpreting model reasoning beyond performance metrics. Ultimately, our work suggests that UID-inspired measures can guide the design of more interpretable and trustworthy reasoning systems.

Limitations

While our study highlights the importance of uniformity in identifying coherent reasoning traces, several limitations and open questions remain.

First, our analysis is primarily restricted to structured reasoning datasets, which may not fully capture model behavior in broader settings such as open-ended dialogue or interactive communication. Notably, our additional prompting analysis reveals a systematic difference between listener-free internal reasoning and listener-oriented explanation: in naive settings, models tend to produce abrupt transitions and uneven intermediate steps with relatively low global variation, whereas introducing a notional listener shifts generation toward smoother local transitions and higher global variance. These observations support our interpretation that the divergence between human and model UID arises from a structural difference between internal reasoning and communicative reasoning. While this analysis provides targeted empirical evidence for a contrast that, to our knowledge, has not been explored previously, future work is needed to evaluate whether this phenomenon generalizes across domains, modalities, and interaction settings.

Second, our methodology focuses on token- and step-level entropy dynamics as proxies for information allocation during generation. Although this perspective offers a tractable and informative view of reasoning structure, it does not provide a mechanistic explanation of why such differences emerge. In particular, we do not explicitly connect the observed UID patterns to the underlying properties of autoregressive decoding, internal information allocation, or training objectives of large language models. A deeper mechanistic account linking generation dynamics to these behavioral signatures would further enrich this line of work and remains an important direction for future research.

Broader Impact and Ethical Considerations

This work contributes to improving the reliability and interpretability of LLM reasoning by introducing information-theoretic diagnostics that characterize how information is distributed and evolves within individual reasoning traces. By treating reasoning as an internal information-flow process rather than solely on final answers or cross-sample agreement, UID-based metrics provide a complementary, sample-efficient signal for identifying coherent versus unstable reasoning trajectories, which may support safer and more trustworthy deploymnet of LLMs in settings such as education, scientific problem-solving, and decision support. At the same time, these metrics are not guarantees of correctness and should not be interpreted as normative standards of good or human-like reasoning. Misuse may arise if UID scores are treated as hard filters, certification signals, or stylistic constraints, potentially leading to overconfidence in incorrect outputs or incentivizing superficial optimization of reasoning traces without genuine improvements in reasoning ability. Moreover, as our analysis shows that effective LLM reasoning diverges from human communicative UID patterns, applying these metrics to evaluate human-facing explanations or across domains beyond those studied may be inappropropriate. We therefore emphasize the UID-based measures are best understood as diagnostic tools operating within a broader evaluation framework that includes answer verification, aggregation methods, and human oversight. Responsible use requires clear communication of their scope, limitations, and probabilistic nature, as well as further study of their behavior across domains, languages, and interaction settings.

Acknowledgments

We thank Ilgee Hong at the Georgia Institute of Technology for helpful discussions, and Chanjoo Jung at Yonsei University for valuable feedback on writing. We also thank Oneline AI for providing GPU resources that supported this research. Minju and Jaehyung are affiliated with the Department of Artificial Intelligence at Yonsei University. This research was supported in part by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2020-II201361, Artificial Intelligence Graduate School Program (Yonsei University); No. RS-2025-25442405, Development of a Self-Learning World Model-Based AGI System for Hyperspectral Imaging.

References

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2025. The unreasonable effectiveness of entropy minimization in llm reasoning.

Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling how long a reasoning model thinks with reinforcement learning. In *Conference on Language Modeling (COLM)*.

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2025. Matharena: Evaluating llms on uncontaminated math competitions. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Siddhant Bhambri, Upasana Biswas, and Subbarao Kambhampati. 2025. Do cognitively interpretable reasoning traces improve llm performance? In *Workshop of Advances in Neural Information Processing Systems (NeurIPS)*.

Eric Bigelow, Ari Holtzman, Hidenori Tanaka, and Tomer Ullman. 2024. Forking paths in neural text generation. In *International Conference on Learning Representations (ICLR)*.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. 2025. Thought anchors: Which llm reasoning steps matter?

Hyungjoo Chae, Yongho Song, Kai Tzu iunn Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. In *Transactions on Machine Learning Research (TMLR)*.

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. A cross-linguistic pressure for uniform information density in word order. In *Transactions of the Association for Computational Linguistics (TACL)*.

MX Collins. 2014. Information density and dependency length as complementary cognitive models. *Journal of Psycholinguistic Research*, 43(5):651–681.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2nd edition. Wiley. The standard textbook explaining entropy as the expected number of bits required to encode a distribution.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Shengyu Feng, Xiang Kong, Shuang Ma, Aonan Zhang, Dong Yin, Chong Wang, Ruoming Pang, and Yiming Yang. 2025. Step-by-step reasoning for math problems via twisted sequential monte carlo. In *International Conference on Learning Representations (ICLR)*.

August Fenk and Gertraud Fenk-Oczlon. 1980. Konstanz im kurzzeitgedächtnis - konstanz im sprachlichen informationsfluß?

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. 2025. Deep think with confidence.

Zitian Gao, Lynx Chen, Haoming Luo, Joey Zhou, and Bryan Dai. 2025. One-shot entropy minimization.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Roscoe: A suite of metrics for scoring step-by-step reasoning. In *International Conference on Learning Representations (ICLR)*.

T. Florian Jaeger and Roger P. Levy. 2006. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hyosoon Jang, Yunhui Jang, Sungjae Lee, Jungseul Ok, and Sungsoo Ahn. 2025. Self-training large language models with confident reasoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable best-of-n selection for large language models via self-certainty. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Max Kuhn and Kjell Johnson. 2013. *Applied Predictive Modeling*. Springer. Discusses entropy in predictive distributions as a signal of model uncertainty.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning.

Dongseok Lee, Jimyung Hong, Dongyoung Kim, and Jaehyung Kim. 2025. Training-free llm verification via recycling few-shot examples.

Zeju Li, Jianyuan Zhong, Ziyang Zheng, Xiangyu Wen, Zhijian Xu, Yingying Cheng, Fan Zhang, and Qiang Xu. 2025. Compressing chain-of-thought in llms via step entropy.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations (ICLR)*.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. Receval: Evaluating reasoning chains via correctness and informativeness. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof qa benchmark. In *Conference on Language Modeling (COLM)*.

Claude E. Shannon. 1948. A mathematical theory of communication.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. 2023. Large language models are in-context semantic reasoners rather than symbolic reasoners.

Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Wilcox, Mario Giulianelli, and Alex Warstadt. 2024. Surprise! uniform information density isn't the whole story: Predicting surprisal contours in long-form discourse. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Skyler Wu, Eric Meng Shen, Charumathi Badrinath, Jiaqi Ma, and Himabindu Lakkaraju. 2023. Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions. In *Workshop of Proceedings of the International Conference on Machine Learning (ICML)*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025a. Qwen3 technical report.

Zhaohui Yang, Chenghua He, Xiaowen Shi, Linjing Li, Qiyue Yin, Shihong Deng, and Daxin Jiang. 2025b. Beyond the first error: Process reward models for reflective mathematical reasoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. 2025. Right question is already half the answer: Fully unsupervised llm reasoning incentivization.

Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang, and Huan Liu. 2025a. Is chain-of-thought reasoning of llms a mirage? a data distribution lens. In *Workshop of Advances in Neural Information Processing Systems (NeurIPS)*.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. 2025b. Learning to reason without external rewards.

Chuyue Zhou, Wangjie You, Juntao Li, Jing Ye, Kehai Chen, and Min Zhang. 2023. INFORM: Information eNtropy based multi-step reasoning FOR large language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. 2025. Ttrl: Test-time reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Justification of Segmentation Strategy and Robustness Analyses

A.1 Justification of newline-based segmentation

We segment reasoning traces using paragraph boundaries defined by `\n\n`. This heuristic aligns with how large language models structure free-form reasoning and is widely adopted in prior work. While explicit empirical validation of step segmentation strategies remains limited, Lightman et al. (Lightman et al., 2023) employ newline-separated steps in the PRM800K dataset, where large-scale human annotation suggests that paragraph boundaries often align with intuitive subgoal transitions.

Several recent studies on step-wise reasoning and process-level supervision similarly adopt newline- or paragraph-based segmentation schemes (Feng et al., 2025; Yang et al., 2025b), indicating that this approach is a practical and commonly accepted default. Importantly, our method does not rely on the optimality of any single segmentation strategy; instead, we verify robustness to alternative stepwise segmentation choices in the following sections.

A.2 Robustness to Fixed Token Window Segmentation

To evaluate robustness to non-semantic step definitions, we apply a coarse fixed-length token window segmentation (2048 tokens per step) following Fu et al. (2025). Table 6 reports UID-conditioned performance across AIME, HMMT, and BRUMO for Qwen3-8B, DeepSeek-R1-Distill-Qwen-7B, and DeepSeek-R1-Distill-Llama-8B (seed 42).

Across nearly all model–dataset pairs, the expected ordering patterns persist: *High UID (3σ)* $\leq$ *Low UID (3σ)* and *High UID (var)* $\geq$ *Low UID (var)*. While we observe an isolated deviation for Qwen3-8B on BRUMO under variance-based grouping, the overall trend remains consistent, with local and global UID signals continuing to distinguish higher- and lower-quality reasoning traces.

A.3 Robustness to Semantic Segmentation

We further evaluate robustness under semantic segmentation, where steps are defined using discourse-level markers such as *But*, *Wait*, and *Alternatively*, following the classifications mentioned in Aggarwal and Welleck (2025). As shown in Table 7, most

Table 6: Fixed window segmentation (2048 tokens)

	Qwen3-8B			DS-R1-Distill-Qwen-7B			DS-R1-Distill-Llama-8B		
Setting	AIME	HMMT	BRUMO	AIME	HMMT	BRUMO	AIME	HMMT	BRUMO
Loc. non-uni (3σ)	**0.77**	0.47	0.70	0.40	0.23	**0.57**	**0.40**	0.14	0.33
Loc. uni (3σ)	**0.77**	**0.47**	**0.70**	0.40	0.23	**0.57**	**0.40**	0.14	0.33
Glob. non-uni (var)	0.67	**0.43**	0.63	0.40	**0.30**	**0.60**	**0.40**	**0.27**	**0.43**
Glob. uni (var)	0.63	0.33	**0.67**	0.27	0.20	0.37	0.27	0.20	0.33

Table 7: Semantic Segmentation

	Qwen3-8B			DS-R1-Distill-Qwen-7B			DS-R1-Distill-Llama-8B		
Setting	AIME	HMMT	BRUMO	AIME	HMMT	BRUMO	AIME	HMMT	BRUMO
Loc. non-uni (3σ)	**0.70**	0.33	0.60	0.37	0.17	0.40	0.20	0.20	**0.43**
Loc. uni (3σ)	0.63	**0.47**	**0.67**	**0.47**	**0.27**	**0.50**	**0.43**	0.23	0.40
Glob. non-uni (var)	0.60	**0.43**	**0.67**	0.27	0.20	**0.53**	0.30	**0.27**	0.40
Glob. uni (var)	0.57	0.33	**0.67**	0.43	**0.27**	0.47	0.37	0.20	0.40

settings preserve the expected UID-based orderings under both 3σ- and variance-based groupings.

We observe minor localized deviations, such as Qwen3-8B on AIME under the 3σ grouping and DeepSeek-R1-Distill-Qwen-7B on HMMT under variance-based grouping. However, these deviations are not systematic and do not alter the broader trend: across AIME, HMMT, and BRUMO, and across all three model families, High UID (3σ) rarely outperforms Low UID (3σ), while High UID (var) generally exceeds Low UID (var).

These results indicate that although semantic segmentation introduces slightly more variability—likely due to sensitivity in defining discourse boundaries—the core UID trends remain stable across segmentation strategies.

B What Entropy Does and Does Not Measure

In this work, entropy is used as a tractable proxy for step-level information density, not as a direct measure of semantic progress, factual correctness, or logical soundness. High entropy indicates that the model is distributing probability mass across multiple plausible next-token continuations, while low entropy indicates a more concentrated predictive distribution. However, either regime may arise in both successful and unsuccessful reasoning: elevated entropy can reflect productive exploration of alternatives, but it can also reflect confusion or hallucination; conversely, low entropy can indicate confident convergence, but also premature commitment to an incorrect path. For this reason, we do not interpret entropy magnitude in isolation. Instead, our hypothesis is that the structure of entropy across steps—whether uncertainty evolves smoothly and resolves in an organized way—provides a more reliable signal of reasoning quality than raw entropy alone.

C Details of Preliminary Analyses

C.1 Empirical analyses on various internal signals as a proxy for ID_i

We consider three metrics as a proxy for information density ID_i: (1) log-probability LP_i as a confidence signal, composed from the average token log-probability over step i, (2) entropy H_i as an uncertainty signal, and (3) confidence gap D_i as divergence signal defined as the difference between the log-probability of the current and the previous steps.

Formally, log-probability LP_i of a step is the average token log-probability over step i:

$$LP_i = \frac{1}{b_i - a_i + 1} \sum_{t=a_i}^{b_i} \ell_t.$$

While token-level entropy H_t as

$$H_t = -\sum_{v \in V} p_t(v) \log p_t(v),$$

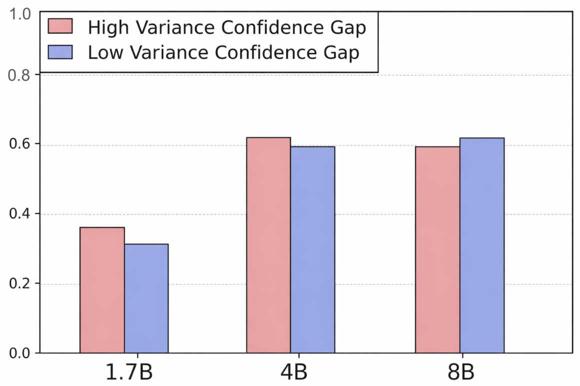
step-level entropy H_i is defined as

$$H_i = \frac{1}{b_i - a_i + 1} \sum_{t=a_i}^{b_i} H_t$$

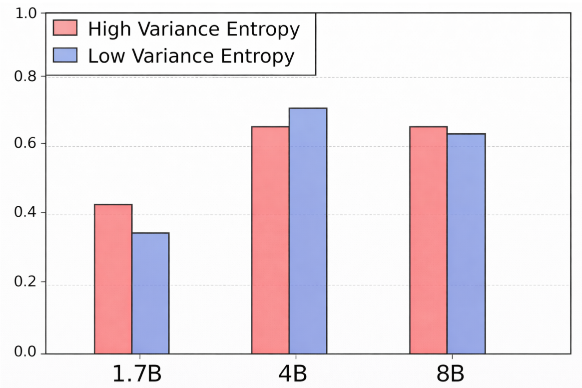
Log-probability gap D_i is defined as

$$D_i = LP_i - LP_{i-1}$$

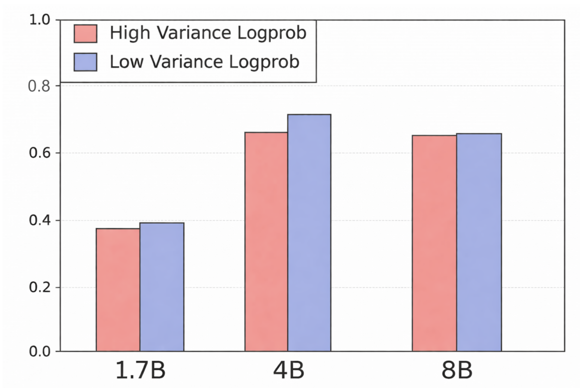
We calculate variance over reasoning traces where each LP_i, H_i, and D_i is used as a proxy of ID_i, and calculate the accuracy of the traces with the highest and lowest variance (i.e., the degree of global uniformity). Results in Figure 4 reveal that entropy measures the difference between highest and lowest trends more profoundly than others across various models.



(a) Acc. of traces with max vs. min var in confidence gap



(b) Acc. of traces with max vs. min var in entropy



(c) Acc. of traces with max vs. min var in log-probability

Figure 4: Empirical results on AIME2025 show that entropy-uniformity is most effective for identifying sound reasoning traces.

C.2 Trace-level Reasoning Step Analyses by Answer Correctness

Figure 5 to Figure 16 are some examples of the case studies performed. We conduct trace-level analysis of reasoning steps of single instances of LLM response to see if we can significantly find patterns only at the token level by looking at entropy peaks where we define peaks as stpes above $u+2\sigma$ where u and σ are, respectively, average and standard deviations of the entropy value of individual steps. We use `Qwen3-8B` on AIME2025 dataset. For both correct and wrong reasoning traces, we (1) identify outlier steps on reasoning traces, and (2) identify their corresponding token-level steps to see if there are a meaningful differences between correct and incorrect traces. We also look at the number of total number of steps of both correct and wrong reasoning traces.

We identify that only qualitively identifying the reasoning traces is not enough to effectively discriminate high-and-low quality reasoning. To be specific, (1) entropy levels in both traces can be quite varying, (2) number of transition words *(i.e. But, Alternatively, Wait)* may appear more at correct traces contrary to our intuition, and (3) the number of steps may be shorter and concise at wrong reasoning traces.

D Sensitivity to Outlier Threshold τ

We initially chose τ to be 2 and 3 as mentioned in our appendix, and reported only τ = 3 as an intuitive "moderate outlier" criterion similar to z-test to avoid expensive hyperparameter tuning. However, to further explore the sensitivity of the results with with respect to the size of τ that captures local uniformity, we also ran a broader sweep with $\tau \in \{2,3,4,5\}$ in Table 15. The results show that increasing τ makes the disruption criterion stricter as fewer changes counted as disruptions, but the overall conclusions are robust: performance trends and the relative behavior of Loc. uni vs. Loc. non-uni do not hinge on a specific choice of k. Therefore, we justify that our τ is a statistical thresholding constant, and not a tuned hyperparameter.

E Implementation Details

E.1 Hyperparameters and GPUs.

We use 2 x H100 GPUs for our main results on Qwen3-8B (thinking mode) and Deepseek-Distill-Qwen-7B, and 4 x RTX A6000 GPUs for all others. Temperature is 0.6, top-p 0.95, and top-k 20, as stated in (Yang et al., 2025a) and (DeepSeek-AI et al., 2025).

E.2 Benchmarks

AIME 2025.[3] The American Invitational Mathematics Examination (AIME) is a prestigious US high school math contest consisting of challenging integer-answer questions. The AIME 2025 benchmark uses problems from the 2025 contests to evaluate a LLM's mathematical reasoning by requiring a single correct integer answer. The test set used in our analysis contains of 30 questions.

BRUMO 2025.[4] The Brown University Math Olympiad (BRUMO) is a mathematics competition for students. The BRUMO 2025 benchmark is built from the problems of the 2025 BRUMO competition, and as part of the MathArena (Balunović et al., 2025) benchmark suite, it consists of 30 final-answer problems. Each requires a unique numeric or closed-form solution. Unlike proof-based contests, BRUMO problems are evaluated on answer correctness.

HMMT 2025.[5] The Harvard-MIT Mathematics Tournament (HMMT) is a renowned competition featuring diverse problems in algebra, geometry, combinatorics, and number theory. The HMMT 2025 benchmark uses newly releassed problems from the February 2025 tournament, providing a broader variety of tasks than AIME. The set used in our analysis contains of 30 questions.

MinervaMath.[6] The Minerva Math benchamrk consists of advanced quantitative problems sourced from university-level STEM courses, including physics, chemistry, and higher mathematics. The set used in our analysis contains of 272 questions.

GPQA-Diamond.[7] GPQA-Diamond (Rein et al., 2023) is the hardest subset of the Graduate-Level Google-Proof Q&A benchmark, consisting of 198 multiple-choice questions in biology, chemistry, and physics. These questions are crafted by domain experts and validated by multiple PhD-level validators to ensure clarity and high difficulty. The Google-proof design means that even with web search, solving them requires deep reasoning rather than lookup.

LSAT-AR.[8] The LSAT Analytical Reasoning (LSAT-AR) benchmark, also known as *logic games*, evaluates a model's ability to perform structured symbolic reasoning under explicit constraints. Each problem presents a set of entities together with a collection of rules, and requires deducing valid arrangements or assignments that satisfy all constraints. LSAT-AR problems are designed to test combinatorial reasoning, constraint satisfaction, and systematic deduction rather than factual knowledge. The benchmark is widely used to assess logical planning and rule-based reasoning abilities in language models.

LSAT-LR.[9] The LSAT Logical Reasoning (LSAT-LR) benchmark consists of short natural-language arguments followed by multiple-choice questions that probe logical validity, assumptions, and implications. Unlike LSAT-AR, LSAT-LR focuses on informal logical reasoning, including identifying flaws, strengthening or weakening arguments, and drawing valid conclusions from given premises. This benchmark evaluates a model's ability to understand argument structure, implicit assumptions, and causal or logical relationships expressed in natural language.

E.3 Baseline Details

We re-implemented all logic using vLLM, unlike some of the codes initially released. Our main baselines are selection methods that leverage LLMs' internal signals.

Self-Certainty. This is the implementation of Kang et al. (2025), where it evaluates LLMs' reasoning by introducing a self-certainty, a confidence score assigned at each reasoning step. Self-certainty captures whether the model is confident in its logical steps, and uses Borda Voting to improve answer selection. Borda Voting ranks outputs by aggregared confidence rather than relying on simple majority voting used in Self-Consistency (Wang et al., 2023).

Highest Confidence. This is a naive implementation that selects the path with the highest oeverall token confidence in the reasoning trace. Similar methods have been introduced in (Jang et al., 2025), where paths for training are selected based on the

[3] https://huggingface.co/datasets/opencompass/AIME2025

[4] https://huggingface.co/datasets/MathArena/brumo_2025

[5] https://huggingface.co/datasets/MathArena/hmmt_feb_2025

[6] https://huggingface.co/datasets/math-ai/minervamath

[7] https://huggingface.co/datasets/Idavidrein/gpqa

[8] https://huggingface.co/datasets/allenai/lsat-ar

[9] https://huggingface.co/datasets/allenai/lsat-lr

traces with the highest confidence.

Lowest Entropy. This is a naive implementation that selects the path with the lowest overall token entropy in the reasoning trace, driven from the idea that entropy itself is a measurement of uncertainty.

F Experiment Details

Details of the main experiment Details of all the experiments carried out on the three seeds (42, 1234, 2025) are in Table 8, 9, and 10.

Details of model size scaling experiment Details of all the experiments carried out on the three seeds (42, 1234, 2025) are in Table 11.

Details of sample size scaling experiment Details of all the experiments carried out on the three seeds (42, 1234, 2025) are in Table 12.

G Additional Analyses

G.1 Complementarity with majority voting

Our primary analysis focuses on *trace-level selection*, where a single reasoning trajectory is chosen based on internal structural signals such as information-flow uniformity. In contrast, *answer-level aggregation* methods—most notably majority voting—operate by comparing multiple independently generated responses and selecting answers based on cross-sample agreement. These two approaches therefore act at different granularities: UID evaluates how information evolves *within* a single trace, whereas aggregation methods exploit consensus *across* traces.

To clarify their relationship, we compare UID-based selectors with majority voting under *matched sampling budgets*, and additionally evaluate combinations of UID with multi-sample aggregation in an extended analysis. In particular, we adopt *Borda voting* as a lightweight aggregation mechanism over trace-level structural signals. Given N sampled traces indexed by $i \in \{1, \ldots, N\}$, each trace is assigned a rank r_i according to a chosen criterion and converted into a weight

$$w_i = (N - r_i + 1)^p,$$

with exponent $p = 0.5$ and $N = 3$ unless otherwise noted. The selected trace is given by $i^* = \arg\max_i w_i$.

We consider two Borda-based variants. *(i) UID-based Borda voting* ranks traces in descending order of their UID scores, evaluating whether information-flow uniformity within a trace is predictive of correctness. *(ii) Spikes/Falls-based Borda voting* ranks traces in ascending order of entropy irregularity, measured as the number of entropy spikes and falls, to examine whether local disruptions in entropy smoothness correlate with performance.

Table 13 reports results across three backbone models. We find that UID-based selection can be competitive with majority voting and often complementary when combined with lightweight aggregation schemes such as Borda voting, consistent with (Kang et al., 2025).

G.2 Qualitative analyses on reasoning traces with low and high UID scores

We qualitatively analyze reasoning traces with the highest and lowest UID scores across the dataset (Figures 17 and 18). Traces that are globally non-uniform exhibit the characteristic ebb and flow of high-quality reasoning, reflecting meaningful shifts in information density. In contrast, low-variance traces reveal a failure mode of superficially uniform but uninformative expansion.

At a local scale, Figure 19 illustrates a reasoning trace that is locally non-uniform yet stuck and repetitive, despite arriving at the correct answer. Conversely, locally uniform traces (Figure 20) avoid such degenerative repetition and maintain coherent progress.

H Prompts for Communicative vs. Naive Prompting

Details of the prompts used for comparing human communication and LLM reasoning are attached in Table 14. We evaluated it on AIME2025 benchmark, sampled by $n = 3$.

I Usage of AI Assistants

In preparing this work, we used AI-based writing assistants to improve sentence structure, correct grammatical errors, and enhance overall readability. These tools were employed solely for language refinement and did not contribute to the development of technical content, research methodology, or experimental analysis. All scientific ideas, results, and conclusions presented in the paper were conceived and authored entirely by the researchers. Use of AI assistance was restricted to editorial purposes and did not affect the originality or intellectual contributions of the work.

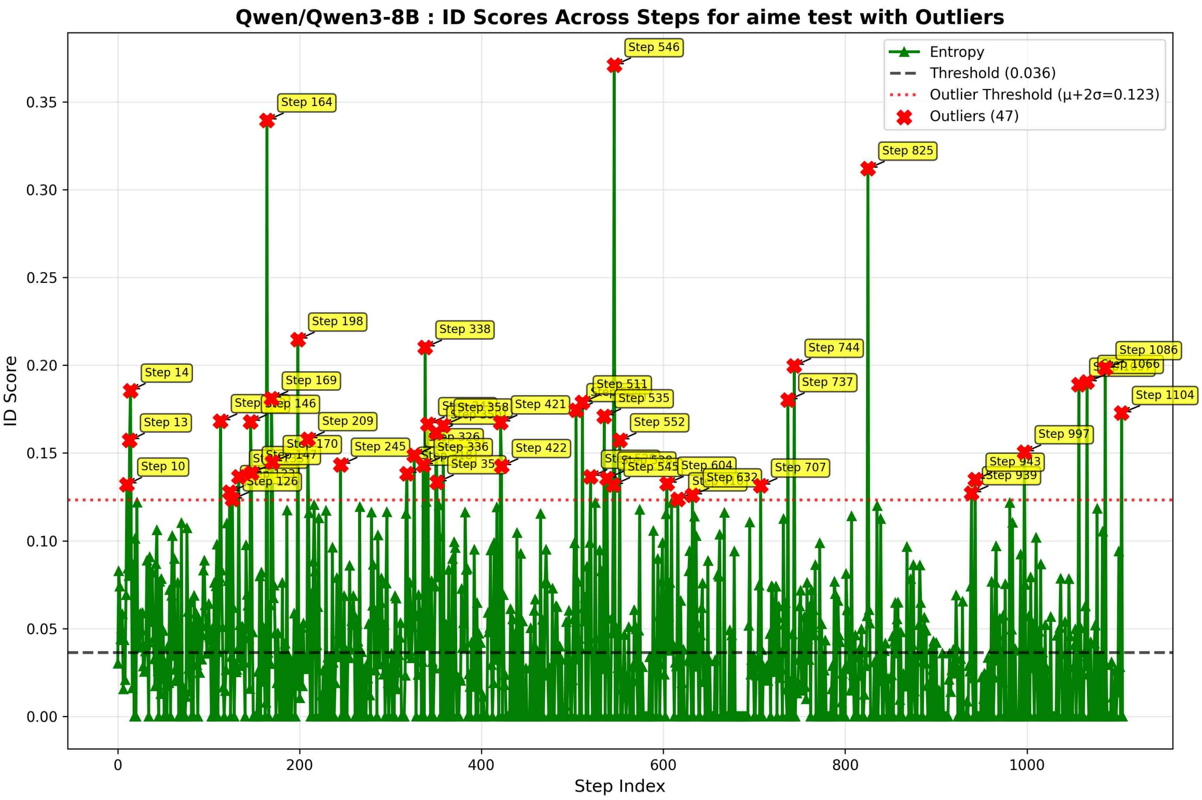


Figure 5: Q6 — Correct Trace Visualization

Outlier Reasoning Steps — step_index & reasoning_text

Dataset: aime • Model: Qwen/Qwen3-8B • Sample ID: 6

STEP INDEX	REASONING TEXT
10	Therefore, the pair containing G will be either of the form XG where X is a letter before G, or GX where X is a letter after G.
13	So, how do we compute this probability?
14	First, let's think about the entire process. The letters are randomly grouped into six pairs. Then each pair is ordered alphabetically. Then the six pairs are sorted alphabetically. The last one is the maximum. So, the question is equivalent to: What is the probability that the maximum pair contains G?
113	Alternatively, maybe there are other cases where the maximum pair contains G. For example, if there is a pair (G, H), but that would require that G is paired with H. However, since H is a high letter, and if H is paired with G, then the pair is (G, H), which is ordered as GH. But if there is another pair that starts with H, like (H, I), then that pair would be larger than GH. Therefore, for GH to be the maximum pair, there must be no pairs starting with H, I, J, K, or L.
123	To find this, we can think about the possible maximum words.
126	But how does G factor into this?
133	But for G to be the highest first letter, there must be no pair starting with a letter higher than G, i.e., H, I, J, K, L.
146	Wait, this is conflicting with our previous thought. What's going on here?
147	Let me take an example. Suppose all high letters (H-L) are paired with low letters (A-F). Let's say:
164	Therefore, the remaining low letter could be any of the 6 low letters (A-F).
169	However, if X is A, then the pair is AG, and the other pairs are between B-F and H-L. The pair starting with F would be (F, something), which is higher than AG.
170	Therefore, the maximum pair is the pair starting with F, which does not contain G.
198	Wait, but earlier we calculated the probability that all high letters are paired with low letters as 16/231. Let me check if that's consistent.
209	Therefore, the probability that all high letters are paired with low letters is (6 * 5!) / [12! / (2^6 * 6!)]
245	So, let's analyze Case 2.

Figure 6: Q6 — Correct Trace Text

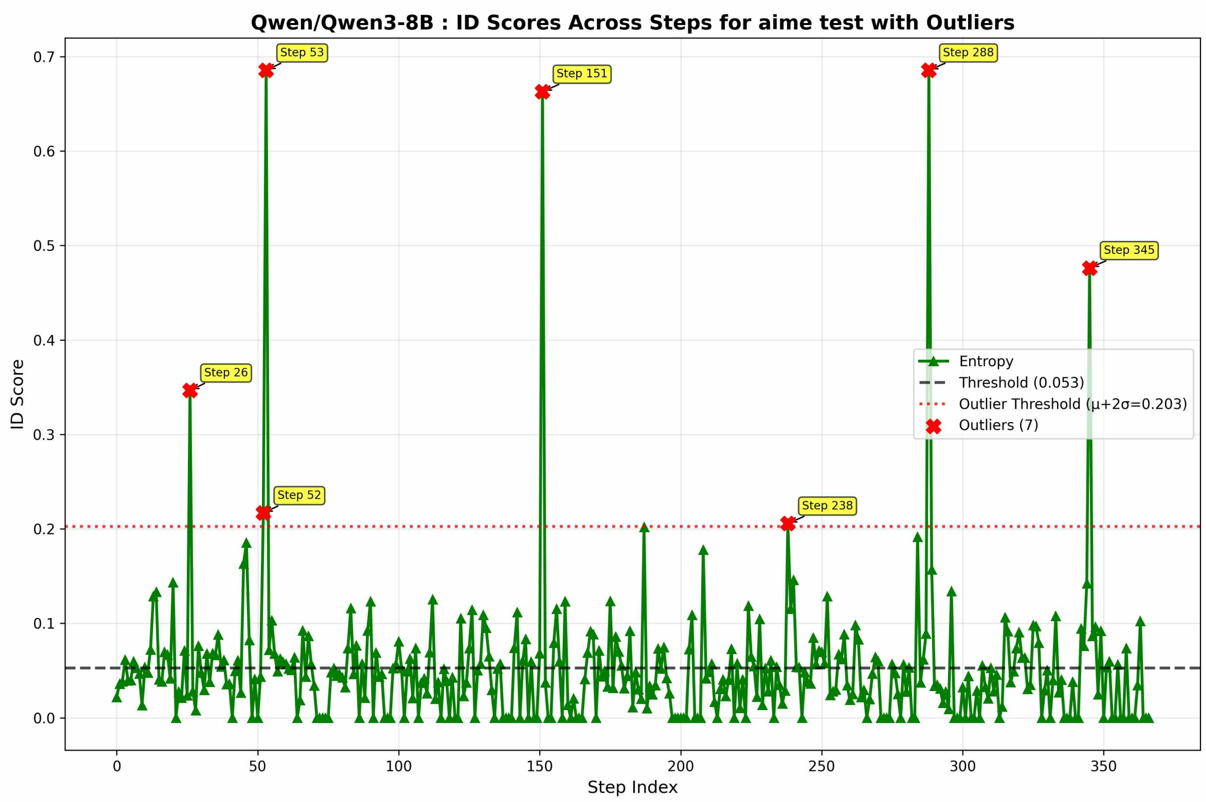


Figure 7: Q6 — Incorrect Trace Visualization

Outlier Reasoning Steps — step_index & reasoning_text

Dataset: aime • Model: Qwen/Qwen3-8B • Sample ID: 6

STEP INDEX	REASONING TEXT
26	As before, the last word is determined by the pair with the maximum smaller letter. So, for the last word to contain G, either:
52	Alternatively, think of the entire process as follows. For Case 1, where G is the maximum smaller letter, meaning that all letters larger than G (H-L) are paired with letters from A-G, and G is paired with one of H-L. Then, the rest of the pairings can be arbitrary as long as they satisfy the above condition.
53	But perhaps another approach is better. Let me think.
151	- All letters except G and M.
238	As before, this is possible only if the number of letters larger than M is even. For M = B, D, F.
288	First, total pairings: 12! / (2^6 * 6!) = (479001600) / (64 * 720) = 479001600 / 46080 = 10395.
345	### Total Number of Pairings

Figure 8: Q6 — Incorrect Trace Text

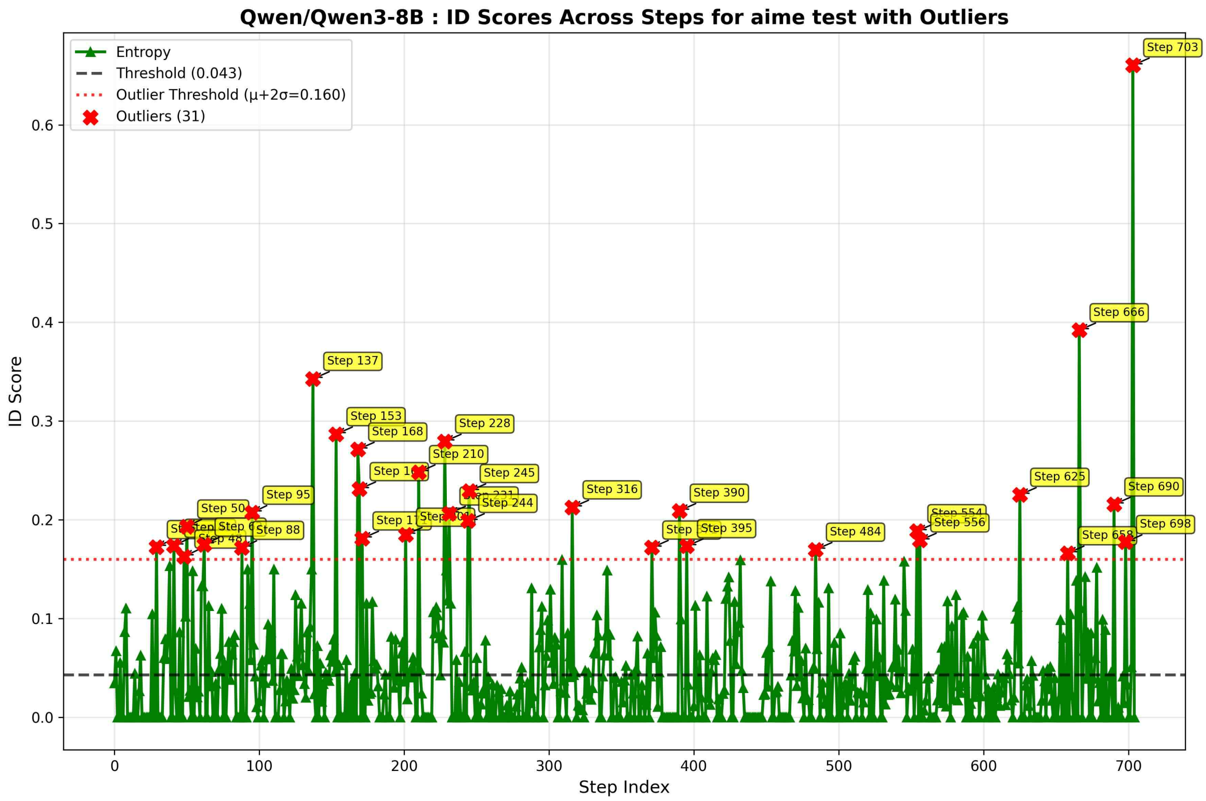


Figure 9: Q11 — Correct Trace Visualization

Outlier Reasoning Steps — step_index & reasoning_text

Dataset: aime • Model: Qwen/Qwen3-8B • Sample ID: 11

STEP INDEX	REASONING TEXT
29	1. (x - y)(1 + z) < 0
41	Let me consider each inequality separately.
48	Therefore, combining these, we need to find the regions where these cases hold.
50	But perhaps I need to analyze the inequalities together.
62	But since x + y + z = 75, maybe we can also express variables in terms of others?
88	Case 2: (x - y) < 0 and (76 - x - y) > 0.
95	But maybe I need to consider the possible combinations.
137	But maybe there's a relation between x and 1 + x?
153	Therefore, y < -1.
168	From x + y > 76, we have x > 76 - y.
169	Therefore, combining:
171	But since x > y, we also have:
201	But x < -1. Also, x + y < 76. Let me see if these can coexist.
210	Subtracting the second inequality from the first:
228	x + 2y > 75.
231	But the problem mentions three disjoint convex regions. So, perhaps there's a third case? Wait, maybe there are other possibilities?

Figure 10: Q11 — Correct Trace Text

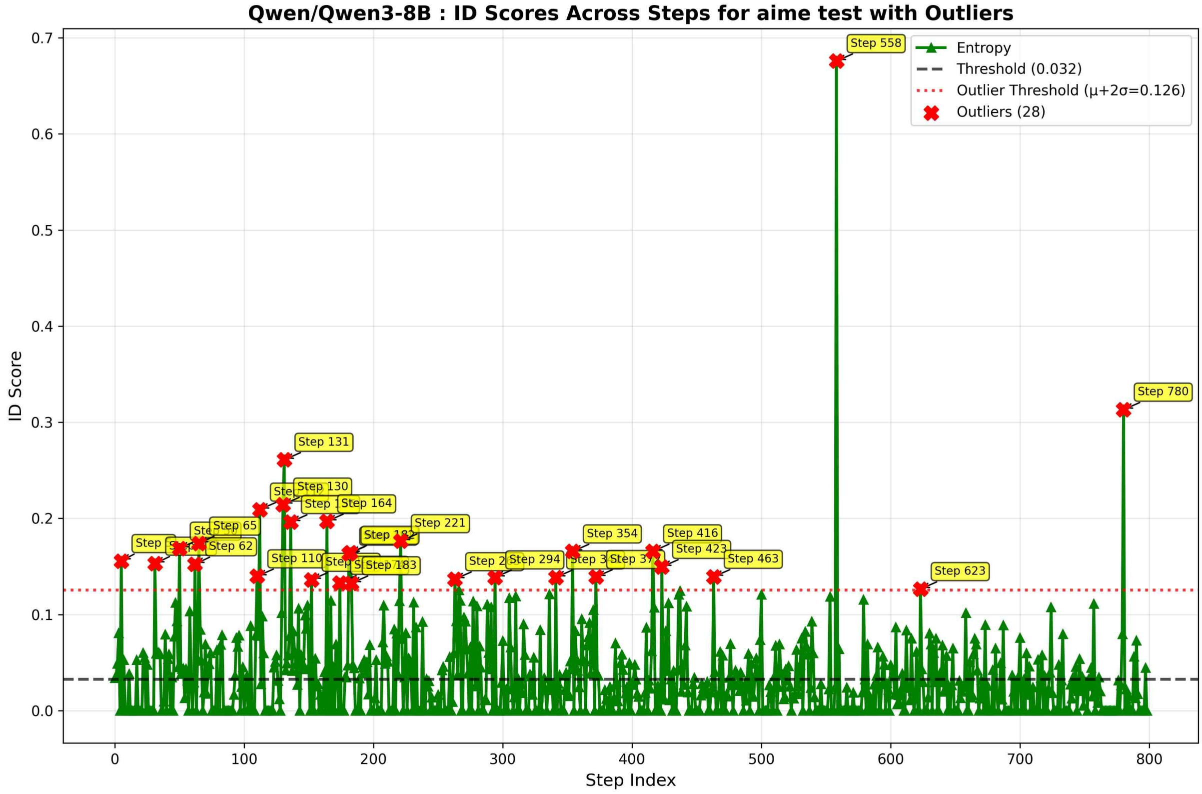


Figure 11: Q11 — Incorrect Trace Visualization

Outlier Reasoning Steps — step_index & reasoning_text

Dataset: aime • Model: Qwen/Qwen3-8B • Sample ID: 11

STEP INDEX	REASONING TEXT
5	Let me rearrange the first inequality:
31	Therefore, combining these, we have combinations of these cases. But since the inequalities are connected (they are both part of the same chain), we need to consider the overlap of the cases.
50	Similarly, for Inequality (2): (x + 2y - 75)(1 + x) < 0. Let me denote C = x + 2y - 75 and D = 1 + x. Then C*D < 0.
62	Alternatively, maybe I can find relations between variables. Let me try to find relations between x, y, z. Let me note that since x + y + z = 75, then 1 + x = 1 + x, and 1 + z = 1 + z. Also, 76 - x - y = 1 + 75 - x - y = 1 + z. Wait, because z = 75 - x - y, so 76 - x - y = 1 + z. So, in Inequality (1), (x - y)(1 + z) < 0. So, that's the same as before.
65	Let me think of Inequality (1): (x - y)(1 + z) < 0. Since z = 75 - x - y, 1 + z = 76 - x - y. So, Inequality (1) is equivalent to (x - y)(76 - x - y) < 0. Similarly, Inequality (2) is (y - z)(1 + x) < 0. Let me think about the possible cases.
110	Wait, maybe not. Let me check if these cases are actually regions on the plane.
112	Alternatively, perhaps there's a mistake in my analysis. Let me check again.
130	Case 2: (x < y and z > -1) and (y > z and x < -1). Therefore, x < y > z, with z > -1 and x < -1. So, this is Case 2.
131	Case 3: (x > y and z < -1) and (y < z and x > -1). Therefore, x > y < z, with z < -1 and x > -1. So, this is Case 3.
136	Case 2: x < y > z, with x < -1 and z > -1. So, x is less than -1, z is greater than -1, and y is between x and z? Wait, no. Wait, in Case 2, we have x < y and y > z. So, x < y > z. Since x < -1 and z > -1, then y can be anything? For example, suppose x is very negative, z is slightly above -1, then y is between x and z? Wait, but if x < y and y > z, then y must be greater than z and greater than x. But z > -1 and x < -1. Therefore, y must be greater than z (which is > -1) and greater than x (which is < -1). Therefore, y > max(z, x). But since z > -1 and x < -1, then y > z. Therefore, in this case, y > z > -1, and x < -1. So, y is greater than z, which is greater than -1, and x is less than -1. Therefore, possible.
152	So, where is the mistake?
164	x = -1 + b > y = 77 + a - b => -1 + b > 77 + a - b => 2b - a > 78.
174	Equation 2: b - 2a > 78

Figure 12: Q11 — Incorrect Trace Text

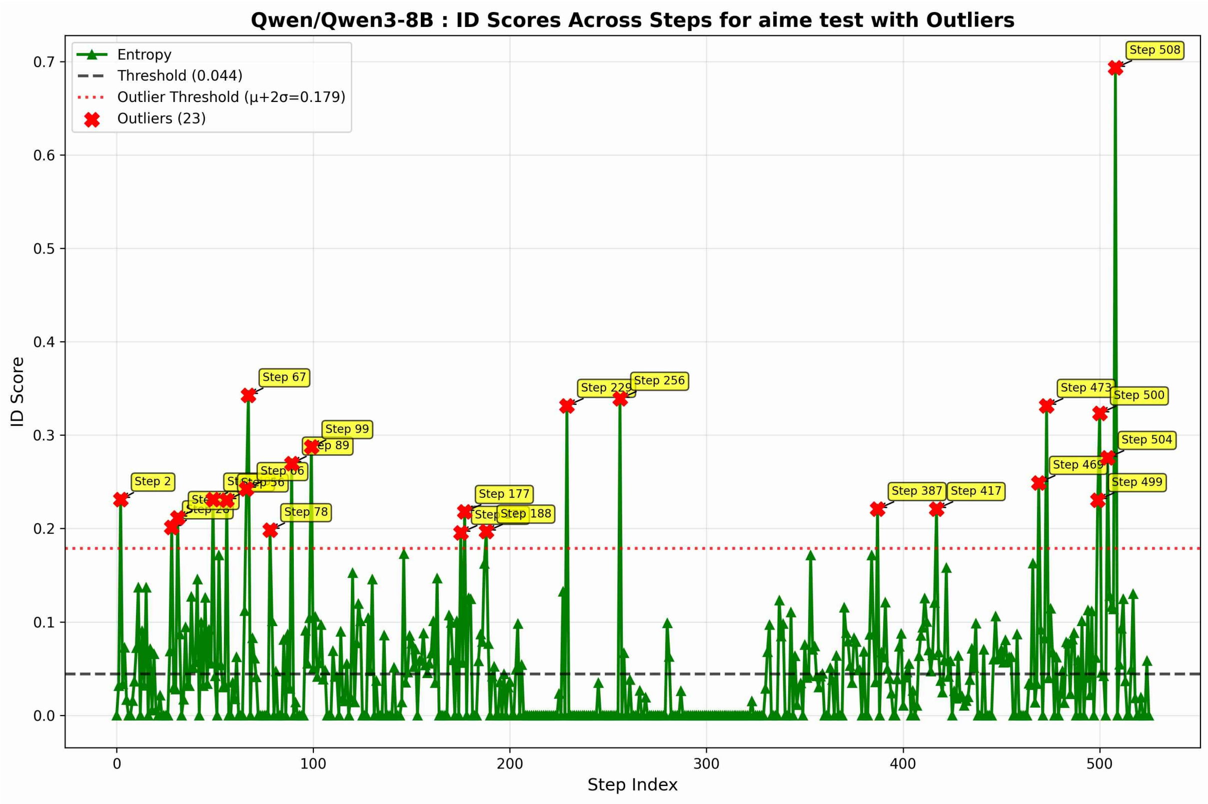


Figure 13: Q17 — Correct Trace Visualization

Outlier Reasoning Steps — step_index & reasoning_text

Dataset: aime • Model: Qwen/Qwen3-8B • Sample ID: 17

STEP INDEX	REASONING TEXT
2	Each unit square is formed by four unit segments. There are four squares in the grid: top-left, top-right, bottom-left, bottom-right. Each of these squares must have exactly 2 red sides and 2 blue sides. The question is how many colorings satisfy this condition.
28	For example, the left vertical segment V1T is part of Square A. The left vertical segment V1B is part of Square C. The middle vertical segments V2T and V2B are part of both Square A and B (V2T is part of Square A and B; V2B is part of Square C and D). Similarly, the middle horizontal segments H2L and H2R are part of both Square A and C (H2L) and Square B and D (H2R). Similarly, the top horizontal segments H1L and H1R are only part of Square A and B respectively, and the bottom horizontal segments H3L and H3R are only part of Square C and D respectively. Similarly, the right vertical segments V3T and V3B are only part of Square B and D respectively.
31	- Top horizontal: H1L, H1R - Right vertical: V3T, V3B - Bottom horizontal: H3L, H3R - Left vertical: V1T, V1B
49	Let me denote:
56	If I = 0, then O = 8. But there are only 8 outer segments, so possible.
66	So, if I can determine the colors of the inner segments, then the outer segments can be determined? Not necessarily, because each outer segment is part of only one square. For example, H1L is part of Square A only. So, if I know the colors of the inner segments of Square A, and the number of red sides needed for Square A, I can determine the color of H1L and V1T. Wait, but there are two outer segments and two inner segments. Let me think.
67	Suppose for Square A, the inner segments V2T and H2L have some colors. Let me denote:
78	- Let c = color of V2T (inner vertical top)
89	Square C:
99	Wait, for example, suppose for Square A, the inner segments c and a are both red. Then, to have exactly two red sides in Square A, the outer segments H1L and V1T must be both blue. Because total red sides would be 2 (from c and a) + 0 (from outer) = 2. Alternatively, if c and a are both blue, then the outer segments must be both red. If one of c or a is red and the other is blue, then the outer segments must have one red and one blue. Wait, but how?
175	Alternatively, think of variables:
177	k_A = number of red in {a,c}
188	k_A = 0 (since a and c are both B)
229	k_C = {a, d} = 1 (a=R, d=B)
256	Then:
387	Wait, earlier we had that O = 8 - 2I. Therefore, O is fixed once I is fixed. However, O is the number of red outer segments. But how does that relate to the constraints?
417	Alternatively, think of the grid as a graph where each square is a node, and edges represent shared segments. However, not sure.
469	Sum_{squares} I_s = 2I (since each red inner segment is counted twice)
473	But since the inner segments are variables as well, it's complex.

Figure 14: Q17 — Correct Trace Text

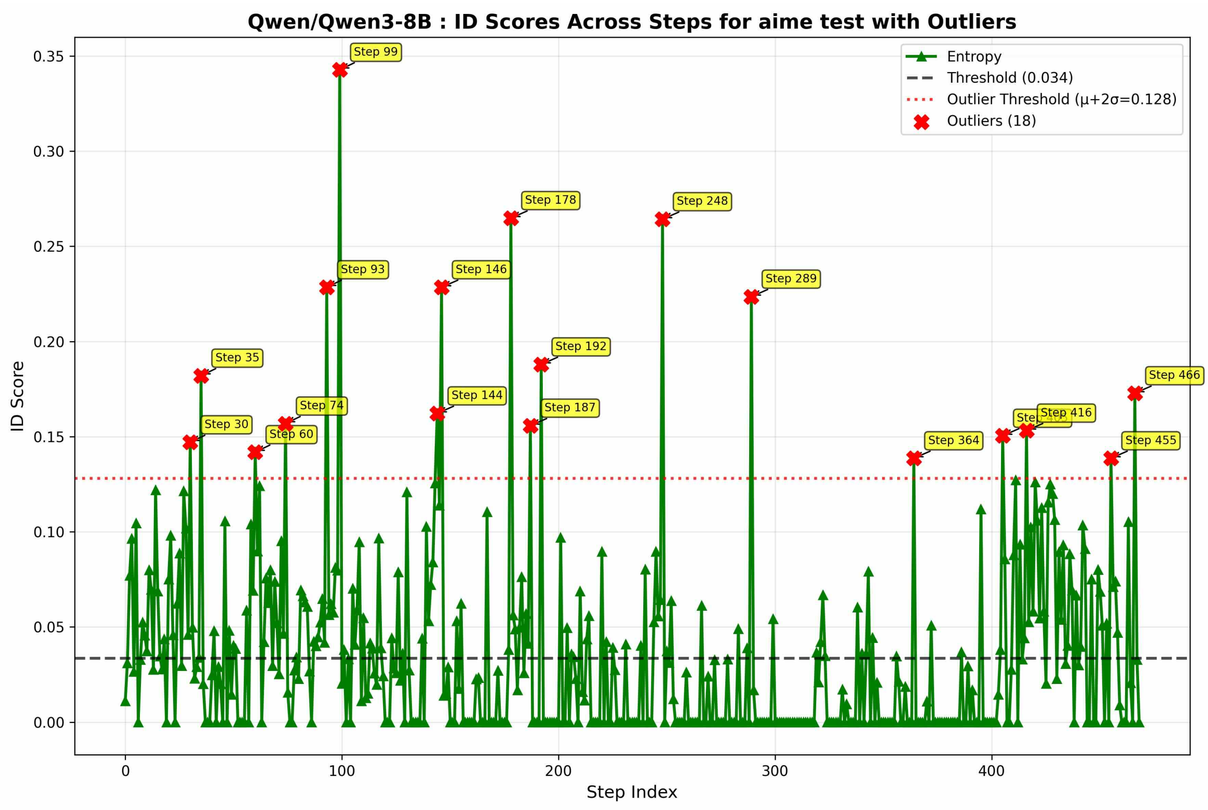


Figure 15: Q17 — Incorrect Trace Visualization

Outlier Reasoning Steps — step_index & reasoning_text

Dataset: aime • Model: Qwen/Qwen3-8B • Sample ID: 17

STEP INDEX	REASONING TEXT
30	There are 3 horizontal lines (top, middle, bottom) and 3 vertical lines (left, middle, right). Each horizontal line has two segments, and each vertical line has two segments.
35	Therefore, each square's sides correspond to specific segments. So, for square A: top side is H1V1V2, right side is V2T, bottom side is H2V1V2, left side is V1T.
60	Alternatively, maybe think of the entire grid and consider that each square has two red sides, so maybe there is some kind of cycle or pattern.
74	So inner segments:
93	Wait, maybe my initial assumption is wrong. If I = 1, then O = 6. So there are 6 red outer segments and 2 blue outer segments. However, if we color one inner segment red, then we need to adjust the outer segments such that each square has exactly two red sides.
99	But each square has two outer sides. For square A, the outer sides are top and left. For square C, the outer sides are left (V1B) and bottom (H3V1V2). Therefore, if we turn one of these outer sides blue for square A and one of these outer sides blue for square C, but we only have two outer segments to turn blue. However, if we turn, say, the top of square A (H1V1V2) blue and the bottom of square C (H3V1V2) blue, then that would satisfy both squares. However, we need to check if this affects other squares.
144	Wait, segment A (H2V1V2) is the bottom of square A and top of square C.
146	So, if both A and B are red, then:
178	Therefore, for square A's outer sides: top (H1V1V2) and left (V1T). Similarly, square B's outer sides: top (H1V2V3) and right (V3T). Square C's outer sides: left (V1B) and bottom (H3V1V2). Square D's outer sides: left (V2B) and bottom (H3V2V3). Therefore, the outer segments for the squares are:
187	For square A: one inner red (bottom) and one outer red (either top or left) => total two red sides.
192	Pair 2: inner segments A (H2V1V2) and C (V2T) are red. Let me analyze.
248	Wait, let me re-examine:
289	For square B: 2 choices (top or right)
364	Total: 2 * 2 = 4.

Figure 16: Q17 — Incorrect Trace Text

Table 8: **Main results.** Results of `Deepseek-R1-Distill-Qwen-7b` on four math benchmarks (AIME2025, BRUNO2025, HMMT2025, and MinervaMath). Performance is reported across seeds 42, 1234, and 2025, with averages and standard deviations (Avg ± Std).

Methods (↓)	AIME2025				BRUMO2025			
	42	1234	2025	Avg ± Std	42	1234	2025	Avg ± Std
Mean Acc.	0.41	0.39	0.41	0.40 ± 0.02	0.55	0.53	0.53	0.54 ± 0.01
Self-Cert.	0.50	0.43	0.50	0.48 ± 0.04	0.50	0.57	0.50	0.52 ± 0.04
High Conf.	0.50	0.43	0.50	0.48 ± 0.04	0.57	0.53	0.47	0.52 ± 0.05
Low Ent.	0.50	0.43	0.50	0.48 ± 0.04	0.60	0.57	0.50	0.56 ± 0.05
UID Metrics (↓)								
Loc. non-uni (2σ)	0.27	0.27	0.27	0.27 ± 0.00	0.47	0.33	0.37	0.39 ± 0.07
Loc. uni (2σ)	0.53	0.47	0.53	0.51 ± 0.04	0.57	0.57	0.57	0.57 ± 0.00
Loc. non-uni (3σ)	0.27	0.27	0.27	0.27 ± 0.00	0.47	0.33	0.37	0.39 ± 0.07
Loc. uni (3σ)	0.57	0.47	0.57	0.53 ± 0.06	0.57	0.57	0.53	0.56 ± 0.02
Glob. non-uni (var)	0.57	0.43	0.57	0.52 ± 0.08	0.60	0.63	0.70	0.64 ± 0.05
Glob. uni (var)	0.33	0.33	0.33	0.33 ± 0.00	0.47	0.43	0.43	0.44 ± 0.02

Methods (↓)	HMMT2025				MinervaMath			
	42	1234	2025	Avg ± Std	42	1234	2025	Avg ± Std
Mean Acc.	0.24	0.23	0.23	0.24 ± 0.00	0.30	0.29	0.30	0.30 ± 0.00
Self-Cert.	0.30	0.27	0.27	0.28 ± 0.02	0.31	0.30	0.31	0.30 ± 0.00
High Conf.	0.27	0.27	0.27	0.27 ± 0.00	0.31	0.31	0.31	0.31 ± 0.00
Low Ent.	0.27	0.23	0.23	0.24 ± 0.02	0.30	0.31	0.30	0.30 ± 0.00
UID Metrics (↓)								
Loc. non-uni (2σ)	0.13	0.20	0.20	0.18 ± 0.04	0.27	0.27	0.27	0.27 ± 0.00
Loc. uni (2σ)	0.30	0.30	0.30	0.30 ± 0.00	0.30	0.32	0.30	0.31 ± 0.01
Loc. non-uni (3σ)	0.13	0.20	0.20	0.18 ± 0.04	0.26	0.27	0.26	0.26 ± 0.01
Loc. uni (3σ)	0.30	0.30	0.30	0.30 ± 0.00	0.31	0.32	0.31	0.31 ± 0.01
Glob. non-uni (var)	0.30	0.23	0.23	0.26 ± 0.04	0.30	0.31	0.30	0.30 ± 0.01
Glob. uni (var)	0.20	0.13	0.13	0.16 ± 0.04	0.28	0.28	0.28	0.28 ± 0.00

Table 9: **Main results.** Results of `Qwen3-8B` on four math benchmarks (AIME2025, BRUMO2025, HMMT2025, and MinervaMath). Performance is reported across seeds 42, 1234, and 2025, with averages and standard deviations (Avg ± Std).

Methods (↓)	AIME2025				BRUMO2025			
	42	1234	2025	Avg ± Std	42	1234	2025	Avg ± Std
Mean Acc.	0.66	0.66	0.68	0.67 ± 0.01	0.67	0.71	0.67	0.68 ± 0.02
Self-Cert.	0.63	0.63	0.63	0.63 ± 0.00	0.70	0.73	0.70	0.71 ± 0.02
High Conf.	0.60	0.60	0.60	0.60 ± 0.00	0.63	0.67	0.60	0.63 ± 0.03
Low Ent.	0.60	0.60	0.60	0.60 ± 0.00	0.63	0.70	0.60	0.64 ± 0.04
UID Metrics (↓)								
Loc. non-uni (2σ)	0.63	0.63	0.63	0.63 ± 0.00	0.53	0.70	0.67	0.63 ± 0.07
Loc. uni (2σ)	0.67	0.67	0.73	0.69 ± 0.03	0.67	0.73	0.70	0.70 ± 0.03
Loc. non-uni (3σ)	0.63	0.63	0.63	0.63 ± 0.00	0.53	0.70	0.67	0.63 ± 0.07
Loc. uni (3σ)	0.67	0.67	0.73	0.69 ± 0.03	0.67	0.73	0.70	0.70 ± 0.03
Glob. non-uni (var)	0.70	0.70	0.70	0.70 ± 0.00	0.53	0.63	0.67	0.61 ± 0.06
Glob. uni (var)	0.67	0.67	0.63	0.66 ± 0.02	0.60	0.77	0.67	0.68 ± 0.07

Methods (↓)	HMMT2025				MinervaMath			
	42	1234	2025	Avg ± Std	42	1234	2025	Avg ± Std
Mean Acc.	0.44	0.41	0.43	0.43 ± 0.01	0.34	0.34	0.33	0.34 ± 0.01
Self-Cert.	0.50	0.47	0.53	0.50 ± 0.03	0.34	0.35	0.34	0.34 ± 0.01
High Conf.	0.50	0.40	0.37	0.42 ± 0.06	0.33	0.34	0.33	0.33 ± 0.01
Low Ent.	0.50	0.40	0.40	0.43 ± 0.05	0.34	0.34	0.32	0.33 ± 0.01
UID Metrics (↓)								
Loc. non-uni (2σ)	0.43	0.37	0.40	0.40 ± 0.03	0.35	0.34	0.32	0.34 ± 0.01
Loc. uni (2σ)	0.53	0.40	0.47	0.47 ± 0.05	0.34	0.34	0.33	0.34 ± 0.01
Loc. non-uni (3σ)	0.43	0.37	0.40	0.40 ± 0.03	0.35	0.34	0.32	0.34 ± 0.01
Loc. uni (3σ)	0.53	0.43	0.47	0.48 ± 0.04	0.34	0.34	0.33	0.34 ± 0.01
Glob. non-uni (var)	0.50	0.47	0.43	0.47 ± 0.03	0.35	0.33	0.32	0.33 ± 0.01
Glob. uni (var)	0.43	0.40	0.40	0.41 ± 0.02	0.35	0.35	0.31	0.34 ± 0.02

Table 10: **Main results.** Results of `Deepseek-R1-Distill-Llama-8b` on four math benchmarks (AIME2025, BRUMO2025, HMMT2025, and MinervaMath). Performance is reported across seeds 42, 1234, and 2025, with averages and standard deviations (Avg ± Std).

Methods (↓)	AIME2025				BRUMO2025			
	42	1234	2025	Avg ± Std	42	1234	2025	Avg ± Std
Mean Acc.	0.32	0.34	0.34	0.33 ± 0.01	0.41	0.37	0.41	0.40 ± 0.02
Self-Cert.	0.33	0.33	0.33	0.33 ± 0.00	0.37	0.33	0.33	0.34 ± 0.02
High Conf.	0.37	0.37	0.33	0.36 ± 0.02	0.40	0.33	0.33	0.36 ± 0.03
Low Ent.	0.33	0.40	0.30	0.34 ± 0.04	0.43	0.30	0.37	0.37 ± 0.06
UID Metrics (↓)								
Loc. non-uni (2σ)	0.33	0.33	0.40	0.36 ± 0.03	0.40	0.37	0.37	0.38 ± 0.02
Loc. uni (2σ)	0.37	0.37	0.37	0.37 ± 0.00	0.37	0.33	0.40	0.37 ± 0.03
Loc. non-uni (3σ)	0.30	0.40	0.37	0.36 ± 0.04	0.40	0.33	0.33	0.36 ± 0.03
Loc. uni (3σ)	0.33	0.40	0.43	0.39 ± 0.04	0.50	0.37	0.47	0.44 ± 0.06
Glob. non-uni	0.40	0.33	0.37	0.37 ± 0.03	0.40	0.30	0.37	0.36 ± 0.04
Glob. uni	0.10	0.33	0.37	0.27 ± 0.12	0.47	0.33	0.40	0.40 ± 0.06

Methods (↓)	HMMT2025				MinervaMath			
	42	1234	2025	Avg ± Std	42	1234	2025	Avg ± Std
Mean Acc.	0.20	0.19	0.22	0.20 ± 0.01	0.23	0.23	0.23	0.23 ± 0.00
Self-Cert.	0.20	0.17	0.20	0.19 ± 0.02	0.22	0.22	0.23	0.22 ± 0.01
High Conf.	0.23	0.20	0.20	0.21 ± 0.02	0.23	0.23	0.22	0.23 ± 0.01
Low Ent.	0.20	0.20	0.17	0.19 ± 0.02	0.21	0.23	0.22	0.22 ± 0.01
UID Metrics (↓)								
Loc. non-uni (2σ)	0.23	0.23	0.23	0.23 ± 0.00	0.22	0.25	0.21	0.23 ± 0.02
Loc. uni (2σ)	0.17	0.20	0.17	0.18 ± 0.02	0.22	0.22	0.22	0.22 ± 0.00
Loc. non-uni (3σ)	0.17	0.20	0.23	0.20 ± 0.03	0.24	0.21	0.24	0.23 ± 0.01
Loc. uni (3σ)	0.27	0.20	0.27	0.24 ± 0.04	0.24	0.22	0.24	0.23 ± 0.01
Glob. non-uni	0.13	0.20	0.20	0.18 ± 0.04	0.21	0.21	0.21	0.21 ± 0.00
Glob. uni	0.30	0.27	0.23	0.27 ± 0.03	0.24	0.24	0.22	0.23 ± 0.01

Table 11: **Model Size Analysis.** Results of `Qwen3` models (1.7B, 4B, 8B) on AIME2025. Performance is reported across seeds 42, 1234, and 2025, with averages and standard deviations (Avg ± Std).

Methods (↓)	1.7B				4B				8B			
	42	1234	2025	Avg ± Std	42	1234	2025	Avg ± Std	42	1234	2025	Avg ± Std
Mean Acc.	0.33	0.34	0.37	0.35 ±0.02	0.66	0.65	0.64	0.65 ±0.01	0.66	0.66	0.68	0.67 ±0.01
Self-Cert.	0.40	0.47	0.47	0.45 ±0.04	0.77	0.70	0.73	0.73 ±0.03	0.63	0.63	0.63	0.63 ±0.00
High Conf.	0.33	0.33	0.43	0.37 ±0.06	0.67	0.57	0.63	0.62 ±0.05	0.60	0.60	0.60	0.60 ±0.00
Low Ent.	0.33	0.33	0.43	0.37 ±0.06	0.73	0.57	0.60	0.63 ±0.09	0.60	0.60	0.60	0.60 ±0.00
UID Metrics (↓)												
Loc. non-uni (2σ)	0.20	0.30	0.23	0.24 ±0.05	0.53	0.57	0.53	0.54 ±0.02	0.63	0.63	0.63	0.63 ±0.00
Loc. uni (2σ)	0.43	0.43	0.37	0.41 ±0.04	0.70	0.73	0.70	0.71 ±0.02	0.67	0.67	0.73	0.69 ±0.04
Loc. non-uni (3σ)	0.20	0.30	0.23	0.24 ±0.05	0.53	0.57	0.53	0.54 ±0.02	0.63	0.63	0.63	0.63 ±0.00
Loc. uni (3σ)	0.43	0.43	0.37	0.41 ±0.04	0.70	0.70	0.67	0.69 ±0.02	0.67	0.67	0.73	0.69 ±0.04
Glob. non-uni	0.37	0.30	0.43	0.37 ±0.07	0.70	0.63	0.63	0.66 ±0.04	0.70	0.70	0.70	0.70 ±0.00
Glob. uni	0.30	0.37	0.33	0.33 ±0.03	0.60	0.73	0.67	0.67 ±0.07	0.67	0.67	0.63	0.66 ±0.02

Table 12: **Sample Size Analysis.** Results across different sampling strategies (`Sample by 3`, `5`, and `10`) on `AIME2025`. Performance is reported across seeds 42, 1234, and 2025, with averages and standard deviations (Avg ± Std).

Methods (↓)	Sample by 3				Sample by 5				Sample by 10			
	42	1234	2025	Avg ± Std	42	1234	2025	Avg ± Std	42	1234	2025	Avg ± Std
Mean Acc.	0.67	0.67	0.67	0.67±0.00	0.66	0.66	0.68	0.67±0.01	0.68	0.67	0.69	0.68±0.01
Self-Cert.	0.70	0.70	0.70	0.70±0.00	0.63	0.63	0.63	0.63±0.00	0.57	0.63	0.67	0.62±0.04
High Conf.	0.63	0.63	0.63	0.63±0.00	0.60	0.60	0.60	0.60±0.00	0.57	0.53	0.60	0.57±0.03
Low Ent.	0.63	0.63	0.63	0.63±0.00	0.60	0.60	0.60	0.60±0.00	0.50	0.57	0.60	0.56±0.04
UID Metrics (↓)												
Loc. non-uni (2σ)	0.63	0.63	0.63	0.63±0.00	0.63	0.63	0.63	0.63±0.00	0.50	0.50	0.60	0.53±0.05
Loc. uni (2σ)	0.73	0.73	0.73	0.73±0.00	0.67	0.67	0.73	0.69±0.03	0.70	0.73	0.73	0.72±0.02
Loc. non-uni (3σ)	0.63	0.63	0.63	0.63±0.00	0.63	0.63	0.63	0.63±0.00	0.50	0.50	0.60	0.53±0.05
Loc. uni (3σ)	0.73	0.73	0.73	0.73±0.00	0.67	0.67	0.73	0.69±0.03	0.70	0.73	0.73	0.72±0.02
Glob. non-uni	0.70	0.70	0.70	0.70±0.00	0.70	0.70	0.70	0.70±0.00	0.70	0.67	0.73	0.70±0.03
Glob. uni	0.70	0.70	0.70	0.70±0.00	0.67	0.67	0.63	0.66±0.02	0.63	0.63	0.63	0.63±0.00

Table 13: **UID-based selection with response-level aggregation.** We use Borda Voting (majority voting under matched sampling budgets) adopted from Kang et al. (2025) on AIME 2025.

Method (↓)	`Qwen3-8B`	`DS-R1-Distill-Qwen-7B`	`DS-R1-Distill-Llama-8B`
Majority Voting	0.67	0.43	0.27
UID Metrics (↓)			
Loc. uni	**0.73**	0.43	**0.40**
Loc uni + Borda	**0.73**	0.43	0.37
Glob. non-uni	0.70	**0.47**	0.33
Glob. non-uni + Borda	0.70	**0.47**	0.33

Settings	**Content**

Communication Prompting	`Please answer the following math question as if you are explaining it to a listener.`
	`Communicate your reasoning clearly and naturally.`
	`You should provide your final answer in the format \boxed{YOUR_ANSWER}.`
	`\n\n Question:\n{question}\n`

Naive Prompting	`Please answer the following math question.`
	`You should provide your final answer in the format \boxed{YOUR_ANSWER}.`
	`\n\n Question:\n{question}\n`

Table 14: Prompts used for Communicative vs. Naive Prompting.

```
<think>\nOkay, so I need to find all positive integers n such that n + 2 divides the product 3(n + 3)(n\u00b2 + 9), and then sum all those n's. Hmm, let me think about how to approach this.

\n\nFirst, let me restate the problem in mathematical terms. We need to find all positive integers n where (n + 2) divides 3(n + 3)(n\u00b2 + 9). In other words, 3(n + 3)(n\u00b2 + 9) divided by (n + 2) should be an integer. So, mathematically, we can write this as:\n\n(n + 2) | 3(n + 3)(n\u00b2 + 9)\n\n Which means that 3(n + 3)(n\u00b2 + 9) mod (n + 2) = 0.

\n\nI remember that when dealing with divisibility conditions, sometimes substituting variables or using modular arithmetic can help. Maybe I can use the fact that if n \u2261 -2 mod (n + 2), then I can substitute n = -2 into the expression to find the remainder. Let me try that.

\n\n Let me consider the expression 3(n + 3)(n\u00b2 + 9) modulo (n + 2). If I substitute n \u2261 -2, then:\n\nFirst, compute each part modulo (n + 2):\n\nn \u2261 -2, so n + 3 \u2261 -2 + 3 = 1.

\n\n Similarly, n\u00b2 + 9 \u2261 (-2)\u00b2 + 9 = 4 + 9 = 13.\n\n Therefore, the entire expression becomes 3 * 1 * 13 = 39. Therefore, the remainder when 3(n + 3)(n\u00b2 + 9) is divided by (n + 2) is 39.

\n\n But since (n + 2) divides the product, this remainder must be zero. Therefore, 39 must be congruent to 0 modulo (n + 2). Which implies that (n + 2) divides 39.

\n\nSo, n + 2 is a positive divisor of 39. Since n is a positive integer, n + 2 must be an integer greater than or equal to 3 (since n \u2265 1, so n + 2 \u2265 3). Therefore, we need to find all positive divisors of 39 that are \u2265 3, then subtract 2 to get n, and then sum all such n's. (. . .)
```

Figure 17: Reasoning trace with the highest variance score, with rich reasoning. `math_equal` is True

```
Then:\n\nFrom 3x_{k+1} = x_k + z_k - 1. \n\nBut z_{k+1} = 1/x_{k+1} = 3 / (x_k + z_k - 1). \n\nNot sure if helpful. \n\nAlternatively, maybe consider the recurrence relation for z_k:\n\nIf z_k = 1/x_k, then:\n\nFrom 3x_{k+1} = x_k + z_k - 1, so x_{k+1} = (x_k + z_k - 1)/3. Then z_{k+1} = 3 / (x_k + z_k - 1). \n\nBut since z_k = 1/x_k, maybe express in terms of z_k:\n\nLet me write x_k = 1/z_k. Then:\n\nz_{k+1} = 3 / (1/z_k + z_k - 1) = 3 / ( (1 + z_k\u00b2 - z_k)/z_k ) = 3z_k / (z_k\u00b2 - z_k + 1). \n\nSo z_{k+1} = 3z_k / (z_k\u00b2 - z_k + 1). \n\nHmm. Not sure if this helps. But maybe there's a pattern here. Let me check for the first few terms.\n\nGiven that x\u2081 = 25/11, so z\u2081 = 11/25. \n\nThen z\u2082 = 3*(11/25) / [ ( (11/25)^2 - (11/25) + 1 ) ] \n\nLet me compute denominator:\n\n( (121/625) - (11/25) + 1 ) = (121/625 - 275/625 + 625/625 ) = (121 - 275 + 625)/625 = (471)/625. \n\nTherefore z\u2082 = 3*(11/25) / (471/625) = (33/25) * (625/471) = (33 * 25)/471 = 825 / 471. Simplify: divide numerator and denominator by 3: 275 / 157. Which is the reciprocal of x\u2082. Indeed, x\u2082 = 157/275, so z\u2082 = 275/157. \n\nSo z\u2082 = 275/157. Then z\u2083 = 3*z\u2082 / (z\u2082\u00b2 - z\u2082 + 1). Let me compute that. \n\nFirst compute z\u2082\u00b2: (275/157)^2 = 75625 / 24649. Then subtract z\u2082: 75625 / 24649 - 275/157. Let me convert 275/157 to denominator 24649: 275/157 = (275 * 157)/24649. Wait, 157*157 = 24649, so 275/157 = (275*157)/24649. Wait, but that would be 275*157. Let me compute that:\n\n275 * 157. Let me compute 275*(150 + 7) = 275*150 + 275*7 = 41250 + 1925 = 43175. Therefore, 275/157 = 43175 / 24649. \n\nSo z\u2082\u00b2 - z\u2082 + 1 = (75625 - 43175 + 24649)/24649. Let me compute numerator: 75625 - 43175 = 32450; 32450 + 24649 = 57099. So denominator is 24649. Therefore z\u2082\u00b2 - z\u2082 + 1 = 57099 / 24649. \n\nTherefore z\u2083 = 3*z\u2082 / (57099 / 24649) ) = 3*(275/157) * (24649 / 57099). Let me compute this. \n\nFirst, note that 24649 is 157\u00b2. And 57099 was from previous steps. Wait, earlier when computing x\u2083, we had 57099 / 129525, but maybe not. Anyway, let me compute:\n\nz\u2083 = 3 * (275/157) * (24649 / 57099). \n\nNote that 24649 = 157\u00b2, so:\n\nz\u2083 = 3 * 275 * 157\u00b2 / (157 * 57099) ) = 3 * 275 * 157 / 57099. \n\nSimplify numerator and denominator:\n\nLet me compute 3*275 = 825. So numerator is 825 * 157. Let me compute that:\n\n825 * 157. Let me do 800*157 + 25*157 = 125600 + 3925 = 129525. \n\nSo numerator is 129525. Denominator is 57099. Therefore z\u2083 = 129525 / 57099. Simplify this fraction. Let me check GCD(129525, 57099). \n\nUsing Euclidean algorithm:\n\nGCD(129525, 57099). \n\nDivide 129525 by 57099: 2 times, remainder 129525 - 2*57099 = 129525 - 114198 = 15327.\n\nGCD(57099, 15327). \n\nAs before, earlier steps showed that GCD(57099, 15327) was 3. Wait, earlier when we had GCD(57099, 129525) was 3. Let me check again:\n\nWait, GCD(129525, 57099) = GCD(57099, 15327). Then GCD(57099, 15327). \n\nEarlier steps showed that GCD(57099, 15327) was 3. So divide numerator and denominator by 3:\n\n129525 / 3 = 43175, 57099 / 3 = 19033. Therefore z\u2083 = 43175 / 19033. Which is the reciprocal of x\u2083. Indeed, x\u2083 = 19033/43175, so z\u2083 = 43175/19033. \n\nSo this seems like the pattern is that z_{k} is the reciprocal of x_k.
```

Figure 18: Reasoning trace with the lowest variance score, showing that the model is stuck. `math_equal` is False

n\n**Another possibility**: b=24, d=-54. 24*(-54)=-1296.\n\na + c=12.\n\nad + bc= a*(-54) + c*24= -54a +24c= -432.\n\nDivide by 6: -9a +4c= -72.\n\nFrom a + c=12 \u2192 c=12 -a.\n\nSubstitute into -9a +4(12 -a) =-72 \u2192 -9a +48 -4a =-72 \u2192 -13a= -120 \u2192 a=120/13. Not integer. **Not good.\n\nAnother possibility:** b= -36, d=36. Then b*d= -36*36= -1296.\n\na + c=12.\n\nad + bc= a*36 + c*(-36)=36a -36c= -432 \u2192 36(a - c)= -432 \u2192 a - c= -12.\n\nAlso, a + c=12.\n\nSolving:\n\na + c=12.\n\na - c= -12.\n\nAdding: 2a=0 \u2192 a=0, c=12.\n\nNow, check ac + b + d=0*12 + (-36) +36=0, which is not 144. So, no.\n\nAnother possibility: b= 144, d= -9. 144*(-9)= -1296.\n\na + c=12.\n\nad + bc= a*(-9) + c*144= -9a +144c= -432.\n\nDivide by 9: -a +16c= -48.\n\nAlso, a + c=12.\n\nSo, a=12 -c.\n\nSubstitute into -a +16c= -48: -(12 -c) +16c= -48 \u2192 -12 +c +16c= -48 \u2192 17c= -36 \u2192 c= -36/17. Not integer.\n\nAnother try: b= 54, d= -24. 54*(-24)= -1296.\n\na + c=12.\n\nad + bc= a*(-24) + c*54= -24a +54c= -432 \u2192 divide by 6: -4a +9c= -72.\n\nFrom a + c=12 \u2192 a=12 -c.\n\nSubstitute into -4(12 -c) +9c= -72 \u2192 -48 +4c +9c= -72 \u219213c= -24 \u2192 c= -24/13. **Not integer.\n\nHmm, perhaps I should try** b= 48, d= -27. 48*(-27)= -1296.\n\na + c=12.\n\nad + bc= a*(-27) + c*48= -27a +48c= -432 \u2192 divide by 3: -9a +16c= -144.\n\nAlso, a=12 -c.\n\nSubstitute: -9(12 -c) +16c= -144 \u2192 -108 +9c +16c= -144 \u219225c= -36 \u2192c= -36/25. **Not integer.\n\nThis is getting too time-consuming. Maybe I should try a different approach.\n\nAlternatively, perhaps there is a substitution I can make.**\n\nWe have from equation 1: (h -6)^2 +k^2=36.\n\nAnd from equation 2: (h +6)(k +6)=72.\n\nLet me let u = h +6 and v =k +6.\n\nThen, equation 2 becomes u*v=72.\n\nEquation 1 becomes (h -6)^2 +k^2=36.\n\nExpress h and k in terms of u and v:\n\nh = u -6, k= v -6.\n\nSo, equation 1 becomes:\n\n(u -6 -6)^2 + (v -6)^2 =36 \u2192 (u -12)^2 + (v -6)^2=36.\n\nSo, equation 1: (u -12)^2 + (v -6)^2=36.\n\nAnd equation 2: u*v=72.\n\nSo, we have:\n\n(u -12)^2 + (v -6)^2=36.\n\nAnd u*v=72.\n\nWe can expand equation 1:\n\nu\u00b2 -24u +144 +v\u00b2 -12v +36=36 \u2192 u\u00b2 + v\u00b2 -24u -12v +180=36 \u2192 u\u00b2 +v\u00b2 -24u -12v= -144.\n\nBut since u*v=72, perhaps we can find u and v such that u*v=72 and u\u00b2 +v\u00b2=24u +12v -144.\n\nAlternatively, we can write u\u00b2 +v\u00b2 =24u +12v -144.\n\nBut from u*v=72, perhaps express v=72/u.\n\nSubstitute into u\u00b2 + (72/u)^2 =24u +12*(72/u) -144.\n\nMultiply both sides by u\u00b2 to eliminate denominators:\n\nu^4 + 5184 =24u^3 +12*72 u -144u\u00b2.\n\nSimplify:\n\nu^4 +5184 =24u^3 +864u -144u\u00b2.\n\nBring all terms to one side:\n\nu^4 -24u^3 +144u\u00b2 -864u +5184=0.\n\nHmm, this seems complicated, but perhaps we can factor it.\n\nLet me try possible rational roots. Possible roots are factors of 5184 over 1, so possible roots include \u00b11, \u00b12, ..., up to \u00b15184.\n\nTesting u=12:\n\n12^4=20736, -24*12^3= -24*1728= -41472, 144*12^2=144*144=20736, -864*12= -10368, +5184.\n\nSo, 20736 -41472 +20736 -10368 +5184.\n\nCompute step by step:\n\n20736 -41472= -20736.\n\n-20736 +20736=0.\n\n0 -10368= -10368.\n\n-10368 +5184= -5184 \u22600.\n\nu=6:\n\n6^4=1296, -24*6^3= -24*216= -5184, 144*36=5184, -864*6= -5184, +5184.\n\nSo, 1296 -5184 +5184 -5184 +5184.\n\nCompute:\n\n1296 -5184= -3888.\n\n-3888 +5184=1296.\n\n1296 -5184= -3888.\n\n-3888 +5184=1296 \u22600.\n\nu= -6:\n\n(-6)^4=1296, -24*(-6)^3= -24*(-216)=5184, 144*36=5184, -864*(-6)=5184, +5184.\n\nSo, 1296 +5184 +5184 +5184 +5184.\n\nWait, wait, no. Wait, the equation is u^4 -24u^3 +144u^2 -864u +5184=0.\n\nPlugging u=-6:\n\n(-6)^4=1296.\n\n-24*(-6)^3= -24*(-216)=5184.\n\n144*(-6)^2=144*36=5184.\n\n-864*(-6)=5184.\n\n+5184.\n\nSo, sum: 1296 +5184=6480, +5184=11664, +5184=16848, +5184=22032. Not zero.\n\nu=8:\n\n8^4=4096, -24*512= -12288, 144*64=9216, -864*8= -6912, +5184.\n\nCompute:\n\n4096 -12288= -8192.\n\n-8192 +9216=1024.\n\n1024 -6912= -5888.\n\n-5888 +5184= -704 \u22600.\n\nu=9:\n\n9^4=6561, -24*729= -17496, 144*81=11664, -864*9= -7776, +5184.\n\nCompute:\n\n6561 -17496= -10935.

Figure 19: Reasoning trace with the highest number of spikes and falls, with a lot of redundancy and confusions. `math_equal` is True

Okay, so I've got this math problem to solve. It says: Find the sum of all integer bases \\(b > 9 \\) for which \\(17_b \\) is a divisor of \\(97_b \\). Hmm, let me try to figure this out step by step.\n\nFirst, I remember that when numbers are written in a base \\(b \\), each digit represents a power of \\(b \\). So, for example, \\(17_b \\) would be equal to \\(1 \\times b^1 + 7 \\times b^0 \\) in decimal, which is \\(b + 7 \\). Similarly, \\(97_b \\) would be \\(9 \\times b^1 + 7 \\times b^0 \\), so that's \\(9b + 7 \\) in decimal.\n\nThe problem is asking for bases \\(b > 9 \\) where \\(17_b \\) divides \\(97_b \\). That means \\(b + 7 \\) should be a divisor of \\(9b + 7 \\). So, mathematically, I can write this as:\n\n\\(b + 7 \\) divides \\(9b + 7 \\).\n\nNow, when a number \\(a \\) divides another number \\(c \\), it means that \\(c \\) divided by \\(a \\) leaves no remainder. Another way to put that is \\(c \\mod a = 0 \\). So, I can write:\n\n\\(9b + 7 \\mod (b + 7) = 0 \\).\n\nAlternatively, \\(9b + 7 \\) divided by \\(b + 7 \\) should result in an integer. Maybe I can perform the division to see what that gives.\n\nLet me try dividing \\(9b + 7 \\) by \\(b + 7 \\) using polynomial long division or substitution. Let's see, if I think of it as a division:\n\n\\((9b + 7) \\div (b + 7) \\).\n\nLet me set it up:\n\nHow many times does \\(b + 7 \\) go into \\(9b + 7 \\)? Well, \\(b \\) goes into \\(9b \\) exactly 9 times. So, I multiply \\(b + 7 \\) by 9, which gives \\(9b + 63 \\), and subtract that from \\(9b + 7 \\).\n\nSo, subtracting: \\((9b +7) - (9b + 63) = 0b + (7 - 63) = -56 \\).\n\nSo, the division gives 9 with a remainder of -56.

Figure 20: Reasoning trace with the lowest number of spikes and falls, with simple and concise reasoning. `math_equal` is True

Table 15: τ**-sweep results (**τ **= 1–5).** Performance on AIME2025, BRUMO2025, HMMT2025, and MinervaMath (MM). Results are from seed 42.

	DS-R1-Distill-Qwen-7B				DS-R1-Distill-Llama-8B				Qwen3-8B			
Methods (↓)	AIME	BRUMO	HMMT	MM	AIME	BRUMO	HMMT	MM	AIME	BRUMO	HMMT	MM
τ **= 2**												
Loc. non-uni	0.23	0.40	0.30	0.27	0.33	0.47	0.17	0.26	0.70	0.70	0.47	0.27
Loc. uni	0.43	0.43	0.27	0.30	0.37	0.50	0.17	0.27	0.73	0.67	0.30	0.30
τ **= 3**												
Loc. non-uni	0.23	0.40	0.17	0.28	0.43	0.43	0.20	0.27	0.70	0.70	0.50	0.28
Loc. uni	0.40	0.47	0.27	0.30	0.33	0.53	0.13	0.28	0.70	0.70	0.37	0.30
τ **= 4**												
Loc. non-uni	0.33	0.50	0.20	0.29	0.43	0.43	0.23	0.28	0.67	0.73	0.47	0.29
Loc. uni	0.43	0.47	0.20	0.31	0.40	0.57	0.20	0.29	0.60	0.67	0.37	0.31
τ **= 5**												
Loc. non-uni	0.30	0.43	0.23	0.30	0.43	0.50	0.20	0.29	0.67	0.70	0.43	0.30
Loc. uni	0.40	0.50	0.20	0.32	0.37	0.53	0.17	0.29	0.63	0.67	0.33	0.32