Revisiting the Uniform Information Density Hypothesis in LLM Reasoning Traces

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large language models (LLMs) often solve problems using step-by-step Chain-of-Thought (CoT) reasoning, yet these intermediate steps are frequently unfaithful or hard to interpret. Inspired by the Uniform Information Density (UID) hypothesis in psycholinguistics – which posits that humans communicate by maintaining a stable flow of information – we introduce entropy-based metrics to analyze the information flow within reasoning traces. Surprisingly, across three challenging mathematical benchmarks, we find that successful reasoning in LLMs is globally non-uniform: correct solutions are characterized by uneven swings in information density, in stark contrast to human communication patterns. This result challenges assumptions about machine reasoning and suggests new directions for designing interpretable and adaptive reasoning models.

1 Introduction

Chain-of-Thought (CoT) reasoning has emerged as a central technique for improving large language models (LLMs) on complex reasoning tasks [1–3]. By generating step-by-step rationales, CoT allows models to decompose problems and produce more interpretable outputs [4, 5]. However, recent studies have highlighted the fragility of this approach [6]. Specifically, despite generating longer reasoning traces, LLMs often fail to generalize, and their intermediate steps can be logically inconsistent or incoherent [7]. This raises an important question: *how can we tell when LLMs are reasoning effectively, rather than merely generating superficially coherent text?*

Human communication provides a potential clue. A psycholinguistic theory suggests that effective communication relies on a uniform flow of information[8, 9], where ideas are expressed at a stable rate to match human cognitive processing limits. When information is delivered too unevenly, understanding breaks down. We hypothesize that a similar principle applies to LLM reasoning: just as humans produce language with balanced information flow, effective reasoning traces may exhibit comparable uniformity. To explore this link, we draw on cognitive science and psycholinguistics; for instance, Bhambri et al. [10] shows that reasoning paths interpretable to humans are also easier for models to generate and learn, suggesting a shared structure between human cognition and machine reasoning. To illustrate, a well-reasoned math solution might show consistent step-level progress, where each step builds smoothly on the previous one, while an incoherent solution might jump between overly trivial and overly complex steps.

We analyze the information flow of LLM-generated reasoning traces on challenging mathematical benchmarks. We define per-step measurements of information density, and examine by answer correctness. Then, we introduce three complementary metrics that quantify the uniformity entire reasoning trace, using entropy-based per-step measurement. Our experiments reveal a clear pattern: unlike human communication, reasoning traces with low global uniformity tend to produce correct answers. This suggests that effective reasoning balances local uniformity and low global uniformity.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Overall, our contributions are threefold:

- ◦ To our knowledge, we are the first to introduce information-theoretic metrics for quantifying reasoning structure at both the step and trace level.
- ◦ We find that reasoning patterns characterized by low global uniformity, correlate with reasoning success on challenging mathematical reasoning benchmarks.
- ◦ We show that deviations from such patterns can serve as an internal signal for predicting failure cases, enabling potential improvements in LLM reasoning and evaluation.

2 Related Work

2.1 Fragility of CoT and the role of individual reasoning steps

CoT prompting improves reasoning but remains fragile [1, 6]. Small, seemingly irrelevant perturbations in the reasoning chain can sharply reduce accuracy [11, 12], suggesting that models often produce the appearance of reasoning rather than logically sound traces [7]. 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 [13]. This raises doubts about the necessity and faithfulness of these step-by-step explanations. Another line of recent 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 [14]. Attribution methods and their frameworks identify and highlight these critical steps, emphasizing the need to understand how individual steps shape outcomes[4, 15, 16]. Despite these advances, there remains no clear interpretation of what constitutes a truly good reasoning pattern.

2.2 Intrinsic signals in LLM reasoning

Research on LLM reasoning has increasingly turned to internal model signals to gain insights into how reasoning unfolds. Many approaches use these signals to improve performance, such as using self-consistency [17], self-certainty [18, 19], or confidence to refine outputs, or using entropy-based measures to encourage diverse reasoning paths [20–23]. We shift focus from controlling reasoning with internal signals to understanding it through their structure. We ground our analysis in long-standing psycholinguistic theory to understand the structure of reasoning itself by revealing how information is introduced, transformed, and propagated through the reasoning process. Our step-level focus provides a deeper understanding of what constitutes a coherent reasoning trace, going beyond prior approaches that emphasize performance gains over interpretability.

3 Exploring the UID Hypothesis in Reasoning Models

3.1 Background: Uniform information density hypothesis

The Uniform Information Density (UID) hypothesis models language as a signal transmitted through a noisy channel with limited capacity [8, 9]. It posits that speakers aim to convey information efficiently without overwhelming the listener's processing resources. 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, we can define surprise as the unexpectedness of a unit, given its previous context. Formally, surprisal is defined as:

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

where $P(u_n|u_{<n})$ is the probability of seeing unit utterance u_n after the earlier sequence $\mathbf{u}_{<n} = [u_1, \ldots, u_{n-1}]$. High surprisal of the unit denotes that it is very unexpected and hard to process for the person receiving the information, while units with lower surprisal are easier to process. In this sense, surprisal can be perceived as information content. To capture the overall cognitive load of a message, we aggregate this surprisal across all units in the sequence. Given a sequence of utterance $\mathbf{u}$, the total processing effort can be expressed as:

$$\text{ProcessingEffort}(u) \propto \sum_{n=1}^{N} s(u_n).$$

If information is concentrated in a few highly surprising units, the receiver experiences sharp spikes in processing difficulty; if it is too sparse, communication becomes inefficient. The high-level intuition of the UID hypothesis is that the most efficient strategy is to distribute surprisal as evenly as possible across the sequence, maintaining a stable level of processing effort. This tendency has been empirically observed across syllables, words, syntax, and discourse.

While UID has been extensively validated in human language, its implications for machine reasoning remain unexplored. LLMs, or more specifically, recent reasoning models such as Deepseek-R1 [24] and Qwen3 [25] generate CoT traces step-by-step, much like 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 effort:

$$\text{ReasoningEffort}(\mathbf{z}) \propto \sum_{n=1}^{N} s(z_n).$$

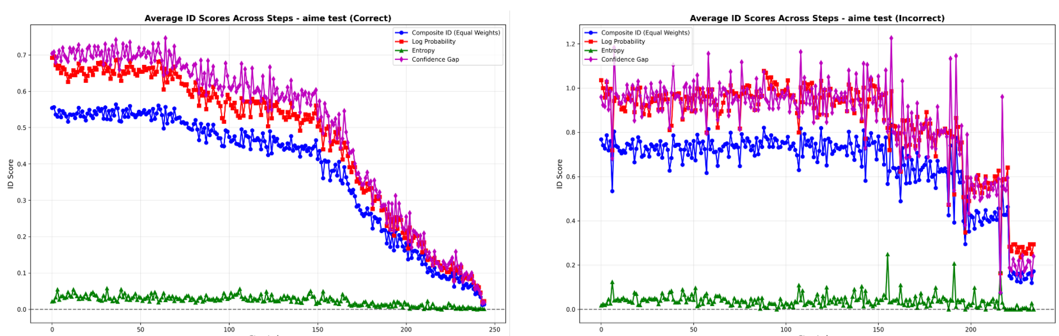
Here, a natural question arises: *does UID hypothesis hold for good reasoning patterns in LLMs?* A smooth, uniform surprisal profile may reflect clear and logical reasoning, while sharp spikes may signal confusion or errors. We extend UID hypothesis beyond psycholinguistics to probe the structure of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

3.2 Preliminary analysis with per-step information density scores of reasoning traces

We start by defining the step-level information density ID_i for a reasoning trace $\mathbf{z} = [z_1, \ldots, z_N]$ with N steps, 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 reasoning steps of a single trace of a reasoning model, Qwen3-8B, by `\n\n`, following Lightman et al. [26]. Then, let $p_t(v)$ be the predictive distribution over the vocabulary at the token position t, and $l_t = \log p_t(x_t)$ the log-probability of the generated token x_t. To characterize ID_i, we consider three metrics over tokens in each step, as defined below.

3.2.1 Three metrics of ID_i

In this work, we consider three metrics for 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 step. Details of the metrics are given in Appendix C.1.



(a) Correct Traces (b) Incorrect Traces

Figure 1: Averaged ID scores ranging of correct and incorrect traces on AIME2025 test set tracked with step-level information density.

3.2.2 Interpretation of the three metrics of ID_i across a reasoning trace

Figure 1 compares the evolution of the three metrics – log-probability LP_i, entropy H_i, and confidence gap D_i – and its composite metric, across reasoning traces for correct and incorrect solutions on AIME2025. For correct traces (Figure 1a), H_i remains consistently low, while LP_i and D_i steadily decrease, forming a smooth trajectory that culminates in a sharp drop of the composite ID_i near the final steps, to ID_i score of 0.0. Incorrect traces (Figure 1b) starts higher, at average of higher than 1.0 ID_i scores and show elevated and unstable LP_i and D_i, with erratic fluctuations and sudden drops.

3.3 Measuring the uniformity of information density in reasoning trace

To measure the uniformity of information density in a reasoning trace, we first clarify what "uniform" means. Prior psycholinguistic theory offers two perspectives [8, 27]. Global uniformity maintains a stable surprisal rate across the trace, while local uniformity smooth, gradual step-level changes.

Grounded in these perspectives, we explore three UID metrics for LLM reasoning traces. (1) Variance measures how much the surprisal values diverges from the mean. High variance means the reasoning process is globally unstable, with large swings in information load across steps. (2) Gini coefficient captures how unevenly the total information is distributed. A high Gini score means a few steps dominate the process, creating potential reasoning bottlenecks. (3) Shannon evenness measures how balanced the information distribution is, normalized to account for sequence length. High Shannon evenness reflects a smooth, well-balanced reasoning process. Together, these metrics distinguish between reasoning where uncertainty (entropy) is globally unstable (high variance) and unevenly concentrated (high Gini, Shannon evenness). Full definitions are in Appendix C.3.

4 Unlike Human Communication, Global Non-uniform Information Distribution Predicts Reasoning Success in LLMs

Table 1: **Main Results**. Accuracy results averaged over three random seeds. The best and second-best scores are bold-faced and underlined, respectively. See Appendix D for more details.

Category	Method	AIME 2025	HMMT 2025	Minerva Math
Baselines	Mean Accuracy	0.673	0.433	0.326
	Self-Certainty	<u>0.689</u>	**0.467**	<u>0.332</u>
	CoT-Decoding	0.678	0.444	0.330
	Highest Confidence	0.633	0.389	0.328
	Lowest Entropy	0.633	0.378	0.331
UID Measurement				
Variance	Highest UID Score (non-uniform)	**0.722**	<u>0.456</u>	**0.342**
	Lowest UID Score (uniform)	0.644	0.433	0.322

Among the ID_i metrics presented, we use entropy to compute the UID score. Among the three UID metrics, global uniformity, measured by variance, emerges as the strongest predictor of reasoning success. Selecting traces with the highest variance (low global uniformity) achieves 0.722 accuracy on AIME and 0.342 on Minerva Math, representing absolute improvements of +4.9% and +1.6% over the best-performing baseline (Self-Certainty: 0.689 on AIME, 0.332 on Minerva Math). On HMMT, high-variance traces reach 0.456 accuracy, which is +2.3% higher than the Mean Accuracy baseline (0.433). These results indicate that reasoning success is closely tied to low global uniformity, where models exhibit large, deliberate swings in information density throughout their thought process rather than maintaining a stable, uniform progression. Variance demonstrates consistent, cross-dataset gains, making it the most reliable signal. Overall, our findings suggest that reasoning is most effective when the model's information flow is globally diverse, allowing it to shift focus dynamically and explore alternative reasoning paths, leading to stronger final answers. Experiment details are in Appendix B.

5 Conclusion

Results show that low global uniformity strongly predicts correct reasoning, while local uniformity exhibits mixed effects. This indicates that while reasoning traces share structural similarities with natural language, their dynamics do not strictly adhere to the UID hypothesis. Instead, effective reasoning appears to rely on irregular, globally non-uniform patterns, reflecting moments of abrupt insight or decisive leaps. These findings highlight that the internal signals embedded in the structure of reasoning traces can offer valuable guidance for model design. Future work could explore how to harness these signals—rather than enforcing strict uniformity—to develop methods that adaptively leverage the natural ebb and flow of reasoning, ultimately improving the robustness and interpretability of reasoning models.

References

[1] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

[2] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. URL https://arxiv.org/abs/2205.11916.

[3] Hyungjoo Chae, Yongho Song, Kai Tzu iunn Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. Dialogue chain-of-thought distillation for commonsense-aware conversational agents, 2023. URL https://arxiv.org/abs/2310.09343.

[4] Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Roscoe: A suite of metrics for scoring step-by-step reasoning, 2023. URL https://arxiv.org/abs/2212.07919.

[5] Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. Receval: Evaluating reasoning chains via correctness and informativeness, 2023. URL https://arxiv.org/abs/2304.10703.

[6] Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang, and Huan Liu. Is chain-of-thought reasoning of llms a mirage? a data distribution lens, 2025. URL https://arxiv.org/abs/2508.01191.

[7] Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025. URL https://arxiv.org/abs/2506.06941.

[8] Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. Revisiting the uniform information density hypothesis, 2021. URL https://arxiv.org/abs/2109.11635.

[9] Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Wilcox, Mario Giulianelli, and Alex Warstadt. Surprise! uniform information density isn't the whole story: Predicting surprisal contours in long-form discourse, 2024. URL https://arxiv.org/abs/2410.16062.

[10] Siddhant Bhambri, Upasana Biswas, and Subbarao Kambhampati. Do cognitively interpretable reasoning traces improve llm performance?, 2025. URL https://arxiv.org/abs/2508.16695.

[11] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2025. URL https://arxiv.org/abs/2410.05229.

[12] Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. Large language models are in-context semantic reasoners rather than symbolic reasoners, 2023. URL https://arxiv.org/abs/2305.14825.

[13] 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. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

[14] Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which llm reasoning steps matter?, 2025. URL https://arxiv.org/abs/2506.19143.

[15] Skyler Wu, Eric Meng Shen, Charumathi Badrinath, Jiaqi Ma, and Himabindu Lakkaraju. Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions, 2023. URL https://arxiv.org/abs/2307.13339.

[16] Eric Bigelow, Ari Holtzman, Hidenori Tanaka, and Tomer Ullman. Forking paths in neural text generation, 2024. URL https://arxiv.org/abs/2412.07961.

[17] 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. Ttrl: Test-time reinforcement learning, 2025. URL https://arxiv.org/abs/2504.16084.

[18] Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty, 2025. URL https://arxiv.org/abs/2502.18581.

[19] Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards, 2025. URL https://arxiv.org/abs/2505.19590.

[20] Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right question is already half the answer: Fully unsupervised llm reasoning incentivization, 2025. URL https://arxiv.org/abs/2504.05812.

[21] Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning, 2025. URL https://arxiv.org/abs/2505.15134.

[22] Zitian Gao, Lynx Chen, Haoming Luo, Joey Zhou, and Bryan Dai. One-shot entropy minimization, 2025. URL https://arxiv.org/abs/2505.20282.

[23] Dongseok Lee, Jimyung Hong, Dongyoung Kim, and Jaehyung Kim. Training-free llm verification via recycling few-shot examples, 2025. URL https://arxiv.org/abs/2506.17251.

[24] 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. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[25] 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. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[26] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

[27] MX Collins. Information density and dependency length as complementary cognitive models. *Journal of Psycholinguistic Research*, 43(5):651–681, October 2014. doi: 10.1007/s10936-013-9273-3. URL https://doi.org/10.1007/s10936-013-9273-3.

[28] Xuezhi Wang and Denny Zhou. Chain-of-thought reasoning without prompting, 2024. URL https://arxiv.org/abs/2402.10200.

A Implementation Details

A.1 Hyperparameters and GPU Setup.

For all our main results, we use Qwen3-8B thinking mode. We set the temperature to 0.6, top-p to 0.95, and top-k to 20, as stated in the Qwen3 Technical Report. We use 4xA6000 GPUs for all our experiments.

B Experiment Setup

B.1 Evaluation and Benchmarks

We use accuracy for evaluation, and use three particularly challenging mathematical benchmarks, AIME2025, HMMT2025, and Minerva Math. We sample each questions five times before the final evaluation.

B.1.1 AIME 2025.

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 evaluat an LLM's mathematical reasoning by requiring a single correct integer answer. The set used in our analysis contains of 30 questions.

B.1.2 HMMT 2025.

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

B.1.3 Minverva Math.

The Minerva Math benchmark 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.

B.2 Baseline Implementation

We re-implemeted all logic using vllm, unlike some of the codes initially released.

B.2.1 Mean Accuracy

This is the mean accuracy of all answers, which are sampled by 5 for each question.

B.2.2 Self-Certainty

This is the implementation of Kang et al. [18], where it first measures the confidence of each sampled answer and select one via borda-voting.

B.2.3 CoT-Decoding

This is the implementation of the path selection strategy used in Wang and Zhou [28]. This method, called CoT-decoding, identifies reasoning paths that contain CoT steps by measuring the model's confidence in the final answer tokens. It computes the average probability margin between the top-1 and top-2 tokens during answer decoding, denoted as Δ. Decoding paths with higher Δ values are strongly correlated with correct CoT reasoning, enabling reliable extraction of CoT paths even when they are not the most probable or majority paths.

B.2.4 Highest Confidence

This selects the path with the highest overall token confidence in the reasoning trace.

B.2.5 Lowest Entropy

This selects the path with the lowest overall token entropy in the reasoning trace.

C Details of ID and UID Operationalizations

C.1 Details of ID_i metrics

Log-probability LP_i of a step is the average token log-probability over step i

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

Given 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

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

Using the three metrics above, we build a composite ID_i score, defined as

$$ID_i = w_{LP} LP_i - w_H H_i + w_D D_i$$

where all weights are equally set as $1/3$ at our current setting.

C.2 Averaged ID scores of correct and incorrect traces on HMMT2025 and Minerva Math

C.3 Mathematical Formulations of UID Operationalizations

Let a reasoning trace z have N steps. Define the (non-negative) information density vector

$$\mathrm{UID}(z) = \mathbf{u} = (ID_1, ID_2, \ldots, ID_N), \quad ID_i \geq 0$$

C.3.1 Operationalizing $UID(z)$ as Variance

To bound $ID_i \in [0, 1]$, u is normalized with min-max normalization to map the non-negative sequence to $[0, 1]$.

Let

$$m = \min_{1 \le i \le N_{\min}} \mathrm{ID}_i, \quad M = \min_{1 \le i \le N_{\max}} \mathrm{ID}_i$$

Then, the normalized ID'_i values are

$$\mathrm{ID}'_i = \frac{\mathrm{ID}_i - m}{M - m}, \quad i = 1, \ldots, N.$$

and their corresponding vector form for $UID'(z) = \tilde{\mathbf{u}} = (\mathrm{ID}'_1, \ldots, \mathrm{ID}'_N)$:

Define

$$S = \sum_{i=1}^{T} ID'_i, \qquad \mu = \frac{1}{T}\sum_{i=1}^{T} ID'_i, \qquad p_i = \frac{ID'_i}{S} \ \text{(when } S > 0\text{)}$$

Then, the population variance of the entries are

$$\mathrm{Var}(\tilde{\mathbf{u}}) = \frac{1}{T}\sum_{i=1}^{T} \left(ID'_i - \mu\right)^2$$

C.3.2 Operationalizing $UID(z)$ as Gini Coefficient

Sort ID_i values from smallest to largest, where the sorted $ID'_1 \le \cdots \le ID'_N$ Then, the Gini coefficient can be calculated as

$$G(\mathbf{u}) = \frac{1}{\mu T}\sum_{i=1}^{T}(2i - T - 1)\, ID_i, \quad (\mu > 0).$$

C.3.3 Operationalizing $UID(z)$ as Shannon Evenness

First compute Shannon entropy of the probability normalization $p_i = \frac{ID_i}{S}$.

$$H(\mathbf{u}) = -\sum_{i=1}^{T} p_i \ln p_i \quad (S > 0),$$

with maximum $H_{\max} = \ln N$. Then, Shannon evenness can be calculated as

$$J'(\mathbf{u}) = \frac{H(\mathbf{u})}{\ln T} \in [0, 1].$$

D Additional Experiment Results

D.1 Main Results with Different Seeds

While other measures of local uniformity such as the Gini coefficient and Shannon evenness also show competitive performance, their effectieveness is limited and more dataset-dependent.

D.2 Scaling models amplifies the role of global non-uniformity

As shown in Table 3, scaling model size from 1.7B to 8B reveals a clear trend where variance becomes an increasingly strong predictor of reasoning success, outperforming all baselines at 8B.

Table 2: **Main results across various seeds**. Performance based on accuracy across three mathematical benchmarks: AIME 2025, HMMT 2025, and Minerva Math. The final sub-column will report the mean and standard deviation across seeds. The best and second-best scores are bold-faced and underlined, respectively.

Category	Method	AIME 2025				HMMT 2025				Minerva Math			
		Seed 42	Seed 1234	Seed 2025	Avg	Seed 42	Seed 1234	Seed 2025	Avg	Seed 42	Seed 1234	Seed 2025	Avg
Baselines	Mean Accuracy	0.680	0.680	0.660	0.673	0.453	0.420	0.427	0.433	0.329	0.325	0.324	0.326
	Self-Certainty	0.700	0.633	**0.733**	0.689	0.433	0.500	0.467	0.467	0.346	0.331	0.320	0.332
	CoT-Decoding	0.667	0.667	0.700	0.678	0.500	0.400	0.433	0.444	0.335	0.335	0.320	0.330
	Highest Confidence	0.667	0.600	0.633	0.633	0.400	0.367	0.400	0.389	0.349	0.320	0.316	0.328
	Lowest Entropy	0.667	0.600	0.633	0.633	0.367	0.367	0.400	0.378	0.349	0.320	0.324	0.331
Three Measures of UID													
Variance	Highest UID Score (non-uniform)	0.700	0.733	0.733	**0.722**	0.467	0.433	0.467	0.456	0.338	0.338	0.349	**0.342**
	Lowest UID Score (uniform)	0.633	0.667	0.633	0.644	0.433	0.433	0.433	0.433	0.335	0.319	0.313	0.322
Gini Coefficient	Highest UID Score (non-uniform)	0.667	0.667	0.633	0.656	0.433	0.333	0.467	0.411	0.338	0.316	0.320	0.325
	Lowest UID Score (uniform)	0.667	0.700	0.667	0.678	0.433	0.433	0.367	0.411	0.324	0.320	0.346	0.330
Shannon Evenness	Highest UID Score (uniform)	0.700	0.667	0.667	0.678	0.433	0.367	0.400	0.400	0.320	0.324	0.331	0.325
	Lowest UID Score (non-uniform)	0.633	0.700	0.600	0.644	0.500	0.500	0.433	**0.478**	0.335	0.320	0.320	0.325

Table 3: **Performance across different model sizes.** Performance across Qwen3-1.7B, 4B, and 8B on AIME 2025. Each model is evaluated with three random seeds (42, 1234, 2025). The last column within each model block shows the average across seeds. The best and second-best scores are bold-faced and underlined, respectively.

Baselines		1.7B				4B				8B			
		Seed 42	Seed 1234	Seed 2025	Avg	Seed 42	Seed 1234	Seed 2025	Avg	Seed 42	Seed 1234	Seed 2025	Avg
	Mean Accuracy	0.367	0.353	0.353	0.358	0.680	0.653	0.667	0.667	0.680	0.680	0.660	0.673
	Self-Certainty	0.367	0.400	0.400	0.389	0.633	0.767	0.667	0.689	0.700	0.633	0.733	0.689
	CoT-Decoding	0.333	0.300	0.300	0.311	0.767	0.633	0.700	0.700	0.667	0.667	0.700	0.678
	Highest Confidence	0.333	0.367	0.367	0.356	0.600	0.567	0.633	0.600	0.667	0.600	0.633	0.633
	Lowest Entropy	0.333	0.367	0.367	0.356	0.567	0.567	0.633	0.589	0.667	0.600	0.633	0.633
Three Measures of UID													
Variance	Highest UID Score (non-uniform)	0.433	0.333	0.333	**0.366**	0.667	0.667	0.700	**0.678**	0.700	0.733	0.733	**0.722**
	Lowest UID Score (uniform)	0.267	0.367	0.367	0.334	0.633	0.667	0.633	0.644	0.633	0.667	0.633	0.644
Gini Coefficient	Highest UID Score (non-uniform)	0.300	0.400	0.400	**0.367**	0.667	0.700	0.667	0.678	0.667	0.667	0.633	0.656
	Lowest UID Score (uniform)	0.367	0.300	0.300	0.322	0.700	0.667	0.700	**0.689**	0.667	0.700	0.667	**0.678**
Shannon Evenness	Highest UID Score (uniform)	0.367	0.267	0.267	0.300	0.667	0.567	0.633	0.622	0.700	0.667	0.667	**0.678**
	Lowest UID Score (non-uniform)	0.300	0.400	0.400	**0.367**	0.667	0.667	0.667	**0.667**	0.633	0.700	0.600	0.644