ENTROPY IS NOT FREQUENCY: AUDITING THE 80/20 STRUCTURE OF REASONING TOKENS AT SCALE

Junjie Nian
College of Computer Science and Artificial Intelligence
Fudan University
Shanghai, China
github.com/JunjieNian/80-20-rule-analysis

ABSTRACT

High-entropy minority tokens have been proposed as "forking tokens" that steer chain-of-thought reasoning and concentrate the useful gradient signal in reinforcement learning. We independently audit the observational side of this claim across 51 uncertainty caches spanning four model configurations and seven reasoning or coding benchmarks. The committed aggregate contains 539,466 trajectories, 1,116,425,428 token positions, and 43,529 decoded token types. Our first result is methodological: instance-level counting and token-type averaging answer different questions and produce different tails. The qualitative majority-low-entropy/minority-high-entropy shape is robust, but the 80th-percentile cutoff is not universal; mean entropy across token types has a cutoff of 1.219 in the auditable aggregate, rather than the 0.672 position-level threshold reported for the original corpus. After orienting five cached metrics consistently, their top-20% high-uncertainty type sets have mean pairwise Jaccard overlap 0.6932, with pairwise values from 0.5362 to 0.8636. Entropy and frequency are only weakly associated: Pearson $r = -0.0165$ on raw counts, Spearman $\rho = -0.2959$, and Pearson $r = -0.3464$ after log-transforming frequency. An inverse P99 analysis identifies tokens such as *Let*, *Since*, and *Therefore* that are both very frequent and high entropy, demonstrating that frequency cannot replace uncertainty. We reproduce a distributional pattern, not the original training intervention: no reinforcement-learning experiment is run, so causal claims about gradient masking remain attributable to prior work. The project releases analysis code, compact figures, and an explicit map from every claim to its statistical unit.

1 INTRODUCTION

Long chain-of-thought (CoT) traces contain many locally predictable tokens and a smaller number of positions where the next-token distribution opens into several plausible continuations. The original *Beyond the 80/20 Rule* study calls these positions *forking tokens* and reports two connected findings (Wang et al., 2025). First, more than half of one analyzed corpus has entropy below 10^{-2}, while only 20% of token positions exceed entropy 0.672. Second, applying policy-gradient updates only to the high-entropy minority can match or surpass full-token RLVR training, especially at larger model scales.

The training result is causal and expensive to reproduce. The distributional result is easier to inspect but also easier to misinterpret. Three distinct statements are often collapsed:

1. 20% of *positions* have entropy above the position-level 80th percentile;
2. some decoded *token types* have high average entropy;
3. high-entropy token types are rare or low-frequency.

Only the first statement follows mechanically from percentile thresholding. The second changes the unit of analysis, and the third introduces a separate frequency variable. A common token such as *the*

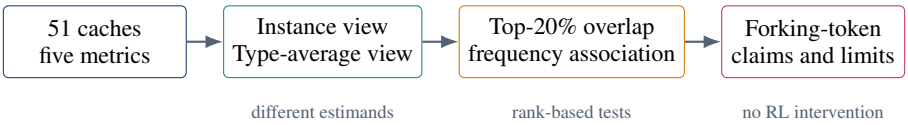

Figure 1: Audit design. The central methodological choice is the unit of aggregation: token positions and token types should not be interpreted interchangeably.

can appear millions of times, populate both uncertainty tails at the instance level, and still have an unremarkable mean. A rare token can have an extreme mean after a handful of observations. Without separating these estimands, word clouds can turn a useful structural observation into a claim about intrinsic word importance.

This paper audits the entropy pattern through two explicitly different views. The *instance view* treats every generated position as an observation and asks which decoded tokens often occur in the high- or low-uncertainty tail. The *type-average view* aggregates every occurrence of a token ID before ranking types. We then compare entropy with four other cached uncertainty signals, quantify the entropy–frequency association, invert the conditioning direction to search for frequent but uncertain types, and perform a small multilingual extension.

Our conclusions are deliberately narrower than the original RLVR result. The minority-high-entropy pattern recurs across a far larger and more heterogeneous cache collection, but its numerical threshold shifts. Five metrics agree substantially but not perfectly. Frequency has a weak negative rank association with entropy and cannot serve as a proxy. These findings clarify what can be concluded from cached generations alone and what still requires intervention.

2 Background

2.1 Token entropy and reasoning forks

For the model distribution p_t over vocabulary items at generation position t, Shannon entropy is

$$H_t = -\sum_j p_{t,j} \log p_{t,j}. \tag{1}$$

Low entropy indicates that probability mass is concentrated on a small set of continuations; high entropy indicates several plausible next tokens. In a CoT trace, a high-entropy position can coincide with choosing a variable, revising a plan, introducing a case, or moving between discourse states. It can also reflect lexical ambiguity, tokenization, noise, or domain shift. Entropy is therefore a candidate marker of a reasoning fork, not a semantic label for one.

Wang et al. (2025) operationalize forking tokens as a high-entropy fraction of each training batch and show that restricting RLVR gradients to this fraction preserves or improves benchmark performance. Their 32B model gains 7.71 points on AIME 2024 and 11.04 points on AIME 2025 over full-gradient DAPO. Conversely, training only the lowest-entropy 80% causes substantial degradation. Our work does not rerun these experiments; it tests whether the descriptive entropy structure and token identities remain stable in a broader cache collection.

2.2 Alternative uncertainty signals

The cache stores Entropy, Confidence, Gini, LogProb, and SelfCert at each token position. Entropy and Gini increase with uncertainty in the stored convention; Confidence, LogProb, and SelfCert increase with certainty. All comparisons orient the metrics before selecting a high-uncertainty set: top values for Entropy and Gini, bottom values for the remaining three. Neuron Agreement Decoding uses several of these external signals as baselines while motivating richer internal activation statistics (Chen et al., 2025).

Agreement between metrics can be measured either as correlation of continuous values or overlap between selected sets. The latter directly matches a fixed token budget. For equal-sized sets A and B,

Jaccard overlap is

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{2}$$

3 Data and Estimands

3.1 Cache collection

The analysis covers 51 caches from four model configurations: DeepSeek-R1-0528-Qwen3-8B, DeepSeek-R1-Distill-Llama-8B, Qwen3-4B-Instruct-2507, and Qwen3-4B-Thinking-2507. The seven task groups are AIME 2024, AIME 2025, GPQA, HumanEval, LiveCodeBench, MBPP, and DeepSeek-8B-S1K. Together, the metadata record 539,466 trajectories.

The auditable aggregate `token_stats_entropy.json` contains 43,529 decoded token types whose occurrence counts sum to 1,116,425,428 positions. This total supersedes an earlier README estimate of "70M+" positions. The aggregate was generated by concatenating token arrays from the 51 caches and incrementing one count for every token occurrence.

The collection is large but not a balanced factorial design. Not every model–dataset pair contributes the same number of caches or trajectories. Aggregate statistics describe this corpus; they do not estimate an equally weighted population of models and tasks.

3.2 Instance-level view

The full method concatenates metric values and token IDs across a cache group, computes the 20th and 80th percentiles over *positions*, and counts how often each decoded token appears in the uncertainty and certainty tails. It answers:

> *Which decoded tokens are frequently observed at high-uncertainty positions?*

A high-frequency token can appear in both tails. This is not a contradiction; its uncertainty is context dependent.

3.3 Token-type-average view

For token type v with occurrences $i = 1, \ldots, n_v$, the average metric is

$$\bar{m}_v = \frac{1}{n_v} \sum_{i=1}^{n_v} m_{v,i}. \tag{3}$$

The type-average method ranks the 43,529 values $\bar{m}_v$ and selects 8,705 types in each 20% tail. It answers:

> *Which token types have unusually high average uncertainty across their observed contexts?*

This view reduces domination by raw frequency but creates a different bias: rare types have unstable averages and are more likely to retain extreme values.

3.4 Forward and inverse conditioning

We call the standard type-average analysis *forward*: select high-entropy types, then inspect their frequencies. The inverse analysis first selects high-frequency types, then ranks them by average entropy. The P99 inverse set contains 436 types occurring at least 236,074 times. Comparing the two views tests whether apparent forking tokens are merely rare.

4 RESULTS

4.1 THE SHAPE REPRODUCES; THE NUMERICAL CUTOFF DOES NOT

The original study analyzes approximately 10^6 positions from Qwen3-8B on AIME 2024/2025 and reports that 50.64% fall below entropy 10^{-2}, while the highest 20% begin above 0.672 (Wang et al., 2025). Our caches reproduce the broad heavy concentration near low entropy with a long upper tail across task and model groups. However, a fixed threshold does not transfer.

In the committed type-average aggregate, the entropy distribution over 43,529 types has 20th percentile 0.135, median 0.594, and 80th percentile 1.219. These are means over types, not raw position values, so they should not equal the original 0.672. Even within one estimand, cutoffs vary with model, dataset, tokenizer, and decoding configuration. The reproducible object is a rank or budget, not one universal entropy constant.

Table 1: Why two valid 80th percentiles need not agree.

Source	Statistical unit	P_{80}	Interpretation
Original study	token positions in its corpus	0.672	position-level high-entropy budget
This audit	mean entropy for 43,529 types	1.219	type-level high-entropy budget

4.2 FIVE METRICS SELECT OVERLAPPING, NON-IDENTICAL TOKEN SETS

Table 2 reports pairwise overlap of high-uncertainty type sets. Mean Jaccard across the ten metric pairs is 0.6932. Entropy aligns most strongly with SelfCert (0.8636) and Gini (0.8269), while Confidence and LogProb have the lowest overlap (0.5362). All values are well above the expected overlap of two independent 20% subsets, but none is one.

Table 2: Jaccard overlap between equal-sized top-20% high-uncertainty token-type sets (8,705 types per metric). Metrics are oriented so that every set denotes uncertainty.

	Entropy	Confidence	Gini	LogProb	SelfCert
Entropy	1.000	0.732	0.827	0.617	0.864
Confidence		1.000	0.666	0.536	0.707
Gini			1.000	0.610	0.787
LogProb				1.000	0.586
SelfCert					1.000

The corresponding certainty sets have mean Jaccard 0.6749. Entropy–Gini certainty overlap is 0.9319, whereas Confidence–LogProb is 0.5001. Thus entropy is a reasonable representative uncertainty signal, but claims of metric redundancy depend on tail, aggregation, and selection definition.

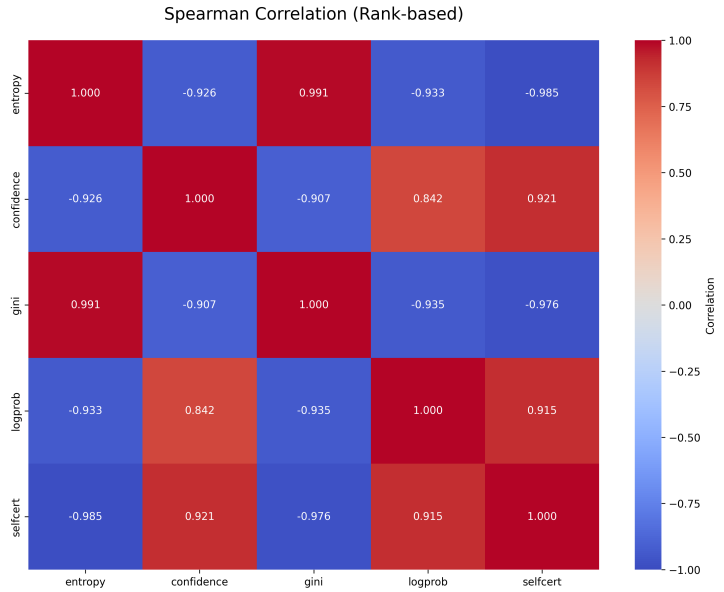

Figure 2: Spearman correlations among token-type average metrics in a curated aggregate. Negative signs occur when one cached metric increases with uncertainty and the other increases with certainty. Set-overlap analysis explicitly reorients these directions.

4.3 Frequency cannot substitute for entropy

Across token types, entropy has essentially no raw linear association with count ($r = -0.0165$), a weak negative rank association ($\rho = -0.2959$), and a somewhat stronger relationship with log count ($r = -0.3464$). The difference between Pearson and Spearman reflects the extreme right skew of token frequency.

Table 3: Entropy–frequency association by dataset and model. Every subgroup has negative Spearman correlation, but the magnitude remains modest.

Dataset	Types	ρ	Model	Types	ρ
AIME 2024	16,036	−0.228	DeepSeek-R1–Qwen3-8B	30,114	−0.308
AIME 2025	16,467	−0.176	R1-Distill-Llama-8B	25,302	−0.264
DeepSeek-8B-S1K	23,531	−0.351	Qwen3-4B-Instruct	17,904	−0.331
GPQA	27,048	−0.285	Qwen3-4B-Thinking	23,038	−0.312
HumanEval	18,512	−0.247			
LiveCodeBench	25,465	−0.271			
MBPP	25,922	−0.306			

The direction is consistent: frequent types tend to be more certain. The predictive value is weak. Frequency explains neither context-specific branching nor the entropy of a particular occurrence.

4.4 Inverse P99 analysis recovers frequent forking tokens

Among the 436 P99 frequency types, high-entropy examples include discourse transitions (*So*, *Let*, *Actually*, *Alternatively*, *Since*, *Therefore*), programming tokens (`.User`, `ndef`, `.combine`), and multilingual or subword fragments. These types are counterexamples to the shortcut "high frequency implies low entropy."

The forward high-entropy word list contains 100 decoded strings and 93 unique strings after trimming whitespace. Forty of these 93 occur in the P99 inverse set, for 43% overlap. Every forward-only

normalized token falls below the P99 count threshold; many lie between P95 and P99. The remaining mismatch therefore reflects the conditioning rule rather than a string-key bug.

Table 4: Selected frequent, high-entropy token types from the normalized forward–inverse overlap. Counts and means are from the documented aggregate analysis.

Decoded token	Count	Mean entropy	Functional reading
So	7,878,521	1.101	consequence / transition
Let	4,982,693	1.345	introduce variable or assumption
Actually	1,499,004	1.004	revision / correction
Alternatively	1,077,355	1.267	branch to another plan
Since	922,267	1.433	premise / justification
Therefore	639,886	1.569	conclusion / transition

The functional readings are hypotheses about typical use, not manual labels for every occurrence. The cache does not establish that each high-entropy *Therefore* marks a valid reasoning branch. It shows that some high-frequency connective types retain high uncertainty across diverse contexts.

4.5 A SMALL MULTILINGUAL EXTENSION

Two multilingual caches contain 35,450 decoded types in the type-average analysis. Their mean-entropy 80th percentile is 1.36, compared with 1.22 in the 51-cache aggregate, and the P99 inverse word cloud contains reasoning transitions alongside Chinese, Indonesian, French, and other script fragments. This is evidence that the pipeline can process multilingual traces, not a comparative language study: two caches are insufficient to separate model, task, language, and tokenizer effects.

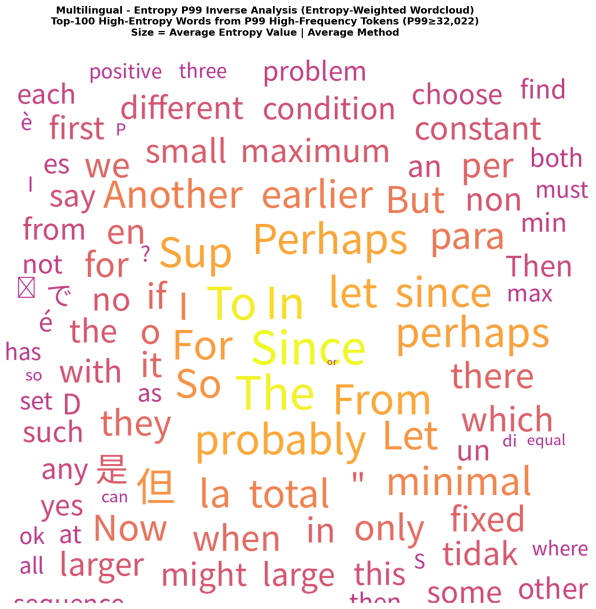

Figure 3: Exploratory multilingual P99 inverse word cloud. Size denotes mean entropy among very frequent decoded types. Mixed scripts and duplicated case forms illustrate why token-ID-level analysis is preferable to treating strings as linguistic words.

5 Discussion

5.1 The 80/20 rule is an allocation rule

Selecting the highest-entropy 20% is always possible by definition. What makes the original result substantive is not the percentile itself, but the intervention: those positions carry enough learning signal that discarding the other 80% of policy-gradient terms preserves or improves performance (Wang et al., 2025). Our audit supports the descriptive premise that uncertainty is concentrated and structured. It does not independently support the training conclusion.

This distinction also explains why a fixed threshold should not be canonized. If the practical objective is a 20% gradient budget, the quantile should adapt to the current batch or corpus. If the objective is a calibrated entropy threshold, one must demonstrate stability across models and tasks. The values 0.672 and 1.219 belong to different data distributions and different units.

5.2 Token identity is not token function

A decoded string is not a reasoning role. *Let* can introduce a variable, begin a proof plan, appear inside quoted text, or be forced by syntax. Its average entropy compresses these contexts. Instance-level bimodality is therefore not noise to eliminate; it is evidence that the same type can serve both deterministic grammar and uncertain planning.

A stronger forking-token analysis would model context directly. Candidate features include sentence position, preceding discourse markers, local entropy slope, hidden-state change, and whether the continuation changes the mathematical subgoal. Manual or verifier-backed labels could test whether entropy predicts genuine plan branching rather than lexical alternatives.

5.3 Metric agreement is budget dependent

Set overlap depends on selecting exactly 20% of types. A smaller top-k emphasizes extreme tails and can reduce stability; a larger set mechanically raises overlap. Continuous correlations and fixed-budget Jaccard answer different questions. The current results justify entropy as one useful summary, not as the unique uncertainty measure.

6 Limitations

No intervention. The largest limitation is categorical: no RLVR training is run. We cannot claim that the identified tokens cause reasoning gains or that masking other gradients is beneficial in these models.

Corpus construction. The 51 caches are heterogeneous and unequally weighted. Summing them gives a large descriptive corpus, not an unbiased estimate over model and task families. Some token positions may be correlated through repeated prompts or sampling.

Tokenizer and decoded strings. Analysis begins at token IDs but many reports aggregate by decoded string. Leading whitespace, casing, byte fragments, and shared strings across IDs can create false matches or false separations. The 43% forward–inverse overlap is reported only after a documented whitespace normalization.

Rare-type instability. Type averages with very small n_v can be extreme. A minimum-count filter, shrinkage estimator, or hierarchical model would better separate stable uncertainty from sampling variance.

Threshold and multiple analysis choices. P20/P80, P95, and P99 are interpretable but post hoc choices. Word-cloud filtering, string normalization, and grouping by model or dataset introduce additional degrees of freedom. Future work should preregister the primary estimand and report sensitivity curves.

Metric provenance. The repository consumes cached Confidence, Gini, LogProb, and SelfCert values rather than recomputing every distribution from logits. Their exact calibration depends on the upstream cache implementation. Direction is audited; equivalence of scale is not assumed.

7 Reproducibility Statement

The public repository contains scripts for both estimands, five-metric set overlap, entropy–frequency correlation, forward–inverse comparison, multilingual analysis, and all figures shown here. Raw caches and regenerated result directories are excluded because of their size. The numerical claims in this manuscript were checked against the committed analysis reports and the local aggregate `token_stats_entropy.json`; the latter contains 43,529 keys and counts summing to 1,116,425,428.

To prevent legacy summaries from overriding auditable evidence, the revised README changes three claims: it replaces "70M+ token instances" with the aggregate count, replaces a universal 0.672 threshold with an estimand-specific statement, and replaces the earlier 0.974 metric-overlap headline with the full top-20% Jaccard matrix.

AI writing and coding assistants supported repository restructuring, prose revision, LaTeX authoring, and consistency checks. The cache-generation code, analysis scripts, saved reports, and numerical results originate from the author's project. The author verified the aggregate counts and reconciled conflicting legacy documentation against the auditable result files.

8 Conclusion

Reasoning-token entropy has a robust long tail, and frequent connectives can occupy its high-uncertainty region. But entropy is not frequency, a token position is not a token type, and a percentile is not a universal threshold. Across more than one billion cached token positions, five uncertainty signals select substantially overlapping but non-identical type sets, while frequency remains only a weak proxy for entropy. These results strengthen the descriptive foundation of forking-token research and narrow its interpretation. The next decisive test is not another word cloud: it is a controlled intervention that links context-sensitive token roles to learning dynamics and verified reasoning outcomes.

References

Kang Chen, Yaoning Wang, Kai Xiong, Zhuoka Feng, Wenhe Sun, Haotian Chen, and Yixin Cao. Do LLMs signal when they're right? evidence from neuron agreement. *arXiv preprint arXiv:2510.26277*, 2025. URL https://arxiv.org/abs/2510.26277.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.01939.

A Top-20% Certainty Overlap

Table 5: Jaccard overlap between equal-sized high-certainty token-type sets.

	Entropy	Confidence	Gini	LogProb	SelfCert
Entropy	1.000	0.562	0.932	0.813	0.690
Confidence		1.000	0.547	0.500	0.552
Gini			1.000	0.823	0.681
LogProb				1.000	0.650
SelfCert					1.000