CHARACTERIZING THE DECISION BOUNDARIES OF TRUNCATED SAMPLING IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models are usually sampled through a truncation rule, which keeps a few of the most probable next tokens and discards the rest, so the rule decides which continuations a model can produce at all. Yet rules are compared only by generating text and scoring it, and a benchmark score, one number per corpus, cannot show which token a rule removed or where. Our starting observation is that every rule we study keeps the m most probable tokens for some m, so at each position a rule reduces to one integer, its cut, on a single sorted order of the vocabulary. We build on it a reference-conditional diagnostic: the model reads a benchmark's reference answer token by token, and at every position we record whether each rule's cut deletes the reference token and how much probability it keeps beyond it, which compares all rules exactly on identical positions. The same representation yields two propositions, stating when a concentrated distribution forces a threshold rule to keep a single token and which rules temperature cannot move, and a contrast that prices a deletion with the prompt held fixed. Across eleven rule families, four benchmarks and two instruction-tuned models, the threshold rules top-p, min-p and top-$n\sigma$ at common settings delete the reference on question answering about twice as often as the count rule top-k. On TriviaQA, top-p 0.95 deletes it at 0.358 of scored answer positions against 0.164 for top-k 50, a gap that a fifty-token floor closes, yet their accuracies differ by less than 0.01 (0.594 against 0.590), and on the 674 prompts where only top-p deletes, it fails no more often (difference -0.0045, 95% interval $[-0.0094, 0.0000]$). Rules that benchmark scores cannot tell apart thus differ systematically in what they leave reachable, and the diagnostic locates that difference, though not its cost, giving a direct criterion for choosing a rule when a specific continuation must stay reachable.

1 INTRODUCTION

A language model generates text one token at a time, and a *truncation rule* decides which tokens it may sample: the rule keeps some of the most probable tokens, renormalizes their probabilities and samples from them. Truncation keeps sampled text coherent (Holtzman et al., 2020), both models we study ship with a rule enabled, and we measure eleven rule families, five of them recent. Accounts of why truncation helps (Hewitt et al., 2022; Finlayson et al., 2024) do not say which rule to use.

Rules are compared by generating text and scoring it on a benchmark (Holtzman et al., 2020; Nguyen et al., 2025; Tang et al., 2025), and the resulting rankings shift with task, model, tuning budget and seed (Wiher et al., 2022; Shi et al., 2024; Song et al., 2025); min-p's reported advantage over top-p (Nguyen et al., 2025) did not survive a reanalysis at a matched budget (Schaeffer et al., 2025a). A score cannot settle the question, because a rule acts at every position while a score is one number per corpus, so it cannot say where two rules acted differently. On GSM8K, top-p 0.95 and top-k 50 generate different text for most prompts, yet each one's accuracy lies inside the other's interval (Figure 3).

Approach. We measure what a rule removes, position by position (Figure 1). Every rule we study keeps a *rank prefix*, the m most probable tokens for some m, so at each position it is one integer, its *cut*: the decision boundary between kept and removed tokens. With *teacher forcing*, the model reads

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

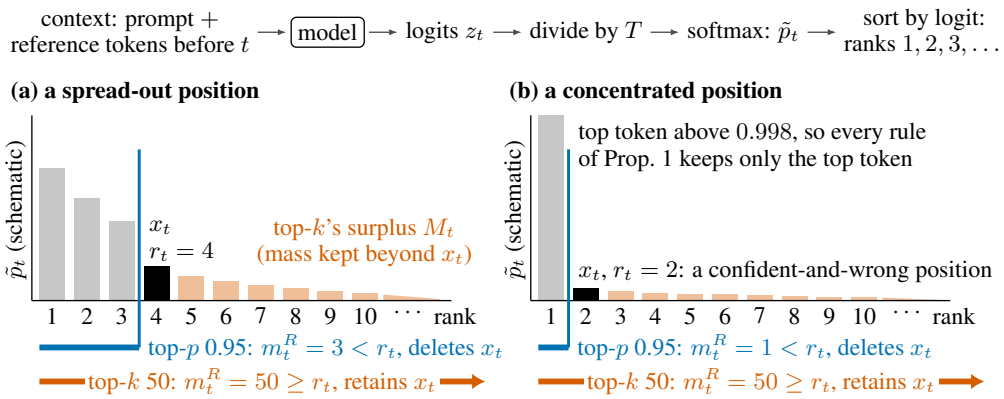


Figure 1: How the diagnostic reads one position (schematic; bar heights are illustrative). **(a)** A rule keeps a prefix of the sorted order, so it is one integer, its cut m_t^R; a cut left of the reference's rank r_t deletes the reference x_t, and the probability kept beyond x_t is the surplus M_t. **(b)** At a concentrated position every threshold rule that Proposition 1 covers keeps only the top token while top-k keeps k, so where the reference is not the top token (a *confident-and-wrong* position) only top-k keeps it.

a benchmark's reference answer one token at a time, so all rules are evaluated at the same positions and contexts. A rule *deletes* the reference when its cut falls short of the reference token's rank, and its *surplus* is the probability it keeps beyond that token.

Closest prior work. Four studies also condition on a reference: Ravfogel et al. (2023) test top-p alone on human-written text; Li et al. (2026) teacher-force answers to calibrate new cutoffs rather than compare published rules; Zhou et al. (2025) compare five rules against a trie of attested continuations, counting tokens rather than probability; and Garces Arias et al. (2026) find that 8 to 18 of every 100 human-written tokens fall outside common truncation boundaries. New here are one sorted order for eleven published families, the mass a rule keeps beyond the reference, and the cost of a deletion with the prompt held fixed.

Contributions. We study rank-prefix rules on the answer spans of four benchmarks, with two instruction-tuned models of 7 and 8 billion parameters (a third, reasoning model enters one check), and propose no new rule.

1. **A diagnostic for every rank-prefix rule** (Section 2). Applying two rules together keeps the shorter prefix, so a count rule cannot undo a threshold rule's deletion while a *token floor*, a minimum number of kept tokens, can; Proposition 1 gives when a concentrated distribution forces a one-token cut, and Proposition 2 which rules temperature cannot move.
2. **Threshold rules delete the reference more often than the count rule** (Sections 4.1 to 4.3), on question answering about twice as often as top-k 50. Concentration explains part of the difference; the larger part arises at non-concentrated positions. A token floor of 50 brings top-p to top-k's deletion rate and, in generated text, to top-k's responses.
3. **What a deletion costs** (Section 4.4). With the prompt held fixed, the deleting configuration fails no more often in all but one comparison, and in a closed-set task the deleted items were ones the untruncated model almost never answered correctly. Deletion rates locate where rules differ; the accuracies of top-p 0.95 and top-k 50 differ by less than 0.01.

2 Measuring Truncation Against a Reference

2.1 Rules as cuts on one sorted order

Table 2 (Appendix A) lists every symbol. At position t the model outputs logits $z_t \in \mathbb{R}^{|V|}$ over the vocabulary V; we write $p_t = \mathrm{softmax}(z_t)$ and $\tilde{p}_t = \mathrm{softmax}(z_t/T)$ for a temperature $T > 0$, where $T < 1$ sharpens the distribution. We sort V by descending logit, breaking ties by ascending

token id, and write $\mathrm{rank}_t(v)$ for the place of token v, so the top token has rank 1; sorting logits rather than float32 probabilities keeps apart tail tokens whose probabilities round alike.

A rule R keeps a set $\mathcal{N}_t^R \subseteq V$ and renormalizes $\tilde{p}_t$ over it (Table 3 gives each definition and the swept values). The *count rule* top-k (Fan et al., 2018) keeps the k most probable tokens. In every other family the position decides how many tokens survive; we call these *threshold rules*, since all but min-k keep the tokens that pass a threshold. With top-k, the *established families* are top-p (Holtzman et al., 2020), which keeps the fewest top-ranked tokens whose probabilities sum to at least p; min-p (Nguyen et al., 2025), which keeps tokens with probability at least α times the top token's; top-$n\sigma$ (Tang et al., 2025), which keeps tokens whose logit, divided by T, lies within n standard deviations of the largest; and ϵ- and η-sampling (Hewitt et al., 2022), which keep tokens with probability at least ϵ and at least $\min(\epsilon, \sqrt{\epsilon}\, e^{-H(\tilde{p}_t)})$, with H the entropy in nats. The five *recent rules* (two are preprints) set their threshold from the position's entropy or logits: top-H (Baghaei Potraghloo et al., 2025) stops before the first token that would raise the renormalized entropy above a fixed fraction of the full entropy; top-b (Halder & Dabre, 2026) is min-p with α set by the entropy; min-k (Ding et al., 2026) reads the cut off the gaps between sorted logits; p-less (Tan et al., 2026) thresholds at the collision probability $\sum_v \tilde{p}_t(v)^2$; and ANTS (Dia, 2026) is top-$n\sigma$ with a width of γ times the entropy of p_t. ANTS and min-k read the logits before temperature, the others $\tilde{p}_t$.

Every family keeps a *rank prefix*: whenever it keeps a token, it keeps every higher-ranked one (argued per definition in Appendix B.2 and confirmed against brute force). At position t a rule is therefore one integer, its *cut* $m_t^R = |\mathcal{N}_t^R|$. Rules that carry a state across positions (Basu et al., 2021) or rank tokens by another criterion (Meister et al., 2023; Li et al., 2023; Su et al., 2022) fall outside this representation.

2.2 Deletion and surplus

A *reference* is a continuation fixed independently of every rule, here a benchmark's answer or worked solution; its token at t is the *reference token* x_t, with rank $r_t = \mathrm{rank}_t(x_t)$. With $C_t[m] = \sum_{v:\,\mathrm{rank}_t(v)\le m} \tilde{p}_t(v)$ the probability of the m top-ranked tokens, the *deletion indicator*, the *signed surplus* and the *surplus* of rule R at t are

$$D_t(R) = \mathbf{1}\big[m_t^R < r_t\big], \qquad G_t(R) = C_t[m_t^R] - C_t[r_t], \qquad M_t(R) = \max\big(0,\, G_t(R)\big). \tag{1}$$

A rule deletes the reference exactly when its cut stops before r_t, the size of the smallest prefix containing x_t. After a deletion the sampler cannot emit x_t in this context, although another tokenization of the same answer may remain reachable. The surplus is the probability kept on tokens ranked below x_t; it separates rules that both keep the reference, in probability where Zhou et al. (2025) count tokens. Surplus is not a cost, since tokens beyond the reference may lead to another correct answer, so deletion and surplus are never combined; clipping at zero keeps a deletion from lowering a rule's average surplus (Appendix B).

The construction rests on four checkable conditions: **(A1)** T and every logit are finite, so every token has positive probability; **(A2)** each rule keeps a rank prefix; **(A3)** ranks are one-to-one, which the tie-break guarantees; and **(A4)** the sampler draws from $\mathcal{N}_t^R$, renormalized, after the reference's own context, which teacher forcing guarantees. A1 fails only in the closed-set experiment (Section 4.4), whose mask sets logits to $-\infty$ and which reads no probability mass. Because $C_t[m]$ never decreases in m, a deletion gives $G_t(R) \le 0$, so no deleted position has a positive surplus. The converse would need exact arithmetic: in float32 the running sum saturates deep in the tail and $G_t(R)$ reads zero at many deletions (Appendix H), so the implementation reads deletion from cut and rank, never from mass.

2.3 What the construction implies

Intersection. Rank prefixes of one order are nested, so two rules applied together keep the shorter prefix, of length $\min(m_t^R, m_t^{R'})$, and delete whenever either does: adding a count rule never undoes a threshold rule's deletion. A *token floor* of k tokens keeps $\max(m_t^R, k)$ tokens and so can undo one (Section 4.2).

Concentration. At a *concentrated* position the model puts almost all its probability on one token, so a rule whose probability threshold exceeds what the other tokens carry keeps the top token alone,

while a count rule still keeps k. Proposition 1 makes this exact below 0.01 nats, the entropy below which we call a position concentrated (proof in Appendix C).

Proposition 1 (Concentration forces a one-token cut)**.** *If $H(\tilde{p}_t) < 0.01$ nats, the top token has probability above 0.998 and every other token has probability below 0.002. Hence a rule that keeps $\{v : \tilde{p}_t(v) \geq \theta_t\}$ for a probability threshold θ_t with $0.002 \leq \theta_t \leq \max_v \tilde{p}_t(v)$ keeps only the top token, and so does top-p with $p \leq 0.998$.*

Of the swept settings, the condition covers every top-p below 1, min-p with $\alpha \geq 0.002/0.998$, every top-b, both forms of p-less and η-sampling at $\epsilon = 0.002$, but not top-$n\sigma$, ANTS, min-k, top-H or smaller ϵ (Table 9). Since a count rule keeps k tokens everywhere, the two kinds must disagree wherever a concentrated position's reference is not the top token, a *confident-and-wrong* position.

Temperature. Dividing every logit by T never reorders tokens, so a rule whose test depends only on the order, or on logit differences measured in units of their own spread, ignores T. Proposition 2 sorts the families by this test, so it says which results at $T = 0.7$ hold at every temperature (proof in Appendix C).

Proposition 2 (Which rules temperature cannot move)**.** *In exact arithmetic and for every $T > 0$: (i) the sorted order, and hence every rank r_t, is the same at every temperature; (ii) top-k and top-$n\sigma$ keep the same tokens at every temperature, and so do ANTS and min-k, which select before temperature is applied; (iii) min-p with threshold α at temperature T keeps what min-p with threshold α^T keeps at $T = 1$; (iv) top-p, ϵ-sampling, η-sampling, top-H, top-b and p-less keep different tokens at different temperatures, and for top-p no change of p undoes a change of T.*

The invariance of top-$n\sigma$, ANTS and min-k is claimed in their papers (Tang et al., 2025; Dia, 2026; Ding et al., 2026); the full partition, clause (iii) and the counterexamples behind clause (iv) are ours, and Section 4.5 checks the partition.

2.4 Estimands

Position level. A rule's *deletion rate* and *mean surplus* are the means of $D_t(R)$ and $M_t(R)$ over a benchmark's scored positions. Both depend only on $\tilde{p}_t$, x_t and $\mathcal{N}_t^R$, so rules are compared on identical positions, and models are never pooled. Because the hyperparameters of different families are not comparable as numbers, Appendix G.3 also compares families under a *surplus cap*, representing each by its lowest-deletion configuration whose mean surplus fits under the cap.

Prompt level. A deletion and a hard prompt both make a wrong answer more likely, so their pooled association mixes the two; the *discordant contrast* holds the prompt fixed. Let $d_i(R) = 1$ when R deletes the reference at some content position of prompt i (Section 3), and let $f_i(R)$ be the failure rate of R's generated responses to prompt i. Then

$$\Delta(R, R') = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \big(f_i(R) - f_i(R')\big), \qquad \mathcal{P} = \{i : d_i(R) = 1,\ d_i(R') = 0\}. \tag{2}$$

The prompts in $\mathcal{P}$, where only R deletes, are *discordant*, and $\Delta > 0$ means the deleting configuration fails more often. Δ measures swapping the whole rule on those prompts, not only the deleted token. A comparison counts as equivalent to zero when its interval lies inside $\pm\delta$; the margin $\delta = 0.05$ is a convention fixed in the analysis code before any interval was read.

Cost. One forward pass and one sort per position serve all K configurations, each then taking one pass over the sorted vocabulary: $O(N(|V| \log |V| + K|V|))$ time beyond the N forward passes, against a full generation run per configuration for generate-and-score.

3 Experimental Setup

Models and benchmarks. The primary model is Qwen2.5-7B-Instruct (Qwen et al., 2024) (*Qwen*). Llama-3.1-8B-Instruct (Grattafiori et al., 2024) (*Llama*) has a similar size but a different training corpus and tokenizer, and DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025) enters only the long-chain check of Section 4.5. The benchmarks, each with a checkable answer span, are

GSM8K (Cobbe et al., 2021), MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024), NQ-open (Kwiatkowski et al., 2019; Lee et al., 2019) and TriviaQA (Joshi et al., 2017): test splits for mathematics and validation splits for question answering, whose test answers are unpublished, with NQ-open and TriviaQA subsampled to 2,000 prompts at seed 13 (Appendix D).

References and scored positions. The scored span is the benchmark's final answer, preceded by the worked solution in mathematics and by the prompt alone in question answering, so no context contains generated text. Because a chat model opens its reply with a sentence, each reference follows a fixed *carrier* phrase: "The answer is" for question answering and "Final answer:" for mathematics. The answer's first token is often a *separator*, such as a leading space, which dominates any rate it enters (Table 6), so we score the remaining *content positions*. Every other protocol choice is varied one at a time (Appendix D.1).

Rules and scoring. Each family is swept over a grid bracketing its papers' operating point (Appendix B.1): 77 configurations, including the untruncated baseline and eight with a token floor. The 45 configurations of the six established families, baseline and floored ones included, form the *established sweep*; the other 32 belong to the recent rules. Every cell is scored at $T = 0.7$, Qwen's shipped temperature, and at $T = 1.0$; rates are at $T = 0.7$ unless stated. Teacher-forced scoring runs one example at a time in float32 with TF32 (reduced-precision matrix multiplication) disabled, on NVIDIA L40S GPUs (Appendix H).

Generation experiment. Thirteen configurations generate responses to GSM8K's 1,319 prompts and the 2,000 TriviaQA prompts at seeds 13, 42 and 2026: seven established (top-k 20 and 50, top-p 0.90 and 0.95, top-p 0.95 + floor 50 or 200, and untruncated sampling) and six recent-rule operating points (top-H 0.4, top-b 0.3, min-k 3, p-less, ANTS 5 and ANTS 32). Responses are sampled at $T = 0.7$ in bfloat16, up to 400 new tokens on GSM8K and 96 on TriviaQA, with Qwen's default repetition penalty of 1.05 before every rule. Within each of the two groups, configurations share each batch's random seed, so their texts differ only where their rules keep different tokens (Appendix F.1). The two floored configurations truncated before temperature, so their teacher-forced counterpart is the $T = 1.0$ column (Appendix F.1). A response is correct under *alias scoring* if it contains any accepted answer at word boundaries, and under *canonical scoring*, which the discordant contrast uses, only if it contains the canonical answer the diagnostic scores.

Closed-set experiment. On the 2,000 TriviaQA prompts, generation is restricted to four literal candidates, the correct answer and three answers of other prompts, so deleting the correct one is fatal. Accuracy is the exact probability of the correct candidate along its token path, with the candidate mask applied first, then temperature, then the rule (Appendix F.3).

Statistics. Every interval is a 95% percentile bootstrap over whole prompts, with 10,000 resamples at seed 13. A resample carries all of a prompt's positions and seeds, and one set of draws serves all configurations, so differences are paired; intervals are not corrected for multiple comparisons. Two biases are stated rather than removed. A surplus-cap representative is selected on the positions it is scored on, which biases its rate low; re-selection on held-out prompt halves bounds that bias (Appendix G.3). And M_t is zero wherever a rule deletes, which favors deleting rules under a cap, so capped comparisons are paired with the uncapped rates of Table 1.

4 Results

4.1 Threshold rules delete the reference more often than the count rule

On question answering the three threshold rules of Table 1 delete the reference about twice as often as top-k 50; on mathematics top-k 50 deletes one MATH-500 position and none on GSM8K. The gap is not confined to references the model ranks low: among references in the model's top ten, top-p 0.95 still deletes at 20.01% $[18.90, 21.11]$ of NQ-open positions and top-k 50 at exactly zero (Appendix E.1). As the intersection argument predicts, applying both kinds together does not close the gap: Qwen's shipped top-k 20 moves its shipped top-p 0.8's NQ-open rate of 43.52% $[42.33, 44.71]$ by 0.03 points (Appendix E.2). The split replicates on Llama (Table 8) and at $T = 1.0$, where the NQ-open ratio of the two rates is 1.8 rather than 2.3.

Table 1: Reference deletion rate on Qwen2.5-7B-Instruct at $T = 0.7$: the proportion of content positions at which a rule removes the reference token (lower means the reference is kept more often). All columns read identical positions against identical references. The last column adds a fifty-token floor to top-p (Section 4.2). Rates in every table are proportions; the text quotes some as percentages. Intervals for the top-p 0.95 column on GSM8K and TriviaQA are in Table 6.

Benchmark	Positions	top-p 0.95	min-p 0.05	top-$n\sigma$ 1	top-k 50	top-p 0.95, ≥ 50
GSM8K	3,027	0.011	0.011	0.010	0.000	0.000
MATH-500	2,105	0.119	0.119	0.117	0.000	0.000
NQ-open	7,544	0.390	0.405	0.392	0.173	0.172
TriviaQA	5,824	0.358	0.366	0.354	0.164	0.164

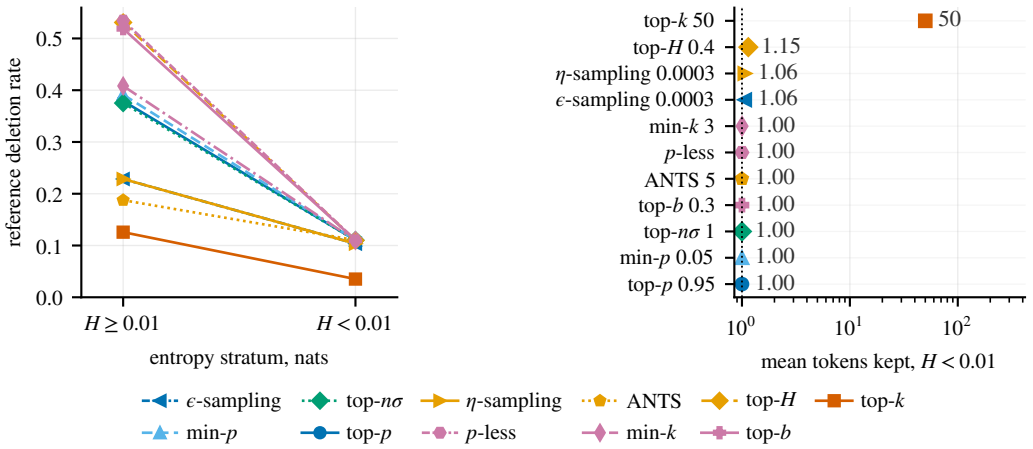


Figure 2: Deletion and cut size by family, Qwen2.5-7B-Instruct, $T = 0.7$, one line or marker per family at its published operating point (ANTS at $\gamma = 5$). **Left:** deletion rate (lower keeps the reference more often) on content positions in two entropy strata, non-concentrated ($H \geq 0.01$ nats) and concentrated ($H < 0.01$), as the unweighted mean over the four benchmarks; lines only join a family's two strata. **Right:** mean number of tokens kept at concentrated positions, log scale, averaged the same way. Every threshold rule keeps about one token there; top-k keeps 50.

4.2 Concentration explains part of the difference

At concentrated positions Proposition 1 is the mechanism: a threshold rule whose probability threshold is at least 0.002 keeps only the top token and a count rule keeps k, so the two disagree at the confident-and-wrong positions. Concentrated positions make up most mathematics positions and just under half of question-answering ones, and the reference is usually their top token (Appendix D.2). The rules the proposition does not cover also keep one token or barely more there at their published settings (Figure 2, right; Table 9). The larger part of the gap, however, lies at non-concentrated positions (Figure 2, left), where threshold rules also keep few tokens: over all NQ-open content positions top-p 0.95 keeps a mean of 4.45 (Appendix B.1). On Llama, whose strata are tabulated, the non-concentrated stratum likewise carries most of the question-answering gap (Table 8).

Most recent rules cut where top-p cuts at concentrated positions. At the 19,074 concentrated positions, pooled over both temperatures, three of the five recent rules cut exactly where top-p does on 100.00% of positions, min-k on 99.86% and top-H on 81.03%; η-sampling against the ϵ-sampling rule it extends gives the same answer (Appendix E.5). ANTS is measured as fixed arms at both ends of its range $\gamma \in [5, 32]$, not as its controller that moves γ between rollouts; at $\gamma = 32$ its large mean cut there comes from a small share of positions (Table 9).

A token floor crosses the divide. A threshold rule under a floor of k tokens cuts at $\max(m_t^R, k)$. The `transformers` library offers this floor as the `min_tokens_to_keep` argument of

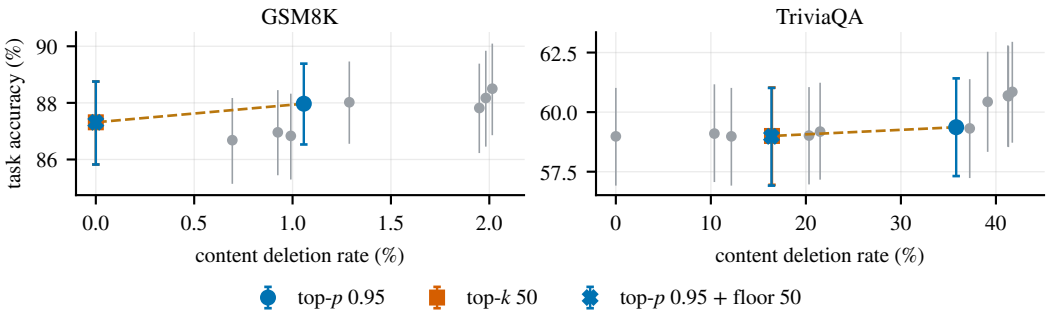


Figure 3: Task accuracy of the thirteen generated configurations against their teacher-forced deletion rate on content positions at $T = 0.7$, Qwen2.5-7B-Instruct, three seeds pooled; bars are 95% bootstrap intervals. Three configurations are highlighted and the other ten are grey (accuracies in Table 10). The floored arm's generation truncated before temperature (Section 3). Deletion rates spread widely while all accuracy intervals overlap.

`TopPLogitsWarper`, fixed at one outside beam search, so it is reachable only by constructing that processor directly (Appendix E.2). With a fifty-token floor, top-p behaves like a count rule on every benchmark (Table 1, last column).

4.3 Generated text splits the same way

Under the shared seed, the configurations that keep the tail form one *tail-keeping class*: top-k 50 matches untruncated sampling on 99.97% of GSM8K responses, and the floored top-p matches top-k 50 on all 3,957. Unfloored top-p 0.95 leaves the class, differing from top-k 50 on 59.82% of GSM8K responses and 7.80% of TriviaQA's, where the floored pair agrees on 99.37% (Appendix F.1). Generated text thus splits the configurations as the diagnostic does, while accuracy does not (Figure 3).

4.4 What a deletion costs

Responses to prompts on which a configuration deletes the reference fail more often, by $+0.272$ $[0.231, 0.311]$ under top-p 0.95 on TriviaQA, but a hard prompt and a harmful deletion both produce that association; the discordant contrast of Equation 2 separates them. On the 674 TriviaQA prompts where top-p 0.95 deletes and top-k 50 does not, $\Delta = -0.0045$ $[-0.0094, 0.0000]$: the deleting rule fails slightly less, not more. Against the floored top-p, whose generation truncated before temperature, $\Delta = -0.0049$ $[-0.0099, 0.0000]$. Because top-p 0.95 and top-k 50 emit identical text on most TriviaQA responses (Section 4.3), Δ averages over many prompts where the deletion changed nothing.

The other pairs agree (Appendix F.2). Among the seven established configurations, 32 of 34 intervals contain zero, and both exceptions favor the deleting rule; of all 169 comparisons, 30 exclude zero, and in 29 of them the deleting rule fails less, the exception resting on 8 prompts. Every TriviaQA interval lies inside $\pm\delta$, so a cost of δ is excluded there, though a smaller one is not; the 10 GSM8K comparisons outside the margin, over 31 to 38 prompts, are too few to decide.

Canonical scoring counts an answer in other words as a failure, and the closed-set arm removes that route: on the 42 items where top-p 0.95 deletes the correct candidate, the untruncated model had answered correctly with probability 0.0042, so the deletion changes accuracy there by -0.0042 $[-0.0070, -0.0018]$ and the corpus by almost nothing. The loss grows, and top-p's corpus lead vanishes, when the rule precedes the mask (Appendix F.3). In this arm a deletion marks an item the model had largely lost; its cost is bounded, not shown to be zero.

4.5 Robustness

Which references: matched at their closest rates, top-p, min-p and top-$n\sigma$ delete nearly the same references, and top-k's deletions overlap theirs less (Table 11). *Temperature:* on positions scored

at both temperatures, no cut of a rule that Proposition 2 calls invariant moves, apart from a float32 rounding residual in top-$n\sigma$ (Table 12). *Matched budget:* at equal mean surplus the two kinds coincide only where both delete nothing (Appendix G.3). *Protocol:* carrier and prefix move every level, and where both rules are reported the ordering holds (Appendix D.1). *Long generations:* on the chains of thought that DeepSeek-R1-Distill-Qwen-7B generates itself, top-p 0.95 deletes at least once within a thousand tokens with probability 0.9982, top-k 50 with 0.0168 (Appendix G.4).

5 Related Work

Evaluating rules. Downstream evaluation scores a rule by human judgement, by MAUVE (Pillutla et al., 2021) or by quality–diversity curves (Caccia et al., 2020; Zhang et al., 2021), none of which names the token a rule removed. The min-p reanalysis (Schaeffer et al., 2025a), which prompted a call for a refutations track (Schaeffer et al., 2025b), and a null for temperature on problem solving (Renze, 2024) rest on the same end-to-end scores. Teacher-forced perplexity and uncertainty estimation (Kuhn et al., 2023; Farquhar et al., 2024; Kadavath et al., 2022) work per position but report probability or confidence, not whether a sampler can reach the reference, and accounts of why truncation helps (Finlayson et al., 2024; Chen et al., 2025; Nadeem et al., 2020) reason about the whole distribution.

Reference-conditional measurement. Section 1 compares the four closest studies. Unlike Garces Arias et al. (2026), whose reference is human text, ours is an answer whose correctness is checkable, which lets Section 4.4 ask what a deletion costs. They find ϵ- and η-sampling settings that exclude fewer human-written tokens than top-p 0.9 while keeping fewer tokens on average; at our concentrated positions those thresholds keep about one token, as top-p does (Figure 2), and the floor that moves top-p across counts tokens instead. Constrained decoding applies a mask whose order relative to the rule matters (Willard & Louf, 2023; Geng et al., 2023; Beurer-Kellner et al., 2024), and the confident-and-wrong rate relates to calibration (Guo et al., 2017; Desai & Durrett, 2020; Holtzman et al., 2021).

6 Discussion and Limitations

Implication. A deployment that needs a specific continuation to stay reachable, such as a required format string, can read the deletion rate directly, and the existing `min_tokens_to_keep` floor gives a threshold rule count-rule behaviour without a new rule.

What a deletion shows. A deletion makes one designated continuation unreachable; it does not show that the model could not have answered another way. Every level depends on prefix and carrier (Table 6), and on TriviaQA the reference ranks higher without the carrier (Appendix D.1). Qwen's repetition penalty is not order-preserving; the sweep reads logits before it and every generation arm applies it, so it is held fixed rather than measured. No checkpoint publishes a decontamination report, and memorization moves references towards the head of the order, so every level is a lower bound; memorization also changes how many positions are concentrated (Appendix D).

Scope of the evidence. The accuracy null covers two benchmarks, one model and answer spans. The closed-set arm is an artificial four-candidate task with a 97% untruncated baseline: it shows that the deleted items were near-hopeless, not that no constrained task is harmed; its corpus contrast depends on the order of mask and rule, and it omits ϵ- and η-sampling and the recent rules. The two floored generation arms match the diagnostic at $T = 1.0$ rather than 0.7 (Appendix F.1). ANTS enters the main comparisons at $\gamma = 5$; its deletion rate at $\gamma = 32$ is not compared with top-k's here.

Untested boundaries. Open-ended generation, the regime truncation and min-p were designed for (Holtzman et al., 2020; Nguyen et al., 2025), is not covered. All three checkpoints are dense decoder-only transformers of 7 to 8 billion parameters, so other architectures, scales and vocabularies are open, and the size of the concentrated regime is itself a checkpoint property (Appendix E.1). Quantized weights, speculative decoding and production batching lie outside the float32 unbatched diagnostic, and no outcome beyond correctness is measured.

7 CONCLUSION

Whether a reference answer token survives each rule's cut sorts the published truncation families into two kinds that end-to-end scores do not separate: threshold rules keep few tokens, often exactly one, where a count rule keeps k, and generated text splits the same way. With the prompt held fixed the deletions cost little measurable accuracy. A deletion rate therefore reports where a rule cuts, which a benchmark score cannot, rather than what the cut costs.

AI USE STATEMENT

In this work, we used generative AI tools to formulate mathematical claims and assist in writing their proofs, to provide feedback on the methodology and experiments, to implement methods in code, and to help interpret results. We did not use generative AI tools to generate synthetic datasets or to assist with translation, and qualitative or thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to create and edit software code and figures, to search for, summarize and format the related literature, to suggest the paper's structure, and to draft, proofread and edit the text and organize its LaTeX. We have reviewed all AI-assisted work. Every rule implementation is checked by unit tests against hand-computed examples and against brute force over explicit token sets, each proved property is also checked by a unit test, and a script requires every number quoted in the text to match a stored result. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI. The final revision of the manuscript used an LLM assistant (Claude, Anthropic) to edit and rewrite text, restructure the presentation of the formal results, and revise figure captions and layout.

ETHICS STATEMENT

This paper measures published checkpoints, used without fine-tuning, on published benchmarks; it collects no data and involves no human subjects or annotation. GSM8K and MATH-500 contain no personal data; NQ-open and TriviaQA are built from public web text whose questions name public figures, and all four are used unmodified under their licences (Appendix D). The supplementary material releases the models' responses on the sampled prompts and a few hundred benchmark items as worked examples, under those licences. The foreseeable risk is reading a deletion rate as a cost, which Sections 4.4 and 6 state it is not.

Broader impacts. The diagnostic makes visible a choice deployments make by default. A team could tighten a rule believing a low deletion rate licenses it; it does not, and neither the rate nor the floor of Section 4.2 is offered as a target. Naming the token a rule removes also makes the measurement usable on references beyond benchmark answers, such as a safety string or a refusal that must stay reachable, a use neither tested here nor covered by these levels. A rule that keeps one token where the distribution is concentrated removes rarer continuations first; this paper measures that against annotated answers only, not against dialect, register or language.

REPRODUCIBILITY STATEMENT

The anonymized code in the supplementary material comprises the measurement code, its configuration (seed, resample count, precision flags, sample sizes, budgets and generation grid), the unit tests, and the result files behind every table, figure and quoted number, with one command that reruns the tests, the prose-number check and the table and figure rebuild on CPU without a model download. Every table body is generated from stored results, and the prose-number check requires each numeric literal in the text, appendices included, to match a stored value or a declared constant; it confirms that a quoted number occurs among the results, not that it was quoted for the intended quantity. Each rule's definition, grid and conventions are in Appendix B. The assumptions are stated in Section 2 and the proofs of both propositions are in Appendix C. The benchmarks, splits, subsampling and reference protocols are in Appendix D. The unit tests, the checks run against the real model, and the precision, hardware and determinism settings are in Appendix H.

REFERENCES

Erfan Baghaei Potraghloo, Seyedarmin Azizi, Souvik Kundu, and Massoud Pedram. Top-H decoding: Adapting the creativity and coherence with bounded entropy in text generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 28482–28513, 2025.

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. Mirostat: A neural text decoding algorithm that directly controls perplexity. In *International Conference on Learning Representations (ICLR)*, 2021.

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. Guiding LLMs the right way: Fast, non-invasive constrained generation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 3658–3673, 2024.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. Language GANs falling short. In *International Conference on Learning Representations (ICLR)*, 2020.

Sijin Chen, Omar Hagrass, and Jason M. Klusowski. Decoding game: On minimax optimality of heuristic text generation strategies. In *International Conference on Learning Representations (ICLR)*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, 2020.

Ousmane Amadou Dia. Adaptive nucleus truncation for long-form reasoning. *arXiv preprint arXiv:2606.13982*, 2026.

Yuanhao Ding, Meimingwei Li, Esteban Garces Arias, Matthias Aßenmacher, Christian Heumann, and Chongsheng Zhang. Min-k sampling: Decoupling truncation from temperature scaling via relative logit dynamics. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 14932–14948, 2026.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 889–898, 2018.

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

Matthew Finlayson, John Hewitt, Alexander Koller, Swabha Swayamdipta, and Ashish Sabharwal. Closing the curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2024.

Esteban Garces Arias, Nurzhan Sapargali, Christian Heumann, and Matthias Aßenmacher. The truncation blind spot: How decoding strategies systematically exclude human-like token choices. In *Proceedings of the 19th International Natural Language Generation Conference (INLG)*, 2026.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 10932–10952, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025.

Deepon Halder and Raj Dabre. Top-*b*: Entropic regulation of relative probability bands in autoregressive language processes. *arXiv preprint arXiv:2603.14567*, 2026.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS)*, 2021.

John Hewitt, Christopher D. Manning, and Percy Liang. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, 2022.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7038–7051, 2021.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1601–1611, 2017.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, et al. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6086–6096, 2019.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 12286–12312, 2023.

Xueyan Li, Guinan Su, Mrinmaya Sachan, and Jonas Geiping. Sample smart, not hard: Correctness-first decoding for better reasoning in LLMs. In *International Conference on Learning Representations (ICLR)*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023.

Moin Nadeem, Tianxing He, Kyunghyun Cho, and James Glass. A systematic characterization of sampling algorithms for open-ended language generation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, pp. 334–346, 2020.

Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *International Conference on Learning Representations (ICLR)*, 2025.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4816–4828, 2021.

Qwen, An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Shauli Ravfogel, Yoav Goldberg, and Jacob Goldberger. Conformal nucleus sampling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 27–34, 2023.

Matthew Renze. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7346–7356, 2024.

Rylan Schaeffer, Joshua Kazdan, and Yegor Denisov-Blanch. Min-p, max exaggeration: A critical analysis of Min-p sampling in language models. *arXiv preprint arXiv:2506.13681*, 2025a.

Rylan Schaeffer, Joshua Kazdan, Yegor Denisov-Blanch, Brando Miranda, Matthias Gerstgrasser, Susan Zhang, Andreas Haupt, Isha Gupta, Elyas Obbad, Jesse Dodge, Jessica Zosa Forde, Francesco Orabona, Sanmi Koyejo, and David Donoho. Position: Machine learning conferences should establish a "Refutations and Critiques" track. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. A thorough examination of decoding methods in the era of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8601–8629, 2024.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 4195–4206, 2025.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 21548–21561, 2022.

Runyan Tan, Shuang Wu, and Phillip Howard. p-less sampling: A robust hyperparameter-free approach for LLM decoding. In *International Conference on Learning Representations (ICLR)*, 2026.

Chenxia Tang, Jianchun Liu, Hongli Xu, and Liusheng Huang. Top-$n\sigma$: Eliminating noise in logit space for robust token sampling of LLM. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 10758–10774, 2025.

Gian Wiher, Clara Meister, and Ryan Cotterell. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012, 2022.

Brandon T. Willard and Rémi Louf. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*, 2023.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pp. 25–33, 2021.

Table 2: Notation.

Symbol	Meaning
V, $\lvert V\rvert$	vocabulary (the tokenizer's exposed entries) and its size
t, $T > 0$	a teacher-forced position; the sampling temperature
$z_t \in \mathbb{R}^{\lvert V\rvert}$	logits at t
$p_t = \mathrm{softmax}(z_t)$	unit-temperature distribution
$\tilde{p}_t = \mathrm{softmax}(z_t/T)$	temperature-scaled distribution, read by every rule except ANTS and min-k
x_t, $\mathrm{rank}_t(v)$, r_t	reference token; position of v in the tie-broken descending order; $\mathrm{rank}_t(x_t)$
R, $\mathcal{N}_t^R$, m_t^R	a rule, the set it retains, and its cut $\lvert\mathcal{N}_t^R\rvert$
$C_t[m]$	mass of the leading m tokens of $\tilde{p}_t$
$D_t(R)$, $G_t(R)$, $M_t(R)$	deletion indicator, signed surplus, surplus
$H(\cdot)$, $h(\cdot)$, $\sigma(\cdot)$	entropy and binary entropy in nats; population standard deviation
θ_t	a probability threshold: the rule retains $\{v : \tilde{p}_t(v) \geq \theta_t\}$
w_t, n_t	top-b's bandwidth; ANTS's width in standard deviations
$d_i(R)$, $f_i(R)$	whether R deletes on prompt i; failure rate of its responses to i
$\Delta(R, R')$, $\mathcal{P}$, δ	discordant contrast, its prompt set, equivalence margin
N, K	scored positions; configurations

Yuxuan Zhou, Margret Keuper, and Mario Fritz. Balancing diversity and risk in LLM sampling: How to select your method and parameter for open-ended text generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 26352–26365, 2025.

A Notation

Hyperparameters keep their own papers' letters (Table 3), except that top-H's fraction is written β here to keep it apart from min-p's α; η-sampling's parameter is ϵ, as in its paper.

B Rules, Grids and Conventions

The untruncated baseline and eight floored composites complete the 77 configurations: top-p at 0.90 and 0.95 under floors of 10, 50 and 200 tokens, and min-p 0.05 and top-$n\sigma$ 1 under a floor of 50. With the baseline, they belong to the established sweep of Section 3.

Conventions. Top-p with $p \geq 1$ is special-cased, because in float32 over Qwen's $151{,}665$ entries the running sum reaches exactly 1 while thousands of tail tokens remain. Top-$n\sigma$ uses the population deviation ($\mathrm{ddof} = 0$) where its reference implementation and `transformers` use $\mathrm{ddof} = 1$, a factor $\sqrt{|V|/(|V|-1)}$ that moves no cut; a zero or non-finite spread falls back to the maximum alone, and in the closed-set arm the deviation is taken over the entries the grammar admits. No rule may return an empty set. Vocabulary means the tokenizer's exposed entries, $151{,}665$ for Qwen2.5-7B-Instruct and $128{,}256$ for Llama-3.1-8B-Instruct, not Qwen's $152{,}064$ padded output columns.

Why the surplus is clipped. Averaging the signed surplus would reward deleting. Within one *cell*, a fixed model, benchmark, prefix protocol, carrier, scope, alias policy and temperature, $\overline{G_t(R)} = \overline{C_t[m_t^R]} - \overline{C_t[r_t]}$ and the second term does not depend on the rule, so the signed average is retained mass shifted by a constant and falls with every deletion. Deletion is already counted on its own axis, so $M_t = \max(0, G_t)$ counts only the mass spent beyond the reference.

B.1 Operating points and coverage

Top-p 0.95 and top-k 50 are common deployment settings, and Qwen's shipped pair is measured beside them (Appendix E.2). Min-p 0.05 and ϵ- and η-sampling at $\epsilon = 0.0003$ lie inside their papers' reported ranges; top-$n\sigma$ 1, top-H at $\beta = 0.4$, top-b 0.3 and min-k at $\tau = 3$ are the points their papers fix; p-less has no hyperparameter. ANTS publishes no single γ, its controller drawing from a log-spaced grid on $[5, 32]$ between rollouts, so both ends are fixed arms. The grids bracket

Table 3: The eleven families: what each retains, the values swept, and the operating point shown wherever one point per family appears. $\tilde{p} = \tilde{p}_t$; ANTS and min-k read the pre-temperature logits z.

Family	Retains	Swept values of its parameter	Point
top-k	the k most probable tokens, k clamped to $[1, \lvert V\rvert]$	5, 10, 20, 50, 100, 200	50
top-p	the shortest prefix with $C_t[m] \geq p$, boundary token kept	0.80, 0.85, 0.90, 0.95, 0.98, 1.00	0.95
min-p	$\tilde{p}(v) \geq \alpha \max_u \tilde{p}(u)$	0.5, 0.2, 0.1, 0.05, 0.02, 0.01, 0.005, 0.001	0.05
top-$n\sigma$	$z_v/T \geq \max_u z_u/T - n\,\sigma(z/T)$	0.25, 0.5, 1, 1.5, 2, 3, 4	1
ϵ-sampling	$\tilde{p}(v) \geq \epsilon$	0.0009, 0.0006, 0.0003, 0.0001	0.0003
η-sampling	$\tilde{p}(v) \geq \min(\epsilon, \sqrt{\epsilon}\, e^{-H(\tilde{p})})$	0.002, 0.0009, 0.0006, 0.0003, 0.0001	0.0003
top-H	greedy prefix whose renormalized entropy stays $\leq \beta H(\tilde{p})$	0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9	0.4
top-b	$\tilde{p}(v) \geq (1 - w_t) \max_u \tilde{p}(u)$, $w_t = b(1 + H(\tilde{p})/\log \lvert V\rvert)$	0.005, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5	0.3
min-k	a cut index read off sorted logit gaps, with fallback τ (Appendix B.2)	0.3, 1, 3, 10, 30, 100, 300	3
p-less	$\tilde{p}(v) \geq \sum_u \tilde{p}(u)^2$, and a normalized variant	none	none
ANTS	$z_v > \max_u z_u - n_t\,\sigma(z)$, $n_t = \gamma H(p_t)$	0.5, 1, 2, 5, 10, 20, 32, 64	5, 32

these points: min-p reaches a factor of 50 below and 5 above its paper's 0.05 to 0.1, top-H covers the range its paper searched, and top-b stops at the 0.5 above which its Definition 1 is not well posed. No rule was tuned against any quantity reported here, and no grid was extended after a rate was seen.

The teacher-forced sweep runs all eleven families on every benchmark. The generation arm omits min-p, top-$n\sigma$, ϵ-sampling and η-sampling, and the closed-set arm runs only top-k, top-p, min-p and top-$n\sigma$ beside the untruncated model. The other omissions are for compute, since each arm decodes or masks every prompt afresh; ϵ- and η-sampling were added after both arms had run, so they are measured under teacher forcing alone.

No published number is reused: no rule's paper reports whether a reference survived at a position, so every rule is re-implemented from its definition under the protocol of Section 3. One objection is that top-k 50 keeps fifty tokens where top-p 0.95 keeps a mean of 1.03 on GSM8K content positions and 4.45 on NQ-open's, so the separation merely counts kept tokens. That difference in kept tokens is the finding (Section 4.2), and two results bound the objection: at a matched surplus budget the kinds still separate wherever both truncate (Appendix G.3), and matched at their closest rates top-k's deletions overlap the threshold rules' less than those rules overlap each other (Appendix G.1).

B.2 Rules that compute the threshold from the position

Section 2.1 needs one property of a rule, that it retains a rank prefix. A membership test monotone in $\tilde{p}_t$ gives one, and so does a rule that names the cut index outright. A criterion not monotone in $\tilde{p}_t$ does not, which puts locally typical sampling (Meister et al., 2023) outside; Mirostat (Basu et al., 2021) is outside too, since it carries a target surprisal across positions. Each rule below is argued from its paper's definition and checked against brute force at every swept value, and all five retain prefixes.

Top-H. Baghaei Potraghloo et al. (2025) pose truncation as maximizing retained mass subject to the renormalized entropy staying at most $\beta H(\tilde{p}_t)$. That problem is NP-hard and its optimum need not be a prefix, since skipping a token can buy headroom for two cheaper ones. Their Algorithm 1, the rule that ships, scans in descending order and stops at the first token whose addition breaches the bound, so it returns a prefix, and top-H is measured as that scan. With $S_t[j]$ the partial sum of $-\tilde{p}_t \log \tilde{p}_t$, the renormalized entropy of the leading j tokens is $(C_t[j] \log C_t[j] + S_t[j])/C_t[j]$, so the scan costs two cumulative sums. Their termination theorem assumes every probability positive, which A1 supplies exactly; a float32 tail that underflows to zero would stop the entropy rising and

let the scan run to the vocabulary's end. That never happens here: at none of the 717,006 scored positions and seven settings does the scan reach the end, the longest cut being 41,170 tokens.

ANTS. Dia (2026) keeps every token whose pre-temperature logit lies strictly within $n_t = n_0 + \gamma H(p_t)$ population standard deviations of the largest: top-$n\sigma$'s geometry with a width set by the position's uncertainty. A logit threshold is monotone in $\tilde{p}_t$, so the set is a prefix. Three parts of the method are not measured: the Thompson-sampling controller that moves γ between rollouts, so each γ is a fixed arm; the no-truncation fallback, which is the full-vocabulary baseline already swept; and the clip of n_t to $[n_{\min}, n_{\max}]$, for which the paper states no bound, leaving the one-token floor as the only guard. These experiments use $n_0 = 0$, with the entropy taken at unit temperature.

Top-b, min-k and p-less. Halder & Dabre (2026)'s rule is min-p with a fraction set by the position, so it retains a prefix for min-p's reason; its Definition 1 requires $w_t \leq 1$ while the entropy factor reaches two, which confines b to $(0, 0.5]$. Ding et al. (2026) sort the logits $z_{(1)} \geq \cdots \geq z_{(|V|)}$, form the position-weighted decay $\omega_i = (z_{(i)} - z_{(i+1)})/(i\rho)$ with range $\rho = z_{(1)} - z_{(|V|)}$, and cut at $\max(\arg\max_i \omega_i, \lfloor \tau/\rho \rfloor)$ clamped to $[1, |V|]$, ties breaking to the earliest index. Naming a cut index makes it a prefix although it reads gaps between neighbouring entries. Its ratio of gaps is temperature-invariant and its fallback τ/ρ is not, so the range is read off the raw logits, as its Algorithm 1 computes it. Tan et al. (2026) set the threshold to the collision probability $\sum_v \tilde{p}_t(v)^2$ with no hyperparameter, and a normalized variant subtracts the chance of an incorrect draw divided by the ways to be wrong; both are probability thresholds and both are measured. Their bound placing the threshold below the top token holds in real arithmetic and can fail by a rounding unit on a near-uniform position, so the one-token floor applies to it as to every rule.

C PROOFS

Both propositions are stated in Section 2.3.

Proof of Proposition 1. Let $H(\tilde{p}_t) \leq 0.01$ nats, which covers every concentrated position, and let $q = \max_v \tilde{p}_t(v)$. Grouping the top token against the rest cannot raise the entropy, so $H(\tilde{p}_t) \geq h(q)$, and $h(q) \leq 0.01$ holds only for $q \leq 0.00131$ or $q \geq 0.99869$. The lower branch is impossible, because an entropy is at least the negative logarithm of its largest probability, which there exceeds 6.6 nats. Hence $q \geq 0.99869 > 0.998$, every other token carries at most $0.00131 < 0.002$, and a rule retaining $\{v : \tilde{p}_t(v) \geq \theta_t\}$ with $\theta_t \geq 0.002$ keeps the top token alone.

The threshold clears 0.002 for min-p, whose $\theta_t = \alpha q$, once $\alpha \geq 0.002/0.998$. It clears it for top-b, whose $\theta_t = (1 - w_t)q$, once $w_t \leq 1 - 0.002/0.998$, which $b \leq 0.5$ guarantees because then $w_t \leq 0.5\,(1 + 0.01/\log|V|)$. It clears it for p-less, whose threshold is the collision probability $L = \sum_u \tilde{p}_t(u)^2 \geq q^2 > 0.996$, and for the normalized variant, whose threshold $L - (1 - L)/(|V| - 1)$ is at least $2L - 1 > 0.992$. It clears it for ϵ- and η-sampling only at $\epsilon \geq 0.002$, since η-sampling's threshold equals ϵ at this entropy. Top-p is not a probability threshold, but it reaches p on the top token alone whenever $p \leq 0.998$.

Four rules escape the argument. Top-$n\sigma$ and ANTS retain every logit within a multiple of the standard deviation of the maximum, a threshold of $q\,e^{-n\sigma/T}$ that low entropy does not bound, because a long tail of very negative logits makes σ large at any concentration: at $z = (0, -10, -100)$ and $T = 1$ the entropy is 0.0005 nats and $\sigma = 44.97$, so top-$n\sigma$ at $n = 1$ keeps two tokens. Min-k names a cut index rather than applying a threshold. Top-H admits a second token whenever the renormalized entropy of the leading pair stays under $\beta H(\tilde{p}_t)$, and a heavy tail inflates $H(\tilde{p}_t)$ without moving the leading pair. What these four keep, and what ϵ- and η-sampling keep below 0.002, is measured in Table 9.

Proof of Proposition 2. (i) $z \mapsto z/T$ is strictly increasing, so in exact arithmetic it preserves the order and every tie, hence r_t; in float32 the division can merge two logits into a tie, which the token-id convention breaks. (ii) Top-k reads only the order. For top-$n\sigma$, dividing all logits by T divides their standard deviation by T, so $z_v/T \geq \max_u z_u/T - n\,\sigma(z/T)$ reduces to $z_v \geq \max_u z_u - n\,\sigma(z)$. ANTS (Dia, 2026, Eq. 1) and min-k (Ding et al., 2026, Alg. 1) select on z before any temperature and then apply a softmax of z/T to what they kept, so T never reaches the

Table 4: Benchmarks as loaded. Where *Scored* is below *Rows*, the prompts are a permutation of the split at seed 13 truncated to 2,000. *Fingerprint*: first eight digits of the run-time content hash. No dataset row is dropped from the teacher-forced cells; only the model-prefix ablation excludes examples (Table 5).

Benchmark	Split	Rows	Scored	Fingerprint	Licence
GSM8K	test	1,319	1,319	`59ec1b7f`	MIT
MATH-500	test	500	500	`f3124375`	MIT
NQ-open	validation	3,610	2,000	`0d657f65`	CC BY-SA 3.0
TriviaQA	validation	17,944	2,000	`b024cd02`	Apache 2.0

selection. Neither is scale-free, since ANTS's width $\gamma H(p_t)$ and min-k's fallback $\lfloor \tau/(\max_u z_u - \min_u z_u) \rfloor$ both move under $z \mapsto cz$. (iii) Min-p keeps v when $\exp\big((z_v - \max_u z_u)/T\big) \geq \alpha$, that is $z_v - \max_u z_u \geq T \log \alpha$, which is the $T = 1$ criterion at threshold α^T. It is not scale-free, since scaling every difference by $c \neq 1$ moves only the left side.

(iv) Each rule named reads a normalized probability or an entropy, neither a function of logit differences alone, so (ii) does not apply, and one witness per rule shows its retained set moving with T. As $T \to 0$, distinct logits concentrate $\tilde{p}$ on the top token, so ϵ-sampling, η-sampling, top-b (whose threshold ratio tends to $1 - b$) and p-less (whose collision probability tends to one) keep only the top token; it suffices that each keeps two tokens at some T. On the logits $(0, -1)$ the second token has probability $s(T) = 1/(1 + e^{1/T})$, rising towards $1/2$. It is kept by ϵ-sampling once $T \geq 1/\ln(1/\epsilon - 1)$, for any $\epsilon < 1/2$, and by η-sampling no later, whose threshold never exceeds ϵ. Top-b keeps it once the ratio $e^{-1/T}$ to the top token, which tends to one, reaches $1 - w_t \leq 1 - b < 1$; the partition function cancels in that ratio, yet T enters through it and through $H(\tilde{p})$. p-less keeps two tokens on the logits $\ln(0.4, 0.35, 0.25)$ at $T = 1$, since $0.35 \geq 0.4^2 + 0.35^2 + 0.25^2 = 0.345$. For top-$H$, on the logits $(0, -1, -1, -1)$ the ratio of the leading pair's renormalized entropy to $H(\tilde{p})$ tends to $1/3$ as $T \to 0$ and to $\ln 2/\ln 4 = 1/2$ as $T \to \infty$, so at $\beta = 0.4$ top-H keeps two tokens at low temperature and one at high. The measured counts behind Table 12 are nonzero at every truncating setting of these six rules. For top-p, suppose $p' = p'(p, T)$ made top-p at T retain what top-p' retains at $T = 1$ on every distribution. Take $Q_1 = (0.5, 0.5)$, $Q_2 = (0.5, 0.4, 0.1)$ and $p = 0.5$. At $T = 1$ both have leading mass 0.5, so any p' cuts both after one token or both after more. At $T = 2$, Q_1 is unchanged while Q_2 flattens to a leading mass of 0.4271 and needs two, so no p' reproduces both.

D Data and Reference Protocols

Splits and selection. TruthfulQA-gen (Lin et al., 2022) is excluded because its free-form answers have no checkable span. Nothing is fitted, so no split is held out; every split is the release's own, and the subsample is the permutation the generation arm also uses, so the arms join on the same prompts. Four choices touched the scored data, each declared: the carrier, picked on a 60-example pilot from the same sample, after which all four candidates were rescored on the full sample (Appendix D.1); the concentration threshold, a convention; the margin $\delta = 0.05$, fixed before any interval was read; and the budget representative, selected on the positions it is scored on (Section 3).

Contamination and external resources. All four benchmarks predate the three checkpoints, none of which publishes a decontamination report, so memorization can only lower the deletion rates (Section 6). No other corpus, retrieval index, synthetic data or supervision enters, and the checkpoints are used as published, at the revisions the released run metadata records.

Metrics. *Accuracy* is alias or canonical scoring as defined in Section 3, applied after case and whitespace normalization, and *failure* is its complement. The closed-set arm reports the exact probability of emitting the gold instead.

Table 5: Measurement cells scored with the carrier at $T = 0.7$ (counts are identical at $T = 1.0$). *Prefix*: gold, prompt-only or model. *Scored*: examples yielding an accepted span. *Excl.*: examples rejected with a machine-readable reason, only under the model prefix. *Scope*: answ. is the answer span, rati. the whole rationale. *Aliases*: canon is the canonical answer only, all is every accepted alias. *Positions*: scored reference positions, the separator included. *Span*: mean reference tokens per scored span; under all aliases an example contributes one span per alias. *rank 1*: fraction of positions at which the reference is the top token.

Benchmark	Prefix	Scope	Aliases	Scored	Excl.	Positions	Span	rank 1
Llama-3.1-8B-Instruct								
NQ-open	prompt-only	answ.	canon	2,000	0	7,167	3.58	0.558
TriviaQA	prompt-only	answ.	canon	2,000	0	5,785	2.89	0.697
GSM8K	gold	answ.	canon	1,319	0	2,786	2.11	0.779
MATH-500	gold	answ.	canon	500	0	2,068	4.14	0.697
Qwen2.5-7B-Instruct								
NQ-open	prompt-only	answ.	all	2,000	0	14,245	4.03	0.519
TriviaQA	prompt-only	answ.	all	2,000	0	50,092	4.11	0.389
GSM8K	prompt-only	answ.	canon	1,319	0	4,344	3.29	0.366
MATH-500	prompt-only	answ.	canon	500	0	2,501	5.00	0.473
NQ-open	prompt-only	answ.	canon	2,000	0	8,191	4.10	0.530
TriviaQA	prompt-only	answ.	canon	2,000	0	5,904	2.95	0.580
GSM8K	gold	answ.	canon	1,319	0	4,344	3.29	0.726
MATH-500	gold	answ.	canon	500	0	2,501	5.00	0.784
GSM8K	model	answ.	canon	903	416	2,905	3.22	0.740
MATH-500	model	answ.	canon	210	290	871	4.15	0.728
GSM8K	gold	rati.	canon	1,319	0	115,225	87.36	0.789
MATH-500	gold	rati.	canon	500	0	109,641	219.28	0.815

D.1 Reference protocols

Three prefix protocols were run: *gold*, the worked solution followed by the carrier; *prompt-only*, the prompt followed by the carrier; and *model*, the target's own generation truncated at its answer cue, which depends on decoding and is an ablation. Gold is primary for mathematics and prompt-only for question answering. A secondary scope scores the human-written solution token by token on mathematics, independent but off-policy and also an ablation. Aliases are deduplicated on a case- and whitespace-normalized key, at most eight per example.

Four carriers were scored on both question-answering benchmarks, and all four are reported. Judged by the ninetieth-percentile reference rank, where lower is nearer the head, the carrier used, "The answer is", is best on NQ-open at 351, against 403 for "Answer:", 534 for none and 901 for "It is". On TriviaQA no carrier is best at 98, against 167 for "Answer:", 447 for the carrier used and 4,861 for "It is". A carrier seems to help where the answer is a span a model would not open its turn with, and to hurt where it is a name the model would. Removing the carrier moves the levels far more than the comparison: on TriviaQA the ratio of top-p 0.95's rate to top-k 50's rises from 2.2 to 2.7, and on GSM8K under the prompt-only prefix top-p 0.95 rises from 0.2574 to 0.7275 and top-k 50 from 0.0003 to 0.3796, a bare answer opening a chat turn being far off-distribution. Table 6 varies each remaining protocol choice one at a time.

D.2 Separator and concentrated strata

Splitting content positions at the concentration threshold and adding the separator stratum leaves the ordering of Table 1 unchanged (full table in the supplementary material). Concentrated positions make up 0.9392 of GSM8K's content positions, 0.8162 of MATH-500's, 0.4370 of NQ-open's and 0.4806 of TriviaQA's. The reference is the top token at 0.9930 of the concentrated content positions on GSM8K, 0.9424 on MATH-500, 0.8116 on NQ-open and 0.8117 on TriviaQA. The complement is the confident-and-wrong rate. On GSM8K separator positions top-k 50 deletes at 0.1222 $[0.1048, 0.1405]$, against top-p 0.95's rate in Table 6.

Table 6: Deletion rate under top-p 0.95 on content positions unless a row names another stratum. The rows through the Llama row each change one thing against the GSM8K primary cell in the first row; the two TriviaQA rows are that benchmark's primary cell and the same cell scored against every accepted alias. *surplus*: mean surplus; *retained*: mean retained mass. The last row changes two things, the prompt-only prefix and no carrier; its one-change contrast is 0.2574 to 0.7275 (Appendix D.1).

Variant	Positions	deletion	[95% CI]	surplus	retained
primary (gold rationale, carrier, T=0.7)	3,027	0.011	$[0.007, 0.014]$	0.003	0.999
temperature T=1.0	3,027	0.010	$[0.006, 0.013]$	0.005	0.999
separator tokens only	1,317	0.732	$[0.708, 0.756]$	0.017	0.993
non-concentrated positions only	184	0.065	$[0.033, 0.102]$	0.047	0.990
prefix: prompt-only, no rationale	3,027	0.257	$[0.242, 0.273]$	0.072	0.984
prefix: the model's own generation	2,003	0.046	$[0.035, 0.057]$	0.001	1.000
scope: the whole gold rationale	102,758	0.147	$[0.145, 0.150]$	0.023	0.994
second model family (Llama-3.1-8B)	1,469	0.003	$[0.001, 0.007]$	0.019	0.994
TriviaQA, canonical answer	5,824	0.358	$[0.347, 0.370]$	0.022	0.989
TriviaQA, every accepted alias	49,771	0.542	$[0.534, 0.551]$	0.020	0.987
prompt-only prefix, carrier removed	3,027	0.727	$[0.712, 0.743]$	0.017	0.990

E Supplementary Teacher-Forced Results

E.1 Where the references sit

Within the primary protocol, the reference's ninetieth-percentile rank is 351 on NQ-open against 4 on GSM8K, and 44.03% of top-p 0.95's NQ-open deletions and 45.73% of its TriviaQA ones fall beyond rank 50. Restricted to references in the model's top ten, top-p 0.95 still deletes at 0.2001 $[0.1890, 0.2111]$ on NQ-open, 0.1689 $[0.1580, 0.1799]$ on TriviaQA, 0.1124 $[0.1003, 0.1249]$ on MATH-500 and 0.0102 $[0.0067, 0.0140]$ on GSM8K, and top-k 50 at exactly zero on all four. A question-answering level therefore reflects where the references sit; the separation does not. The concentrated share of GSM8K content positions is 0.4010 under Llama against Qwen's in Appendix D.2, so the size of that regime is a checkpoint property.

E.2 Deployed settings and the floor

The models ship $T = 0.7$ with top-p 0.8 and top-k 20 (Qwen2.5-7B-Instruct) and $T = 0.6$ with top-p 0.9 (Llama-3.1-8B-Instruct). Qwen's top-p 0.8 deletes at 0.4352 $[0.4233, 0.4471]$ of NQ-open content positions, more than top-p 0.95 in Table 1, so the main text reads conservatively. The shipped composite deletes at 0.4354 there, a paired difference before rounding of 0.0003 $[0.0000, 0.0007]$, and on mathematics the two are identical. `transformers` applies top-p after top-k, to the renormalized top-20 set, which scales the cumulative mass by $1/C_t[20]$ and can shorten the top-p cut even where top-p 0.8 alone keeps twenty or fewer. That rescaling can also only add deletions, so the composite measured here bounds the shipped stack's deletion rate from below. A co-applied count rule can only shorten a threshold rule's cut, whereas a floor can only lengthen it, so under the token floors of Table 7 deletion never rises and surplus and retained mass never fall.

E.3 The second model

On Llama-3.1-8B-Instruct, never pooled with Qwen, every qualitative statement of Sections 4.1 and 4.2 holds (Table 8): the threshold rules delete three to four times as often as top-k 50 on question answering. Every nonzero content rate is lower than Qwen's, which the higher rank-1 fractions of Table 5 explain on three benchmarks but not on MATH-500. The separator stratum inverts, MATH-500 carrying Llama's highest separator rate where GSM8K carried Qwen's, a statement about tokenizers rather than rules.

Table 7: Top-p 0.95 under a token floor on content positions, Qwen2.5-7B-Instruct, $T = 0.7$: deletion rate, mean surplus and mean retained mass; "none" is the unfloored rule. A floor of k tokens cuts at $\max(m_t^R, k)$.

Benchmark	none	floor 10	floor 50	floor 200
deletion rate				
GSM8K	0.011	0.000	0.000	0.000
MATH-500	0.119	0.007	0.000	0.000
NQ-open	0.390	0.239	0.172	0.125
TriviaQA	0.358	0.229	0.164	0.122
mean surplus				
GSM8K	0.0029	0.0034	0.0034	0.0034
MATH-500	0.0104	0.0113	0.0113	0.0113
NQ-open	0.0267	0.0306	0.0318	0.0324
TriviaQA	0.0222	0.0254	0.0266	0.0271
mean retained mass				
GSM8K	0.9994	1.0000	1.0000	1.0000
MATH-500	0.9979	1.0000	1.0000	1.0000
NQ-open	0.9870	0.9958	0.9981	0.9992
TriviaQA	0.9892	0.9965	0.9986	0.9994

Table 8: Deletion rate on Llama-3.1-8B-Instruct at $T = 0.7$ by stratum, in the layout of Table 1.

Benchmark	Positions	top-p 0.95	min-p 0.05	top-$n\sigma$ 1	top-k 50	top-p 0.95, ≥ 50
content tokens (all entropies)						
GSM8K	1,469	0.003	0.003	0.003	0.000	0.000
MATH-500	1,672	0.064	0.062	0.065	0.000	0.000
NQ-open	6,524	0.323	0.343	0.340	0.109	0.108
TriviaQA	5,705	0.207	0.219	0.215	0.057	0.056
$H(\tilde{p}_t) < 0.01$ *nats*						
GSM8K	589	0.003	0.003	0.003	0.000	0.000
MATH-500	835	0.010	0.010	0.010	0.000	0.000
NQ-open	1,965	0.047	0.047	0.047	0.014	0.014
TriviaQA	2,405	0.049	0.049	0.049	0.003	0.003
$H(\tilde{p}_t) \geq 0.01$ *nats*						
GSM8K	880	0.003	0.003	0.003	0.000	0.000
MATH-500	837	0.118	0.114	0.119	0.000	0.000
NQ-open	4,559	0.442	0.471	0.466	0.150	0.148
TriviaQA	3,300	0.322	0.343	0.337	0.096	0.095
separator tokens (all entropies)						
GSM8K	1,317	0.172	0.181	0.190	0.000	0.000
MATH-500	396	0.449	0.465	0.472	0.000	0.000
NQ-open	643	0.165	0.176	0.173	0.053	0.053
TriviaQA	80	0.113	0.113	0.113	0.000	0.000

E.4 What each rule keeps at concentrated positions

Table 9 reads the threshold condition of Proposition 1 against the sweep at the $19{,}074$ concentrated content positions, pooled over both temperatures. All 23 configurations the condition covers keep exactly one token at every such position, as do top-$n\sigma$ at $n \leq 1.5$ and ANTS at $\gamma \leq 5$, which it does not cover.

E.5 Adaptivity isolated: η- against ϵ-sampling

Of the two rules of Hewitt et al. (2022), η-sampling is ϵ-sampling with an adaptive term, keeping tokens above $\min(\epsilon, \sqrt{\epsilon}e^{-H})$, so the pair isolates adaptivity. The term binds only when $H > \frac{1}{2}\log(1/\epsilon)$, which is 4.06 nats at $\epsilon = 0.0003$. That high-entropy band holds 503 of the $41{,}880$

Table 9: What each family keeps where $H(\tilde{p}_t) < 0.01$, over the 19,074 concentrated content positions, both temperatures pooled. *Threshold*: whether Proposition 1 covers the setting. $\Pr[m_t^R > 1]$: share of those positions keeping more than one token; the last column is its range over the family's truncating grid, top-p 1.00 and floored composites excluded. Means pool temperatures and so can differ from Figure 2.

Family	setting	threshold ≥ 0.002	$\Pr[m_t^R > 1]$	mean m_t^R	over the grid
top-k	50	no, a count rule	1.0000	50.00	1.0000–1.0000
top-p	0.95	yes, $p \leq 0.998$	0.0000	1.00	0.0000–0.0000
min-p	0.05	yes, $\alpha \geq 0.002/0.998$	0.0000	1.00	0.0000–0.0059
top-b	0.3	yes, every $b \leq 0.5$	0.0000	1.00	0.0000–0.0000
p-less	–	yes, no parameter	0.0000	1.00	0.0000–0.0000
ϵ-sampling	0.0003	no, needs $\epsilon \geq 0.002$	0.0665	1.07	0.0095–0.1450
η-sampling	0.0003	no, needs $\epsilon \geq 0.002$	0.0665	1.07	0.0000–0.1450
top-$n\sigma$	1	no, a logit spread	0.0000	1.00	0.0000–0.3203
ANTS	5	no, a logit spread	0.0000	1.00	0.0000–0.0519
ANTS	32	no, a logit spread	0.0198	92.64	0.0000–0.0519
min-k	3	no, a cut index	0.0014	1.00	0.0014–1.0000
top-H	0.4	no, a prefix entropy	0.1897	1.36	0.0024–0.7394

scored positions, both temperatures pooled. Inside it η-sampling lowers the deletion rate from ϵ-sampling's 0.2883 to 0.2286, yet the band carries only 1.62% of ϵ-sampling's deletions, so the corpus rate moves from 0.2138 to 0.2131 and the two rules agree on 98.80% of cuts. The adaptive term never reaches the concentrated region, where the two thresholds coincide.

F Behavioural Arms

F.1 Generation arm

Prompts of similar length are batched together, and every configuration in a group draws from the same per-batch seed. The seven established configurations ran on NVIDIA RTX A6000 GPUs and the six recent-rule configurations on NVIDIA L40S GPUs, so generated text is compared only within each group. Accuracy intervals overlap across configurations (Table 10), so the table supports only the statement that the floor costs no accuracy this corpus could detect.

The pooled association is confounded by difficulty. TriviaQA failure runs 0.2136 where neither top-p 0.95 nor its floored form deletes, 0.4392 where only the unfloored rule does and 0.6597 where both do. For every truncating configuration except the two ANTS arms, deleting on $1{,}920$ to $4{,}752$ of $6{,}000$ responses, the risk difference lies between $+0.27$ and $+0.33$; the ANTS arms, which delete on many correctly answered responses, sit at $+0.008$ $[-0.037, +0.054]$ for $\gamma = 32$ and $+0.037$ $[-0.004, +0.077]$ for $\gamma = 5$. On GSM8K it runs from $+0.032$ to $+0.141$ over 63 to 171 deleted responses, too noisy to read. A difficulty marker and a causal effect both produce such a difference, which is why Section 4.4 holds the prompt fixed.

Order of truncation and temperature in the floored arms. The diagnostic applies every rule to $\tilde{p}_t$, after temperature. The two floored generation arms, top-p 0.95 with a floor of 50 or 200 tokens, were built by passing `TopPLogitsWarper` with `min_tokens_to_keep` as a user-supplied logits processor, and the library runs such a processor ahead of its own temperature step. Those two arms therefore truncated the $T = 1.0$ distribution and then sampled at $T = 0.7$, and their teacher-forced counterpart is the $T = 1.0$ column of the floored rule rather than the $T = 0.7$ one. The other eleven arms truncate after temperature, as the diagnostic does. Statements that pair a floored arm with a $T = 0.7$ rate, in Figure 3, in the floored contrast of Section 4.4 and in the deletion difference of Appendix F.1, inherit this mismatch; the headline contrast against top-k 50 does not.

Which configurations emit the same text. The tail-keeping class of Section 4.3 comprises top-k 20 and 50, top-p 0.95 under floors of 50 and 200, and untruncated sampling. On GSM8K every pair inside the class emits identical text on at least 99.72% of responses: the top-p 0.95 + floor 200 and untruncated sampling agree on all of them, and top-k 20 and 50 on 99.75%. On TriviaQA the class

Table 10: Task accuracy per generated configuration, three seeds pooled, 95% bootstrap intervals; *mean tokens*: mean response length in tokens. ≥ 50 marks a fifty-token floor and top-p 1.00 is untruncated sampling. The top-p 0.95 + floor 200 is omitted, agreeing with floor 50 to four decimals on both benchmarks.

Benchmark	Configuration	n	accuracy	[95% CI]	mean tokens
GSM8K	ANTS 32	3,957	0.8668	$[0.8514, 0.8817]$	232.5
GSM8K	ANTS 5	3,957	0.8683	$[0.8529, 0.8832]$	232.3
GSM8K	min-k 3	3,957	0.8696	$[0.8544, 0.8845]$	232.0
GSM8K	p-less	3,957	0.8850	$[0.8686, 0.9009]$	231.3
GSM8K	top-b 0.3	3,957	0.8782	$[0.8623, 0.8939]$	232.2
GSM8K	top-H 0.4	3,957	0.8817	$[0.8645, 0.8984]$	232.6
GSM8K	top-k 20	3,957	0.8731	$[0.8582, 0.8875]$	233.1
GSM8K	top-k 50	3,957	0.8731	$[0.8582, 0.8875]$	233.2
GSM8K	top-p 0.90	3,957	0.8802	$[0.8655, 0.8946]$	231.8
GSM8K	top-p 0.95	3,957	0.8797	$[0.8653, 0.8939]$	232.4
GSM8K	top-p 0.95, ≥ 50	3,957	0.8731	$[0.8582, 0.8875]$	233.2
GSM8K	top-p 1.00	3,957	0.8731	$[0.8582, 0.8875]$	233.2
TriviaQA	ANTS 32	6,000	0.5910	$[0.5707, 0.6117]$	17.6
TriviaQA	ANTS 5	6,000	0.5918	$[0.5717, 0.6123]$	17.6
TriviaQA	min-k 3	6,000	0.6043	$[0.5833, 0.6253]$	17.4
TriviaQA	p-less	6,000	0.6085	$[0.5872, 0.6295]$	17.8
TriviaQA	top-b 0.3	6,000	0.6067	$[0.5853, 0.6277]$	17.7
TriviaQA	top-H 0.4	6,000	0.6070	$[0.5855, 0.6280]$	17.8
TriviaQA	top-k 20	6,000	0.5902	$[0.5697, 0.6105]$	17.7
TriviaQA	top-k 50	6,000	0.5900	$[0.5695, 0.6103]$	17.7
TriviaQA	top-p 0.90	6,000	0.5932	$[0.5723, 0.6138]$	17.7
TriviaQA	top-p 0.95	6,000	0.5937	$[0.5732, 0.6142]$	17.7
TriviaQA	top-p 0.95, ≥ 50	6,000	0.5898	$[0.5692, 0.6102]$	17.7
TriviaQA	top-p 1.00	6,000	0.5898	$[0.5692, 0.6102]$	17.7

is looser, at least 98.40% identical, or 99.30% without top-k 20: top-k 20 and 50 agree on 99.02%, and top-k 50 and untruncated sampling on 99.30%.

Pairs a score cannot separate. Every pair whose accuracy difference the benchmark cannot resolve is listed, with its deletion and surplus differences, in the supplementary material. The pairs with exactly zero accuracy difference all lie inside the tail-keeping class, and on TriviaQA that zero difference hides a deletion difference of 0.1638 between top-p 0.95 + floor 50 and untruncated sampling.

F.2 Discordant comparisons

A comparison is one configuration pair on one benchmark under one scoring, restricted to the discordant prompts $\mathcal{P}$ of Equation 2; it reports Δ, failure under the deleting configuration minus failure under the retaining one. The two of Section 4.4's 34 that exclude zero are both on TriviaQA: top-k 20 against top-p 0.90, at -0.0074 $[-0.0143, -0.0006]$, and against top-p 0.95, at -0.0071 $[-0.0130, -0.0013]$. They share a top-k arm and most of their prompts, so they are one finding, reported uncorrected; under alias scoring none of the 34 excludes zero. Of the 34, 24 lie inside the margin, 13 of them with zero-width intervals because both rules produced identical outcomes; ten of those 13 pair two members of the tail-keeping class, and the other three rest on 2 to 7 prompts. The intervals outside the margin reach a half-width of 0.0614, which is underpowering rather than evidence of an effect; the 11 GSM8K comparisons rest on 7 to 38 prompts, against 2 to 1,512 for the 23 TriviaQA ones. Over all 169 comparisons, the deleting configuration fails less by at most 0.0897, for ANTS at $\gamma = 5$ against top-b 0.3 on GSM8K, where top-b deletes. The one comparison in which it fails more is min-k at $\tau = 3$ against top-p 0.95 on 8 GSM8K prompts, where top-p deletes, at $+0.1667$. Of the 169, 91 are equivalent at the margin, 29 of them with zero-width intervals.

F.3 CLOSED-SET ARM

Each of the 2,000 TriviaQA prompts receives four literal candidates: its gold and three distractors drawn at the sample's seed from other prompts' golds, redrawn if one equals the gold or the alias matcher would accept it. An item whose candidates tokenize non-uniquely is dropped. The candidates appear in random order, and generation is constrained to the trie of their token sequences. In the primary order the candidate mask is applied, then the temperature, then the rule, and the survivors are renormalized at every node, so the gold's probability is the exact product along its path. The mask leaves at most four live tokens per node, so no top-k with $k \geq 5$ removes anything and every count-rule arm coincides with the untruncated model. Accuracy is the mean over items of the gold's probability; the untruncated model scores 0.9704, and every measured rule scores at least that.

Reversed order. Running the rule before the mask, on the same items, changes only what the rule sees; the untruncated model again scores 0.9704. Its probability of answering the discordant items correctly rises to 0.0140 on the 43 items where top-p 0.95 deletes and top-k 50 does not, and to 0.0396 on the 49 where top-p 0.8 deletes and top-k 20 does not, against 0.0042 and 0.0179 in the primary order. The corpus lead shrinks to $+0.0001$ $[-0.0004, +0.0004]$ and $+0.0003$ $[-0.0006, +0.0011]$, both intervals now containing zero, so a threshold rule's lead on this corpus, like the 0.0042 of Section 4.4, holds in the primary order alone. On the discordant items a deletion zeroes the gold's probability, so only the magnitude informs: -0.0140 $[-0.0333, -0.0030]$ and -0.0396 $[-0.0672, -0.0168]$ here, against -0.0042 and -0.0179 in the primary order. The reversed order also creates items both configurations delete, 1 of 2,000 against top-k 50 and 3 against top-k 20, where a deletion is fatal whichever rule causes it.

G ROBUSTNESS CHECKS

G.1 WHICH REFERENCES EACH RULE DELETES

Two rules at one rate can delete different references. Table 11 pairs configurations matched at their closest achievable rates and compares the Jaccard index of their deleted position sets with its value under independence, on non-concentrated content positions, since a rate near zero or one forces overlap arithmetically. The twelve matches among top-p, min-p and top-$n\sigma$, within 0.0068 in rate, overlap more than the four matches of top-k against an unfloored threshold rule within 0.02, and all lie far above independence, so the families differ partly in which references they remove, not only in how many. Support is uneven: three GSM8K rows rest on 8 to 12 deleted positions and are exact counts rather than estimates, every other printed row rests on 45 or more, and some unprinted ANTS and min-k matches on GSM8K rest on as few as 11.

G.2 TEMPERATURE INVARIANCE ON PAIRED POSITIONS

Proposition 2 assigns each rule a clause from its definition, and the assignment is falsifiable: a position scored at both temperatures has the same prompt, span index and reference token, so an invariant rule must return the same cut at both. No reference rank changes across the $20{,}940$ paired positions, as clause (i) requires; the single rank disagreement this study records is between batched and unbatched scoring (Appendix H). Table 12 counts the cuts that moved. Top-$n\sigma$'s invariance is exact in real arithmetic, but its criterion is evaluated on logits divided by T in float32, so its count is of positions at which a threshold fell inside the rounding, not of violations.

G.3 THE MATCHED-BUDGET FRONTIER

The frontier asks how little a rule must delete within a budget on mean surplus. Figure 4 draws it on GSM8K and NQ-open over the established sweep; the figures below cover all four benchmarks and all 77 configurations, on non-concentrated content positions at $T = 0.7$, and a family with no configuration inside a budget is absent rather than extrapolated. At a budget of 0.05 the two kinds stay apart on question answering with non-overlapping intervals, the count rule deleting less: top-k 5 deletes at 0.4544 $[0.4392, 0.4697]$ on NQ-open against the best unfloored threshold rule, top-H at $\beta = 0.9$, at 0.5114 $[0.4962, 0.5269]$, and top-k 20 deletes at 0.3028 $[0.2882, 0.3178]$ on TriviaQA

Table 11: Overlap of deleted position sets between rules of two families matched at the closest deletion rates the two grids attain, non-concentrated content positions, $T = 0.7$. *Gap*: difference of the matched rates; *indep.*: Jaccard index under independence; *union*: positions either rule deletes, the index's whole support.

Benchmark	Rule a	Rule b	rate a	rate b	gap	Jaccard	indep.	union
GSM8K	min-p 0.02	top-p 0.98	0.049	0.049	0.000	1.000	0.025	9
GSM8K	min-p 0.01	top-$n\sigma$ 1.5	0.043	0.038	0.005	0.875	0.021	8
GSM8K	top-$n\sigma$ 1	top-p 0.95	0.060	0.065	0.005	0.917	0.032	12
MATH-500	min-p 0.02	top-p 0.98	0.328	0.328	0.000	0.969	0.196	129
MATH-500	min-p 0.2	top-$n\sigma$ 0.5	0.463	0.465	0.003	0.983	0.302	181
MATH-500	top-k 5	top-$n\sigma$ 3	0.096	0.098	0.003	0.667	0.051	45
MATH-500	top-$n\sigma$ 0.5	top-p 0.8	0.465	0.470	0.005	0.967	0.305	184
MATH-500	min-p 0.001	top-k 5	0.181	0.096	0.085	0.507	0.067	71
MATH-500	top-k 5	top-p 0.98	0.096	0.328	0.233	0.291	0.080	127
NQ-open	top-$n\sigma$ 0.5	top-p 0.8	0.624	0.627	0.003	0.967	0.455	2,701
NQ-open	min-p 0.2	top-p 0.8	0.631	0.627	0.004	0.972	0.459	2,708
NQ-open	min-p 0.2	top-$n\sigma$ 0.5	0.631	0.624	0.007	0.985	0.457	2,685
NQ-open	top-k 100	top-$n\sigma$ 4	0.222	0.213	0.009	0.642	0.122	1,126
NQ-open	min-p 0.005	top-k 5	0.482	0.454	0.028	0.792	0.306	2,220
NQ-open	top-k 5	top-p 0.98	0.454	0.498	0.044	0.735	0.312	2,331
TriviaQA	top-$n\sigma$ 0.5	top-p 0.8	0.571	0.572	0.001	0.978	0.400	1,749
TriviaQA	min-p 0.2	top-p 0.8	0.575	0.572	0.003	0.978	0.402	1,754
TriviaQA	min-p 0.2	top-$n\sigma$ 0.5	0.575	0.571	0.004	0.993	0.402	1,740
TriviaQA	top-k 20	top-$n\sigma$ 3	0.303	0.307	0.004	0.738	0.180	1,061
TriviaQA	min-p 0.001	top-k 5	0.402	0.408	0.006	0.824	0.254	1,344
TriviaQA	top-k 5	top-p 0.98	0.408	0.470	0.062	0.752	0.279	1,516

Table 12: Cuts that moved between $T = 0.7$ and $T = 1.0$ on paired positions, summed over each family's settings. *Settings* include floored composites: top-p's 12 are six grid values and six composites, and min-p's 9 and top-$n\sigma$'s 8 each include one. *Clause*: the part of Proposition 2 the definition assigns; *pre-temperature*: the rule reads logits before temperature; p-less$_{\text{norm}}$ is the normalized variant of p-less.

Rule	Settings	Clause	Predicted	Paired positions	Cuts that moved
no truncation	1	(ii)	no truncation	20,940	0
top-k	6	(ii)	scale-free	125,640	0
top-$n\sigma$	8	(ii)	scale-free	167,520	3
ANTS	8	(ii)	pre-temperature	167,520	0
min-k	7	(ii)	pre-temperature	146,580	0
min-p	9	(iii)	reparameterized	188,460	38,758
top-p	12	(iv)	not invariant	251,280	31,035
ϵ-sampling	4	(iv)	not invariant	83,760	41,259
η-sampling	5	(iv)	not invariant	104,700	49,529
top-H	7	(iv)	not invariant	146,580	45,516
top-b	8	(iv)	not invariant	167,520	3,434
p-less	1	(iv)	not invariant	20,940	897
p-less$_{\text{norm}}$	1	(iv)	not invariant	20,940	897

against min-k at $\tau = 300$, at 0.3709 $[0.3557, 0.3870]$. On GSM8K no count rule fits that budget, and the threshold families order min-k at $\tau = 100$ first, at 0.0489 $[0.0215, 0.0815]$, then top-$n\sigma$ 1 at 0.0598 $[0.0275, 0.0958]$ and top-p 0.95 at 0.0652 $[0.0326, 0.1022]$. The kinds read the same rate only where it is zero, on GSM8K at the two loosest budgets.

Selecting on the scored positions makes the chosen rate optimistic. When each family's best configuration is chosen on half the prompts and re-read on the other half, over 20 halvings in both directions, the largest optimism of any fully supported family at the 0.05 budget is 0.0087 on GSM8K, 0.0074 on MATH-500, 0.0027 on TriviaQA and 0.0006 on NQ-open. Re-selection moves the difference between the top-p and top-k representatives by 0.0002 on NQ-open and by 0.0023 on TriviaQA, far less than the differences themselves, 0.1186 and 0.1857.

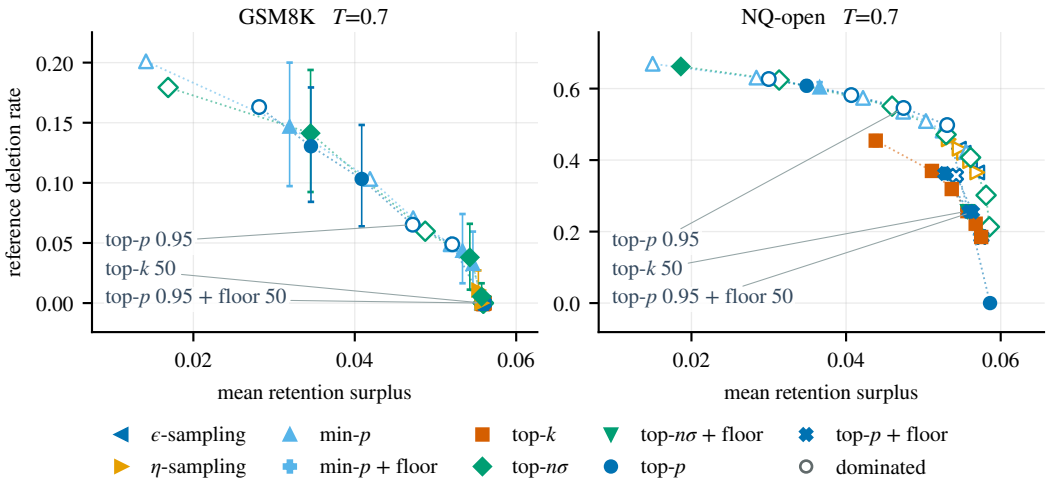


Figure 4: Deletion rate (lower keeps the reference more often) against mean surplus (axis label: retention surplus), Qwen2.5-7B-Instruct, non-concentrated content positions, established sweep, $T = 0.7$. Filled markers are Pareto-efficient and hollow ones dominated, with 95% intervals; dotted lines join one family in hyperparameter order, and nothing between markers was evaluated.

Read the other way, with a budget on deletion, one of four fixed in the run configuration, the frontier is almost flat. The cheapest surplus over the established sweep is 0.061 at every budget on MATH-500, top-k's. On NQ-open and TriviaQA it is 0.059 and 0.053, met only by the untruncated baseline and top-p 1, which renames it; only GSM8K varies, over 0.047 to 0.056 with overlapping intervals. Of the recent rules only ANTS meets these budgets on question answering, at $\gamma = 64$ for the tightest and $\gamma = 20$ for the loosest, always at the untruncated surplus to three decimals. Per-budget and per-configuration rows are in the supplementary material.

G.4 A REASONING MODEL'S OWN LONG CHAINS

ANTS is argued for at budgets of 8K to 32K tokens (Dia, 2026), whereas the answer spans here run from a few tokens to a few hundred, so this arm measures deletion over long chains directly rather than projecting per-position rates. DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025) generated chains of thought for 250 MATH-500 prompts without truncation at a budget of 8,192 tokens, of which the first 4,096 were scored: 624,279 positions over chains of median length 2,454. An untruncated reference is independent of every rule under test, though not of the model.

Deletion does not concentrate at either end of a chain. Top-p 0.95 deletes at 0.0085 of first-decile positions and 0.0051 of last-decile ones, peaking at 0.0113 near the seventh decile, and every rule traces the same shallow arc. Against the model's own chain, top-k 50 deletes at 10 of the 624,279 positions. Projected from the per-position rates on human-written rationales, the probabilities of at least one deletion within a thousand tokens would be 0.9929 for top-k 50 and 1.0000 for top-p 0.95, against the measured values of Section 4.5, because a human-written solution contains tokens the model ranks low and its own trajectory does not. The projection errs most where the rate is low: η-sampling at $\epsilon = 0.0003$ also projects to 1.0000 against a measured 0.192. The family ordering survives the change of reference and the levels do not; this arm scores deletion, not accuracy.

H IMPLEMENTATION AND DETERMINISM

Tests and checks. Unit tests cover the tie convention, the reference rank, every rule family against a hand-computed example and against brute force over explicit token sets at every swept setting, the rank-prefix property, Proposition 1 at every setting it covers, the sign derivation of Section 2.2 and the saturated sum that defeats its converse, each clause of Proposition 2, the top-p edge cases, the impossibility of an empty nucleus and the floored composites. Three checks run against the real

model. Decoded reference spans must equal the reference string, which catches a protocol scoring one position early. Scoring a span alone and in a batch must agree within 10^{-4} in reference mass; the largest disagreement is 2.77×10^{-5}. A brute-force recomputation with each rule as an explicit token set reproduced every deletion decision in $3{,}976$ rule comparisons at 142 positions from 40 spans, with a largest mass or surplus disagreement of 2.02×10^{-5}; ϵ- and η-sampling, added later, are covered by the unit tests. Every result row records protocol, scope, alias handling and carrier, every unscorable example carries a machine-readable reason, and both arms draw prompts under one permutation at one seed, so they join on the prompt identifier.

Precision. Scoring the same spans alone and inside a batch disagrees by up to 0.69 in a logit under bfloat16, a factor of 2.0 in a probability ratio, against 2.9×10^{-4} in float32 with TF32 off, so under bfloat16 a deletion rate would depend on which examples shared its batch. Float32 still bounds two quantities. Where the reference sits deep in the tail, the cumulative sum saturates at one and $G_t(R)$ reads exactly zero, at $10{,}633$ of the $174{,}238$ deletions of the established sweep on the primary cells' content positions at $T = 0.7$ and at $15{,}471$ of $341{,}674$ over all 77 configurations; this is why deletion is read from cut and rank. The brute-force recomputation found one cut differing by one token, affecting no deletion decision, and one rank differing between batched and unbatched scoring, so a cut at a probability boundary is reproducible to a token, not exactly.

Hardware and determinism. Teacher-forced scoring ran on NVIDIA L40S GPUs (47 GiB, 142 streaming multiprocessors, compute capability 8.9) on a 32-core host, under CUDA 13.0, PyTorch 2.13, `transformers` 5.14 and Python 3.11. Independent cells are spread over 4 such GPUs, and no cell spans two. For scoring, weights are float32 with TF32 disabled on both the matrix-multiply and cuDNN paths; the generation arm samples in bfloat16 on the GPUs named in Appendix F.1. PyTorch's deterministic-algorithms mode is enforced, cuDNN runs in deterministic, non-benchmarking mode, and the cuBLAS workspace configuration is fixed before the first CUDA call; every operation used has a deterministic CUDA implementation. The printed figures carry no creation date, so regenerating them reproduces them byte for byte.

Scale. The cost bound of Section 2.4 is exercised at $20{,}940$ positions from $5{,}819$ examples, all 77 configurations resolved from one forward pass per span. The generation arm decodes the responses counted in Table 10 for each of its thirteen configurations, and the closed-set arm masks and rescores $2{,}000$ prompts per configuration. Wall-clock time was not instrumented, so no compute total is claimed.