TRUNCATION SAMPLERS ARE NESTED: COMPARING DECODING RULES IS A MEASUREMENT PROBLEM

Anonymous authors
Paper under double-blind review

ABSTRACT

Settling one comparison between two truncation samplers has cost six thousand A100-hours, and nothing says in advance how far apart two samplers can be. We observe that almost every deployed truncation rule keeps a prefix of the probability-sorted vocabulary, so any two such rules are nested at every decoding step. Nesting yields a closed-form per-step distance and a coupling under which both samplers emit identical text until they first disagree. Accumulated over a generation, this gives a separation certificate that bounds the difference between two samplers on any bounded outcome, is computable from logs a comparison already produces, and equals their sequence-level total variation distance exactly when neither rule ever overtakes the other. Two consequences follow without measurement: exactness is fragile, and certifying a small difference requires the two retained masses to agree to very high precision at every step. On six open-weight models over four math and science benchmarks, the certificate for every hyperparameter grid we tested is at least 0.998, so a bound alone rarely settles a comparison, although two members of one family can be far closer. What remains useful is the coupling, which gives an unbiased paired estimator that cuts variance 1.2–$3.8\times$ against prompt-level pairing. Read against benchmark sizes at the default pair's variance and our point estimate of its between-prompt component, resolving a one-point difference needs 8 to 32 samples per prompt on GSM8K, MATH-500 and GPQA-Diamond and about 210 on AIME 2025; at that component's upper confidence bound GPQA-Diamond needs more than 32 and AIME 2025 cannot be resolved at any sample count. Choosing a truncation sampler is a measurement problem at least as much as a design one.

1 INTRODUCTION

Training-free samplers decide how to draw from a language model's next-token distribution, and the dominant family truncates. Top-k keeps a fixed number of tokens (Fan et al., 2018), top-p a fixed cumulative mass (Holtzman et al., 2019), and a further series of rules from the last five years adapt the cutoff to the shape of the distribution (Hewitt et al., 2022; Basu et al., 2020; Nguyen et al., 2025; Tang et al., 2024; Tan et al., 2026; Baghaei Potraghloo et al., 2026; Halder & Dabre, 2026; Dia, 2026; Ding et al., 2026; Davoodi et al., 2026). Roughly a dozen are in use, and choosing among them is expensive and contested. A re-examination of min-p found its reported advantage does not survive once swept hyperparameters are equalized across samplers (Schaeffer et al., 2025), at a cost of six thousand A100-hours, and that cost is not specific to one paper. Reasoning benchmarks swing by double digits across seeds (Hochlehnert et al., 2025), benchmark variance is under-reported (Madaan et al., 2024; Biderman et al., 2024), and run-to-run variation can misrank systems (Messing, 2026).

Two things make these comparisons harder than they need to be. First, nothing says in advance how far apart two truncation rules can be, so a study cannot be sized before it is run and a reported gap cannot be checked against what the samplers could have produced. Second, every protocol that makes two rules comparable does so by equalizing a scalar summary, whether hyperparameter grid volume (Schaeffer et al., 2025), estimated risk level (Zhou et al., 2025), or expected retained mass (Kempton & Burrell, 2025). To our knowledge none has been shown to be the quantity that governs how far apart the rules can be, so a matched comparison rests on an untested assumption.

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.

Most samplers above, all but Top-W and locally typical sampling, retain a *prefix* of the vocabulary sorted by probability, so at any decoding step the candidate sets of any two are nested, meaning one contains the other. The narrower rule's distribution is then exactly the wider rule's conditioned on the narrower set, so the distance between them is a ratio of two numbers decoding already computes. The two rules can also be run so they emit identical tokens until the first draw falls in the band between the cutoffs (Figure 1). That coupling accumulates the per-step distance into a bound on any bounded outcome, computable from logs a comparison already produces. Written in log form it also answers the second problem, because it is a dispersion while the matching criterion is a mean.

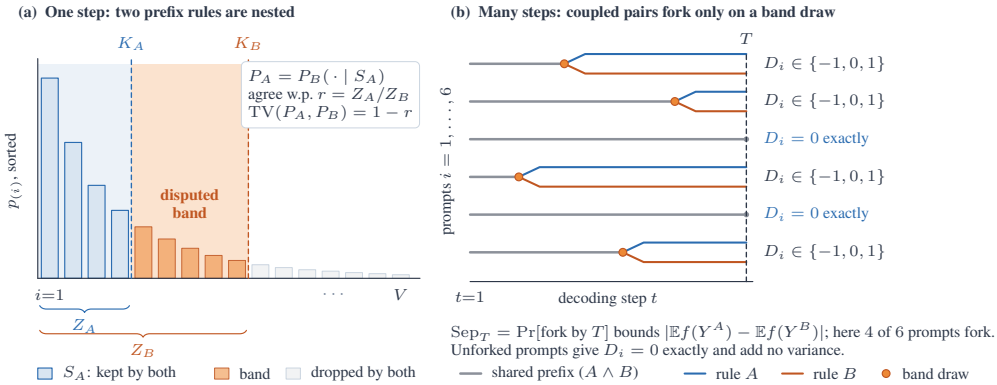


Figure 1: **Nesting gives a free per-step distance and a coupling.** **(a)** Two prefix rules cut the sorted vocabulary at $K_A \le K_B$ and disagree only on the shaded band. Drawing from B, both emit the draw if it lands in S_A, otherwise A redraws, so they agree with probability $r = Z_A/Z_B$. **(b)** Coupled pairs share tokens until a draw lands in the band, then fork. An unforked prompt gives $D_i = f(Y_i^A) - f(Y_i^B) = 0$ exactly, and the fraction that forks estimates Sep_T.

Measured on real models, the certificate is large. It is at least 0.998 between the narrowest and widest members of every hyperparameter grid we tested, which is expected once a grid contains untruncated and near-greedy members, and it stays large between rules matched on operating point, which matching does not prevent, since it equalizes a mean while separation is governed by a dispersion. Under the maximal coupling such samplers diverge on almost every generation, and for the grid extremes, which never flip, this is also their exact total variation distance. Their accuracies barely differ: on the eight pairs where we measured both, every accuracy gap is at most 2.4% of its certificate. A bound cannot settle such a comparison in advance, so it has to be measured, and nesting helps there too. The same coupling gives a paired estimator whose variance is bounded by the certificate, and reading its measured variances against benchmark sizes shows which benchmarks can resolve a one-point difference at all. Choosing a truncation sampler is a measurement problem at least as much as a design one.

Contributions. (1) We show that prefix truncation rules are nested, which gives an exact per-step distance and a rejection-free maximal coupling between any two of them (Lemma 2). (2) Accumulating it gives the *separation certificate*, which bounds the difference on any bounded outcome, is exact iff no trajectory flips which rule is narrower, and covers an entire hyperparameter grid from one decoding run (Theorem 3, Corollary 4). (3) Two structural results say when a certificate can be exact or small: exactness decays geometrically once the narrower rule can change (Proposition 5), and a sufficient per-step budget with a dispersion lower bound bracket when a small certificate is possible (Corollary 6). (4) Measured on seven open-weight models (six for the grid, plus a distilled reasoner for long horizons), the certificate is large on every grid we tested, the coupling cuts estimator variance 1.2–$3.8\times$, and the resulting sample-size arithmetic shows which benchmarks can resolve a one-point difference (Sections 4 and 5).

2 Prefix Truncation Rules

Fix a vocabulary $\mathcal{V}$ of size V. At step t with context $x_{<t}$ the model gives $p_t \in \Delta^{V-1}$; write $p_{(1)} \ge \cdots \ge p_{(V)}$ for the probabilities in descending order.

Table 1: **Most deployed truncation samplers return a cutoff index in the sorted order,** so the retained set is a prefix and any two such rules are nested at every step; Top-W and locally typical sampling, marked below the rule, do not. The nine rows above the gap are the family used in Section 4: six marked • each contribute six hyperparameter values, and three marked single (basic, p-less and Top-H, none swept over multiple settings) contribute one member each.

Sampler	Cutoff rule	Prefix	Swept
basic (no truncation)	$K = V$	✓	single
top-k (Fan et al., 2018)	$K = k$	✓	•
top-p (Holtzman et al., 2019)	smallest K with $Z_K \geq p$	✓	•
ϵ (Hewitt et al., 2022)	keep $p_{(i)} \geq \epsilon$	✓	•
η (Hewitt et al., 2022)	keep $p_{(i)} \geq \min\{\epsilon, \sqrt{\epsilon}e^{-H}\}$	✓	•
min-p (Nguyen et al., 2025)	keep $p_{(i)} \geq p_{\text{base}}\, p_{(1)}$	✓	•
top-$n\sigma$ (Tang et al., 2024)	keep $\ell_{(i)} \geq \ell_{(1)} - n\,\mathrm{sd}(\ell)$	✓	•
p-less (Tan et al., 2026)	collision entropy, no hyperparameter	✓	single
Top-H (Baghaei Potraghloo et al., 2026)	bounded-entropy cutoff at its default	✓	single
min-k (Ding et al., 2026)	index at the local logit-decay cliff	✓	
mirostat (Basu et al., 2020)	index from a perplexity controller	✓	
top-W (Davoodi et al., 2026)	prefix of a scan over embedding geometry, not p	×	
locally typical (Meister et al., 2023)	orders by $\lvert -\log p_{(i)} - H\rvert$	×	

Every rule above the gap returns a prefix of the sorted order, most by comparing a threshold against $p_{(i)}$, which is monotone in i; min-k and mirostat return an index directly. Across 997,465,872 logged cutoffs there are zero out-of-range cutoffs and zero non-nested pairs (mirostat, not implemented, is classified from its description). Top-W orders tokens by a score that mixes embedding geometry with log-probability, and locally typical sampling omits the top token on 8.8% of our logged steps at $\tau = 0.2$ (Appendix D).

Definition 1 (Prefix rule)**.** A *prefix rule* maps a context and a next-token distribution to a cutoff index $K(x_{<t}, p_t) \in \{1, \ldots, V\}$, retains $S_K = \{(1), \ldots, (K)\}$ and samples from $P_K(i) = p_{(i)}/Z_K$, where $Z_K = \sum_{i \leq K} p_{(i)}$ is the *retained mass*. The order is sorted by descending probability and breaks ties by token index, and all rules under comparison must use the same order.

Two prefixes of one total order are always nested, so for any prefix rules A, B and any step, $S_A \subseteq S_B$ or $S_B \subseteq S_A$, and every result below rests on this. Table 1 writes the deployed samplers in this form.

Comparing rules also needs a common notion of aggressiveness, since any sampler can behave like greedy or untruncated decoding and raw hyperparameters are not comparable across families. Following Kempton & Burrell (2025), we match on the operating point $\rho(K) = \mathbb{E}[Z_K]$, the expected retained mass over a teacher-forced calibration corpus of 200 held-out GSM8K prompts (Appendix E). Rules are operating-matched at level ρ when their hyperparameters give equal operating points.

3 Separation Theory

Nesting becomes a distance, the distance becomes a bound, and the coupling behind the bound becomes an estimator. Section 3.1 derives the certificate for a pair and for a whole grid, Section 3.2 asks when it can be exact or small, and Section 3.3 turns the coupling into an estimator. Proofs are in Appendix A.

3.1 Nesting gives an exact distance

Write A, B for two prefix rules acting on the same p_t, with retained masses Z_A, Z_B and $r = \min\{Z_A, Z_B\}/\max\{Z_A, Z_B\}$, defined by a minimum and a maximum since which rule is narrower can change from step to step.

Lemma 2 (Nesting)**.** *At any step, $S_A \subseteq S_B$ or $S_B \subseteq S_A$. Relabel so $S_A \subseteq S_B$. Then* (i) *$P_A = P_B(\cdot \mid S_A)$, hence $\mathrm{TV}(P_A, P_B) = 1 - Z_A/Z_B = 1 - r$; and* (ii) *drawing $X \sim P_B$ and setting $Y = X$ if $X \in S_A$, else drawing $Y \sim P_A$ afresh, gives marginals (P_A, P_B) with $\mathbb{P}(Y \neq X) = 1 - r$, so this coupling is maximal.*

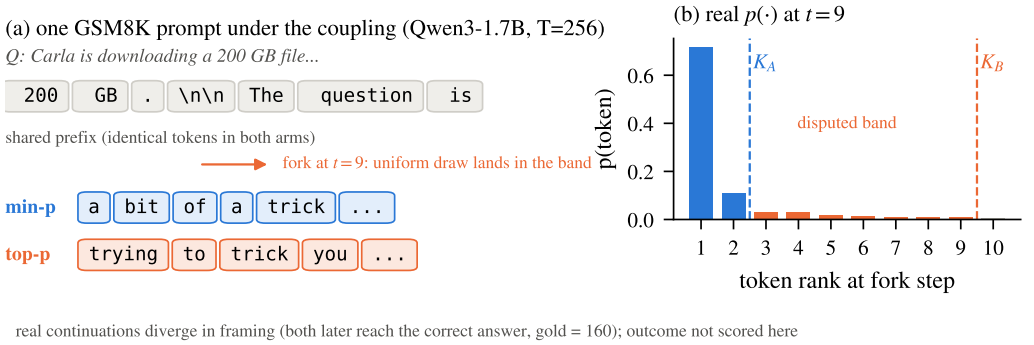


Figure 2: **The coupling on one GSM8K prompt** (Qwen3-1.7B, min-p versus top-p, T=256). **(a)** The rules share tokens until step 9, where the draw lands in the band: top-p keeps it, min-p redraws, and both continuations later reach the correct answer. **(b)** The next-token distribution at the fork, with each rule's cutoff.

The whole distance lives in the disputed band between the cutoffs. Part (ii) is the standard maximal coupling for nested supports (Lindvall, 2002): draw from the wider rule, and if the draw lands in the narrower set both samplers emit it, otherwise the narrower rule redraws and the two continue independently (Figure 2 shows one such fork on a real prompt). Because the narrower rule can be A at some steps and B at others, the shared prefix follows the *pointwise-narrower* rule $A \wedge B$, $K_{A\wedge B} = \min\{K_A, K_B\}$, not A.

Theorem 3 (Separation certificate)**.** *Let f map a generated sequence to $[0,1]$ and set*

$$\mathrm{Sep}_T(A,B) \;=\; 1 - \mathbb{E}_{y\sim A\wedge B}\Big[\textstyle\prod_{t=1}^{T} r_t(y_{<t})\Big].$$

(i) Sep_T *is the probability that a coupled pair diverges within T steps, and $|\mathbb{E}f(Y^A) - \mathbb{E}f(Y^B)| \le \mathrm{Sep}_T$; for f bounded in $[a,b]$ the bound is $(b-a)\mathrm{Sep}_T$.* (ii) $\mathrm{TV}(P_A^{(T)}, P_B^{(T)}) \le \mathrm{Sep}_T$, *with equality iff no trajectory in the common support ever flips which rule is narrower.*

Part (i) is the instrument, one number that no bounded metric can exceed; part (ii) says when it is tight rather than merely valid, which Proposition 5 makes precise. Computing it needs only the retained masses of both rules along one $A \wedge B$ walk per prompt (Appendix E), so it is available from any decoding run's own logs. The same construction extends from a pair to a whole grid.

Corollary 4 (One number for a whole grid)**.** *Let $\mathcal{G}$ be a finite family of prefix rules, for instance every sampler in a study crossed with its hyperparameter grid, and let $\mathcal{G}^\wedge, \mathcal{G}^\vee$ be its pointwise-narrowest and pointwise-widest bounding rules, both prefix rules but not necessarily members of $\mathcal{G}$ itself. All of $\mathcal{G}$ can be coupled simultaneously, so for every $A, B \in \mathcal{G}$ and every f bounded in $[0,1]$, $|\mathbb{E}f(Y^A) - \mathbb{E}f(Y^B)| \le \mathrm{Sep}_T(\mathcal{G}^\wedge, \mathcal{G}^\vee) =: \mathrm{Sep}_T(\mathcal{G})$, computed by the same procedure above with $A = \mathcal{G}^\wedge$, $B = \mathcal{G}^\vee$. When $\mathcal{G}$ is a chain under pointwise inclusion its extremes are themselves members, so no smaller scalar bounds every pair (Appendix B gives the metric that attains it).*

Coupling the family at once is what yields a common scalar, since bounding each pair separately would put every pair under its own measure. One run under $\mathcal{G}^\wedge$ gives $\mathrm{Sep}_T(\mathcal{G})$, but certifying an individual pair still needs a run under that pair's own $A \wedge B$ or under one of its members (Proposition 8), since $\mathcal{G}^\wedge$'s trajectories do not cover the region the pairwise expectation integrates over.

3.2 When can a certificate be exact or small?

The log form connects the certificate to existing theory. Write $g_t = \log Z_{A,t} - \log Z_{B,t}$ for the per-step log-gap between the two rules' retained masses. With $\Delta_T = -\mathbb{E}_{A\wedge B}[\sum_t \log r_t]$,

$$\Delta_T = \mathbb{E}_{A\wedge B}\Big[\textstyle\sum_{t\le T} |g_t|\Big] \;\ge\; \mathbb{E}_{A\wedge B}\Big[\big|\textstyle\sum_{t\le T} g_t\big|\Big], \tag{1}$$

with equality iff no trajectory flips. The right side is the difference of the two rules' *log* local normalization distortions (Kempton & Burrell, 2025) (for a single rule, $\prod_t Z_t^{-1}$), so Δ_T is the

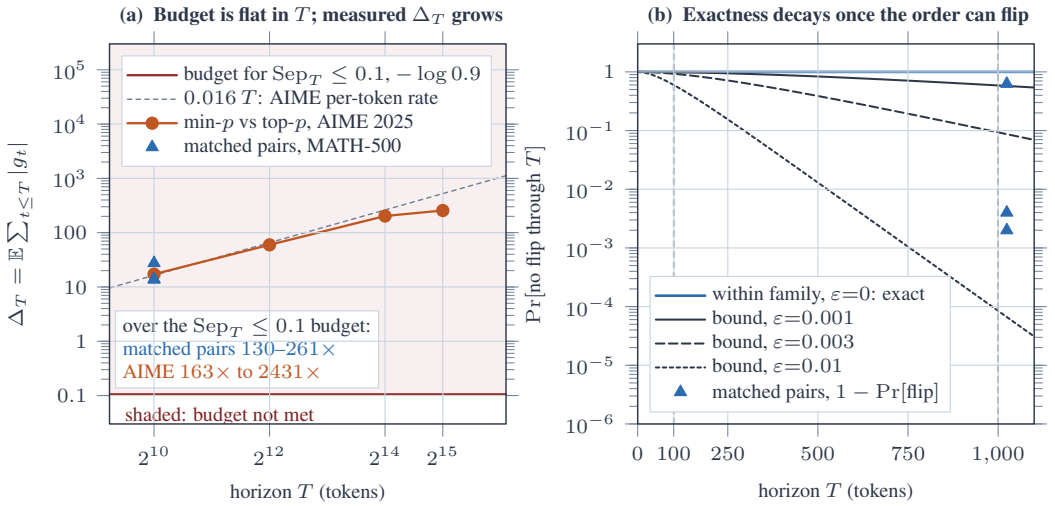


Figure 3: **Measurements against the two structural results.** Markers are measured, solid lines are theory. **(a)** The sufficient budget of Corollary 6(a) for $\mathrm{Sep}_T \le 0.1$ is flat in T; measured Δ_T exceeds it 130–261× for the matched pairs of Table 4 (triangles, T=1024) and 163–2431× for the AIME series (circles; gray dashes extend its per-token rate, $0.016\,T$). **(b)** Proposition 5's bound on the no-flip probability for three ε, against the matched pairs' measured $1 - \Pr[\text{flip}]$ at T=1024. Within a family (ε=0) exactness holds at every horizon; at ε=0.05 the bound is 3.1×10^{-23} by T=1024.

accumulated *absolute* log-gap while theirs is the *net* log-gap, and the two coincide exactly when the certificate is exact. Equation (1) is also a dispersion, whereas the operating point of Section 2 is a mean, and equalizing a mean does not constrain a spread.

Proposition 5 (Exactness is fragile)**.** *Write* $s_t = \mathrm{sign}(g_t) \in \{-1, 0, +1\}$ *along an* $A \wedge B$ *trajectory and call step* t decisive *if* $s_t \neq 0$*. A trajectory* flips *if it has decisive steps of both signs; ties never break constancy, so Theorem 3(ii) is exact precisely when no trajectory in the common support flips. Suppose that from every reachable prefix,* $s_{t+1} = +1$ *and* $s_{t+1} = -1$ *each have conditional probability at least* $\varepsilon > 0$ *(ties may absorb the rest of the mass). Then* $\mathbb{P}(\text{no flip through } T) \le 2(1-\varepsilon)^T - (1-2\varepsilon)^T$*, and this is sharp.*

Exactness therefore survives only when one rule stays weakly narrower throughout on every trajectory in the common support, as in a monotone sweep within one family. Any pair whose order can change loses it geometrically in the horizon (3.1×10^{-23} at $\varepsilon = 0.05$, $T = 1024$). Exactness and fairness pull against each other: matching operating points equalizes $\mathbb{E}[Z]$ across the two rules, exactly the condition that makes neither sign dominant.

Corollary 6 (Per-step budget)**.** *Let* $S = \sum_{t\le T}|g_t|$ *along an* $A \wedge B$ *trajectory, so* $\Delta_T = \mathbb{E}[S]$*.* (a) *If* $\Delta_T \le -\log(1-\gamma)$ *then* $\mathrm{Sep}_T \le \gamma$*: a time-averaged per-step gap of* $\frac{1}{T}\sum_{t\le T}\mathbb{E}[|g_t|] \le -\log(1-\gamma)/T$ *suffices to certify* $\mathrm{Sep}_T \le \gamma$*, with no assumption that the per-step terms are identically distributed.* (b) *For every* $s > 0$*,* $\mathrm{Sep}_T \ge (1 - e^{-s})\mathbb{P}(S \ge s)$*; taking* $s = \Delta_T/2$ *and applying Chebyshev's inequality gives* $\mathrm{Sep}_T \ge \left(1 - e^{-\Delta_T/2}\right)\left(1 - 4\mathrm{Var}(S)/\Delta_T^2\right)$*.*

Part (a) is sufficient but does not explain why a certificate is large, since Δ_T can vastly exceed the budget while Sep_T stays small if S is heavy-tailed. Part (b) closes that gap: when S concentrates around its mean, Sep_T itself must be large. The budget of part (a) is demanding. Ruling out a 0.1 difference needs the two retained masses to agree to a mean relative gap of 1.0×10^{-4} per step over 1,024 tokens, and 3.2×10^{-6} over 32,768. Retained masses sit near 1 and move with context, so two different rules rarely hold that precision for a thousand consecutive steps, and nothing we measured comes within two orders of magnitude of the budget (Figure 3).

3.3 FROM COUPLING TO ESTIMATOR

Proposition 7 (Coupled estimator). *Let f map a generated sequence to $[0,1]$ and $D_i = f(Y_i^A) - f(Y_i^B)$ over n coupled pairs. Then $\hat{\Delta} = \frac{1}{n}\sum_i D_i$ is unbiased for $\mathbb{E}f(Y^A) - \mathbb{E}f(Y^B)$, the same estimand independent runs target, and $\mathrm{Var}(D) \le \mathbb{E}[D^2] \le \mathrm{Sep}_T$. Write $\sigma_A^2 = \mathrm{Var}(f(Y^A))$ and $\sigma_B^2 = \mathrm{Var}(f(Y^B))$ for the outcome variances under independent sampling from each rule. When $\mathrm{Sep}_T > 0$, $\mathrm{Var}(D_{\mathrm{indep}})/\mathrm{Var}(D_{\mathrm{coupled}}) \ge (\sigma_A^2 + \sigma_B^2)/\mathrm{Sep}_T$.*

Every prompt where the pair never diverges contributes an exact zero, so the estimator pays variance only where the samplers disagree. This is the common-random-numbers argument (Glasserman & Yao, 1992) applied to a coupling nesting makes explicit; what is new is that Sep_T, computable from the comparison's own logs or, by Proposition 8, from a single rule's logs, bounds the variance ratio without a separate calibration run. The guarantee holds against independent runs, and prompt-level pairing is a stronger baseline for which we claim no theory. The ratio bound exceeds 1 only when $\sigma_A^2 + \sigma_B^2 > \mathrm{Sep}_T$, so for a binary outcome ($\sigma_A^2 + \sigma_B^2 \le 1/2$) it guarantees no improvement once $\mathrm{Sep}_T > 1/2$.

Assumptions. Both rules must be prefix rules acting on the same p_t under the same sorted order, so temperature and every other logit processor is shared; every experiment here uses temperature 1.0, applied before truncation (unequal temperatures: Appendix B). Outcomes are bounded and white-box decoding control is required. A maximum over a swept comparison is a multiple comparison, bounded by Corollary 4 but still needing correction when reporting intervals.

4 EXPERIMENTS

We check the harness against Theorem 3, then ask how large the certificate is, whether it is exact, whether the coupling reduces variance, and whether it tracks the real gap.

Setup. Open-weight models a single GPU can serve: Qwen3 at 0.6B, 1.7B, 4B and 8B (Yang et al., 2025), Gemma 3 4B (Team et al., 2025) and gpt-oss-20b (Agarwal et al., 2025) for an architecture contrast, and DeepSeek-R1-Distill-Qwen-1.5B (Guo et al., 2025) for long chain of thought. Outcomes are automatically verifiable, so no judge enters the measurement: MATH-500 (Hendrycks et al., 2021), GPQA-Diamond (Rein et al., 2023), GSM8K (Cobbe et al., 2021), and AIME 2025 with the generation cap at $32{,}768$ tokens. Six of the nine samplers marked in Table 1 are each swept over six log-spaced values, an equal-size grid following the hyperparameter-volume point of Schaeffer et al. (2025), and the other three contribute one member each, giving a grid $\mathcal{G}$ of 39 members and 741 pairs. All decoding, including every estimator arm, runs in one custom loop over HuggingFace Transformers (Wolf et al., 2019) at batch size one, since next-token logits drift with batch size. Certificates use the top 2048 logged probabilities and exclude steps whose cutoff would exceed that, an exclusion the grid certificate never needs. All decoding totals 64.5 H200-GPU-hours, under two percent of the six thousand A100-hours of Schaeffer et al. (2025). Appendix E gives the full grid, the exclusion analysis and the compute accounting.

Apart from the gate and the five-seed estimator results, every quantity below comes from a single configuration without a reported interval, so those comparisons are descriptive rather than inferential, and we name each confound where it appears.

4.1 RESULTS

Does the harness measure what the theorem says? Theorem 3(i) predicts the certificate equals the coupled divergence rate, so the harness can be checked before it is trusted. On Qwen3-1.7B/GSM8K (min-p=0.05 versus top-p=0.95, n=200) the certificate is 0.722 against a measured divergence rate of 0.73 (95% CI $[0.67, 0.79]$) at T=256, and 0.945 against 0.935 (CI $[0.90, 0.965]$) at T=1024. The certificate falls inside the bootstrap interval at both horizons, and three controls pass (Appendix F).

How large is the certificate? Across nine model and task configurations (Qwen3 0.6B/4B/8B on GSM8K, Qwen3-1.7B on GSM8K, GPQA-Diamond, MATH-500 and AIME 2025, and Gemma 3 4B and gpt-oss-20b on MATH-500) the grid certificate $\mathrm{Sep}_T(\mathcal{G})$ is at least 0.998 (0.998064 at the

Table 2: **The coupled estimator's variance reduction against prompt-level pairing.** GSM8K rows are Qwen3-1.7B with five seeds, cluster bootstrap intervals and Holm correction across the three pairs; the MATH-500 and GPQA-Diamond rows replicate the default pair.

Task	Pair	Reduction	95% CI, significance
GSM8K	min-p(0.05) vs top-p(0.95)	$3.8\times$	$[3.1, 4.9]$, significant after Holm
GSM8K	top-k(10) vs top-p(0.70)	$1.3\times$	$[1.1, 1.5]$, significant after Holm
GSM8K	ϵ(0.1) vs η(0.0003)	$1.3\times$	$[1.2, 1.6]$, significant after Holm
MATH-500	min-p(0.05) vs top-p(0.95)	$1.7\times$	$[1.2, 2.4]$, bootstrap $p = 0.023$
GPQA-Diamond	min-p(0.05) vs top-p(0.95)	$18.7\times$	$[8.0, 62.1]$, bootstrap $p = 0.16$

minimum, 95% CI $[0.997832, 0.998296]$ over seeds), falling on Qwen3-1.7B/GSM8K, and above 0.9997 on every other configuration (Figure 5, Appendix F). Two qualifications matter. The maximum comes from comparisons against p-less and against wide-top-p and small-ϵ endpoints, so this is a statement about hyperparameter-volume-controlled sweeps rather than about truncation rules in general. And on the seven Qwen3 configurations $K_{\mathcal{G}^\wedge}$ equals one on 99.3% of $9{,}006{,}080$ logged steps, so the certificate is measured along essentially greedy paths rather than typical trajectories of any single member rule. Rules matched on operating point stay far apart too (Table 4), consistent with Corollary 6(a): their retained masses miss the sufficient budget by two orders of magnitude. Within a family the picture differs, as the budget allows: a monotone top-k sweep from 10 to 320 gives certificates as low as 0.10.

Is the certificate exact? Exactness decides whether the number is a property of the samplers or an artifact of the bound, and Theorem 3(ii) makes it checkable from the same logs. Within a family the flip rate is exactly 0 everywhere, so exactness holds in practice, while across families it is 0.15–0.36 on raw grids, and 0.998, 0.996 and 0.374 for the three matched pairs of Table 4. Proposition 5 predicts this ordering, since matching makes neither order dominant. We read the contrast as consistent with that prediction rather than establishing it, because the raw and matched rates come from different pairs and possibly different horizons, and a flip rate is interpretable only at a stated horizon (Figure 3b). A sweep over five matched operating points shows a moderate relationship in the direction Equation (1) predicts (Appendix F).

Longer generations make it worse. Every step of a long reasoning chain is another chance to diverge. On DeepSeek-R1-Distill-Qwen-1.5B/AIME 2025 with the default pair (min-p=0.05, top-p=0.95), read at four horizons as prefixes of one T=32,768 run, Δ_T is 17.2 at T=1024, 59.7 at 4096, 202.8 at 16384 and 256.2 at 32768, and Sep_T, computed directly over the 30 prompts, is 0.9988 at T=1024 and rounds to 1.00 at the three longer horizons. It is already above 0.5 at the shortest cap, the falsification condition specified in advance for the certificate's usefulness as a smallness guarantee. At T=32,768 the per-step budget for a 0.1 certificate is 3.2×10^{-6}, and Δ_T exceeds it by a factor of 2431. Since most generations finish before the longer caps, Appendix F also reports Δ_T per realized token.

Does the coupling reduce variance? The coupling is useful whether the certificate is large or small. On GSM8K, a single run at each of four model scales gave variance reductions of $1.2\times$ (Qwen3-0.6B), $1.8\times$ (4B), $1.9\times$ (8B) and $2.1\times$ (1.7B) relative to prompt-level pairing, and Table 2 reports the five-seed and cross-benchmark results. The GPQA-Diamond estimate is large but not significant, likely because near-guessing accuracy destabilizes a variance ratio. None of the five-seed pairs reaches $\mathrm{Sep}_T < 1/2$ (the lowest is 0.71), so Proposition 7's guarantee does not apply and these reductions are empirical; two simpler ways of sharing randomness fall significantly short of the maximal coupling (Appendix F). A negative control, top-p 0.98 versus top-p 0.99 ($\mathrm{Sep}_T = 0.437$), is the pair where the guarantee applies: with $\sigma_A^2 = 0.240$ and $\sigma_B^2 = 0.241$ it guarantees only $1.10\times$, while the measured reduction is about $12.0\times$, so the guarantee is far from tight there.

Does the certificate predict the gap? A bound can be correct and still useless if it never tracks the real gap. Table 5 (Appendix F) lists eight pairs on Qwen3-1.7B/GSM8K at T=256, three cross-family and five within-family including the negative control, with Sep_T from 0.17 to 0.99. Every measured

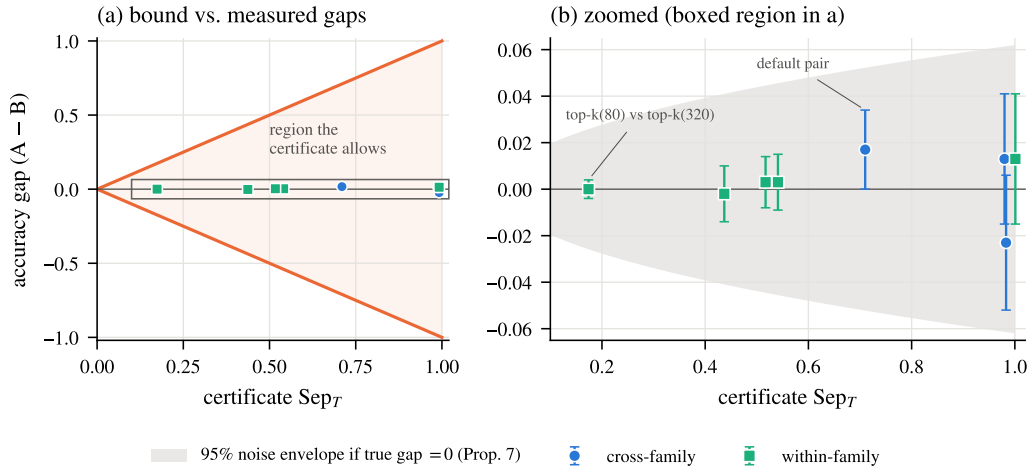


Figure 4: **Certificates are large, accuracy gaps are small** (the eight pairs of Table 5). **(a)** Theorem 3 bounds the expected gap by the certificate (cone). **(b)** The boxed region with 95% CIs. The gray band, $\pm 1.96\sqrt{\mathrm{Sep}_T/n}$ with n=1,000, is the zero-gap spread implied by Proposition 7; it is conservative in using a variance bound and anti-conservative in treating 200 prompts $\times$ five seeds as independent. Points near Sep_T=0.99 are offset by ± 0.009.

gap sits far inside its certificate (Figure 4; a tightness ratio, gap over Sep_T, of at most 0.024). On GSM8K no generation stops before the cap, so part of the measured divergence is text written after the answer, which the grader sees only through a conflicting second answer (Appendix B). Sep_T correlates with the absolute gap at Pearson $r = 0.85$ (95% CI $[0.35, 0.97]$), but this is not on its own evidence of prediction: Proposition 7 gives $\mathrm{Var}(D) \leq \mathrm{Sep}_T$, so pure sampling noise with every true gap equal to zero already induces a positive correlation. Against a simulated null that draws each pair's gap from a zero-mean Gaussian at its own measured standard error, the null mean of r is 0.54 and the observed $r = 0.85$ sits at its 94.5th percentile (one-sided $p = 0.055$); standardizing each gap by its own standard error still gives $r = 0.68$ against a null mean of 0 (two-sided $p = 0.056$). Both sit at the conventional significance boundary, so at eight pairs we read the result as inconclusive, neither a demonstration nor a null artifact. Whether a small certificate ever appears between two *cross-family* rules remains open.

5 What This Means for Benchmark Comparisons

The estimator's variances have a consequence that follows from the coupled arm alone. At one sample per prompt, detecting a one-point difference between two samplers needs roughly 787–16,379 prompts under coupling and 14,686–23,727 under prompt-level pairing, across the pairs of Table 2. Both lower ends come from GPQA-Diamond, whose 18.7$\times$ reduction is not significant, so the representative reduction is the 1.3–3.8$\times$ of the other four pairs. At one sample per prompt every benchmark, AIME 2025 included, falls short of at least one measured requirement.

Drawing k samples per prompt changes this. For the default pair on GSM8K the between-prompt variance τ^2 is statistically indistinguishable from zero (bootstrap 95% upper bound 9.3×10^{-4} over prompts), so the required prompt count falls close to proportionally with k (Figure 6, Appendix F). At the point estimate, GSM8K reaches the requirement by k=8, MATH-500 by k=32, and GPQA-Diamond almost exactly at k=32, though not at τ^2's upper bound. AIME 2025 is still 6.6$\times$ short at k=32 and would need $k \approx 210$, while at the upper bound no k suffices, and this extrapolation assumes the τ^2 measured on GSM8K rather than on AIME 2025. At the default pair's variance, coupling plus resampling removes the constraint for the larger benchmarks; whether it does for the smallest one is not settled by what we measured (Appendix F).

What to report. Three quantities cost one decoding run each: every sampler's operating point with its flip rate, Sep_T for each compared pair beside the measured gap, flagging any gap above it, and the per-step retained masses, which make every certificate recomputable.

6 Related Work

Local normalization distortion. The closest prior work is Kempton & Burrell (2025), who isolate a *local normalization distortion*, the reciprocal of the product of retained masses along a generated sequence, show it pulls decoding away from the quality-diversity frontier, and find that its variance across sequences rather than its mean is what matters. Their matching criterion is the operating point we adopt, and their comparison is between a rule and its own globally normalized counterpart. Ours is between two deployed rules, and Equation (1) shows the two coincide exactly when no trajectory flips.

Truncation rules and matched comparisons. Holtzman et al. (2019) established truncation as the remedy for degeneration and Hewitt et al. (2022) gave it a generative reading; most later rules differ only in which statistic sets the cutoff. Among protocols that match samplers on a scalar summary, Zhou et al. (2025) is closest in spirit, matching at equal estimated risk levels, but without coupling, so obtaining no bound on an end-task difference. Existing theory explains why truncation helps rather than how much two truncations differ (Hewitt et al., 2022; Finlayson et al., 2024; Chen et al., 2025).

Coupling. Coupling through shared randomness to reduce variance is standard (Glasserman & Yao, 1992; Yang & Nelson, 1991; Lindvall, 2002; Bavarian et al., 2016); what is new is not the construction but that truncation rules make the maximal coupling closed-form and rejection-free at every step. Pairing on shared prompts is established practice (McNemar, 1947; Dror et al., 2018) and documented to be underpowered (Card et al., 2020). Estimating sequence-level total variation between autoregressive distributions is generally hard (Price et al., 2026); ours is a structured case where it is computed rather than estimated. Appendix C covers speculative decoding, which on a nested pair reduces to exactly this coupling, and further related lines.

7 Limitations and Conclusion

Limitations. The theory covers prefix rules on a shared base distribution, excluding locally typical sampling, Top-W, contrastive decoding, reordering rules and cross-model comparison; a certificate rules differences out, never in, and both certificate and estimator degrade as the divergence probability approaches one. Sequence-level total variation is estimable from a run under either rule alone (Proposition 8) but has not yet been computed on a real log, so Theorem 3(ii)'s equality is checked directly only on synthetic models, with measured flip rates as indirect evidence. The grid's 741 pairwise certificates are not yet decoded. The predictive-validity check uses eight GSM8K pairs on one model and horizon, five within-family, and its null-calibrated test lands at the significance boundary. Every task is short-answer mathematics or science, whereas truncation samplers were designed for open-ended generation, so our certificates may be large partly because reasoning chains are not the regime these rules target, and on GSM8K, where no generation stops before the cap, part of each certificate reflects text written after the answer.

Conclusion. Nesting makes the distance between two truncation samplers free to compute. The certificate is large while accuracy gaps are small, so a comparison has to be measured rather than bounded. The useful question is not only which truncation rule is best but whether a given experiment, on a given benchmark, was ever sized to tell.

Reproducibility Statement

Every result in Section 3 is proved in Appendix A, with its assumptions stated alongside it, and the supporting results in Appendix B are stated with their sources and are checked numerically where applicable. The supplementary material contains the reference implementation of every quantity defined in Section 3 and the two scripts behind the checks of Appendix G, both of which run on

CPU with NumPy alone. Section 4 names models, datasets and statistical procedures. All datasets are public and all models are open-weight. Code and per-step retained-mass logs will accompany the camera-ready under an anonymized URL, and both are runnable from a clean checkout with the commands in `REPRODUCE.md`. Model checkpoints are pinned to a specific revision, not `main`, for every model reported, Qwen3-0.6B/1.7B/4B/8B, Gemma 3 4B, gpt-oss-20b and DeepSeek-R1-Distill-Qwen-1.5B. GSM8K (MIT), GPQA-Diamond (CC-BY-4.0), AIME-2025 (Apache-2.0) and MATH-500 (MIT, via the underlying Hendrycks et al. MATH dataset) are each pinned to a specific dataset revision. Every generation uses five seeds (0–4) where seeds are reported at all, and one fixed prompt template per task with no chat formatting, few-shot examples or system prompt, so the certificate is measured against the model's base per-step distribution. Where a falsification condition is stated for a result, such as the threshold in Section 4.1 for the certificate's usefulness as a smallness guarantee, it was specified in advance of running the corresponding experiment, not registered with an external, timestamped service, and we say "specified in advance" rather than "pre-registered" for that reason.

Ethics Statement

This work involves no human subjects, no personal data and no model training. It studies evaluation methodology for existing open-weight models on public benchmarks. One consideration deserves note. An audit tool applied to published work can be misused to overstate criticism, and a certificate bounds what two samplers can differ by but does not establish that any method lacks value.

Use of Generative AI

Generative AI tools were used substantially throughout this work: in framing the theory and formulating and proving its claims, in implementing the code and running the experiments, in interpreting results, in reviewing related literature, and in drafting and editing the text. Every formal claim is checked by the synthetic verification suite of Appendix G, which caught several errors during preparation, and a human author independently re-derived Lemma 2, Theorem 3 and Proposition 7 and checked the coupled decoder against Lemma 2(ii). Every number about a real language model comes from a logged decoding run. The human authors take responsibility for the final content of this work.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Avinash Amballa, Aditya Parashar, Aditya Vikram Singh, Jinlin Lai, and Benjamin Rozonoyer. Quasi-random multi-sample inference for large language models. *Frontiers in Probabilistic Inference: Learning meets Sampling*, 2025.

Esteban Garces Arias, Nurzhan Sapargali, Christian Heumann, and Matthias Aßenmacher. The truncation blind spot: How decoding strategies systematically exclude human-like token choices. *arXiv preprint arXiv:2603.18482*, 2026.

Erfan Baghaei Potraghloo, Seyedarmin Azizi, Souvik Kundu, and Massoud Pedram. Top-h decoding: Adapting the creativity and coherence with bounded entropy in text generation. *Advances in Neural Information Processing Systems*, 38:28482–28513, 2026.

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R Varshney. Mirostat: A neural text decoding algorithm that directly controls perplexity. *arXiv preprint arXiv:2007.14966*, 2020.

Mohammad Bavarian, Badih Ghazi, Elad Haramaty, Pritish Kamath, Ronald L Rivest, and Madhu Sudan. Optimality of correlated sampling strategies. *arXiv preprint arXiv:1612.01041*, 2016.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9263–9274, 2020.

Sijin Chen, Omar Hagrass, and Jason Klusowski. Decoding game: On minimax optimality of heuristic text generation strategies. In *International Conference on Learning Representations*, volume 2025, pp. 83323–83342, 2025.

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

Arash Gholami Davoodi, Navid Rezazadeh, Seyed Pouyan Mousavi Davoudi, and Pouya Pezeshkpour. Geometry-aware decoding with wasserstein-regularized truncation and mass penalties for large language models. *arXiv preprint arXiv:2602.10346*, 2026.

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 (Volume 1: Long Papers)*, pp. 14932–14948, 2026.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1383–1392, 2018.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, 2018.

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*, volume 2024, pp. 12295–12313, 2024.

Daniel Gareev, Thomas Hofmann, Ezhilmathi Krishnasamy, and Tiago Pimentel. Local and global decoding in text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14577–14597, 2024.

Paul Glasserman and David D Yao. Some guidelines and guarantees for common random numbers. *Management Science*, 38(6):884–908, 1992.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 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. *arXiv preprint arXiv:2103.03874*, 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.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. *arXiv preprint arXiv:2504.07086*, 2025.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2019.

Daniel Israel, Guy Van den Broeck, and Aditya Grover. Accelerating diffusion llms via adaptive parallel decoding. *Advances in neural information processing systems*, 38:52870–52888, 2026.

Thomas Kempton and Stuart Burrell. Local normalization distortion and the thermodynamic formalism of decoding strategies for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, September 2025.

Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International conference on machine learning*, pp. 3499–3508. PMLR, 2019.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International conference on machine learning*, pp. 19274–19286. PMLR, 2023.

Michael Y Li, Anthony Zhan, Kanishk Gandhi, Noah D Goodman, and Emily B Fox. Quasimotto: Quasi-monte carlo test-time scaling. *arXiv preprint arXiv:2607.01179*, 2026.

Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002.

Lovish Madaan, Aaditya K Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

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

Solomon Messing. Hidden measurement error in llm pipelines distorts annotation, evaluation, and benchmarking. *arXiv preprint arXiv:2604.11581*, 2026.

Minh 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*, volume 2025, pp. 70333–70366, 2025.

Eric Price, Kevin Tian, Zhiyang Xun, and Yusong Zhu. Total variation distance estimation in autoregressive models. *arXiv preprint arXiv:2607.19510*, 2026.

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

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Joseph Rowan, Truong Buu Phan, and Ashish Khisti. List-level distribution coupling with applications to speculative decoding and lossy compression. *Advances in Neural Information Processing Systems*, 38:5576–5619, 2026.

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*, 2025.

Chloe H Su, Zhe Ye, Samuel Tenka, Aidan Yang, Soonho Kong, and Udaya Ghai. Learning adaptive llm decoding. *arXiv preprint arXiv:2603.09065*, 2026.

Runyan Tan, Shuang Wu, and Phillip Howard. p-less sampling: A robust hyperparameter-free approach for llm decoding. In *International Conference on Learning Representations*, volume 2026, pp. 119506–119549, 2026.

Chenxia Tang, Jianchun Liu, Hongli Xu, and Liusheng Huang. Top-$n\sigma$: Not all logits are you need. *arXiv preprint arXiv:2411.07641*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Luke Vilnis, Yury Zemlyanskiy, Patrick Murray, Alexandre Tachard Passos, and Sumit Sanghai. Arithmetic sampling: parallel diverse decoding for large language models. In *International conference on machine learning*, pp. 35120–35136. PMLR, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Wei-Ning Yang and Barry L Nelson. Using common random numbers and control variates in multiple-comparison procedures. *Operations Research*, 39(4):583–591, 1991.

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 (Volume 1: Long Papers)*, pp. 26352–26365, 2025.

A PROOFS

Throughout, at the step in question $K_A \le K_B$ without loss of generality, so $S_A \subseteq S_B$, $Z_A = \sum_{i \le K_A} p_{(i)}$, $Z_B = \sum_{i \le K_B} p_{(i)}$ and $r = Z_A/Z_B$. Both rules act on the same p under the same sorted order. Since $K \ge 1$ and $p_{(1)} > 0$, every retained mass is strictly positive.

Lemma 2, nesting. S_A and S_B are prefixes of the same total order, so the one with the smaller cutoff index is contained in the other. □

Lemma 2(i). For $i \in S_A$, $P_B(i \mid S_A) = P_B(i)/P_B(S_A) = (p_{(i)}/Z_B)/(Z_A/Z_B) = p_{(i)}/Z_A = P_A(i)$, and both sides vanish off S_A. For the distance,

$$2\,\mathrm{TV}(P_A, P_B) = \sum_{i \in S_A} \left| \tfrac{p_{(i)}}{Z_A} - \tfrac{p_{(i)}}{Z_B} \right| + \sum_{i \in S_B \setminus S_A} \tfrac{p_{(i)}}{Z_B} = Z_A\Big(\tfrac{1}{Z_A} - \tfrac{1}{Z_B}\Big) + \tfrac{Z_B - Z_A}{Z_B} = 2\Big(1 - \tfrac{Z_A}{Z_B}\Big),$$

using $Z_A \le Z_B$ to drop the absolute values. □

Lemma 2(ii). The B-marginal holds by construction. For $i \in S_A$, $\mathbb{P}(Y = i) = \mathbb{P}(X = i) + \mathbb{P}(X \notin S_A)P_A(i) = \tfrac{p_{(i)}}{Z_B} + (1 - \tfrac{Z_A}{Z_B})\tfrac{p_{(i)}}{Z_A} = \tfrac{p_{(i)}}{Z_A} = P_A(i)$, and $\mathbb{P}(Y = i) = 0$ off S_A. When $X \in S_A$ we set $Y = X$, and when $X \notin S_A$ we have $Y \in S_A$ and $X \notin S_A$, so $Y \ne X$ necessarily. Hence $\mathbb{P}(Y \ne X) = \mathbb{P}(X \notin S_A) = 1 - r = \mathrm{TV}(P_A, P_B)$ exactly, and since TV is the minimum of $\mathbb{P}(Y \ne X)$ over all couplings, this one is maximal. □

Theorem 3(i). Run the coupling of Lemma 2(ii) at every step. Given agreement through $t-1$, Lemma 2(ii) gives agreement at step t with probability $r_t(y_{<t})$, and by Lemma 2(i) the emitted token is then distributed under the narrower rule at that step, which is $P_{A\wedge B}$. Since the narrower rule may be A at some steps and B at others, the shared prefix is distributed as an $A \wedge B$ trajectory and not in general as an A trajectory. Chaining, $\mathbb{P}(\text{no divergence in } T) = \mathbb{E}_{A\wedge B}[\prod_{t\le T} r_t]$, so the divergence probability equals Sep_T. Each arm follows its own rule marginally at every step, by Lemma 2(ii) before divergence and by construction after, so Y^A and Y^B have the laws of full A- and B-trajectories. On the no-divergence event $Y^A = Y^B$, hence $|\mathbb{E}f(Y^A) - \mathbb{E}f(Y^B)| = |\mathbb{E}[(f(Y^A) - f(Y^B))\mathbf{1}[\text{div}]]| \le (b-a)\mathrm{Sep}_T$ for $f \in [a,b]$. □

Theorem 3(ii). First, y lies in $\mathrm{supp}(A) \cap \mathrm{supp}(B)$ iff every y_t lies in $S_{A,t} \cap S_{B,t} = S_{A\wedge B,t}$, so the common support is exactly $\mathrm{supp}(A \wedge B)$. The coupled sequences differ exactly on the divergence event, so $\mathbb{P}(Y^A \neq Y^B) = \mathrm{Sep}_T$, and as TV is the minimum of $\mathbb{P}(Y^A \neq Y^B)$ over couplings, $\mathrm{TV} \le \mathrm{Sep}_T$.

For equality, write $a_t = P_A(y_t \mid y_{<t})$ and $b_t = P_B(y_t \mid y_{<t})$, so $A(y) = \prod_t a_t$, $B(y) = \prod_t b_t$, and $\min\{a_t, b_t\} = p_{y_t}/Z_{\max,t}$ on the common support. Then $1 - \mathrm{Sep}_T = \sum_{y\in\mathrm{supp}(A\wedge B)} \prod_t \min\{a_t, b_t\}$ and $1-\mathrm{TV} = \sum_y \min\{A(y), B(y)\}$, the latter having zero terms off the common support. Termwise $\min\{\prod a_t, \prod b_t\} \ge \prod \min\{a_t, b_t\}$, which re-proves $\mathrm{TV} \le \mathrm{Sep}_T$. Since every term of positive weight must agree, fix y and let $S = \{t : a_t < b_t\}$, $L = \{t : a_t > b_t\}$. If $L = \emptyset$ then $\prod \min = \prod a \le \prod b$ and the two agree, and symmetrically if $S = \emptyset$. If both are nonempty then $\prod \min$ is strictly smaller than both products, because it replaces a_s by $b_s < a_s$ for $s \in L$ and b_t by $a_t < b_t$ for $t \in S$, all factors being strictly positive. Since $a_t < b_t$ iff $Z_A > Z_B$, both sets nonempty is exactly a flip. □

Corollary 4. Pointwise minima and maxima of finitely many prefix rules are prefix rules. At a step, $S_{\mathcal{G}^\vee}$ equals the retained set of whichever member attains the maximum, so drawing $X \sim P_{\mathcal{G}^\vee}$ is realizable, every $R \in \mathcal{G}$ with $X \in S_R$ emits X, and every other draws from its own P_R. Each marginal is correct by the computation in Lemma 2(ii) with $\mathcal{G}^\vee$ in place of B, for each R separately and regardless of how the remaining draws are correlated. All rules agree iff $X \in S_{\mathcal{G}^\wedge}$. If so all emit X, and if not, the member attaining the pointwise minimum redraws inside $S_{\mathcal{G}^\wedge}$ while some member emits $X \notin S_{\mathcal{G}^\wedge}$. Conditional on agreement the common token is $P_{\mathcal{G}^\vee}(\cdot \mid S_{\mathcal{G}^\wedge}) = P_{\mathcal{G}^\wedge}$ by Lemma 2(i), and the argument of Theorem 3(i) applies. □

Proposition 5. Track two states along a trajectory that has not yet flipped: U, no decisive step yet, and D, every decisive step so far shares one sign. From D with sign c, the next step flips the trajectory (ends it) with conditional probability at least ε (the opposite sign), so D survives n further steps with probability at most $(1-\varepsilon)^n$. From U, the next step is decisive with some probability $q \ge 2\varepsilon$ and, being the first decisive step, cannot itself flip anything; let v_n be the largest probability of not having flipped after n steps starting from U. Conditioning on the first step, $v_n = \max_{q\ge 2\varepsilon} \left[q\,(1-\varepsilon)^{n-1} + (1-q)\,v_{n-1}\right]$, since a decisive first step moves to D with $n-1$ steps left to flip (probability of surviving those $\le (1-\varepsilon)^{n-1}$, attained by splitting q evenly between the two signs so the survival bound is tight regardless of which sign was drawn) and a tied first step stays in U with $n-1$ steps left. By induction $v_{n-1} \ge (1-\varepsilon)^{n-1} \ge 0$, so the bracket is non-increasing in q (its q-derivative is $(1-\varepsilon)^{n-1} - v_{n-1} \le 0$) and the maximum over $q \ge 2\varepsilon$ sits at the smallest feasible value, $q = 2\varepsilon$, giving $v_n = 2\varepsilon(1-\varepsilon)^{n-1} + (1-2\varepsilon)v_{n-1}$ with $v_0 = 1$. Unrolling this recursion gives the closed form $v_T = 2(1-\varepsilon)^T - (1-2\varepsilon)^T$, which one checks satisfies the same recursion and boundary condition by direct substitution. The process that puts mass exactly ε on each sign while in U, and exactly ε on the opposite sign once in D, attains every inequality above, so the bound is sharp. □

Corollary 6. Here $\mathrm{Sep}_T = 1 - \mathbb{E}[e^{-S}]$. *(a)* $x \mapsto e^{-x}$ is convex, so Jensen gives $\mathbb{E}[e^{-S}] \ge e^{-\Delta_T}$ and hence $\mathrm{Sep}_T \le 1 - e^{-\Delta_T}$. Requiring the right side to be at most γ gives $\Delta_T \le -\log(1-\gamma)$, and since $\Delta_T = \mathbb{E}[S] = \sum_{t\le T} \mathbb{E}[|g_t|] = T \cdot \frac{1}{T}\sum_{t\le T} \mathbb{E}[|g_t|]$ with no assumption that the per-step terms are identically distributed, this is exactly a bound on their time average, giving the per-step form. This bound is close to tight for small γ, and at $\gamma = 0.1$ the two sides differ by 5%. *(b)* Pointwise, $1 - e^{-S} \ge (1-e^{-s})\mathbf{1}[S \ge s]$ for every $s \ge 0$, since $1 - e^{-S} \ge 0$ and is increasing in S; taking expectations gives $\mathrm{Sep}_T \ge (1-e^{-s})\mathbb{P}(S \ge s)$. At $s = \Delta_T/2$, Chebyshev's inequality gives

$\mathbb{P}(S < \Delta_T/2) \le \mathbb{P}(|S - \Delta_T| \ge \Delta_T/2) \le 4\mathrm{Var}(S)/\Delta_T^2$, so $\mathbb{P}(S \ge \Delta_T/2) \ge 1 - 4\mathrm{Var}(S)/\Delta_T^2$, which substituted above gives the stated bound. □

Proposition 7. Unbiasedness holds because Lemma 2(ii) preserves both marginals. Since $D = 0$ off the divergence event and $f \in [0, 1]$, $|D| \le \mathbf{1}[\mathrm{div}]$, so $\mathrm{Var}(D) \le \mathbb{E}[D^2] \le \mathrm{Sep}_T$. For independent runs $\mathrm{Var}(f(Y^A) - f(Y^B)) = \sigma_A^2 + \sigma_B^2$, giving the ratio whenever $\mathrm{Sep}_T > 0$. □

Proposition 8. On $\mathrm{supp}(A) \cap \mathrm{supp}(B)$, the $A \wedge B$ walk agrees with the A walk exactly when every step lands in $S_{A\wedge B} = S_A \cap S_B$, which for $y \in \mathrm{supp}(B)$ happens for every step of the A walk since $S_A \subseteq S_B$ or $S_B \subseteq S_A$ at each step (Lemma 2). So, writing $r_t = \min(Z_{A,t}, Z_{B,t}) / \max(Z_{A,t}, Z_{B,t})$ as in Theorem 3,

$$(A \wedge B)(y) \prod_t r_t = \prod_t \frac{p(y_t)}{\max(Z_{A,t}, Z_{B,t})} = A(y) \prod_t \frac{Z_{A,t}}{\max(Z_{A,t}, Z_{B,t})}$$

for $y \in \mathrm{supp}(A) \cap \mathrm{supp}(B)$, and both sides are 0 for $y \notin \mathrm{supp}(B)$ since then some step has $y_t \notin S_B$ so the $A \wedge B$ walk could not have reached y. Summing over y and using Theorem 3(i)'s $1 - \mathrm{Sep}_T = \sum_y (A \wedge B)(y) \prod_t r_t$ gives the first identity. For the second, $B(y)/A(y) = \prod_t p(y_t)/Z_{B,t} \cdot Z_{A,t}/p(y_t) = \prod_t Z_{A,t}/Z_{B,t}$ on $\mathrm{supp}(A) \cap \mathrm{supp}(B)$, wherever this ratio is defined, and $B(y) = 0$ for $y \in \mathrm{supp}(A) \setminus \mathrm{supp}(B)$; so $\mathrm{TV} = \sum_y (A(y) - B(y))_+ = \mathbb{E}_{y\sim A}[(1 - B(y)/A(y))_+]$, which is the stated form once $B(y)/A(y)$ is written out. Both integrands are products of terms in $(0, 1]$, truncated by an indicator or a positive part, so they lie in $[0, 1]$, and A is the sampling measure in both expectations, so a single run under A estimates either quantity with the usual Hoeffding guarantee for a bounded statistic. □

B Supporting Results

Four facts Section 3 refers to are recorded here briefly rather than as full statements, since none is novel to this paper. When $K_A(p) \le K_B(p)$ at every reachable context, the supremum in Theorem 3(ii) is attained at $f^\star(y) = \mathbf{1}[y \in \mathrm{supp}(A)]$, which we verify by exact enumeration in Appendix G. Unequal temperatures are covered by the standard construction for non-nested but still totally ordered supports. If a shared logit transformation is applied before truncation but the two samplers use different temperatures, retained sets stay nested since temperature preserves the sorted order, but Lemma 2(i)'s equality fails because $P_A(i)/P_B(i)$ now depends on i, so the explicit no-rejection coupling no longer applies directly. What survives is that the per-step distance between the two distributions is still computable in closed form at no extra query cost, and Theorem 3 holds with $1 - r_t$ replaced by that distance, with the expectation taken along the induced overlap measure rather than along $A \wedge B$. When both rules are driven by one shared uniform random variable through the inverse CDF of their truncated distributions, as arithmetic and quasi-Monte-Carlo samplers do (Vilnis et al., 2023; Li et al., 2026), the resulting coupling is valid but not maximal, so its divergence probability is at least $1 - r$, though the certificate of Theorem 3 still bounds the outcome difference since that bound is a property of the marginals and holds under any coupling. Finally, Theorem 3 as stated fixes T; with EOS, let τ be the step at which either rule first emits it (inclusive), a stopping time with respect to the shared $A \wedge B$ trajectory since whether step t's emitted token is EOS is observable at step t. Replacing T by $\tau \wedge T$ throughout, so the product stops at $\tau \wedge T$ rather than running to a fixed T regardless of EOS, gives $\mathrm{Sep}_{\tau\wedge T}$ and the same two parts of Theorem 3 with the same proof, since the induction it relies on never used T being non-random, only that it upper-bounds the number of steps taken. Our own certificate logs already use this stopping time, since `decode.py` logs the EOS-emitting step and then stops (Section 4 reports the resulting `eos_rate` for the long-horizon AIME series), so every Sep_T reported in this paper is already $\mathrm{Sep}_{\tau\wedge T}$ in this sense. The variance-reduction estimator's coupled, prompt-paired and independent arms do not currently share this stopping rule: they run the full T steps regardless of EOS, and grading does not simply take the first or the last answer the completion states. Math tasks are graded with `math_verify`, which parses every `\boxed{}` expression in the completion and scores it correct only if exactly one distinct value appears and it matches the gold answer, so a completion that states an answer, reconsiders, and states a different one anywhere later scores incorrect regardless of which value is right; GPQA is graded by the first standalone option letter found, which does take the first answer. So outcome noise from text generated after a would-be stopping point is real, but its effect is task-dependent rather

than uniformly "extra sampling after the answer is ignored." This is a real difference between the certificate and the estimator, not merely a presentational gap, and is not resolved here.

Proposition 8 (Pairwise certificate and total variation from one rule's run)**.** *For prefix rules A and B and $y \sim P_A^{(T)}$,*

$$1 - \mathrm{Sep}_T(A, B) = \mathbb{E}_{y\sim A}\Big[\mathbf{1}[y \in \mathrm{supp}(B)] \prod_{t\le T} \frac{Z_{A,t}}{\max(Z_{A,t}, Z_{B,t})}\Big],$$

$$\mathrm{TV}\big(P_A^{(T)}, P_B^{(T)}\big) = \mathbb{E}_{y\sim A}\Big[\Big(1 - \mathbf{1}[y \in \mathrm{supp}(B)] \prod_{t\le T} \frac{Z_{A,t}}{Z_{B,t}}\Big)_+\Big].$$

Both integrands lie in $[0, 1]$, so a Monte Carlo estimate from a single rule's decoding run admits a Hoeffding interval.

A run under A alone therefore gives both the pairwise certificate and the exact sequence-level total variation for every B compared against it, not only the pointwise-narrower member of a pair: a grid of n swept rules needs n runs for all $\binom{n}{2}$ pairwise quantities in either direction, rather than one run per pair under that pair's own $A \wedge B$. This does not change Sep_T or TV themselves, only which decoding run recovers them; the trajectory-level coupling of Theorem 3(i) still needs the $A \wedge B$ walk, since it is the joint law of (Y^A, Y^B), not just their marginals' distance, that the one-rule identity does not reconstruct. Applying it needs no new *sampling*, but it is not free of engineering cost: the estimator's independent-arm runs in this paper (Section 4) save only the final token sequence, not the per-step retained masses $Z_{A,t}$ the identity is stated in terms of, so recovering it for those runs needs a further pass, either logging retained masses on a fresh independent-arm decode or a teacher-forced forward pass over the already-sampled sequences, not a re-read of data already on disk.

C Extended Related Work

Evaluation of truncation. Kempton & Burrell (2025) characterize top-k, nucleus and temperature sampling as equilibrium states and show the local normalization distortion is large. Nesting is what makes their object usable for comparing two rules, because the intractable global constant that forces Kempton & Burrell (2025) to work with paired ratios cancels structurally and the maximal coupling becomes explicit. Gareev et al. (2024) compare local and global decoding empirically, again within a rule rather than across rules. Zhou et al. (2025) match at equal estimated risk levels via a prefix tree, comparing distributional quantities on open-ended text. Arias et al. (2026) ask what truncation removes relative to human text and Ravfogel et al. (2023) calibrate top-p per entropy bin, both distinct goals.

Speculative decoding. On a nested pair, speculative decoding's accept-or-resample-from-the-residual rule (Leviathan et al., 2023; Rowan et al., 2026) reduces to exactly the coupling of Lemma 2(ii): with $P_A = P_B(\cdot \mid S_A)$, the acceptance probability $\min(1, P_A(x)/P_B(x))$ is 1 on S_A and 0 outside it, so acceptance is the deterministic set membership Lemma 2(ii) uses rather than a stochastic ratio, and the residual $\propto \max(0, P_A - P_B)$ is P_A itself. Speculative decoding's target use, drafting from a cheap model to skip evaluations of an expensive one, is unrelated to ours, comparing two truncation rules of the same model for measurement, but the mechanism is the same textbook construction specialized to nested supports rather than a new one; the contribution here is showing that truncation rules make this coupling closed-form and rejection-free at every step, not designing it.

Other couplings. Adaptive parallel decoding accepts to first disagreement (Israel et al., 2026), and watermarking couples generation to a key (Kuditipudi et al., 2023). A parallel line fixes the truncation rule and redesigns the joint law of k parallel draws (Vilnis et al., 2023; Kool et al., 2019; Amballa et al., 2025), bounded by Li et al. (2026) via a union-bound argument, and the two axes compose.

Context-varying rules. Context-varying rules (Dia, 2026; Su et al., 2026) select a prefix rule per context from a policy rather than a fixed hyperparameter; treating that policy's range as the family $\mathcal{G}$ in Corollary 4 gives a single scalar bounding every comparison the policy can induce, from one decoding run under its pointwise-narrowest member.

D PREFIX-CLASS AUDIT

The class assumption behind Table 1 is confirmed on $997{,}465{,}872$ real logged cutoffs, pooled across every decoding log in our implementation's output (grid walks, pairwise walks and the ρ sweep, not just the swept-family grid walks alone): zero out-of-range cutoffs, zero non-nested pairs, among the rules implemented here. Mirostat is not yet implemented, so its prefix-ness is inferred from its published description rather than checked directly.

Top-W and locally typical sampling are the two rules outside our guarantees. Top-W's cutoff follows a score $\varphi_i = f_i + \lambda \log p_i$ mixing token-embedding distance f_i with log-probability (Davoodi et al., 2026), a different total order from the probability-sorted one Definition 1 requires, so its retained set is a prefix of that order but not of ours. Its optimality argument yields a crop that is a single token or a prefix of an order mixing probability with token-embedding geometry, so Top-W, like locally typical sampling, lies outside the prefix class.

Locally typical sampling can omit the top token, since Meister et al. (2023)'s rule sorts tokens by $|-\log p_{(i)} - H|$ ascending and keeps the shortest prefix of that order reaching cumulative mass τ, which need not include the token nearest probability 1. At $p = (0.4, 0.3, 0.3)$, $H \approx 1.09$ nats, the two 0.3 tokens sit closer to H than the 0.4 token does, so at $\tau \leq 0.6$ the retained set is the two 0.3 tokens and the top token is dropped. Comparing $p_{(1)}$ against e^{-H} directly is not a measurement of this: since $-\log p_{(1)}$ is the minimum surprisal over the support, $H = \mathbb{E}[-\log p_{(i)}] \geq -\log p_{(1)}$ on every distribution, with equality only when every token with positive mass is equally likely, so that comparison holds by construction and says nothing about whether the mass-based rule retains the top token. Implementing that rule and counting how often it omits the top token, pooled over the seed-0 grid-walk log of each of the nine model/task configurations of Section 4 ($2{,}798{,}071$ steps total, per-configuration rates from 2.5% to 24.3%), gives a pooled omission rate of 8.8% at $\tau = 0.2$, a commonly used operating point for this rule.

E EXPERIMENTAL DETAILS

Calibration corpus. The operating point $\rho(K) = \mathbb{E}[Z_K]$ is measured over a teacher-forced calibration corpus of 200 held-out GSM8K prompts, each extended with its known correct solution (the gold answer) as the continuation, so the model's next-token distribution is logged against real, on-distribution text rather than its own generations.

Computing a certificate. Computing $\mathrm{Sep}_T(A, B)$ from a decoding run follows directly from the coupling of Lemma 2: for each prompt, walk T steps under $A \wedge B$, at each step accumulate $\log(Z_{\min}/Z_{\max})$ in log space over the retained masses of A and B, and average $1 - e^{L}$ over prompts with a bootstrap interval. No component beyond the retained masses and the pointwise-narrower draw is needed, which is why the certificate is computable from any decoding run's own logs.

Grid. Each of the six tunable families is swept at six log-spaced points, alternating $1\times$ and $3\times$ each power of ten: top-$k \in \{10, 20, 40, 80, 160, 320\}$, top-$p \in \{0.70, 0.80, 0.90, 0.95, 0.98, 0.99\}$, $\epsilon \in \{3{\times}10^{-4}, 1{\times}10^{-3}, 3{\times}10^{-3}, 1{\times}10^{-2}, 3{\times}10^{-2}, 1{\times}10^{-1}\}$, η over the same six values, min-$p \in \{0.01, 0.02, 0.05, 0.10, 0.20, 0.30\}$, and top-$n\sigma \in \{0.5, 0.8, 1.0, 1.5, 2.0, 3.0\}$. Basic (untruncated) and p-less have no hyperparameter, and top-H is used at its default (top-n=100, α=0.4) rather than swept, so the three contribute one member each, giving $\mathcal{G}$ 39 members and $\binom{39}{2} = 741$ pairs.

Batch size. Implementation is a custom decoding loop over HuggingFace Transformers, run at batch size one throughout, since a batch-invariance check found next-token logits drifting by up to 0.94 between batch size 1 and batch sizes 8 and 64, so certificates are batch-size dependent. Every estimator arm (coupled, prompt-paired, independent) runs the same custom decoding loop at batch size one, and none uses a batched path or vLLM (Kwon et al., 2023), so the batch-invariance drift cannot have confounded any comparison in Section 4.1.

Top-2048 exclusion. Certificates use the top 2048 logged probabilities, and steps whose cutoff would exceed that are excluded consistently across every certificate drawn from the same log. Excluding a step drops a factor of at most 1 from the retained-mass product rather than the true,

smaller factor, so this is anti-conservative in principle: the size of the understatement depends on how far each excluded step's true r_t sits below 1, not merely on how many steps are excluded, so the excluded-step fraction alone is not itself a bound on the error. The grid certificate needs no such exclusion at all. $\mathcal{G}^\vee$'s cutoff is basic's, the untruncated rule, whose retained mass is exactly 1 by construction regardless of how much of the tail is logged, so its true value is used directly; summing only the logged head instead would understate this mass by up to 9% on the steps we checked, making any certificate computed that way nominally anti-conservative. $\mathcal{G}^\wedge$'s cutoff is 1 on essentially every step, since top-$k=10$ alone already bounds it at 10, and 1 or 10 never approaches the logged 2048. An exclusion flag set whenever *any* of the grid's 39 members needs the tail, rather than whenever $\mathcal{G}^\wedge$ or $\mathcal{G}^\vee$ themselves do, would report excluded-step fractions as high as 6.8% on gpt-oss-20b/MATH-500; with the check the grid certificate actually needs, no step is excluded on any of the nine configurations.

Compute. All decoding runs to date total 64.5 H200-GPU-hours, plus under one CPU-hour for prompt preparation and the logs-only experiments, against the six thousand A100-hours of Schaeffer et al. (2025), so certifying every family we tested cost under two percent of one baseline's training-adjacent search.

F Additional Results

Gate controls. On Qwen3-1.7B/GSM8K (T=256, n=200, pair min-p=0.05 versus top-p=0.95) every per-prompt first-step cutoff matched exactly between an independent coupled read and an independent solo read of the same prompt ($200/200$), the identical-pair control showed zero divergence ($0/200$), and the greedy-versus-untruncated control diverged on every prompt ($200/200$). The comparison against the point prediction, that the certificate equals the probability a real simulated coupled decoding run diverges, uses a percentile bootstrap over the per-prompt divergence indicator with resampling unit prompt and n=200.

The pairwise distribution over the grid. The full pairwise distribution over the grid's 741 pairs is not yet reported. A pairwise certificate can be read off a run under either member alone (Proposition 8), so the 741 pairs need at most 39 new single-rule runs rather than 741 pair-specific ones, cheaper but still undecoded. The existing $\mathcal{G}^\wedge$-walked grid logs do not substitute for this, since Proposition 8 needs a run under one specific member, not under the grid's pointwise-narrowest rule, for which a $\mathcal{G}^\wedge$ run remains not conservative. Of the 741 grid pairs, only the default one (min-p versus top-p at default hyperparameters) has that log so far, separately from the matched-operating-point pairs of Table 4: certificate 0.945, 95% CI $[0.931, 0.959]$ over 200 prompts, computed by the machinery that would produce the full distribution once the remaining 740 pairs are decoded.

Why the flip-rate contrast is not established. Three reasons qualify the raw-versus-matched flip-rate contrast of Section 4.1. The raw and matched rates were measured on different pairs and possibly different horizons, so the comparison is not paired. Per-trajectory flip rate is monotone in T, so a horizon difference alone could produce it. And within Table 4 the flip rate is monotone in matching quality, with the worst-matched pair flipping least, consistent with the mechanism but also meaning the three points sit on a dose-response curve at one operating point and all share top-p as an arm.

Matched operating-point sweep. On Qwen3-1.7B/GSM8K, min-p versus top-p at five matched operating points $\rho \in \{0.82, 0.87, 0.90, 0.95, 0.98\}$, where ρ<0.82 is unreachable for min-p and η on this model's real distributions since their measured floors are 0.72 and 0.80, the per-step minority-sign rate is 0.018–0.021 and the trajectory flip rate 0.70–0.93, against 0.014 and 0.99 raw. The same sweep for η versus min-p, the matched pair not involving top-p, gives per-step rates 0.012–0.029 and trajectory rates 0.67–0.98. The correlation between $\log(|\mathbb{E}[g_t]|/\mathrm{sd}[g_t])$ and the per-step minority-sign rate across all eleven conditions, raw plus the ten matched, is -0.53. Matching lowers the mean gap relative to its spread, in the direction Equation (1) predicts, but a correlation of this size across eleven points is a moderate relationship rather than a clean collapse onto one curve.

Long-horizon details. The four AIME horizons are read from one T=32,768 decoding run as true prefixes, so all four numbers come from the same trajectories, and Sep_T is computed directly as

$1 - \frac{1}{n}\sum_i e^{-S_i}$ over the 30 prompts rather than from the Jensen bound $1 - e^{-\Delta_T}$. It is monotone non-decreasing in T, as the theory requires once the same trajectories are reused rather than resampled at each horizon. Generations do not all reach the cap: $\mathrm{eos_rate}$ (the fraction of the 30 prompts finished by that horizon) rises from 0 at T=1024 to 0.07, 0.47 and 1.00 at the three longer caps, so we report Δ_T per realized token (0.017, 0.015, 0.016, 0.016 respectively) rather than per capped token, which would understate the rate once most generations finish early. A shorter-horizon run (Qwen3-1.7B/GSM8K, T=256 through 1024) confirms the methodology where every generation reaches the cap, $\mathrm{eos_rate} = 0$. Table 4 reports the operating-point-matched certificates in full.

Coupling details. The maximal coupling cuts variance 3.8× against prompt-level pairing on the default pair. The shared-seed baseline and a shared-uniform-coupling ablation (Vilnis et al., 2023; Li et al., 2026) have 3.9× and 2.0× the maximal coupling's variance on the same pair, both intervals excluding 1, so these two particular ways of sharing randomness, not sharing randomness as such, fall short of its reduction. Shared Gumbel-max noise, drawing one Gumbel vector per step and taking each rule's argmax over its own retained set, is itself a realization of the same maximal coupling for nested prefixes (the two argmaxes agree with probability exactly Z_A/Z_B, verified against the closed form to 4 significant figures on a synthetic distribution), but was not run here. The negative control has small coupled-arm outcome variance (0.040) and a near-zero mean paired gap (−0.002), because both rules answer most prompts the same way regardless of how often their tokens diverge, not because the rules are similar as samplers; its measured reduction is $(\sigma_A^2 + \sigma_B^2)$/coupled variance ≈ 12.0×, with σ_A^2, σ_B^2 measured on the independent arms. For GPQA-Diamond the interval $[8.0, 62.1]$ excludes 1 while the bootstrap $p = 0.16$ is not significant even before any correction across the sweep. These need not agree: the interval is a percentile interval of the bootstrap ratio itself, while the p-value tests against a bootstrap distribution re-centered at its own mean, a different, more conservative construction under skew, and this ratio's bootstrap distribution is heavily right-skewed here. The reduction is therefore not an artifact of the one saturated benchmark it was first measured on, since MATH-500 replicates it with bootstrap $p = 0.023$.

Predictiveness. In Table 5, a within-family certificate as low as 0.17 is not surprising: a monotone top-k sweep can reach Sep_T=0.10, since nothing in the theory requires two members of one family to be close. Under the null the measured gap's own standard error scales with $\sqrt{\mathrm{Sep}_T}$, which by itself induces a positive Sep_T-versus-|gap| correlation with no real prediction behind it; standardizing each gap by its standard error is what that null noise mechanism should remove but a real predictive relationship should not. `code/predictiveness_null_check.py` reproduces both null statistics.

Table 3: **Prompts needed to resolve a one-point difference with k coupled samples per prompt** (default pair, GSM8K variance decomposition). Benchmark sizes: GSM8K 1,319, MATH-500 500, GPQA-Diamond 198, AIME 2025 30.

k	$\tau^2 = 0$ (point estimate)	$\tau^2 = 9.3 \times 10^{-4}$ (upper bound)
1	6,318	6,392
8	790	863
32	197	271
≈ 210	within AIME 2025's 30 (untested)	103
any k	no floor	never below about 73

Benchmark-size derivation. The k-sample requirements of Table 3 decompose the five-seed coupled outcomes for the default pair on GSM8K into within- and between-prompt variance (σ_w^2 and τ^2); Appendix G describes the script that reproduces them from the raw per-seed outcomes. The k=8 point-estimate figure, 790, is close to but not the same number as GPQA-Diamond's own k=1 coupled requirement (787): the two are unrelated quantities, computed from different pairs' variance decompositions, that happen to be numerically close. At the upper confidence bound, the required count never falls below about 73 however large k is taken, since τ^2 alone then accounts for more than AIME 2025's 30 prompts can resolve.

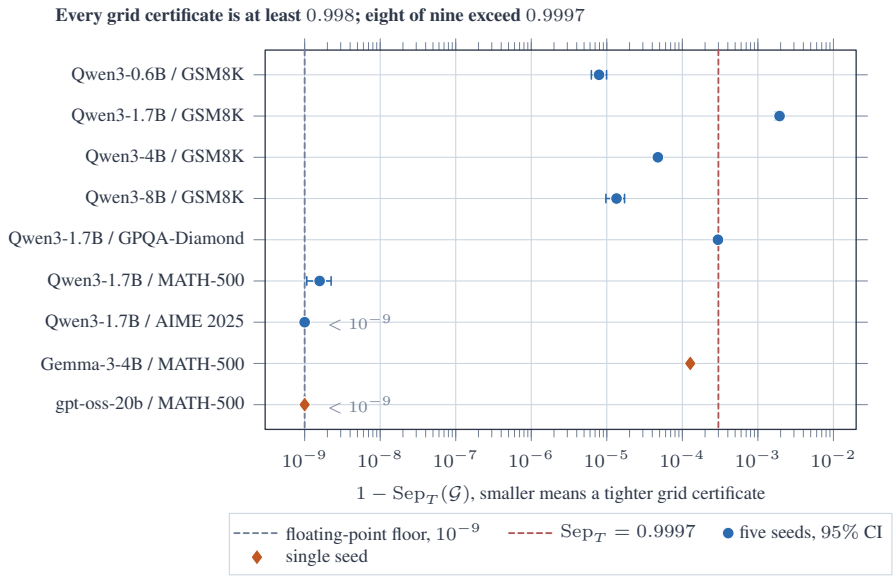


Figure 5: **The grid certificate on all nine configurations,** as $1 - \mathrm{Sep}_T(\mathcal{G})$ on a log axis, so a point further right is a looser certificate. Circles are five-seed estimates with 95% CIs, diamonds single-seed estimates. Two values fall below floating-point resolution and are drawn at 10^{-9}. Only Qwen3-1.7B/GSM8K (0.998064) lies right of the $\mathrm{Sep}_T = 0.9997$ line.

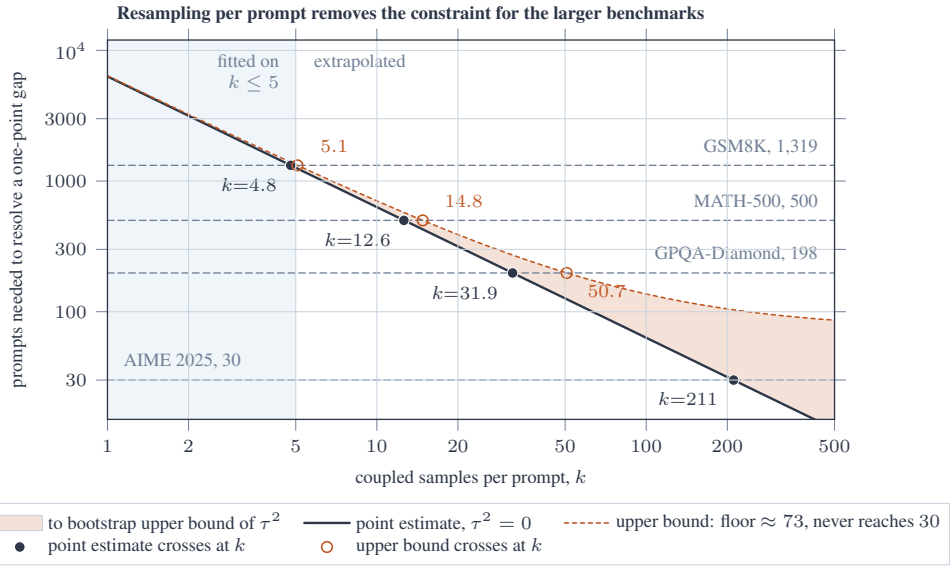


Figure 6: **Prompts needed to resolve a one-point difference with k coupled samples per prompt** (default pair, GSM8K variance; Table 3). The solid line is the point estimate $\tau^2 = 0$; the band reaches the upper bound of τ^2, where the requirement never falls below about 73. Markers show where each curve meets a benchmark's size (dashed). The model is fitted on five samples per prompt, so everything beyond $k = 5$ is extrapolated, with GSM8K's τ^2 used for every benchmark.

G Mechanical Verification of the Theory

The supplementary code checks every formal claim numerically on synthetic models, and is written to falsify rather than confirm. It verifies prefix membership and nesting for the swept rules of Table 1 (including ties and boundary cases), Lemma 2(i) to machine precision and the marginals and maximality of Lemma 2(ii) by Monte Carlo, and the bounds of Proposition 7 against adversarially

Table 4: **At a matched operating point the certificate stays large and stops being exact** (Qwen3-8B, MATH-500, T=1024). Rows are ordered by how closely the pair was matched; the flip-rate paragraph above discusses the dose-response reading of the flip-rate column.

Pair, matched to $\rho = 0.95$	achieved ρ	Sep_{1024}	Δ_{1024}	Pr[flip]
min-p vs top-p	0.954	0.993	13.67	0.998
top-$n\sigma$ vs top-p	0.957	0.994	13.93	0.996
η vs top-p	0.978	0.998	27.55	0.374

Pr[flip] is the fraction of trajectories on which the nesting order flips; wherever it is positive the certificate is a bound, not an equality. Single runs without intervals, and flip rate grows with T, so this column is comparable only at T=1024. Sep at three decimals fixes $1 - \mathrm{Sep}$ to one significant figure, so compare rows by Δ. Qwen3-8B/MATH-500 is not one of the nine grid configurations, so it has no grid certificate; the grid certificate never flips by construction ($\mathcal{G}^\wedge$ is narrower than $\mathcal{G}^\vee$ at every context) and is therefore exact.

Table 5: **The eight pairs behind the predictiveness check** (Qwen3-1.7B, GSM8K, T=256, five seeds). Family marks whether the pair compares two different truncation rule types (cross) or the same rule at two hyperparameter settings (within), including the negative control.

Pair	**Family**	Sep_T	gap	gap 95% CI
min-p(0.05) vs top-p(0.95)	cross	0.710	0.017	$[0.000, 0.034]$
top-k(10) vs top-p(0.70)	cross	0.992	-0.023	$[-0.052, 0.006]$
ϵ(0.1) vs η(0.0003)	cross	0.989	0.013	$[-0.015, 0.041]$
top-p(0.98) vs top-p(0.99)	within	0.437	-0.002	$[-0.014, 0.010]$
top-k(10) vs top-k(320)	within	0.541	0.003	$[-0.009, 0.015]$
top-k(10) vs top-k(80)	within	0.517	0.003	$[-0.008, 0.014]$
top-k(80) vs top-k(320)	within	0.174	0.000	$[-0.004, 0.004]$
min-p(0.01) vs min-p(0.30)	within	0.992	0.013	$[-0.015, 0.041]$

gap is the mean signed accuracy difference over 1,000 coupled-arm observations (200 prompts $\times$ five seeds); the certificate bounds |gap|. Observations lie in $\{-1, 0, 1\}$, so means and bootstrap percentiles are multiples of 0.001, which is why two rows share gap and CI (0.013, $[-0.015, 0.041]$) although their Sep_T and mean absolute gaps (0.203 vs 0.205) differ. The default pair's 0.710 differs from the gate's 0.722 at T=256 by Monte Carlo variation across decoding sessions, within the gate's interval. These certificates are computed online against the full softmax, so no top-2048 exclusion applies.

constructed outcomes. Exact enumeration over all sequences of context-dependent synthetic models confirms that the certificate's expectation must be taken along $A \wedge B$ (substituting $\mathcal{G}^\wedge$ is not conservative), that Sep_T equals the sequence-level total variation exactly on non-flipping pairs and exceeds it otherwise, the equality condition of Equation (1), the dominance of the grid certificate over every pairwise certificate, Proposition 8, the attaining metric of Appendix B, the Gumbel-max agreement rate Z_A/Z_B, and the locally typical counterexample of Appendix D. A second script checks Proposition 5 against its exact three-state computation and both parts of Corollary 6 against the exact distribution of S. All 28 checks pass, and a deliberately incorrect variant of Equation (1) fails as it should. These checks show that theory and code agree on synthetic inputs; no number from them is reported as a property of a language model.

Two real-model computations have their own scripts. `code/benchmark_size_analysis.py` reproduces the benchmark-size requirements and the within/between-prompt decomposition with its bootstrap upper bound on τ^2, and `code/aime_horizons_as_prefixes.py` computes the four AIME certificates as prefixes of one T=32,768 run, along which Sep_T is monotone in T.