SAMPLER GAINS SHRINK TO NOISE UNDER EQUAL TUNING BUDGETS: THE WINNER'S CURSE IN LLM DECODING EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Truncation samplers such as top-p, min-p and typical sampling are often reported to beat one another by sweeping the proposed sampler, reporting its best score and comparing it with untuned defaults. We show that this practice carries a quantifiable winner's curse. Modeling tuning as random search with budget K over paired evaluation noise, we express the inflation of best-of-K reporting and the spurious gap between identical methods tuned with unequal budgets, and show which part of a tuned-vs-untuned gain a dev/test split removes; on seed-replicate runs, whose truth is known, measured inflation runs at 0.81–$1.10\times$ the formulas. Our equal-budget audit (EBET) gives every sampler family the same budget, selects on dev items and reports on held-out ones, across eleven model–task tiers on GSM8K and MATH-500 (1,218,657 generations in all). On these benchmarks temperature, top-k, top-p and typical are indistinguishable at their tuned optima: equivalence tests on the Qwen tiers exclude pooled differences of 3.5 points or more (smaller ones are not resolved), and no family beats greedy past the protocol's resolution floor. Under our min-p prior, which puts most of its temperatures above its rivals' caps, min-p trails top-p in all eight Qwen tiers; redrawn within their range, or with p paired to T as its guidelines advise, it ties ($-.001$ and $.000$ pooled), though neither arm excludes a deficit as large as the original. Re-analyzed in a stylized published style (best of 24 against an untuned default), the same records show gains of $+1.4$ to $+8.8$ points in every tier, by our estimate mostly selection bias. All code and per-item generation records will be made public at `https://github.com/xxx/xxx` upon acceptance.

1 INTRODUCTION

How should a language model pick its next token? Recent decoding research has produced a run of truncation samplers: top-k (Fan et al., 2018), nucleus (top-p) (Holtzman et al., 2020), typical (Meister et al., 2023b), η/ϵ (Hewitt et al., 2022), and min-p (Nguyen et al., 2025), each introduced with evaluations showing it beats its predecessors. These comparisons steer deployments: inference stacks ship the winners as defaults and practitioners copy published hyperparameters; the recent dispute over min-p's reported gains (Schaeffer et al., 2025) shows the field lacks a shared standard.

We argue that the dominant evaluation practice has a structural flaw, hard to see from any single paper's main table: *unequal tuning budgets combined with best-score reporting*. A new sampler is swept over its grid; the best-scoring configuration goes in the main table; baselines appear at conventional defaults (e.g. top-p 0.95 at $T{=}1$) or over smaller grids. The number used to *select* the winner is often the number *reported*: the min-p evaluation displays the hyperparameters achieving the best results per method (Nguyen et al., 2025) and a broad decoding comparison reports per-dataset best hyperparameters without stating a dev/test split (Shi et al., 2024). This is best-of-K selection on a noisy estimate (Fig. 1a), and it inherits the selection bias known as the winner's curse or optimizer's curse (Smith & Winkler, 2006; Zöllner & Pritchard, 2007; Sculley et al., 2018).

This paper makes the winner's curse of sampler evaluation calculable, checks the formulas where the truth is known, and runs the equal-budget audit they call for:

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.

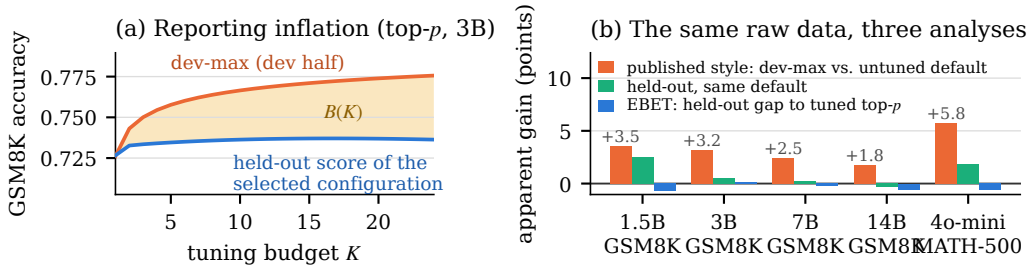

Figure 1: **The winner's curse in sampler evaluation.** (a) One family pool (top-p, Qwen2.5-3B, GSM8K, 325 items): the best of K configurations is chosen on a random half of the items and scored on the other half. From $K{=}1$ to 24 the dev-max score (dev size $n/2$) rises 4.9 points and the held-out score of the chosen configuration 1.0; the shaded gap is selection bias $B(K)$, 3.9 points at $K{=}24$. (b) The same generations analyzed three ways, in accuracy points, averaged over families, on the five tiers that carry a seed-replicate leg (Table 10 has all eleven). Orange, published style: a family's dev-max over $K{=}24$, no split, minus an untuned top-p 0.95 default (Table 10). Green: the same comparison held out at dev size $n/2$ (Table 14's tuning term), which removes the selection bias. Blue, EBET: the held-out gap to equally tuned top-p instead of the default, over the rivals the tier expresses, min-p included (Table 3).

- **Theory (§3).** Because every configuration is scored on the *same* items, evaluation noises are correlated through item difficulty; selection cannot exploit the share ρ of noise common to all configurations. Modeling tuning as random search (Bergstra & Bengio, 2012) with budget K, we show the reported best-of-K score overstates the selected configuration's true quality by $\sigma\sqrt{1-\rho}\,e_K$ (the classical expected maximum of equicorrelated Gaussians) on flat landscapes (all K configurations equally good) with independent, equal-variance Gaussian residuals, where σ is the per-configuration standard error and e_K the expected maximum of K standard normals. We also derive the expected *spurious gap* between two identical methods given unequal budgets: several accuracy points at realistic n, K and accuracy. A decomposition (Prop. 3) separates the effect of tuning from selection bias: held-out reporting removes only the latter, so a comparison of methods also needs matched budgets.
- **Validation (§5.1).** We check Props. 1–2 where the truth is known. On five seed-replicate legs (17 runs of one configuration differing only in seed: a flat landscape), measured inflation and the dev-max gap between two identically distributed pseudo-methods run at 0.81–1.10$\times$ Eqs. (3) and (4).
- **Audit (§5.2–5.4).** EBET (*Equal-Budget Evaluation of Truncation samplers*) gives every family the same budget, $K{=}24$ configurations drawn from priors built around published truncation values (App. B marks which parts are ours), selects on dev items, reports on held-out items, and prints beside each gap the protocol's *minimum resolvable sampler gap* (MRSG, §3.2), a resolution floor that is conservative for claimed differences and lenient for claimed ties. Across eleven (model, task) tiers and 1,218,657 generations, temperature, top-k, top-p and typical are indistinguishable at their tuned optima on these math benchmarks. A permutation test of whether two families' configurations are exchangeable rejects in three of their 61 comparisons (chance predicts 3.1), and pooled over the eight Qwen tiers the other three lie within 0.50 points of top-p. The min-p sampler is the exception, trailing top-p in all eight Qwen tiers: a consistent sign (sign test $p = 0.008$, tiers as independent units) of a pooled size not significant on our conservative single-run yardstick (cluster bootstrap, $p = 0.17$; at the bagged estimate's own spread, $z \approx -3.1$). The span of its evaluation's main results, sampled uniformly with p drawn independently of T (our simplification), puts 17 of its 24 draws above $T{=}1.5$, beyond the caps we set for its rivals. Redrawn at $T \le 1.5$ at the same budget (Table 13), it trails in four of eight ($-.001$ pooled); over its full range with p paired to T following its guidelines, in four of eight (.000; App. F). Neither redrawn arm's gap is resolved from zero, and neither equivalence bound ($\pm.028$, $\pm.028$) excludes the audit arm's deficit.

We do not claim samplers never matter: truncation demonstrably prevents degeneration from the unreliable tail (Holtzman et al., 2020), most visibly at high temperature (Nguyen et al., 2025),

and in an exploratory control giving every rival min-p's temperature range (one tier; App. F) the rules separate. Our claim is about *evaluation*. At tuned optima under equal budgets on these math benchmarks, equivalence tests on the eight Qwen tiers exclude pooled differences of $.035$ or more among four of the five audited families, and no family beats greedy past the MRSG; published-style reporting of the same records yields wins over untuned top-p in every tier, of several points and mostly selection bias at the item counts of most tiers (Table 10, Fig. 1b).

2 Related work

Truncation samplers and their evaluations. Fan et al. (2018) introduced top-k; Holtzman et al. (2020) diagnosed neural text degeneration and proposed nucleus sampling; Basu et al. (2021) controlled perplexity directly; Meister et al. (2023b) derived typical sampling information-theoretically; Hewitt et al. (2022) recast truncation as desmoothing, yielding η/ϵ-sampling; Nguyen et al. (2025) proposed min-p, reporting wins across GPQA, GSM8K and creative writing; Finlayson et al. (2024) and Tang et al. (2025) proposed basis-aware and top-$n\sigma$ truncation. Meister et al. (2023a) ask whether sampling adapters' gains reflect the mechanism; we ask it of the protocol. Nadeem et al. (2020), Ippolito et al. (2019), Wiher et al. (2022) and Zhang et al. (2021) compared decoding methods across settings, the last over small matched grids, without held-out reporting. Shi et al. (2024) give the broadest modern comparison, per-dataset tuned over unequal grids with no split reported, for us an object of study. Renze (2024) finds no significant temperature effect over 0.0–1.0 on its models (our pools are not flat in T; Table 18). Zhou et al. (2025) choose a truncation method and parameter for open-ended generation by a diversity–risk criterion, and Freitag et al. (2023) compare samplers tuned on dev data for MBR decoding, where they differ.

The min-p dispute. Schaeffer et al. (2025) re-evaluated min-p against three samplers on nine models over $T \in [0, 3]$ and found it does not surpass them at equal numbers of swept configurations (best-of-N, no held-out split), besides raising statistical and judge-protocol concerns. We add a validated analytic account of the gaps selection produces and a held-out, equal-budget audit with resolution floors.

Budget-aware reporting and evaluation noise. Dodge et al. (2019) proposed reporting expected validation performance against tuning budget; Tang et al. (2020) showed that estimator's plug-in form is biased and gave the unbiased hypergeometric weighting we build on. Neither models inflation in closed form nor measures the selected configuration's held-out quality (Props. 1, 4). Reimers & Gurevych (2017) argued for score distributions over random seeds and Dror et al. (2018) for proper significance tests; Melis et al. (2018) showed neural-LM architecture gains dissolve once baselines get matched tuning budgets (closest in design to our audit), and Lucic et al. (2018), Li & Talwalkar (2019), Choi et al. (2019), Sivaprasad et al. (2020), Schmidt et al. (2021) and Musgrave et al. (2020) found GAN, architecture-search, optimizer and metric-learning rankings change the same way. Lourie et al. (2024) give confidence bands for budget curves; Gorman & Bedrick (2019) show rankings reverse under resplitting; Demšar (2006) reviews cross-dataset tests, the sign test among them; Jensen & Cohen (2000), Varma & Simon (2006), Tibshirani & Tibshirani (2009) and Smith & Winkler (2006) document (and correct) the optimism of selected estimates; Andrews et al. (2024) give valid intervals for a selected winner; Cooper et al. (2021) show the tuning protocol can decide which algorithm wins; and Singh et al. (2025) document best-of-N private testing on an LLM leaderboard. On the noise side: seed and tuning variance (Henderson et al., 2018; Bouthillier et al., 2021), LLM benchmark variance and the decoding sensitivity of small math benchmarks (Madaan et al., 2024; Hochlehnert et al., 2025), paired SEs (Miller, 2024), minimum detectable effects (Card et al., 2020), and the minimum detectable ability difference (MDAD) of micro-benchmarks (Yauney et al., 2026), of which our MRSG is the budget-aware member.

3 A theory of sampler-evaluation inflation

3.1 Setup

A *sampler family* s (temperature, top-k, top-p, min-p, typical) has configuration space Θ_s; for truncation families a configuration is a pair $\theta = (T, \tau)$ of temperature and truncation parameter. A *search prior* P_s over Θ_s models how the family is tuned; App. B marks which of its ranges are cited

and which are ours. A task supplies items $i = 1, \ldots, n$; configuration θ receives per-item outcomes $X_i(\theta) \in [0, 1]$ (correctness for math, rescaled judge score for open-ended generation), empirical score $\hat{S}(\theta) = \frac{1}{n}\sum_i X_i(\theta)$, true score $\mu(\theta) = \mathbb{E}[X_i(\theta)]$, and noise $\varepsilon(\theta) = \hat{S}(\theta) - \mu(\theta)$.

Paired noise. All configurations are evaluated on the *same* items, so noises are correlated through item difficulty. Decompose $X_i(\theta) = \mu(\theta) + a_i + u_i(\theta)$ with a_i the zero-mean item main effect shared across configurations, $u_i(\theta)$ the zero-mean residual (item×configuration interaction plus generation randomness), and $\bar{a}, \bar{u}(\theta)$ their item means. Then

$$\varepsilon(\theta) = \bar{a} + \bar{u}(\theta), \qquad \sigma^2 := \operatorname{Var}\varepsilon(\theta) = \tfrac{1}{n}\big(\operatorname{Var}(a_i) + \operatorname{Var}(u_i(\theta))\big), \qquad \rho := \operatorname{Var}(\bar{a})/\sigma^2, \quad (1)$$

for n items (in Props. 1–3 the items a score averages over; in Prop. 4 and Def. 1 the pool's items, split into halves), where ρ is the share of the noise common to all configurations (their intraclass correlation when residuals are uncorrelated across configurations). This assumes i.i.d. items, an item effect with unit loading in every configuration and $\operatorname{Cov}(a_i, u_i(\theta))=0$; for binary outcomes σ^2 varies with $\mu(\theta)$, so σ below is a pool average. Since $X_i \in [0, 1]$, ε is $\frac{1}{4n}$-sub-Gaussian and the Gaussian approximation is accurate for moderate n; for binary accuracy $\sigma^2 = \mu(1-\mu)/n$.

Tuning as random search. Draw $\theta_1, \ldots, \theta_K \sim$ iid P_s, evaluate all on dev items, select $\hat{k} = \arg\max_k \hat{S}_{\text{dev}}(\theta_k)$ (dev and test scores are $\hat{S}$ on those items), breaking ties uniformly at random among tied configurations. Two reporting conventions coexist: **(R1) dev-max reporting** publishes $M_K = \hat{S}_{\text{dev}}(\theta_{\hat{k}})$, where the number used to select is the number reported (as far as their reporting shows, Nguyen et al., 2025; Shi et al., 2024); **(R2) test-of-selected** publishes $\hat{S}_{\text{test}}(\theta_{\hat{k}})$ on held-out items.

3.2 Four propositions

Proposition 1 (Dev-max inflation)**.** *The bias of R1 relative to the true quality of the selected configuration is $B(K) := \mathbb{E}[M_K] - \mathbb{E}[\mu(\theta_{\hat{k}})] = \mathbb{E}[\varepsilon(\theta_{\hat{k}})] \geq 0$, with*

$$B(K) \;\leq\; \mathbb{E}\big[\max_k \varepsilon(\theta_k)\big] \;\leq\; \mathbb{E}[\bar{\sigma}]\sqrt{2\ln K}, \quad \bar{\sigma} := \max_k \sigma_k\sqrt{1-\rho_k}, \quad \textit{if every } \bar{u}_k \textit{ is } \sigma_k^2(1-\rho_k)\textit{-sub-Gaussian.} \quad (2)$$

With equal variances $\bar{\sigma} = \sigma\sqrt{1-\rho}$; a pool-average σ does not suffice (App. A), and for binary outcomes the premise holds only in the Gaussian approximation; boundedness alone gives the weaker $\sqrt{\ln K/(2n)}$ (App. A). On a flat landscape ($\mu(\theta_k) \equiv \mu$) the first inequality is an equality, and under the Gaussian approximation with independent, equal-variance residuals its value is

$$B_{\text{flat}}(K) = \sigma\sqrt{1-\rho}\, e_K, \qquad e_K := \mathbb{E}\big[\max \textit{ of } K \textit{ iid } \mathcal{N}(0,1)\big], \quad (3)$$

with e_1=0, e_2=0.564, e_6=1.267, e_{20}=1.867, e_{24}=1.948. For equal-variance Gaussian residuals with non-negative correlations, (3) is an upper envelope; with unequal variances a pool-average σ need not bound it, so on real pools §5.1 checks it empirically.

The shared component $\bar{a}$ has mean zero and does not affect the argmax; only the residual noise, of SD $\sigma\sqrt{1-\rho}$, is exploitable by selection (proof in App. A). *The curse bites hardest when a family's configurations are equivalent*. Where the landscape varies far more than the noise, selection is signal-driven and $B(K)$ is negligible; where tuned configurations sit near a plateau it is mostly noise-driven and reported numbers inflate by up to the full $\sigma\sqrt{1-\rho}\, e_K$ (under Eq. (3)'s conditions; §5.1). GSM8K-style accuracy near 0.5 at n=200 gives $\sigma \approx 0.035$; with ρ=0.52 (the median of Table 12) and K=20 that is ≈4.6 accuracy points.

Proposition 2 (Spurious gap between identical methods)**.** *Let families A and B have identical landscapes and noise (σ, ρ) and share dev items, with R1 reporting and budgets $K_A > K_B$. In the flat-Gaussian regime (flat μ; independent, equal-variance Gaussian residuals; $\rho < 1$),*

$$\mathbb{E}\big[M_{K_A} - M_{K_B}\big] = \sigma\sqrt{1-\rho}\,\big(e_{K_A} - e_{K_B}\big) > 0, \quad (4)$$

and, flat or not, whenever the K_A+K_B dev scores are exchangeable (the same prior on the same landscape, with jointly exchangeable residuals) and ties are broken uniformly among the tied configurations, A "wins" (holds the dev argmax over all K_A+K_B configurations, ties broken uniformly) with probability $K_A/(K_A + K_B) > 1/2$, an upper bound for $\Pr[M_{K_A} > M_{K_B}]$.

The shared item effect cancels; the residual maxima differ in expectation. Under the parameters above, a method swept at K_A=20 against a baseline at K_B=2 settings collects a spurious ≈3-point gap with no true difference on a plateau.

Proposition 3 (Two budget effects; held-out reporting removes only one)**.** *The expected R1-reported gain of tuned family A over an untuned baseline θ_b decomposes as*

$$\mathbb{E}[M_K^A - \hat{S}(\theta_b)] = \underbrace{\mathbb{E}[\mu(\hat{\theta}_A)] - \mu(\theta_b)}_{\textit{tuning term}} + \underbrace{B_A(K)}_{\textit{selection bias}} . \tag{5}$$

Here θ_b is fixed in advance and $\hat{\theta}_A$ is A's dev-selected configuration. R2 reporting removes $B_A(K)$, since test items are drawn independently of the dev items and the draws. The tuning term remains. Given $\bar{a}$, with residuals independent across configurations and of $\bar{a}$, it is nondecreasing in K when $g(s) = \mathbb{E}[\mu(\theta) \mid \mu(\theta) + \bar{u}(\theta) = s]$ is nondecreasing in s (e.g. i.i.d. Gaussian residuals independent of μ; for binary outcomes an assumption, App. A), and can fall otherwise: selecting on a noisy dev score need not improve the expected true score if noise, or its correlation across configurations, varies with μ.

Held-out reporting thus removes the selection bias but not the budget-dependent tuning term: a comparison of methods, rather than of tuning budgets, also needs matched budgets.

Proposition 4 (Unbiased budget curves from an evaluated pool)**.** *Evaluate a pool of N configurations per family once on n items from the task's population and split the items at random into dev and test halves. For any $K \le N$, weighting configurations by their dev rank r (ascending) with the hypergeometric weights of Tang et al. (2020),*

$$w_r(K) = \binom{r-1}{K-1} \Big/ \binom{N}{K}, \qquad \hat{\mathbb{E}}[\text{devmax@}K] = \mathbb{E}_{\text{splits}}\big[\textstyle\sum_{r=K}^{N} w_r(K)\, \text{dev}_{(r)}\big], \tag{6}$$

gives unbiased estimates of the expected dev-max under budget-K random search over the pool at dev size $n/2$, and (our extension) the same weights applied to the rank-r configuration's test-half *score, under the same uniformly tie-broken ranking, give unbiased estimates of the selected configuration's expected true score (items from a superpopulation; App. A). Their gap measures $B(K)$ with no model.*

The test half is independent of the dev ranking, so the second curve is unbiased; scoring the selection instead against each configuration's mean over all n items (its *pool truth*) recovers only part of $B(K)$ (half, with equal disjoint halves; App. A).

Definition 1 (Minimum resolvable sampler gap)**.** *For a protocol $\pi = (n_{\text{dev}}, n_{\text{test}}, K, P_A, P_B)$, let SE_{total} be the standard deviation of the held-out gap $\hat{S}_{\text{test}}(\hat{\theta}_A) - \hat{S}_{\text{test}}(\hat{\theta}_B)$ over item samples and dev/test splits, given the realized pool. $\text{MRSG}_\pi(0.8) := 0.8416 \cdot \text{SE}_{\text{total}}$ is the true gap at which a Gaussian gap estimate of that spread orders the pair correctly with probability 0.8. We estimate SE_{total} by a paired bootstrap that redraws the split, resamples items within the disjoint dev and test halves and re-runs the dev-argmax selection over the realized pool; that overstates one fresh run's spread by about $\sqrt{1.5}$, so a printed MRSG orders a fresh run's pair correctly with probability about 0.85 and omits the prior draw's variance, which the draw bootstrap of App. C approximates.*

MRSG charges a comparison for selection noise, not just item noise. The gap we report, $\hat{\Delta}$, averages 600 bootstrap replicates, and its own spread over fresh items is about $1/3.4$ of SE_{total} (App. C). Every interval, z, equivalence bound and MRSG uses the wider SE_{total}, so all are conservative for $\hat{\Delta}$'s spread (and, by the draw bootstrap, cover the draw's variance here, App. C); the item-conditional permutation test (§3.3) uses the same selection averaged over 600 shared random splits of the items, observed and permuted alike. $|\hat{\Delta}| > \text{MRSG}$ (an *MRSG crossing*) is a screen, not a test: conservative for a claimed difference and lenient for a claimed tie. It fires on about 0.4% of null comparisons for the bagged $\hat{\Delta}$ under a fresh-item null (4.5% between random arms of one prior) but on 30–40% for one unbagged split-and-select run (App. C), so the protocol bags. *Significant* means the 95% criterion on SE_{total}; a permutation test *rejects*.

On flat landscapes $B(K) \propto 1/\sqrt{n}$. The propositions also give a retrospective check (cf. Zöllner & Pritchard, 2007): from a paper's n, accuracy and number $\hat{K}$ of settings displayed, $\hat{\sigma}\sqrt{1-\hat{\rho}}\, e_{\hat{K}}$ estimates what selection alone adds to its dev-max score on a plateau (an overestimate off one only for equal residual variances; App. C).

3.3 THE EBET PROTOCOL

The propositions convert directly into a protocol. **(1)** Search priors follow published ranges where they exist, cited, with every extension marked as the evaluator's (and a restricted-range check for any family whose prior extends past its rivals', §5.2). **(2)** Every family gets the same budget K and dev items; seeds derive from family-specific (configuration, item, replicate) identifiers. **(3)** Selection is on dev; reporting is test-of-selected, bagged over bootstrap replicates, with paired-item bootstrap intervals. **(4)** Budget curves (Prop. 4) for dev-max and test-of-selected are reported; their gap is the winner's-curse diagnostic. Every claimed gap sits beside its MRSG, and every family pair gets a *permutation test*: reassign the two families' $2K$ configurations at random to two arms of K, re-run selection and held-out scoring, and locate the observed gap among the permuted ones. It is distribution-free, but its null is exchangeability given the items drawn, so it tests prior and rule jointly, on these items. **(5)** All per-item outcomes are released.

4 EXPERIMENTAL SETUP

Models and tiers. A *tier* is one (model, task) pair; there are eleven. Four Qwen2.5-Instruct scales (1.5B, 3B, 7B, 14B; Qwen Team, 2024) run both tasks, the eight Qwen tiers that §5.4 pools; Llama-3.1-8B-Instruct (Grattafiori et al., 2024) runs both as a cross-architecture control, on fewer items and so at low resolution (large MRSGs); GPT-4o-mini (OpenAI, 2024a;b) is audited on MATH-500 only (the closed tier: it exposes only temperature and top-p, and "greedy" is a non-deterministic temperature-0 request). Open-weight models run under llama.cpp (Gerganov et al., 2023) at 4-bit (Q4_K_M), with a 325-item prefix of the 1.5B GSM8K tier also at 8-bit as a quantization control. With greedy there are 153 method pairs, 101 between families (61 without min-p). The open-weight sampler chain is temperature-first, as in vLLM (Kwon et al., 2023), which the min-p evaluation used (truncation-first in App. F).

Configuration pool. $K{=}24$ per family, drawn once from each family's prior, with T drawn independently of τ (e.g. min-p: $p \sim \log U[0.01, 0.30]$, $T \sim U[0.7, 3.0]$, the temperatures of its evaluation's main results; App. B), plus greedy and three convention configurations (top-p 0.9/0.95 at $T{=}1$, top-k 40 at $T{=}0.7$) for untuned-baseline practice: 124 per tier (51 on the closed tier), identical across models and tasks, one shared draw seed.

Tasks. GSM8K (Cobbe et al., 2021) (150–1306 items per configuration; at 1.5B the balanced panel of the complete test set) and the harder MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024) (128–325, the whole numeric-answer file where extended; per-tier counts in Table 3), scored by numeric match within $\max(10^{-2}, 10^{-4}|\text{gold}|)$; a 30-prompt creative-writing leg scored 1–10 by an LLM judge at temperature 0 (App. E); and a *flat-landscape* leg on five tiers: 17 seed-replicates of one configuration (top-p 0.9 at $T{=}1$) differing only in generation seed, so the true landscape is flat by construction (App. C). Despite per-record seeds, batching non-determinism (Yuan et al., 2025) moved 12% of scores in the partly regenerated 7B MATH-500 panel (App. C), noise that enters σ. The study comprises 1,218,657 generations (Table 19; terms in Table 2).

Estimation. Every audit analysis uses the *balanced* panel, the items every configuration in a tier was scored on (1306 of 1319 on the full GSM8K tier; for an added arm, those it shares with the pools it is compared with), and every split puts $\lfloor n/2 \rfloor$ items in dev and (up to one more) in test, the protocol's dev size ($n/2$ below). σ and ρ are estimated per pool from per-item outcome matrices (App. C). Budget curves and measured $B(K)$ average Prop. 4 over 400 random splits (2,000 on the replicate legs); gaps, intervals, z, MRSG and equivalence bounds come from 600 paired bootstrap replicates that resample items within disjoint dev and test halves and re-run selection; permutation p-values from 20,000 permutations (2,000 for the added arms, 1,000 in the injection control).

5 RESULTS

5.1 THE THEORY HOLDS WHERE IT MUST

Prop. 1: the inflation formula. Measured $B(K)$ is the dev-max minus the held-out score of what it selects (Prop. 4), so no parameter is fitted to $B(K)$ itself. On the five replicate legs, whose true landscape is flat, it runs at 0.89–1.10× Eq. (3) at $K{=}2$ and 0.81–1.07× at $K{=}17$ (Table 11, App.

Table 1: **Headline comparisons, against top-p unless named.** Held-out gap in accuracy points ($K{=}24$ per arm; greedy and the untuned default $K{=}1$), pooled over the Qwen tiers an arm covers; *aud.*: the audit min-p arm on the same tiers and items (audit order for truncation-first; llama.cpp for the second stack); *neg.*: tiers with a negative gap; single-tier rows give the permutation p. Sign tests treat tiers as independent and are conditional on the one configuration draw (five tiers: smallest p .0625); the empirical cluster p resamples the drawn configurations and items, not the prior draw (§5.4). The restricted (App. C) and added arms (App. F) are exploratory, their p uncorrected; per tier: Tables 3, 13, 16 and App. F (greedy: bagged gaps to the default).

Family / arm	source	tiers	gap	aud.	neg.	test
temperature	audit	8	+0.1	–	3/8	sign $p = 0.73$
top-k	audit	8	−0.5	–	6/8	sign $p = 0.29$
typical	audit	8	+0.3	–	3/8	sign $p = 0.73$
min-p	audit	8	−2.1	–	8/8	sign $p = 0.008$; cluster $p = 0.17$
min-p, $T \leq 1.5$ redraw	restricted arm	8	−0.1	−2.3	4/8	sign $p = 1.00$
min-p, p paired to T	added arm	8	+0.0	−2.3	4/8	sign $p = 1.00$
min-p, truncation first	added arm	5	−1.0	−2.9	3/5	sign $p = 1.00$
min-p, second stack	added arm	1	−5.8	−4.2	–	perm. $p \leq .0005$
top-$n\sigma$, min-p's T range	added family	5	−2.1	−2.9	5/5	sign $p = 0.0625$
greedy vs. untuned top-p 0.95	audit	8	+2.3	–	2/8	sign $p = 0.29$

Fig. 3). On the audit pools, where landscapes are not flat, Eq. (3) was an upper envelope: above measured $B(24)$ in all 11 tiers (tier averages; in 43 of 52 family pools), at 1.10–2.10× (Table 12). Fitted on one half of a tier's items and measured on the other, it sits at 1.02–1.67× measurement (per-tier medians), at or above it in at least 54% of each tier's splits (App. C): approximately conservative out of sample, not a bound. The $\sqrt{1-\rho}$ correction matters for tightness: at the closed tier's $\rho = 0.70$ the uncorrected formula is 1.8× the corrected one.

Prop. 2: identical methods, unequal budgets. A direct test splits each flat leg's replicates into two disjoint pseudo-methods, identically distributed generators, at budgets $K{=}6$ versus $K{=}2$ (App. Fig. 4). Measured mean dev-max gaps sit at 0.86–1.07× Eq. (4) (Table 11); the higher-budget pseudo-method "wins" in 75/74/75/74/76% of resamples, a pipeline check, since exchangeability guarantees 75%. That is 1.7 to 2.8 points of artifact on dev halves of 150–156-item legs from a mild budget asymmetry.

5.2 The equal-budget audit: four families flat, min-p behind under our prior

Table 1 collects the headline comparisons; per-tier leaderboards are in App. Fig. 5 and Table 3.

Four families are indistinguishable, and greedy is never beaten past the MRSG. Of the 153 method pairs, 18 have a held-out gap larger than their MRSG (Table 8): each involves greedy (12, always ahead) or min-p (ten, always behind; four both), none between two of the other four families. The permutation test rejects exchangeability in three of the 61 family pairs without min-p (chance predicts 3.1) but in 15 of the 40 with it. Of the 8 rejections surviving Benjamini–Hochberg at FDR 0.05 (Benjamini & Hochberg, 1995), 7 are min-p comparisons and the other, temperature against typical on Llama MATH-500, recurs on Llama GSM8K (chance at the test's nominal size there, App. C, or a Llama-specific difference); all four surviving Benjamini–Yekutieli (Benjamini & Yekutieli, 2001), valid under any dependence, are min-p deficits (Table 9). The repeated Llama pair coincides with temperature's median selected T of 0.47 on both Llama tiers, below typical's $T{=}0.7$ floor (Table 18): a difference of prior support at least as much as of rule. On the Qwen tiers the four agree to within the pooled bound of §5.4, and their ceilings (oracles over each family's 24 draws) do not differ detectably ($|z| \leq 1.27$; App. C). Dev selection cools temperature and top-k (median T over all tiers 0.52, 0.64) but runs top-p and typical near $T{=}1$ (0.98, 0.97; Table 18). Where truncation acts the rules may differ (one tier, 1.5B GSM8K; exploratory). Pinned at $T{=}1$ the four truncation families separate only through top-k, which trails the other three by .030–.039 (three of six pairs at $p < 0.05$), while the other three lie within .009 of each other, with min-p nominally first; searched

over min-p's range, $T \sim U[0.7, 3.0]$ (which also raises their floors), the four rivals separate in three of six pairs, two past their MRSGs (App. F). *Greedy decoding*, the default of reasoning evaluations (Sprague et al., 2025), where it generally beats sampling (Song et al., 2025), is never beaten past the MRSG; the one-sided 95% bound on any family's pooled lead over it is .014.

The min-p exception moves with its prior. min-p trails top-p at permutation $p < 0.05$ in six of the ten tiers that express both, two after Benjamini–Hochberg (two of the six are the small Llama tiers). Under its audit prior 17 of its 24 draws lie above T=1.5, beyond the rivals' caps (App. B). Redrawn at matched budget with $T \leq 1.5$, against the unchanged top-p pool in all eight Qwen tiers, its pooled deficit on the shared items moves from $-.023$ to $-.001$, negative in four tiers instead of eight (Table 13). Drawn over the full range with p paired to T following its guidelines (App. B; second build), it pools to .000 on eight tiers, negative in four, and beats the audit arm in all eight (sign test $p = 0.008$; cross-build). Neither redrawn arm excludes a deficit as large as the audit arm's: their equivalence bounds are $\pm.028$ and $\pm.028$. Resampling the draw all tiers share, with items fixed (App. C), keeps the audit arm's pooled gap in $[-.032, -.017]$ and the restricted arm's in $[-.008, .005]$ (90%). On 1.5B GSM8K, redrawn outright, min-p trails top-p in four of five draws of its audit prior and ties in one (to within half a point), and in two of five restricted ones (Table 17). How much the range carries depends on the control: at a matched budget of seven in-range draws min-p still trails ($-.018$, against $-.027$ unrestricted; averaged over top-p's 7-subsets, App. C), and search depth and those seven draws cannot be separated. Given min-p's range too (one tier; floors raised to 0.7), typical still leads it by .068 ($p = 0.003$) and top-p by .035 ($p = 0.060$) within the second build.

The controls (App. F; exploratory, uncorrected across arms). On HuggingFace transformers (Wolf et al., 2020) at 16-bit the sign replicates ($-.058$, against $-.042$ under llama.cpp; one tier, not past its MRSG; there top-p's optimum sits below min-p's T=0.7 floor). Rerun truncation-first with the audit's own configurations and seeds (second build), its pooled deficit over five tiers, among them the three largest audit deficits, falls from $-.029$ to $-.010$, partly by regression. Top-$n\sigma$ (Tang et al., 2025), also searched up to T=3.0 (second build), trails top-p at permutation $p < 0.05$ at 1.5B and 3B MATH-500 (the latter where min-p trails most) and is not resolvably behind elsewhere. η and ϵ trail top-p on the second stack ($-.047$, $-.050$) until searched over $T \in [0.3, 1.5]$; typical, whose prior also starts at T=0.7, ties it.

5.3 The same raw data, three analyses

Table 10 (five tiers in Fig. 1b) re-analyzes the identical records in the published style: each family's dev-max over its 24 configurations, with no split, against the untuned default top-p 0.95 at T=1. Averaged over families it shows a gain in every tier, $+1.4$ to $+8.8$ points (a median $+1.2$ against greedy instead; Table 10), and only one family–tier cell fails to show a win, min-p's at 14B MATH-500 (Table 10). Eq. (5) says where the gain comes from. On held-out items at the protocol's dev size, selection bias $B(24)$ is a median 78% of the expected dev-max gain (35–146% by tier, above 100% where the tuned families fall below the default; Table 14) and, with $B(24)$ rescaled to each realized gain's dev size, about 69% of those gains, an approximation that errs high (28% on the near-complete 1.5B GSM8K test set). The remaining tuning term is below one point in five of 11 tiers and negative in two; per family the held-out tuned-vs-untuned gap clears its MRSG in 5 of 52 cells and is significant in none (Table 7). Magnitudes shrink with item count on the one tier tested ($+3.5$ on the near-complete 1.5B GSM8K test set, $+7.6$ on its 150-item prefix). Applied to the min-p evaluation's own GSM8K table, the check of §3.2 leaves its $T \leq 1$ differences of undetermined significance (the larger at $p = .029$ nominally, .12 corrected) and finds its T=1.5 gain about 14 SEs out, a real robustness effect (App. C).

5.4 Pooling the ladder: what eight tiers say that one cannot

Averaging the per-tier bootstrap gaps replicate-wise gives the pooled sampling distribution (Table 6) and lifts one run's median power to call a true five-point gap *significant* from .24 per Qwen tier (Table 4) to .88 (at the pooled min-p–top-p SE). Temperature, top-k and typical pool to within 0.50 accuracy points of top-p, with signs at chance (5–6 of 8, sign-test $p \geq 0.29$). Because non-rejection is not equivalence, we bound it: two one-sided tests at α=0.05 (Schuirmann, 1987), on the wider of the naive SE and the cluster-bootstrap SE described below, reject a pooled difference of .031 or more for each of the three against top-p and .035 or more for all six pairs among the four (per pair in

Table 6). That bound lies inside the interquartile range of the published-style gains (Table 10; App. Fig. 2b) and says nothing about gaps of one or two points.

min-p trails top-p in *all eight* Qwen tiers (pooled $-.021$ against a pooled MRSG of .0135; exact sign test $p = 0.008$ with tiers as units; App. Fig. 2a). The tiers share one configuration draw and one item set per task, so a two-way cluster bootstrap (600 replicates) resamples the drawn configurations and the items jointly across tiers (not the prior draw), re-running selection: the deficit reproduces at $-.022$ (cluster SE .0158 against the naive .0160, $z = -1.41$; calibration in App. C). The claim is thus a consistent sign, not a significant size by our conservative yardstick (at the bagged estimate's own spread plus the draw SD it would sit near $z = -3.1$; App. C): against the cluster-bootstrap null the pooled gap has empirical two-sided $p = 0.17$, and flipping the weakest tier would raise the sign-test p to 0.07.

Can the protocol detect anything? With every temperature configuration raised by a planted δ (a uniform shift; App. C), the permutation test detects δ=.03 in a median 78% of injections per tier on the audit's fixed items but 35% when the items are redrawn too, and δ=.05 in 100% and 85%. Raising only the six best temperature configurations, a difference at the optimum, gives 43% (δ=.03) and 88% (δ=.05) per tier on fresh items. Greedy leads the untuned default top-p 0.95 by .023 pooled (2.2× its pooled MRSG; $z = 1.86$ on the tuned-arm yardstick, but $z = 3.1$ paired over items, as befits two fixed configurations).

6 Discussion and limitations

What is new. The protocol's elements exist in prior work (Cawley & Talbot, 2010; Melis et al., 2018; Dodge et al., 2019; Tang et al., 2020; Yauney et al., 2026); Schaeffer et al. (2025) already compared samplers at equal sweep sizes (without a held-out split); and $\sigma\sqrt{1-\rho}\, e_K$ is the expected maximum of equicorrelated Gaussians (David & Nagaraja, 2003). New are its application to paired decoding evaluation, its validation on seed-replicate legs, and a held-out, equal-budget multi-family audit with resolution floors. The added arms reuse the audit's items after its results were seen (adaptive reuse; Dwork et al., 2015) and are exploratory.

Interfaces and judges. None of the three closed vendors whose models we probed documents min-p or typical; through the API endpoint we used, top-k and min-p fields were accepted without observable effect (top-p passed through; App. C). Our GPT-4o-mini judge at temperature 0 disagrees with itself on 9.2% of identical calls, noise that selection harvests.

Scope. The audit is math only (typical was proposed for story generation and summarization); on open-ended generation our 30-prompt judged leg resolves nothing (App. E). Differences may also be resolvable on diversity-sensitive metrics, under majority voting or pass@k (Wang et al., 2023; Chen et al., 2021), at open-weight scales beyond 14B, and for methods needing a second model or intermediate-layer logits (Li et al., 2023; Chuang et al., 2024); the min-p evaluation's self-consistency and diversity analyses lie outside this audit. The MATH-500 tiers use its numeric-answer items. The open-weight tiers run on llama.cpp at 4-bit, with an 8-bit control (where every permutation rejection is a min-p comparison) and a 16-bit second stack on one tier each. The added arms ran on a second llama.cpp build; regenerating the audit's top-p there moves its tuned score by $-.015$ to .003 across the eight Qwen tiers, the drift that cross-build comparisons carry (App. F). The priors are partly ours (App. B), and the conclusions are conditional on them. Each family's pool is a single draw from its prior; by our draw bootstrap (approximate at K=N) the draw contributes a pooled SD of at most .005 to a pooled gap (per tier up to .021; across genuine redraws on one tier up to .026; Apps. C, F). Unparseable outputs score as failures; scoring them as missing moves the pooled min-p deficit by 3% of its size (App. C).

What should a sampler paper report? The budget K and the space searched, with its sources; disjoint selection and reporting items; the MRSG beside every claimed gap, for a bagged estimate (for one unbagged run, the 95% criterion); the item-conditional permutation p; and raw per-item outcomes. Equal budgets alone do not make a comparison fair: where a prior spends its budget can move a family's standing (§5.2). Nothing in §3 is specific to decoding: any evaluation that reports the configuration it selected on the same benchmark inherits the arithmetic and the remedy. Under such a protocol, on the Qwen tiers, four of the five audited families tie within .035 at their tuned optima; smaller differences are not resolved.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing, editing, and debugging code, including code used to implement our methods and experiments. We did not use generative AI tools for any of the other tasks that require disclosure: in particular, we did not use them to generate synthetic data, to develop theoretical models or conceptual frameworks, to formulate mathematical claims or assist in proofs, to propose or refine hypotheses, to design the research methodology or experiments, to assist with translation, to clean or reformat datasets, or to support data analysis or the interpretation of results. Additionally, we used generative AI tools to edit the manuscript to improve readability and clarity, and to obtain simulated peer-review-style feedback on drafts, which informed revisions to the writing and presentation. All AI-assisted code was reviewed and tested for correctness by the authors, and all AI-suggested edits and feedback were reviewed and adopted at the authors' discretion; the scientific content, claims, and results are entirely the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work audits evaluation methodology using public benchmarks (GSM8K and the MATH-500 subset of MATH, MIT-licensed), open-weight models under their licenses (Qwen2.5-Instruct under Apache-2.0, its 3B model under the Qwen Research License; Llama-3.1-8B under the Llama 3.1 Community License), one proprietary model (GPT-4o-mini, also the creative-leg judge, which therefore scores its own stories on the 4o-mini tier), accessed via its API, and three further hosted models queried via their APIs only for the interface probes of App. C; the thirty story prompts are our own, and llama.cpp (MIT) and Transformers (Apache-2.0) are the inference stacks. No human subjects or personal data are involved. Our critique targets reporting practices, not individuals; the practices we audit are field-wide conventions, and min-p is our main case study because its gains are publicly disputed, and apart from one illustrative check of a published table (App. C) we replicate the practices on our own data rather than re-litigating any single paper's numbers. Potential negative use is limited; we believe transparent noise floors for evaluation reduce, rather than create, harm.

REPRODUCIBILITY STATEMENT

Every accuracy, inflation and test statistic in this paper is recomputable from the supplementary artifacts, and one script (`scripts/build_paper.sh`) rebuilds every table, figure and number computed from generation records without querying a model; the few external inputs (the published table of App. C, interface-probe and smoke-test outcomes, throughput and cost) are marked in the code. The artifacts are per-item generation records (jsonl, with per-record deterministic seeds, parsed answers and flags; the math records ship without the generated text for size), every added arm's records, a manifest of record counts and checksums, SHA-256 hashes of the Qwen weight files, the configuration pools with their generating seeds, analysis code implementing Props. 1–4 and the MRSG, and the serving configuration (llama.cpp server flags, request payloads with the sampler chain, prompts, token limits and answer extraction; App. C). Proofs are in App. A; search priors with citations in App. B; per-tier tables in App. D; judge protocol and its measured noise in App. E (prompt and rubric text in `code/judge.py`). Regenerating the records needs the models and the GPT-4o-mini API; the audit's llama.cpp commit was not recorded; regeneration changed 12% of item scores on the same build (one leg's duplicates), and only 79% agree on a second build on a fresh instance (Apps. C, F). An anonymized repository containing code and all per-item records accompanies the submission as supplementary material and will be made public upon acceptance.

REFERENCES

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on winners. *The Quarterly Journal of Economics*, 139(1):305–358, 2024.

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.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2021.

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)*, 2020.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Dami Choi, Christopher J. Shallue, Zachary Nado, Jaehoon Lee, Chris J. Maddison, and George E. Dahl. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446*, 2019.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

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.

A. Feder Cooper, Yucheng Lu, Jessica Zosa Forde, and Christopher De Sa. Hyperparameter optimization is deceiving us, and how to stop it. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Herbert A. David and Haikady N. Nagaraja. *Order Statistics*. Wiley, 3rd edition, 2003.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

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 (ACL)*, pp. 1383–1392, 2018.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

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)*, 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 (ICLR)*, 2024.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9198–9209, 2023.

Georgi Gerganov et al. llama.cpp: LLM inference in C/C++. `https://github.com/ggml-org/llama.cpp`, 2023.

Kyle Gorman and Steven Bedrick. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2786–2791, 2019.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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 *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 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*, 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. In *Conference on Language Modeling (COLM)*, 2025.

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.

Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3752–3762, 2019.

David D. Jensen and Paul R. Cohen. Multiple comparisons in induction algorithms. *Machine Learning*, 38(3):309–338, 2000.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. 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.

Liam Li and Ameet Talwalkar. Random search and reproducibility for neural architecture search. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 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)*, 2023.

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

Nicholas Lourie, Kyunghyun Cho, and He He. Show your work with confidence: Confidence bands for tuning curves. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024. arXiv:2311.09480.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? A large-scale study. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

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.

Clara Meister, Tiago Pimentel, Luca Malagutti, Ethan G. Wilcox, and Ryan Cotterell. On the efficacy of sampling adapters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023a. arXiv:2307.03749.

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

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1707.05589.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision (ECCV)*, 2020. arXiv:2003.08505.

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 (AACL-IJCNLP)*, 2020. arXiv:2009.07243.

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.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024a.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`, 2024b. Model `gpt-4o-mini`.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 338–348, 2017.

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*, 2025.

Robin M. Schmidt, Frank Schneider, and Philipp Hennig. Descending through a crowded valley – Benchmarking deep learning optimizers. In *International Conference on Machine Learning (ICML)*, 2021.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

D. Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. Winner's curse? On pace, progress, and empirical rigor. In *ICLR Workshop Track*, 2018.

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)*, 2024.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The leaderboard illusion. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

Prabhu Teja Sivaprasad, Florian Mai, Thijs Vogels, Martin Jaggi, and François Fleuret. Optimizer benchmarking needs to account for hyperparameter tuning. In *International Conference on Machine Learning (ICML)*, 2020.

David Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal*, 41 (2):463–501, 1962.

James E. Smith and Robert L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

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 (Volume 1: Long Papers)*, 2025.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. In *International Conference on Learning Representations (ICLR)*, 2025.

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 (Volume 1: Long Papers)*, pp. 10758–10774, 2025.

Raphael Tang, Jaejun Lee, Ji Xin, Xinyu Liu, Yaoliang Yu, and Jimmy Lin. Showing your work doesn't always work. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Ryan J. Tibshirani and Robert Tibshirani. A bias correction for the minimum error rate in cross-validation. *The Annals of Applied Statistics*, 3(2):822–829, 2009.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006.

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

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

Gregory Yauney, Shahzaib Saqib Warraich, and Swabha Swayamdipta. How reliable is language model micro-benchmarking? In *International Conference on Learning Representations (ICLR)*, 2026.

Jiayi Yuan, Hao Li, Xinheng Ding, Wenya Xie, Yu-Jhe Li, Wentian Zhao, Kun Wan, Jing Shi, Xia Hu, and Zirui Liu. Understanding and mitigating numerical sources of nondeterminism in LLM inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.09501.

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.

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.

Sebastian Zöllner and Jonathan K. Pritchard. Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *The American Journal of Human Genetics*, 80(4):605–615, 2007.

A Proofs

A.1 Proposition 1

Write $M_K = \mu(\theta_{\hat{k}}) + \varepsilon(\theta_{\hat{k}})$; taking expectations, $B(K) = \mathbb{E}[\varepsilon(\theta_{\hat{k}})]$. Since $\theta_{\hat{k}}$ is one of the K draws, $\varepsilon(\theta_{\hat{k}}) \le \max_k \varepsilon(\theta_k)$ pointwise, for any landscape. Decompose $\varepsilon(\theta_k) = \bar{a} + \bar{u}_k$ (Eq. 1): $\bar{a}$ is common to all k, so $\max_k \varepsilon(\theta_k) = \bar{a} + \max_k \bar{u}_k$ and $\mathbb{E}[\max_k \varepsilon(\theta_k)] = \mathbb{E}[\max_k \bar{u}_k]$. The residuals $\bar{u}_k$ are mean-zero with variance $(1-\rho)\sigma^2$. Two routes bound their maximum. Distribution-free: each $\varepsilon(\theta_k)$ is $\frac{1}{4n}$-sub-Gaussian (§3.1) and $\mathbb{E}[\max_k \varepsilon(\theta_k)] = \mathbb{E}[\max_k \bar{u}_k]$, so the maximal inequality, which needs no independence, gives $\mathbb{E}[\max_k \bar{u}_k] \le \sqrt{2\ln K/(4n)} = \sqrt{\ln K/(2n)}$; since $(1-\rho)\sigma^2 \le \sigma^2 \le 1/(4n)$, this is never tighter than (2). Under the Gaussian approximation of (1) with $\bar{u}_k \sim \mathcal{N}(0, (1-\rho)\sigma^2)$ independent across k, $\mathbb{E}[\max_k \bar{u}_k] = \sigma\sqrt{1-\rho}\, e_K$ by definition of e_K. With unequal residual variances $(1-\rho_k)\sigma_k^2$ the sub-Gaussian maximal inequality gives $\max_k \sigma_k\sqrt{1-\rho_k}\,\sqrt{2\ln K}$ (given sub-Gaussian proxies equal to those variances), not the plug-in of a pool-average σ, which can fall below $\mathbb{E}[\max_k \bar{u}_k]$ (four residuals of variance $6\tau^2$ and twenty of zero, pool-average variance τ^2, give an expected maximum near 2.56τ, above even $\tau\sqrt{2\ln 24} \approx 2.52\tau$); the bound on real pools is therefore checked, not assumed. With equal variances and non-negative correlations, Slepian's inequality (Slepian, 1962) makes the maximum stochastically smaller than in the independent case, so (3) is an upper envelope; for equicorrelation r it is exactly $\sigma\sqrt{1-\rho}\sqrt{1-r}\, e_K$ with ρ defined against an item effect external to the pool (App. C). On a flat landscape the selection rule $\arg\max_k(\mu + \bar{a} + \bar{u}_k) = \arg\max_k \bar{u}_k$ coincides with the noise argmax, so the first inequality is an equality for any noise distribution, and in the Gaussian case $B_{\text{flat}}(K) = \sigma\sqrt{1-\rho}\, e_K$. Nonnegativity needs neither flatness nor independence: let $k^* = \arg\max_k \mu(\theta_k)$, which depends on the draws only. Since $\hat{S}_{\text{dev}}(\theta_{\hat{k}}) \ge \hat{S}_{\text{dev}}(\theta_{k^*})$ and the dev noise has mean zero given the draws, $\mathbb{E}[M_K] \ge \mathbb{E}[\mu(\theta_{k^*})] = \mathbb{E}[\max_k \mu(\theta_k)] \ge \mathbb{E}[\mu(\theta_{\hat{k}})]$, so $B(K)$ is at least the expected regret of the selection. □

A.2 Proposition 2

With identical flat landscapes and shared dev items the μ terms cancel, and $M_{K_A} - M_{K_B} = (\bar{a} + \max_{k \le K_A} \bar{u}_k^A) - (\bar{a} + \max_{k \le K_B} \bar{u}_k^B)$; the shared item effect cancels exactly, and by linearity the expected difference of the two maxima is $\sigma\sqrt{1-\rho}(e_{K_A} - e_{K_B})$, which for $\rho < 1$ is positive and increasing in K_A for fixed K_B since e_K is strictly increasing. For the win probability: in the continuous flat-Gaussian case $\Pr[M_{K_A} > M_{K_B}] = \Pr[\max_{k \le K_A} \bar{u}_k^A > \max_{k \le K_B} \bar{u}_k^B]$ and ties have probability zero. More generally, if the $K_A + K_B$ dev scores $\mu(\theta_k) + \bar{u}_k$ are exchangeable, which holds when both families draw i.i.d. from one prior with the same noise, and ties are broken uniformly at random among the tied configurations, not by a coin between the two families (accuracies on finitely many items are discrete, and ties are common), the position of the overall maximum after tie-breaking is uniform over the $K_A + K_B$ slots, so A holds the argmax with probability $K_A/(K_A + K_B) > 1/2$ whenever $K_A > K_B$, e.g. 0.75 at K_A=6, K_B=2. This part needs neither flatness nor Gaussianity. The strict event $M_{K_A} > M_{K_B}$ excludes cross-family ties at the maximum, so its probability is at most this, with equality when such ties have probability zero. The flat-leg check of §5.1 breaks ties at random among configurations, and its 74–76% win rates therefore check exchangeability, not σ or ρ. □

A.3 PROPOSITION 3

Add and subtract $\mathbb{E}[\mu(\hat{\theta}_A)]$: $\mathbb{E}[M_K^A - \hat{S}(\theta_b)] = (\mathbb{E}[\mu(\hat{\theta}_A)] - \mu(\theta_b)) + (\mathbb{E}[M_K^A] - \mathbb{E}[\mu(\hat{\theta}_A)])$, and the second term is $B_A(K)$ by definition. Since θ_b is fixed in advance, $\mathbb{E}[\hat{S}(\theta_b)] = \mu(\theta_b)$. Under R2, $\mathbb{E}[\hat{S}_{\text{test}}(\hat{\theta})] = \mathbb{E}[\mathbb{E}[\hat{S}_{\text{test}}(\hat{\theta}) \mid \hat{\theta}]] = \mathbb{E}[\mu(\hat{\theta})]$ since test items are drawn independently of the dev items and the draws, eliminating $B_A(K)$. The tuning term is nondecreasing in K under a condition, not unconditionally (on a flat landscape it is constant). The K pairs $(\mu(\theta_k), \hat{S}_{\text{dev}}(\theta_k))$ share $\bar{a}$, which does not move the argmax, so condition on $\bar{a}$: under joint Gaussianity with $\text{Cov}(a_i, u_i)=0$ and independent residuals, $\bar{u}_k$ is independent of $\bar{a}$, so given $\bar{a}$ the $s_k = \mu(\theta_k) + \bar{u}_k$ are i.i.d., the selected $s_{\hat{k}}$ under budget $K+1$ first-order stochastically dominates that under K, and $\mathbb{E}[\mu(\hat{\theta}_K) \mid \bar{a}] = \mathbb{E}[g(s_{\hat{k}}) \mid \bar{a}]$ with $g(s) = \mathbb{E}[\mu \mid \mu + \bar{u} = s]$, so the term is nondecreasing when g is. Residual noise with a log-concave density (e.g. Gaussian of common variance) independent of μ has monotone likelihood ratio, which makes g nondecreasing for any prior on μ. (The unconditional $\mathbb{E}[\mu \mid \hat{S}_{\text{dev}}{=}d]$ is the wrong object: given that θ_k is the maximum *and* $\hat{S}_{\text{dev}}{=}d$, the split of d between $\bar{a}$ and s_k differs from that for a single draw.) The term can fall: with three configuration types, $\mu{=}.6$ with no residual (probability $\frac{1}{2}$) and $\mu{=}.5$ with residual $\pm.5$ ($\frac{1}{4}$ each), $\mathbb{E}[\mu(\hat{\theta}_K)] = .5 + .1(.75^K - .25^K)$, which is .55, .55, .541 at $K{=}1, 2, 3$ and tends to .5 (a stylized example; scaling both the μ gap and the residual by the same factor keeps its shape at realistic n). For binary accuracy the residual variance varies with μ, so in the audit monotonicity is an assumption, not a consequence. Positive correlation of dev and true scores is not enough: under heteroskedastic noise, selection can favour high-variance configurations and the term can fall. The 19 of 52 family–tier cells summarized in Table 7 (37%; listed per cell in the supplement) where a family's tuned selection scores below the untuned baseline are levels at $K{=}24$ against a fixed top-p convention configuration, not a slope in K, so they mix a within-family tuning effect with a cross-family difference and are consistent with this reading rather than a demonstration of it. □

A.4 PROPOSITION 4

Condition on one split, which fixes a dev and a test score for each of the N pool configurations and hence a dev ranking (ties broken uniformly). A budget-K tuning run selects a uniformly random K-subset of the pool, and its dev-max is the subset's highest-ranked member. The probability that the rank-r configuration (ascending) is the maximum of a uniform K-subset is $\binom{r-1}{K-1}/\binom{N}{K}$: the subset must contain rank r and choose its remaining $K-1$ members among the $r-1$ lower ranks. Weighting dev scores by $w_r(K)$ therefore gives the exact conditional expectation of the dev-max, and weighting the corresponding test scores gives the exact conditional expectation of the selected configuration's test score. The dev half is a sample of $n/2$ items from the task's population, taken as a superpopulation (for accuracy on the finite benchmark itself the test half is the complement of the dev half, and the second curve is biased low by $B(K)/2$), so the first is unbiased for the expected dev-max at that dev size; the test half is independent of the dev half and hence of the ranking, so the second is unbiased for the selected configuration's expected true score. Averaging over random splits keeps both unbiased. The with-replacement plug-in of Dodge et al. (2019) is biased for the same functional, as shown by Tang et al. (2020), whose dev-side weighting we reuse; the test-of-selected extension is immediate from the same conditioning argument. Scoring the selection instead against each configuration's mean over the same items the dev sample is drawn from is *not* unbiased: selection then also picks up the item noise in those means, and a with-replacement dev sample carries that noise twice. With equal disjoint halves the pool mean is $(\text{dev} + \text{test})/2$, so the gap is exactly $B(K)/2$ on any landscape (for n odd, $\lceil n/2 \rceil / n$ of it). With a with-replacement dev sample of n_{dev} from the n items, under the Gaussian model on a flat landscape the curve gap is $1/\sqrt{1 + n_{\text{dev}}/n}$ of $B(K)$ at a fresh dev sample of size n_{dev} ($1/(1 + n_{\text{dev}}/n)$ of the with-replacement design's own dev-max inflation), 0.82 at $n_{\text{dev}}{=}n/2$, and the synthetic check of App. C measures 0.80–0.81. Because pool members are i.i.d. draws from P_s, every K-subset is K i.i.d. draws, so the estimator is also unbiased for budget-K random search from the prior itself. □

B SEARCH PRIORS

Each family's truncation-parameter range is built around values recommended or used for that family in the literature, with our extensions marked below. Temperature is drawn jointly with it, as the min-p evaluation varies both; drawing the two independently is our simplification for every family, and so are the densities and the temperature ranges of families whose proponents did not vary temperature. Temperature: $T \sim U[0.3, 1.5]$, overlapping the sweep of Renze (2024), whose null result is scoped to 0.0–1.0 and who report accuracy degrading above it; the prior deliberately extends past that flat region; both bounds are ours. top-k: $k \sim \log U[2, 200]$ (rounded) with $T \sim U[0.5, 1.5]$; k=10 in Fan et al. (2018); both bounds are ours. top-p: $p \sim U[0.70, 0.99]$ with $T \sim U[0.5, 1.5]$; Holtzman et al. (2020) report p usually in $[0.9, 1)$; the range below 0.9 is ours, and it is where dev selection mostly lands (median selected p below 0.9 in every tier, Table 18). min-p: $p \sim \log U[0.01, 0.30]$ with $T \sim U[0.7, 3.0]$; core values $\{0.05, 0.1\}$, with 0.2–0.3 also tested and higher values (0.5–0.7, "or even close to 1") advised at very high temperatures (Nguyen et al., 2025); the range below 0.05 is ours, and drawing p independently of T with a cap of 0.3 and a log-uniform density is our simplification, which pairs most hot draws with a low p (the prior puts about half its mass below 0.05) and none with the advised high one. The arm of App. F that follows the guidelines instead keeps $T \sim U[0.7, 3.0]$ and pairs p with it: $p \sim \log U[0.05, 0.10]$ for $T \leq 2.0$ and $p \sim U[0.5, 0.7]$ above (the guidelines give no number for "very high"; $T > 2$ is our reading, and the cap at 0.7 is ours). Typical: $\tau \sim U[0.20, 0.95]$ with $T \sim U[0.7, 1.3]$, an interpolation of ours spanning the strongly task-bimodal recommendations (0.2 story generation, 0.95 summarization) of Meister et al. (2023b), neither of them for math. The temperature ranges of top-k, top-p, typical, η and ϵ are our choice, since their proponents did not vary T, and the floors differ (0.3 for the temperature family, 0.5 for top-k and top-p, 0.7 for typical, η, ϵ, min-p and top-$n\sigma$); on math, where no family resolvably beats greedy, a lower floor may help, which App. F probes on one tier by searching η and ϵ over $U[0.3, 1.5]$ (a probe that lowers the floor and raises the cap together). The floor asymmetry works against min-p too, but it is not what produces the audit arm's deficit: with its temperatures capped at 1.5 and its 0.7 floor kept, the redrawn arm's point estimate is already $-.001$ pooled (resolved neither from zero nor from the audit arm's deficit), and since selection cools T for the other families on math, a lower floor would more likely help min-p than hurt it. For min-p, the uniform density on $[0.7, 3.0]$ is also ours, and the share of draws above T=1.5 depends on it (expected 15.7 of 24; the realized draw has 17), and $[0.7, 3.0]$ is the span its evaluation's main results tested (further tests reach 0–5), not an operating range its authors recommend; we treat it as the prior. Every truncation family is therefore a two-dimensional draw, so among them equal K is equal budget over equal search *dimension* (temperature alone is one-dimensional), but not over equal *density* where a family can compete, which is the asymmetry the redraw of §5.2 (Table 13) corrects and which these ranges create: every draw of all four rivals sits at $T \leq 1.5$, against 7 of min-p's 24, because min-p's evaluated range reaches T=3.0 and the rivals' priors, by our choice, stop at or below 1.5. The min-p evaluation itself ran its rivals across the same temperatures (Nguyen et al., 2025), so App. F adds the symmetric control on one tier: each rival keeps its own truncation prior and draws $T \sim U[0.7, 3.0]$, min-p's range, on its own RNG stream. A prior is a claim about where a method works. The four families that come out flat do so under priors that contain their proponents' recommended values where such values exist and extend them, which is as far as the audit's fairness to them goes; and it is why the redrawn and guideline-paired arms, not the audit arm alone, are what the deficit's interpretation rests on. Arms added after the audit (App. F) use fresh RNG seeds and the audit's priors unless stated, and the three families beyond the audit's five use: top-$n\sigma$ (Tang et al., 2025): $n \sim U[0.5, 2.0]$, an interval around one standard deviation of the logits, with $T \sim U[0.7, 3.0]$, min-p's range, since both are promoted as robust at high temperature; η and ϵ (Hewitt et al., 2022): $\eta \sim \log U[2 \times 10^{-4}, 4 \times 10^{-3}]$ and $\epsilon \sim \log U[10^{-4}, 10^{-3}]$, spanning the values of their evaluation, with $T \sim U[0.7, 1.3]$. Convention configurations: top-p 0.9 and 0.95 at T=1, the top-p thresholds of the min-p evaluation's baselines (Nguyen et al., 2025), which it ran at every temperature and we fix at T=1 as a convention; top-k 40 at T=0.7, a setting Holtzman et al. (2020) evaluate, with k=40 also llama.cpp's default. One draw seed (20260803) generates the pool once, family by family from one RNG stream; the identical 124 configurations are used for every model and task, of which the closed tier can express 51.

C IMPLEMENTATION DETAILS

Sampler-chain semantics. llama.cpp applies its sampler chain in request order; its default places temperature *after* truncation, whereas HuggingFace and vLLM apply temperature first. For top-p, min-p and typical the two orders produce different distributions (truncation on the tempered vs. untempered distribution); for top-k and top-$n\sigma$, whose kept sets temperature does not change, they coincide. We pin `samplers` to temperature-first for every request, matching vLLM, which the min-p evaluation's main results used, and current HuggingFace transformers (the order has varied across transformers releases); the closed tier's order is not documented. Requests name only the audited family's samplers: a nominal no-op such as typical=1.0 still sorts the model's full vocabulary (152k tokens for Qwen) per token on the CPU (a measured $2.3\times$ end-to-end slowdown), so omission, which is distribution-identical, is used instead. The request's `samplers` list holds only temperature and the family's own sampler, so llama.cpp's default top-k, top-p and min-p are not applied.

Degenerate-output censoring, and how much of min-p's deficit it is. The chat endpoint returns HTTP 500 when a completion cannot be parsed into the chat template, which happens on the most degenerate generations, at high temperature in every family whose prior reaches it. Such responses are scored as failures and flagged (`degenerate_unparseable`); dropping them instead would delete hot configurations' worst outputs and inflate their scores. 3.3% of the eight audit panels' cells carry the flag: 7.7% of temperature cells, 8.3% of min-p's, 1.2% of top-p's and essentially none elsewhere. The flag is not concentrated on min-p: temperature carries nearly as much of it and pools to within its own standard error of top-p, so the flag alone neither explains nor excludes the deficit. The channel is bounded directly instead, by re-running the pooled comparison on identical items per tier under three valuations of a flagged cell. Scored as failures it is $-.0208$, negative in eight of eight ($p = 0.008$). Scored as *missing*, so that a configuration is judged on the items it did parse (what a regenerate-on-500 arm buys if regeneration behaved like that configuration's other items), it is $-.0214$, still eight of eight ($p = 0.008$), a change of 3%. Scored as *correct* it reverses to $.0546$, but that is a bound rather than a scenario: no regeneration policy beats treating every unparseable completion as a right answer, and an unparseable completion is not one.

Closed-API expressibility probes. On each closed endpoint we sent (i) top-$k = 1$, which forces greedy if honored (output remained stochastic); (ii) min-$p = 0.95$ at T=1.5, which forces near-argmax sampling if honored (output remained diverse). Both were accepted without error and neither changed the output, for all four backends tested (OpenAI, Anthropic, Google and one hosted open-weight model, probed 2026-08). The scope of that is the endpoint we used, not any vendor's documentation (Anthropic and Google document a native top-k on their own parameter sets, which these probes did not use), so the finding is a practitioner's risk, since an endpoint can accept min-p and apply nothing. The endpoint did pass top-p through, the one truncation knob the closed tier uses: before that tier ran, top-$p = 0.01$ at T=1.6 collapsed two different seeds to near-identical completions while top-$p = 1.0$ did not (a smoke test, not part of the released records). The released records agree: above T=1.3 the temperature family's completions lengthen to 387 tokens (371 at $T \leq 1$) and its accuracy falls to .627 (.671), while top-p configurations at the same temperatures stay at 372 tokens and .669 (`code/closed_topp_check.py`). What none of the three closed vendors (OpenAI, Anthropic, Google) documents in its API reference is min-p or typical, and that is what the main text's claim is about. Self-hosted vLLM and llama.cpp servers, and resellers that expose their native parameters, accept `min_p` and fall outside that claim.

Why the split is item-first. A bootstrap that resamples all items with replacement and then cuts the *result* into dev and test halves by position leaves the halves disjoint in position but not in items: an item drawn twice can land once in each half. Simulated at every item count in the study, 39% of test draws are items also present in that replicate's dev half, and the share does not fall with n. That violates Prop. 3's premise, that test items are independent of the selection event, inside the estimator built to measure what happens when it fails. Splitting the items first and resampling within each half restores it, and the difference is measurable: on 3B MATH-500 every held-out level is 1.7 accuracy points lower under the item-first split, the selected configuration otherwise carrying that much of the dev noise that chose it, while the between-family gap barely moves ($-.056 \to -.054$, the overlap being symmetric across two arms selected on the same half) and the paired standard deviation is 27% larger. Every held-out number in this paper uses the item-first split; the pooled z of §5.4 is correspondingly smaller than a position-split estimator would report.

Serving and scoring. Open-weight models (Q4_K_M GGUF files: the official Qwen2.5 releases, with SHA-256 hashes in the supplement, and a Q4_K_M quantization of Llama-3.1-8B-Instruct) are served on one 24GB consumer GPU (CUDA 13.1) by llama.cpp's `llama-server`, a CUDA source build of August 2026 for the audit and its $T \leq 1.5$ redraw (its commit was not recorded) and release b10240 for the arms of App. F, with full GPU offload, 12–16 parallel slots of 1,280 context tokens each, and (for those arms) the host prompt cache capped at 256 MB; every request goes to the chat endpoint with a two-shot prompt that asks for brief work ending in "The answer is N" (`code/tasks.py`), stops at a new "Q:" or at 400 new tokens on GSM8K and 900 on MATH-500 (at 600, 16.7% of 1.5B MATH-500 generations truncated; at these caps min-p's draws above T=1.5 reach the cap in 6.6–25.7% of generations per tier, every family's draws at or below 1.5, greedy included, in at most 10.8%; `code/token_cap.py`; generation lengths show no pile-up below the cap, so the 1,280-token context does not bind), and names its sampler chain explicitly: temperature, then the family's one truncation sampler, with no repetition penalty; greedy is temperature 0. The prediction is the number after "answer is", else the last number in the completion, and it scores as correct within $\max(10^{-2}, 10^{-4}|\text{gold}|)$ of the gold answer. GPT-4o-mini is the `gpt-4o-mini` model accessed via its API in August 2026, with the same prompt and scorer.

Determinism and data hygiene. Every generation carries a deterministic seed derived from (configuration, item, replicate) via CRC32. That fixes the sampler's draw given a probability vector, and it is less than full reproducibility: 15,760 cells generated twice from the same triples under identical seeds and settings disagreed on the *score* 12% of the time (all on the 7B MATH-500 leg; a duplicated cell enters the analysis as the mean of its scores, so a disagreeing pair counts 0.5, and the generation counts of Table 19 count records, duplicates included). Under continuous batching the probability vector depends on which other requests share the batch, since floating-point reductions are not associative, and no seed reaches that. The rate is even across the five families (12–13%), so it is dispersion the protocol already measures as part of σ rather than a confound for any comparison. Item sets are frozen prefix-stable pools: the frozen draw comes from a 250-item GSM8K pool and a 350-item MATH-500 pool filtered to numeric-answer items, and a tier extended past that draw appends the rest of the test set behind it in a fixed order. Item counts can therefore grow or shrink without orphaning a collected record. All released files are listed with record counts and checksums in the supplement's manifest; analysis runs from the released files only. Each flat leg's 17 replicates are the audit's own top-p 0.9, T=1 configuration (replicate 0) and 16 seed replicates; every seed-replicate series is checked against every model's audit-pool item profile (per-item correctness correlated with each pool's item means): the replicates of all five legs match their own model best (the 14B leg's in 16 of 16, $r = .74$); an earlier set of 2,400 14B replicate records, which did not match the 14B pool, was superseded by this leg and is excluded from every analysis and count.

Out-of-sample inflation. Estimating (σ, ρ) on a random half of a tier's items and measuring $B(24)$ on the other half (itself cross-fitted, dev and test quarters disjoint), over 12 splits per family and all eleven tiers, gives flat-envelope ratios with per-tier medians of 1.02–1.67, at or above measurement in at least 54% of each tier's splits. A second test transfers the closed form the way a practitioner without a sweep of their own would: keep a *source* tier's ρ (as this paper reports it), substitute the *target* tier's own σ, and read the envelope off Prop. 1 with nothing refitted on the target. Over all 490 ordered source–target pairs sharing a family, the median predicted/measured ratio is 1.29× (5–95%: 0.88–2.52), never below 0.71×, with 85% of pairs at or above measurement. The envelope therefore mostly survives being read off someone else's fit, which is the only form in which a reader of a published sweep can use it, but it can dip below measurement when the borrowed ρ is too high.

The retrospective check on a published table. The GSM8K column of the min-p evaluation's main table (Nguyen et al., 2025, Table 2, Mistral-7B, 8-shot CoT) shows, at each of five temperatures (0.7–3.0), one top-p setting (0.9) against the better of two min-p settings (0.05, 0.1). At GSM8K's 1,319 test items (the evaluation does not state its count) and accuracy near 0.3, $\sigma \approx .013$, so with ρ=0.52 a min-p row that shows the better of its two settings carries a spurious $\sigma\sqrt{1-\rho}(e_2 - e_1) \approx 0.5$ points by Eq. (4), and one cell's paired gap has a single-run SE near 1.2 points. The reported differences at T=0.7 (−0.9 points) and T=1.0 (+3.2) are then about 0.7 and 2.6 SEs, and the latter falls to about 2.2 once the asymmetry is removed; the differences at T=1.5 (+17.7, top-p at 0.7%) and T=2.0 (+6.2, top-p at 0) are about 14 and 5 SEs, far beyond anything selection produces (at T=3.0 both score 0). By the 95% criterion the T=1.0 difference is nominally significant ($p = .010$, or .029 without the asymmetry); corrected for the table's four temperatures at which the rows differ, $p = .039$ and .12. Every z rests on a ρ borrowed from our Qwen pools: the SE is $\hat{\sigma}\sqrt{2(1-\hat{\rho})}$

with $\hat{\sigma} = \sqrt{0.3 \cdot 0.7/n}$, and over our pools' range of $\hat{\rho}$ (.33–.70) the asymmetry-corrected $T{=}1.0$ difference is 1.8 to 2.9 SEs, so its significance is undetermined without per-item data. At $T{=}1.5$, with both accuracies far from 0.3, that SE is conservative. The check thus splits the published claims into a $T \leq 1$ part of undetermined significance, at temperatures where tuning puts every rival (their median selected T never exceeds 1.12; Table 18), and a real high-temperature robustness effect, which tuned comparisons of the rivals never reach (the symmetric control of App. F degrades temperature and top-k over that range, though it also raises their floors).

The redrawn restricted arm. The four cells of App. D (Table 15) cannot restrict min-p at $K{=}24$, because min-p's audit prior places only 7 of its 24 draws at $T \leq 1.5$; a budget-matched restricted arm has to be redrawn, not subsetted. 24 configurations from $\log U[0.01, 0.30] \times U[0.7, 1.5]$ with their own RNG stream and their own generations, in all eight Qwen tiers, against the *unchanged* top-p pool on the items both were scored on. Budget, items, reference arm and estimator are fixed, so the temperature range and the draw itself are the only differences and Prop. 3's budget confound does not arise.

The audit arm reproduces inside this harness ($-.023$ on the shared items against $-.021$ on each tier's full set, negative in eight of eight, sign test $p = 0.008$). The restricted arm does not: $-.001$, negative in only four of eight (sign test $p = 1.00$). The point estimate loses 95% of its magnitude, but magnitude was never the resolvable part here. What the restriction removes is the one feature of the deficit that *was* statistically notable: that all eight tiers pointed the same way. Read as a statistic in its own right the shift is $+.021$ (bootstrap SD .017), positive in eight of eight tiers (Table 13). Its two nulls agree: the exact sign test across tiers, as in §5.4, gives $p = 0.008$, and permuting all 48 min-p configurations between the arms (2000 permutations; unlike the 600-replicate gaps of §4, the shift and each permuted shift average the same 20,000 random splits) gives $p = 0.037$. Its magnitude clears its own MRSG by about the factor the deficit it explains clears its own ($1.5\times$ against $1.5\times$) and, like that deficit, is not significant by the $\mathrm{SE}_{\mathrm{total}}$ criterion (.021 against a bootstrap SD of .017), although its permutation test rejects.

Why restricting the range works: wasted prior volume. That the shift exists does not say what produced it, and two mechanisms fit. Either the hot tail is harmful, so dropping it raises what min-p can reach; or the hot tail is merely useless, so dropping it raises how much of the budget lands where min-p can compete. Pool truth, each configuration's accuracy on the shared items before any selection, separates them. Dropping the tail does *not* move min-p's ceiling: its best configuration is already one of the $seven$ draws at $T \leq 1.5$ in $seven$ of the eight tiers. What the tail mostly holds is budget. Of the 24 draws from min-p's audit prior, 17 sit above $T{=}1.5$, where the rivals' priors stop, and they drag its mean pool truth far below top-p's (.166 median gap in the means, against .011 once redrawn; at the worst tier its pool mean is .278, against .460 for its own low-T draws). No prior here is efficient: the share of a family's 24 draws landing within 0.03 of its own ceiling runs 25–54%, and min-p's 27% against top-p's 40% is a difference of degree. The tuned gap depends on how many draws are competitive; how badly the rest do (min-p's pool mean sits .192 below its own ceiling, against .033–.101 for the other four, the widest being temperature) drives the permutation test, which reshuffles those draws between arms, not the tuned gap. Redrawing all 24 inside the competitive region recovers a median 94% of min-p's ceiling gap to top-p. The budget curves of Prop. 4 show the same mechanism directly: pooled over the Qwen tiers, the audit arm's held-out gap to top-p is $-.151$ for a single random draw, $-.028$ at $K{=}6$ and $-.022$ at $K{=}24$, while the restricted arm's is $-.015$, $-.005$ and $-.001$. A deficit that shrinks as the budget reaches the competitive draws is what wasted prior volume predicts; a rule that is worse where it competes would keep its gap at every K.

Ceilings: does search efficiency hide a rule difference? If priors differ in efficiency by $3\times$, four families could reach indistinguishable tuned scores while their rules differ, the search gaps masking the rule gaps. The quantity that separates those readings is each family's pool-truth ceiling: the best its own 24 draws reach on the shared items, an oracle selector at the same budget, with no dev/test split. A ceiling is itself a maximum of 24 noisy means, so it is inflated by the same winner's curse, and by more for a family whose best configurations lie closer together, so we report beside it a cross-fitted ceiling that picks each family's best on one item half and scores it on the other. Across the eight Qwen tiers the four non-min-p families' ceilings are as flat as their tuned scores: pooled ceiling gaps of at most .010 (.009 cross-fitted), $|z| \leq 1.27$ on the wider of the naive and tier-cluster standard errors, and one tier-level comparison of the six pairs over eight tiers outside a 95% interval, against the 2.4 expected by chance. All four min-p pairs do separate ($|z|$ 2.5–3.7), every one against min-p: its audit prior is behind at its ceiling as well as at its optimum, which is what a prior that puts

17 of its draws above $T{=}1.5$ predicts, and why redrawing inside the competitive range recovers a median 94% of the ceiling gap, a ratio of two ceiling differences whose denominator runs from .006 upward (median .025), inside the item-noise scale of a single configuration's mean, so it is a direction with a magnitude this study cannot resolve. That last fact is also this comparison's limit: a ceiling is the best of 24 draws from a prior, so indistinguishable ceilings argue against the reading in which unequal search efficiency hides a rule difference, and say nothing against the reading in which a prior fails to cover its own rule's optimum. min-p is the case where that happened, and the only one the redraw corrects.

The same manipulation on a different architecture. Redrawn on Llama-3.1-8B at its own frozen items, min-p's deficit goes from $-.082$ to .015 on MATH-500, the tier where it was largest anywhere in the study and where the permutation test rejects (this harness's splits and seeds give the audit arm $-.082$ where the audit's give $-.076$ on the same items), and from $-.055$ to $-.053$ on GSM8K, where the restriction removes almost nothing. Neither shift is resolvable on its own ($p = 0.07$ and $p = 0.90$, at 128 and 150 items), and two tiers support no sign test: one tier moves as the Qwen tiers do and one does not, so this is consistent with the Qwen result, not a replication of it.

That is this paper's own Prop. 3 applied to a prior rather than to a protocol, and specifically its *legitimate* term: held-out reporting removes $B_A(K)$, the selection bias, but not the fact that, when Prop. 3's monotonicity condition holds, the best of 24 competitive draws beats the best of $seven$ in expectation. Eq. (4) is the dev-max version of the same asymmetry and does not apply here, because EBET reports test-of-selected. A recommended search space that spends most of its volume where its method cannot compete therefore hands that method a real disadvantage that no amount of held-out discipline will remove. On this evidence most of the deficit this audit measures is a property of our audit prior (min-p's evaluated range, sampled with p independent of T), and the residual $-.001$ sits where every other between-family difference in this study sits, below the floor. Not all of it: at a budget of seven, averaged over top-p's 7-subsets, the restriction closes only 33% of the gap, and whether search depth or the particular seven in-range draws account for the difference from $K{=}24$ this study cannot separate; nor does the symmetric control of App. F remove the deficit against typical.

Is the permutation test calibrated or merely conservative? Three rejections among the 61 comparisons that do not involve min-p is what an α of .05 delivers (3.1 expected), but matching the nominal rate is not by itself calibration and a shortfall would not by itself be an absent effect. Those 61 p-values are indistinguishable from uniform (mean 0.47, median 0.47, Kolmogorov–Smirnov $D = 0.10$, $p = 0.49$), which is what a calibrated test gives when all 61 nulls hold; a conservative test facing a few small true effects can look the same, so this is consistency with calibration, not proof of it, and it cannot show an effect absent (non-rejection bounds it only through the two one-sided tests of §5.4). (The comparisons share a configuration pool and an item set within each tier, so the Kolmogorov–Smirnov statistic's own independence assumption is only approximate here; it is reported as a description of the p-value distribution, not as a test.) The 40 involving min-p are nothing like uniform (mean 0.26, median 0.16, $D = 0.42$, $p < 10^{-6}$).

Quantization control. Truncation acts on the tail that 4-bit weights perturb, so holding quantization fixed leaves it confounded with the finding. We re-ran the 1.5B GSM8K pool at 8-bit (Q8_0) on a 325-item prefix and re-scored the 4-bit arm on those same items: both arms are paired item-wise on the prefix, not the audit's 1306, where the 4-bit min-p deficit against top-p is .022 ($p = .032$), not the prefix's .042. Mean accuracy over the pool rises 4.9 points at 8-bit. The permutation test rejects two pairs at 4-bit and three at 8-bit: temperature vs min-p is new there ($+.077$, $p = .009$), none lost. Every rejection at either precision is a min-p comparison, and among the six pairs of the other four the largest gap at either precision is .028 against floors of .045–.050 (each pair's gap is bagged on its own, so pairwise gaps are not exactly additive). min-p's deficit against top-p is wider at the higher precision (from .042, $p = .036$, to .047, $p = .001$). 4-bit weights do not manufacture the min-p exception. For the four-family flatness, neither precision separates any of those six pairs.

Resolution scaling on nested item sets. Eight tiers were re-run at up to $8.7\times$ their frozen item count with the configuration pool, the per-(configuration, item, replicate) seeds and the prompts unchanged, the new items appended behind the preserved prefix. The prefix is therefore an *exact nested subset* of the extended set, and the audit's own estimator can be run on both without refitting anything. Divided by the $\sqrt{n}$ ratio Eq. (1) predicts, σ scales at a median 1.00 (0.97–1.08): its item-count dependence holds on real records. $B(24)$ falls faster (1.07, 0.97–1.75), since ρ and the realized landscape move

with the item set, and the MRSG slightly slower (0.99, 0.89–1.09), since it is 0.8416 times the bootstrap SD of a gap between *selected* configurations and so is not a linear statistic of the item sample. The yardstick shrinks a little less than the inflation it measures: the measured form of §6's answer to whether a bigger benchmark fixes this.

The MRSG screen's null rate. $\mathrm{SE}_{\mathrm{total}}$ is, up to the $\sqrt{1.5}$ excess of Def. 1, the spread of one split-and-select run (with s^2 the per-item variance of the paired outcome difference, the random half contributes s^2/n and the resample $2s^2/n$, against $2s^2/n$ for a fresh half; for a selected pair the dev-side resample also adds selection noise, so this factor is approximate). The statistic the screen is applied to averages 600 such runs, so the split and the selection inside each run are averaged out of it and only its dependence on the items remains, which is why a single run's null crossing rate (40% at an exact SE, about 30% given that excess) is not the screen's. An outer bootstrap over items settles what is: resample the items, run the complete estimator with its item-first split made on the distinct items drawn (so a duplicated item stays in one half), repeat 60 times per pair on the eight Qwen tiers (each ratio carries a Monte Carlo error near 9%). The sampling SD of the bagged gap is $1/3.4$ of $\mathrm{SE}_{\mathrm{total}}$ (10–90% range of the ratio 2.8–4.0), so under a fresh-item null the screen fires on about 0.4% of comparisons (0.1–1.9%). Two fixed-item nulls give higher rates: 4.5% against disjoint halves of one family's pool (624 random half-splits, expected gap zero by construction) and 0.6% under the permutation null of §3.3. The 6 crossings among the 101 family pairs (5.9%) exceed the fresh-item and permutation nulls but not, beyond chance, the widest ($P(X{\geq}6) \approx .30$ at 4.5%); what stands out is that all 6 involve min-p; the remaining crossings of the 18 involve greedy, a $K{=}1$ arm against $K{=}24$ arms that neither null models. The same ratio says the MRSG is conservative for the reported gap by a factor of about 3.4: a comparison under it can still be resolved at the bagged estimate's own precision, which is the reading §5.2's permutation test takes. That precision is conditional on the drawn pools, and by the draw bootstrap (below) the draw contributes a pooled SD of up to .005, so every interval, z and equivalence bound here keeps the single-run $\mathrm{SE}_{\mathrm{total}}$, which covers both whenever the draw SD is below $0.96\,\mathrm{SE}_{\mathrm{total}}$; over every tier and pair it is at most $0.55\,\mathrm{SE}_{\mathrm{total}}$. A practitioner with one unbagged sweep should read the 95% criterion, not the screen. Bagging over bootstrap replicates also makes each replicate's dev selection noisier than a fresh half's; averaged instead over the permutation test's plain random splits (the protocol's dev size $n/2$), the 101 family gaps move by a median .002 (at most .008), and min-p's pooled gap to top-p is $-.020$ against $-.021$. The convention costs power where a gap is real: at the bagged estimate's own spread plus the pooled draw SD, min-p's pooled gap under the audit's prior would sit near $z = -3.1$, not -1.41.

Injected-effect positive control. The between-family nulls are informative only if the same machinery detects a between-family gap of the size at issue when one exists. The tuned-vs-untuned control cannot show that, since its true effect is unknown, so we plant one. In each of the eight Qwen tiers every configuration of temperature, which the permutation test finds exchangeable with top-p in every tier, is raised by a known δ: each incorrect item flips to correct with probability $\delta/(1 - \mathrm{acc})$, so the configuration's expected accuracy rises by δ, each item's in proportion to its error rate. The audit's estimator, its permutation test (1000 permutations) and its MRSG screen are then re-run, 20 injections per δ and tier. The held-out gap recovers 0.92–1.22 of the planted δ. With nothing planted the comparison is the audit's own temperature–top-p pair, which no tier rejects, so $\delta{=}0$ checks the harness, not the test's size (the calibration paragraph above addresses that); at $\delta{=}.02$ it rejects in a median 15% of injections per tier, at .03 in 78%, and at .05 in 100% (at least 75% in every tier). The MRSG screen, built from the wider single-run SE, fires in a median 48% at .03 and 98% at .05. Pooled over the eight tiers, the z-test crosses 1.96 in 0, 75 and 100% of injections at $\delta{=}.02$, .03 and .05. These rates are conditional on the audit's items. Redrawing the items with each injection (20 per case, item-first split on the distinct items drawn; `code/positive_control_uncond.py`), the per-tier test detects $\delta{=}.03$ in a median 35% and $\delta{=}.05$ in 85% (at $\delta{=}0$, 5%), and 43% and 88% when only the six best temperature configurations are raised; the pooled z-test detects $\delta{=}.03$ in 50% (70% for the top-six shift) and $\delta{=}.05$ in 100%. Per tier the protocol reliably sees five-point gaps and pooled it sees three-point gaps about half the time; about gaps of one or two points it says nothing.

Permutation-test size by item count. Two arms drawn from one family's pool are exchangeable by construction, so the rate at which the permutation test rejects between them is its size on the audit's items. Splitting each family's 24 configurations at random into two arms of 12 (arms half the audit's size; 20 splits per family and tier, 1,000 permutations each; `code/perm_size_check.py`), it rejects at $p < 0.05$ in 5.4% of 240 tests at the 128–156-item tiers and 5.5% of 800 at the tiers with

325 items or more: nominal at both sizes, so the small tiers' higher rejection rate (App. D) is not the test over-rejecting.

Cluster-bootstrap calibration. Under a null with the pooled effect removed, all eight tier signs agree .0017–.0100 of the time against the nominal .0078, and mean inter-tier residual correlation is under .014. Both say the item cluster costs the test almost nothing, and the reason is that the quantity pooled is a within-tier paired difference, so the shared item effect cancels before pooling. What the device prices only approximately is the configuration draw: resampling 24 configurations with replacement re-runs the audit at about 15 distinct ones (duplicates carry identical generations, so a resample has less selection noise, which can offset the added draw variance and is consistent with the cluster SE, .0158, sitting at the naive .0160 rather than above it), and a bootstrap of the drawn pool approximates a fresh draw from the prior rather than replacing it (the next paragraph isolates that component, the genuine redraws of App. F check it). That draw is shared by all eight tiers, so the sign test's eight units are eight replications in model, task and items and one in the draw. The calibration also recentres single-run gaps, in which that shared component is diluted by run noise; in the bagged gaps the sign test is applied to it is a larger share, which the calibration does not cover.

Resampling the prior draw. To put the draw into an interval we bootstrap it the way random-search budgets are usually bootstrapped: each family's 24 configurations are resampled with replacement from the configurations its prior produced, with the same resampled indices in every tier, so the pooled statistic carries the draw once, as the audit does. The restricted prior's configurations are the redrawn arm's 24 plus the 7 audit min-p draws at $T \leq 1.5$ (a temperature uniform on $[0.7, 3.0]$ conditioned on $T \leq 1.5$ is uniform on $[0.7, 1.5]$). Over 1000 resamples, with the items fixed and each gap averaged over 400 splits, the pooled draw SD of every between-family gap is at most .005 (an approximate estimate, see below), well inside the single-run $\mathrm{SE}_{\mathrm{total}}$ every interval here uses; the six four-family pairs stay within $[-.015, .012]$ (5–95%); min-p's audit arm is behind top-p in 100% of resamples ($[-.032, -.017]$; all eight tiers negative in 62%) and the restricted arm lies in $[-.008, .005]$ (all eight negative in 0%). Per tier the draw SD reaches .021; at 1.5B GSM8K it is .017 for min-p against top-p, against .014 across the genuine redraws of App. F. The budget-matched cells of Table 15, averaged over top-p's 7-subsets instead of conditioned on one, leave min-p − top-p at $-.018$ (SD over subsets .005) with its seven in-range draws and at $-.027$ with seven unrestricted ones: at that budget the restriction closes 33% of the gap, against 95% at $K{=}24$. A bootstrap runs each arm at about 15 distinct configurations (the restricted arm, resampled from 31, at about 17) and treats the drawn configurations as the prior, so it cannot detect a draw unrepresentative as a whole. The naive bootstrap of a best-of-K statistic is not consistent at $K{=}N$, and searching about 15 distinct draws it leans toward a larger min-p deficit, which widens at smaller budgets ($-.028$ at $K{=}6$ against $-.022$ at $K{=}24$); the genuine redraws are the check on both.

Estimator self-checks. A synthetic-data test suite validates the analysis chain: hypergeometric curves against direct Monte-Carlo subset simulation (agreement to three decimals); the (σ, ρ) estimators against known generating values; and the estimand check behind App. A's pool-truth result: on synthetic populations with a flat true landscape, where $B(K)$ is known from the generating parameters, the cross-fitted curves of Prop. 4 recover 0.98 of it at every $K{=}2$–8 (the reference is the Gaussian closed form; the Monte Carlo standard error over the 240 synthetic data sets is 0.01, and the discreteness of Bernoulli means accounts for 0.003), and curves that score a with-replacement dev sample against each configuration's mean over the same items recover 0.80–0.81. The (σ, ρ) estimators are $\sigma^2 = \overline{v}/n_{\mathrm{dev}}$, with $\overline{v}$ the mean over a pool's configurations of the per-item outcome variance, and ρ the mean between-configuration item covariance over $\overline{v}$. Eq. (3)'s residual-independence assumption is checked directly on the audit pools, but not by centering residuals across a family's own configurations, which forces the mean pairwise correlation to $-1/(K-1)$ whatever the truth. Estimating the item main effect from the configurations *outside* the family leaves the correlation free: it is then small, mean .043 across pools and at most .128 (1 of 52 pools negative). Because ρ is estimated within each pool it already absorbs such equicorrelation ($\hat{\rho} = \rho + r(1-\rho)$), so the envelope needs no further $\sqrt{1-r}$ factor; even applied, a factor of at least 0.93 could not produce ratios of 1.10–2.10, so the slack on the audit pools (Table 12) is mostly signal-driven selection, with unequal variances, heterogeneous correlations and discreteness not separated from it; the envelope's status on real pools is empirical.

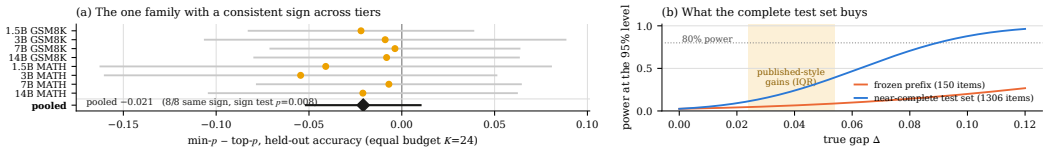

Figure 2: **(a)** min-p against top-p in each of the eight Qwen tiers (95% intervals) and pooled. Intervals use the single-run SE (App. C); no tier's interval excludes zero and one of the eight reaches its MRSG; every tier has the same sign, and the pooled gap, $-.021$, is $1.5\times$ the pooled MRSG of .0135. **(b)** Power of one run of the protocol to declare a true gap at the 95% level, at the frozen item prefix and on the near-complete test set (1.5B GSM8K). The shaded band is the interquartile range of the published-style gains of Table 10.

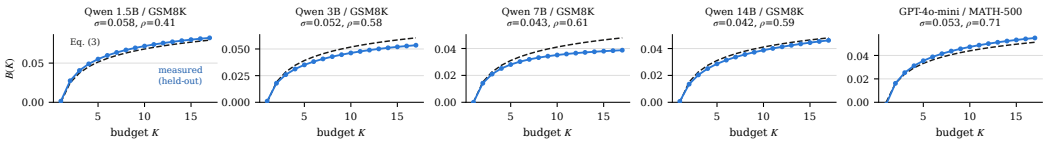

Figure 3: **Measured inflation matches Eq. (3) where the landscape is flat.** Dev-max inflation $B(K)$ on the five seed-replicate legs (17 replicates of one configuration): measured on held-out items (blue; dev-max on a random half minus the selected replicate's score on the other half, Prop. 4) against $\sigma\sqrt{1-\rho}\,e_K$ (dashed), with nothing fitted to the realized replicate means.

Compute. Open-weight generation ran on one shared 24GB consumer GPU; the entire study, including judging, used under US$5 of API spend. Sampling in llama.cpp is CPU-serial per sequence, so multiple server processes per GPU were required to saturate it ($4.1\times$ throughput).

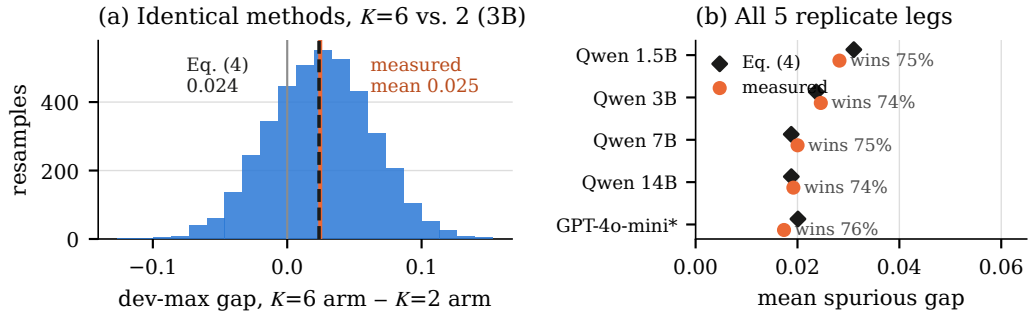

Figure 4: **Identical methods, unequal budgets, predictable "wins."** (a) Distribution of the dev-max gap between two disjoint pseudo-methods built from the same replicate pool (Qwen2.5-3B), budgets $K{=}6$ vs. 2, dev scores on random item halves; the dashed line is Eq. (4). (b) Measured vs. Eq. (4) mean spurious gap on all five replicate legs (* = MATH-500 on the closed tier); the higher-budget side wins 74–76% of resamples, as exchangeability predicts.

D PER-TIER RESULTS

Table 11 collects the flat-leg parameters and validation statistics per tier; Table 12 collects the audit pools' noise parameters and their in-situ measured inflation $B(24)$ against the flat envelope, plus the untuned convention baseline used by the published-style re-analysis of §5.3; Table 8 lists every pairwise comparison among the 153 whose measured gap exceeds the pair's MRSG; Table 9 lists every family pair the permutation test separates, with its p; Table 10 gives the published-style gain for every family in every tier. Rejection rates run opposite to resolution: the permutation test of §3.3 rejects exchangeability in 33% of the 21 comparisons at the 128–156-item tiers and 14% of the 80 at the eight tiers run with 325 items or more. Counted against the crossings each comparison's permutation null expects, the MRSG screen exaggerates that contrast (38.9× its null at the small tiers against 7.8× at the large), because the screen's own null rate rises with item count: the MRSG shrinks with items while the configuration-level spread between two random arms of one prior does not, which is why crossing counts are read against per-comparison nulls rather than a single pooled one. The small tiers are almost all the Llama control, so item count and architecture are confounded in this contrast, and the test's own size is not the explanation: between two arms drawn from one family's pool it rejects in 5.4% of 240 tests at the small tiers and 5.5% of 800 at the others (App. C); on the eight Qwen tiers alone min-p rejects in 10 of its 32 comparisons. Table 13 gives the redrawn $K{=}24$ restricted arm per tier, which is the comparison §5.2 rests on. The subsetting analyses below restrict the same range without redrawing (Table 15 gives every cell per tier). Their budget-matched cell, seven in-range audit draws against seven of top-p's, leaves top-p ahead by .020 on the same eight Qwen tiers where the $K{=}24$ redraw leaves $-.001$ (.018 averaged over top-p's 7-subsets, App. C); search depth and the particular seven draws cannot be separated here, and the draw bootstrap of App. C prices what a single prior draw does to every gap. The temperature-restricted min-p subset, which the redraw supersedes, selects over the 7 of 24 min-p configurations with $T \leq 1.5$: held-out means .389/.581 at 1.5B/3B MATH-500 versus .381/.579 unrestricted, so selection there mostly avoided the high-temperature region already (Table 18). That is a statement about where selection lands, not about where the deficit comes from: the subset restricts the range without restoring the budget inside it, and the redrawn arm of Table 13, which restores both, is what answers the question. Because restricting min-p's temperature and cutting top-p's budget are two different manipulations, the permutation runs all four cells over the 10 tiers where both families are expressible, as mean gap (top-p − min-p) with the number of tiers rejecting at $p < 0.05$: top-p_{24} vs min-p_{24}, .029 (6); top-p_{24} vs min-$p_7^{T\leq 1.5}$, .027 (2); top-p_7 vs min-$p_7^{T\leq 1.5}$, .027 (2); top-p_7 vs min-p_7, .034 (0). Cells A, C and D are budget-matched; C and D differ in the temperature restriction but also in which seven min-p draws and which top-p subset enter, and on this one subset C closes 22% of D's gap over these 10 tiers (averaged over top-p's 7-subsets on the eight Qwen tiers, 33%; App. C). In the second cell one tier, 7B GSM8K, sits at zero to three decimals, so its sign, and with it whether that cell is positive in all eight Qwen tiers, depends on the bootstrap stream. The second cell leaves top-p at 24 while min-p drops to 7: it keeps 93% of the first cell's gap (a difference of .0020), mixing the restriction with a

Table 2: **Terms used in the paper.**

Term	Meaning
family	one sampling rule with its hyperparameters: temperature, top-k, top-p, min-p, typical (and the added top-$n\sigma$, η, ϵ)
configuration	one setting of a family, $\theta = (T, \tau)$
prior	the distribution a family's configurations are drawn from (App. B)
pool	a family's K=24 configurations, drawn once from its prior
tier	one (model, task) pair: eleven in the audit
arm	one pool as a unit of comparison; *audit arm*: a family's pool as drawn for the audit (for min-p, from its audit prior); *restricted arm*: min-p redrawn at $T \leq 1.5$; *added arm*: generated after the audit (App. F)
flat leg	17 seed replicates of one configuration: a pool whose true landscape is flat
dev-max	the best dev score among the configurations searched (R1 reporting)
test-of-selected	the held-out score of the dev-selected configuration (R2 reporting)
$B(K)$	dev-max minus the true (held-out) score of what it selects
pool truth	a configuration's accuracy on the tier's analyzed items (its balanced panel)
ceiling	the best pool truth in a family's pool (an oracle)
$\hat{\Delta}$	the reported held-out gap between two arms, averaged over 600 bootstrap replicates
$\mathrm{SE}_{\mathrm{total}}$	the SD of one split-and-select held-out gap (Def. 1)
MRSG	$0.8416\,\mathrm{SE}_{\mathrm{total}}$; a *crossing* is $\lvert\hat{\Delta}\rvert >$ MRSG
permutation test	reassign two arms' configurations at random and re-run the selection; tests exchangeability given the items
published style	dev-max over 24 configurations, no split, minus an untuned default
draw bootstrap	resample each pool's configurations, jointly across tiers (App. C)
cluster bootstrap	resample tiers' shared items and configurations together (§5.4)

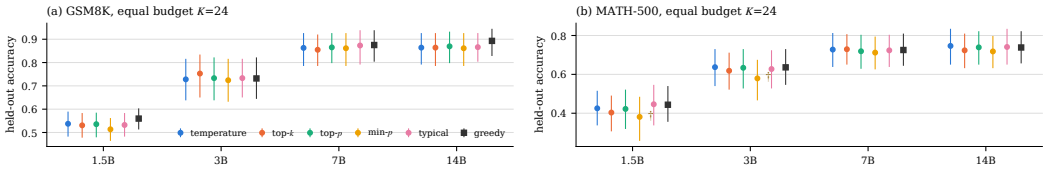

Figure 5: **Equal-budget leaderboards, Qwen tiers.** Held-out accuracy of the selected configuration per family, K=24, with marginal 95% paired-bootstrap intervals (paired comparisons are in Tables 8–9). Greedy, one configuration against 24, is the nominal best in four of the 11 tiers (Table 3); † marks min-p at 1.5B and 3B on MATH-500, whose deficits exceed the per-pair MRSG against two or more methods. Table 8 lists all 18 MRSG crossings among the 153 pairs of all 11 tiers and Table 9 the permutation rejections.

one-sided budget cut, and should not be read as the temperature effect. The restriction also bounds the degenerate-censoring channel of App. C: every censored min-p output at these tiers comes from configurations with $T \geq 1.99$, outside the restricted support. Tables built from separate bootstrap streams can print the same comparison with Monte-Carlo differences in the third decimal (1.5B GSM8K min-p − top-p in Tables 13, 16 and 17, for instance). Full per-tier budget curves, leaderboard intervals, and per-configuration scores are recomputable from the released records.

Table 3: **Equal-budget leaderboard (EBET, $K{=}24$/family): select on dev, report held-out.** Bold marks the nominal best per tier; *items* is the number of benchmark items behind that tier, and MRSG the range of its 15 per-pair resolution floors (3 on the closed tier). The closed tier can only express temperature and top-p, and its greedy column is temperature 0 requested through the API, which is not deterministic (§4); ‡ marks the cross-architecture control (Llama-3.1-8B) with identical configurations on the frozen item prefix. min-p at 1.5B, 3B and Llama-8B on MATH-500 exceeds the per-pair MRSG against two or more methods; Table 8 lists every crossing and Table 9 every permutation rejection.

Task	Model	items	temp.	top-k	top-p	min-p	typical	greedy	MRSG
GSM8K	1.5B	1306	.537	.531	.536	.514	.532	**.560**	.022–.028
	3B	325	.728	**.753**	.733	.724	.733	.732	.033–.043
	7B	325	.863	.855	.865	.861	.873	**.875**	.023–.032
	14B	325	.864	.864	.870	.861	.866	**.893**	.025–.034
MATH-500	1.5B	325	.425	.403	.422	.381	**.446**	.444	.039–.054
	3B	325	**.637**	.619	.634	.579	.627	.636	.035–.046
	7B	325	.728	**.729**	.719	.712	.724	.725	.028–.034
	14B	325	**.746**	.724	.739	.718	.741	.738	.029–.039
	4o-mini	156	.672	—	**.677**	—	—	.665	.044–.049
GSM8K‡	Llama-8B	150	.764	.761	**.779**	.729	.744	.768	.052–.064
MATH-500‡	Llama-8B	128	.494	.484	.502	.427	.474	**.550**	.059–.070

Table 4: Resolution of the protocol per tier. SE is the median, over the tier's between-family pairs, of the paired-bootstrap standard error of a single-run held-out gap ($n_{\text{test}}{=}n/2$); MRSG $= 0.8416$ SE is the gap one run orders correctly with probability 0.8 (about 0.85 at this bootstrap SE, which overstates a fresh run's by about $\sqrt{1.5}$; a true gap equal to it clears the screen half the time); MDE_{95} is the gap needed for 80% power under a two-sided 95% test. The last two columns are the probability that a true gap of $.05$ clears the MRSG screen and that its 95% interval excludes zero in the right direction, for one run with spread SE. The bagged gap the audit reports varies about $1/3$ as much, which sharpens its screen: it fires more often at $\Delta{=}.05$ where the MRSG is below $.05$ and less often where it is above (the two Llama tiers; App. C).

Tier	items	SE	MRSG	MDE_{95}	$\Pr_{\Delta=.05}$(MRSG)	$\Pr_{\Delta=.05}$(95%)
1.5B GSM8K	1306	.031	.026	.087	0.78	0.37
3B GSM8K	325	.048	.040	.134	0.58	0.18
7B GSM8K	325	.035	.029	.097	0.73	0.30
14B GSM8K	325	.037	.031	.103	0.70	0.28
1.5B MATH-500	325	.055	.046	.154	0.53	0.15
3B MATH-500	325	.048	.041	.136	0.58	0.18
7B MATH-500	325	.038	.032	.106	0.69	0.26
14B MATH-500	325	.043	.036	.120	0.63	0.21
4o-mini MATH-500	156	.058	.049	.163	0.51	0.14
Llama-8B GSM8K	150	.071	.060	.200	0.44	0.10
Llama-8B MATH-500	128	.078	.066	.219	0.42	0.09

Table 5: Greedy against each family, pooled over the eight Qwen tiers, by the same replicate-wise averaging as Table 6. Temperature, top-k, typical and top-p are all indistinguishable from greedy at 95% with sign counts at chance; min-p is not ($z = -2.21$), the same direction as its deficit against top-p in Table 6 and larger. Every pooled gap is negative: greedy is nominally ahead of every family. The SE is the single-run one (App. C): conservative for a claimed difference, lenient for a claimed tie. Greedy is a K=1 arm against K=24 arms, an asymmetry that runs against greedy whenever Prop. 3's monotonicity condition holds.

comparison	pooled gap	SE	z	same sign	sign p
temperature vs. greedy	−.0093	.0132	−0.70	5/8	.727
top-k vs. greedy	−.0157	.0139	−1.13	6/8	.289
top-p vs. greedy	−.0107	.0135	−0.79	6/8	.289
min-p vs. greedy	−.0315	.0143	−2.21	8/8	.008
typical vs. greedy	−.0075	.0132	−0.57	5/8	.727

Table 6: Each family against top-p, pooled over the eight Qwen tiers. Pooling averages the per-tier bootstrap gaps replicate-wise (tiers treated as independent; the cluster SE resamples their shared configurations and items but not the prior draw), so the pooled SE is smaller than a single tier's by about $\sqrt{8}$. The sign test is exact but treats tiers as independent, which their shared draw and items make approximate (App. C). Three families pool to within half an accuracy point of top-p with signs at chance; min-p is negative in every tier. The cluster SE is the two-way cluster bootstrap of §5.4, which resamples the shared configuration pool and the shared item ids; z divides the pooled gap by the naive SE, whereas the cluster bootstrap re-runs selection inside each replicate and re-estimates min-p's gap at $-.022$ ($z = -1.41$). The equivalence bounds of §5.4 are $|\text{gap}| + 1.645\,\text{SE}$ on the wider SE available; the six-pair bound also covers the three pairs without top-p (temperature–top-k .031; temperature–typical .027; top-k–typical .033, naive SE; inflated by the largest cluster/naive ratio, 1.05, they give the bound of .035; .033 uninflated).

vs. top-p	pooled gap	SE	cluster SE	z	same sign	sign-test p	MRSG
temperature	.001	.015	.016	+0.10	5/8	0.727	.012
top-k	−.005	.015	.015	−0.34	6/8	0.289	.012
min-p	−.021	.016	.016	−1.30	8/8	0.008	.013
typical	.003	.015	.015	+0.22	5/8	0.727	.012

Table 7: Tuned against untuned: the same machinery applied to each tuned family against the untuned convention baseline, the held-out estimate of Prop. 3's tuning term. Each row gives the range of held-out gaps of the tuned families over the baseline, and how many of them the protocol resolves; §5.4 gives the between-family comparison. These gaps come from their own paired bootstrap stream, so subtracting Table 12's baseline from Table 3 reproduces them only to Monte-Carlo error, at most .008 over every cell.

Tier	tuned − untuned	>MRSG	95% CI excl. 0
1.5B GSM8K	.004–.028	2/5	0/5
3B GSM8K	−.008 to .021	0/5	0/5
7B GSM8K	−.008 to .010	0/5	0/5
14B GSM8K	−.008 to .000	0/5	0/5
1.5B MATH-500	−.008 to .057	1/5	0/5
3B MATH-500	−.019 to .038	1/5	0/5
7B MATH-500	−.001 to .016	0/5	0/5
14B MATH-500	−.030 to −.001	0/5	0/5
4o-mini MATH-500	.010–.016	0/2	0/2
Llama-8B GSM8K	−.017 to .032	0/5	0/5
Llama-8B MATH-500	−.008 to .068	1/5	0/5

Table 8: Every pairwise method comparison (of 153 across the 11 tiers) whose measured held-out gap exceeds the pair's MRSG. Sign convention: gap = first method − second. z is the gap over its paired-bootstrap SE; the pooled tests in Table 6 are what carry a strict claim, not these per-tier z.

Tier	pair	gap	MRSG	z
1.5B GSM8K	top-k vs. greedy	−.029	.024	−1.01
1.5B GSM8K	top-p vs. greedy	−.024	.023	−0.88
1.5B GSM8K	min-p vs. greedy	−.046	.023	−1.70
1.5B GSM8K	typical vs. greedy	−.027	.022	−1.05
14B GSM8K	temperature vs. greedy	−.029	.027	−0.92
14B GSM8K	typical vs. greedy	−.027	.025	−0.90
1.5B MATH-500	top-k vs. greedy	−.040	.039	−0.87
1.5B MATH-500	min-p vs. typical	−.065	.054	−1.02
1.5B MATH-500	min-p vs. greedy	−.063	.047	−1.13
3B MATH-500	temperature vs. min-p	+.058	.044	+1.11
3B MATH-500	top-p vs. min-p	+.054	.045	+1.01
3B MATH-500	min-p vs. typical	−.048	.045	−0.90
3B MATH-500	min-p vs. greedy	−.057	.042	−1.14
Llama-8B MATH-500	temperature vs. min-p	+.068	.065	+0.88
Llama-8B MATH-500	top-k vs. greedy	−.067	.063	−0.89
Llama-8B MATH-500	top-p vs. min-p	+.076	.070	+0.92
Llama-8B MATH-500	min-p vs. greedy	−.124	.064	−1.64
Llama-8B MATH-500	typical vs. greedy	−.076	.059	−1.09

Table 9: Every family-pair comparison whose configurations are *not* exchangeable at $p < 0.05$, from 20,000 permutations of the pair's 48 configurations between the two arms with the audit's selection re-run inside each on the same 600 random item splits (observed and permuted alike), as $(1 + \#\{|\text{null}| \geq |\text{obs}|\})/(1 + R)$ over R permutations, so the smallest attainable p is $1/(R{+}1)$. Sign convention: gap = first − second. * survives Benjamini–Hochberg at FDR 0.05 over all 101 comparisons, † also Benjamini–Yekutieli. Of the 61 comparisons among temperature, top-k, top-p and typical, 3 reject (3.1 expected by chance); of the 40 involving min-p, 15 do, and every one is a min-p deficit. The one Benjamini–Hochberg survivor not involving min-p, temperature against typical on the Llama MATH-500 tier, also rejects on Llama GSM8K: two of the three non-min-p rejections are that pair on the two smallest tiers; the test's size is nominal at those item counts (App. C), so that pair either differs on Llama or rejects by chance.

Tier	comparison	gap	MRSG	perm. p
3B MATH-500	top-p vs. min-p	+.0545	.0455	$<.0001^{*\dagger}$
3B MATH-500	min-p vs. typical	−.0481	.0448	$<.0001^{*\dagger}$
Llama-8B MATH-500	top-p vs. min-p	+.0758	.0695	$<.0001^{*\dagger}$
Llama-8B MATH-500	min-p vs. typical	−.0472	.0625	$<.0001^{*\dagger}$
1.5B GSM8K	min-p vs. typical	−.0186	.0241	$.0012^{*}$
3B MATH-500	top-k vs. min-p	+.0392	.0447	$.0017^{*}$
Llama-8B MATH-500	temperature vs. typical	+.0204	.0608	$.0019^{*}$
1.5B MATH-500	min-p vs. typical	−.0650	.0537	$.0023^{*}$
1.5B MATH-500	top-p vs. min-p	+.0410	.0523	.0052
Llama-8B MATH-500	top-k vs. min-p	+.0572	.0677	.0063
Llama-8B GSM8K	top-p vs. min-p	+.0494	.0643	.0082
3B MATH-500	temperature vs. min-p	+.0578	.0437	.0128
Llama-8B MATH-500	temperature vs. min-p	+.0676	.0646	.0132
Llama-8B GSM8K	temperature vs. typical	+.0207	.0538	.0167
1.5B GSM8K	top-p vs. min-p	+.0220	.0262	.0316
14B MATH-500	top-p vs. min-p	+.0210	.0359	.0349
1.5B MATH-500	temperature vs. min-p	+.0442	.0490	.0368
7B GSM8K	top-k vs. typical	−.0184	.0289	.0420

Table 10: Published-style gain per family and tier: dev-max over the family's $K{=}24$ pool, no split, minus the untuned top-p 0.95 baseline, in accuracy points; greedy itself beats this default by 2.3 points pooled over the Qwen tiers (Table 1). Every family–tier cell shows a win except min-p on 14B MATH-500. The magnitude falls with item count on the one tier tested: re-running the 1.5B GSM8K tier on the near-complete test set cuts its mean gain from $+7.6$ to $+3.5$ points (§5.3). Against the other untuned baselines the audit generated, the tier means are -0.7 to $+11.1$ (median $+3.2$) against top-p 0.9 at $T{=}1$; -0.2 to $+8.5$ (median $+1.8$) against top-k 40 at $T{=}0.7$; -3.0 to $+5.1$ (median $+1.2$) against greedy: how large the published-style gain looks depends on which default is called untuned.

Tier	temp.	top-k	top-p	min-p	typical
1.5B GSM8K	+4.2	+3.8	+4.2	+1.3	+4.1
3B GSM8K	+1.8	+5.2	+3.4	+2.5	+3.1
7B GSM8K	+2.5	+1.8	+2.5	+1.8	+3.7
14B GSM8K	+1.5	+2.2	+2.5	+1.2	+1.5
1.5B MATH-500	+5.8	+3.7	+6.2	+2.5	+9.2
3B MATH-500	+5.8	+4.9	+6.5	+0.3	+5.5
7B MATH-500	+3.7	+3.5	+3.1	+2.2	+3.7
14B MATH-500	+2.5	+0.6	+1.8	−0.6	+2.8
4o-mini MATH-500	+5.8	—	+5.8	—	—
Llama-8B GSM8K	+4.7	+5.3	+8.7	+2.7	+4.7
Llama-8B MATH-500	+9.4	+9.4	+13.3	+3.1	+8.6

Table 11: Seed-replicate (flat) legs: 17 replicates of top-p 0.9 at $T{=}1$, true landscape flat by construction. n is the leg's item count (the frozen prefix, not the audit tier's) and σ is at the dev size $n/2$. r@K = measured $B(K)$ (dev-max minus the held-out score of the selected replicate, Prop. 4) over Eq. (3) at $K{=}2$ and $K{=}17$. P2 = mean spurious gap between disjoint pseudo-methods of 6 and 2 replicates, the dev-max difference Prop. 2 describes (4,000 draws of the two replicate sets and of a dev half, taken without resampling): measured, Eq. (4), and their ratio; *win*: the higher-budget pseudo-method holds the dev argmax (ties broken uniformly across both sets). P2 is dev-side, so it differs from $B(6) - B(2)$ read off App. Fig. 3 by the realized held-out term, which is zero only in expectation. Resimulating each leg with Bernoulli outcomes at its items' observed accuracies (40 synthetic legs each) gives median r@17 of 0.98–1.02, with 10–90% ranges reaching 0.69–1.22: a shortfall such as 7B's is within one leg's sampling spread, not binomial skew.

Tier	n	σ	ρ	r@2	r@17	P2 meas.	Eq. (4)	ratio	win
1.5B GSM8K	150	.058	.41	1.10	1.04	.0282	.0311	0.91	75%
3B GSM8K	150	.052	.58	0.94	0.89	.0246	.0237	1.04	74%
7B GSM8K	150	.043	.61	0.94	0.81	.0200	.0188	1.07	75%
14B GSM8K	150	.042	.59	0.89	0.96	.0192	.0188	1.02	74%
4o-mini MATH-500	156	.053	.71	1.00	1.07	.0174	.0201	0.86	76%

Table 12: Audit-pool statistics per tier, averaged over the tier's family pools: (σ, ρ) at the dev size $n/2$, measured dev-max inflation $B(24)$ (dev-max minus the held-out score of the selected configuration, Prop. 4) against the flat envelope $\sigma\sqrt{1-\rho}\, e_{24}$ and the ratio envelope$/B(24)$ (the inverse of Table 11's r@K), and the untuned convention baseline (top-p 0.95, T=1) that published-style analysis compares against.

Tier	σ	ρ	$B(24)$	envelope	env.$/B$	baseline
1.5B GSM8K	.018	.33	.014	.029	2.10	.510
3B GSM8K	.035	.50	.036	.048	1.34	.732
7B GSM8K	.028	.56	.030	.036	1.20	.862
14B GSM8K	.028	.52	.030	.037	1.23	.871
1.5B MATH-500	.035	.38	.037	.052	1.43	.388
3B MATH-500	.037	.54	.031	.049	1.56	.600
7B MATH-500	.034	.63	.036	.040	1.13	.714
14B MATH-500	.035	.56	.034	.045	1.33	.748
4o-mini MATH-500	.053	.70	.051	.057	1.10	.654
Llama-8B GSM8K	.050	.42	.051	.074	1.46	.747
Llama-8B MATH-500	.057	.43	.073	.083	1.14	.438

Table 13: **min-p redrawn inside its rivals' temperature range, at the audit's own budget.** Held-out gap (min-p arm − top-p pool, so negative is a min-p deficit; Table 9 prints the audit's comparison on each tier's full item set, where it rejects, with the arms the other way round) with min-p drawn from its audit prior $p \sim \log U[0.01, 0.30]$, $T \sim U[0.7, 3.0]$ and from $\log U[0.01, 0.30] \times U[0.7, 1.5]$, both at K=24, both against the same unchanged top-p pool on the items both arms were scored on. *shift* is the restricted column minus the audit column as a statistic in its own right (pooled SE of the restricted gap .016), with p from permuting all 48 min-p configurations between the two arms and recomputing both gaps. The audit arm is negative in every tier and the restricted arm is not; the shift is positive in every tier. *MRSG* is the restricted arm's floor. Per tier the shift reaches the restricted arm's MRSG only at 1.5B GSM8K (reaches it; $p = 0.121$) and 3B MATH-500 (exceeds it; $p = 0.023$); the MRSG is a yardstick here, not the shift's own floor. The pooled shift is what the claim rests on; the draw bootstrap of App. C prices how much of it one draw can move. *p paired to T*: a third min-p arm over the published temperature range with p paired to T following its guidelines (App. B), generated on the second build (App. F); in that column only, * marks a permutation $p < 0.05$ (none reaches it). ‡: the same two arms on Llama-3.1-8B at its frozen item prefix (App. C); not in the pooled rows.

Tier	items	audit	restricted	shift	perm. p	MRSG	p paired to T
1.5B GSM8K	325	−.040	.010	.050	.121	.050	.014
3B GSM8K	325	−.009	.001	.010	.582	.041	−.002
7B GSM8K	325	−.005	−.004	.002	.896	.028	.003
14B GSM8K	325	−.008	−.006	.002	.689	.029	−.006
1.5B MATH-500	325	−.036	.006	.041	.109	.044	.001
3B MATH-500	325	−.055	−.011	.043	.023	.041	−.015
7B MATH-500	325	−.008	.005	.013	.210	.033	.009
14B MATH-500	325	−.020	−.011	.009	.416	.037	−.004
pooled	—	−.023	−.001	.021	.037	—	.000
tiers with gap < 0 / shift > 0	—	8/8	4/8	8/8	—	—	4/8
Llama-8B GSM8K‡	150	−.055	−.053	.002	.896	.060	–
Llama-8B MATH-500‡	128	−.082	.015	.097	.069	.076	–

Table 14: The Eq. (5) decomposition of the published-style gain, per tier, in accuracy points. *realized* is Table 10's family-mean gain (dev size n, one draw). The next columns come from the cross-fitted budget curves of Prop. 4 at the protocol's dev size $n/2$: expected dev-max gain over the untuned baseline, its selection-bias part $B(24)$ (Table 12), and the held-out tuning term, which add up by construction; *share* is bias over expected gain (median 78%, §5.3). *at* n prices the realized gain with $B(24)/\sqrt{2}$, the flat-landscape scaling $B \propto 1/\sqrt{n}$ of §3.2 taken to dev size n, an approximation (median 69%, quoted in §5.3); $B(24)$ falls slightly faster than $1/\sqrt{n}$ (App. C), so it errs high. Shares above 100% mark tiers whose held-out tuning term is negative. The tuning term here uses the halves as drawn; the leaderboard (Table 3) resamples items within each half, which makes dev selection noisier and reads the term lower in 9 of 11 tiers, by up to 0.6 points. The style is stylized at $K{=}24$ against one default; at $K{=}6$, nearer the grids of published tables (the one checked in App. C compares two settings with one), $B(6)$ is a median 67% of the expected dev-max gain of 2.9 points.

Tier	n	realized	expected	$B(24)$	tuning term	share	at n
1.5B GSM8K	1306	+3.5	+3.9	1.4	+2.5	35%	28%
3B GSM8K	325	+3.2	+4.2	3.6	+0.5	87%	80%
7B GSM8K	325	+2.5	+3.3	3.0	+0.3	91%	85%
14B GSM8K	325	+1.8	+2.6	3.0	−0.4	114%	119%
1.5B MATH-500	325	+5.5	+6.7	3.7	+3.0	55%	47%
3B MATH-500	325	+4.6	+5.5	3.1	+2.4	57%	48%
7B MATH-500	325	+3.2	+4.3	3.6	+0.8	83%	78%
14B MATH-500	325	+1.4	+2.3	3.4	−1.1	146%	169%
4o-mini MATH-500	156	+5.8	+7.0	5.1	+1.9	73%	63%
Llama-8B GSM8K	150	+5.2	+6.6	5.1	+1.5	78%	69%
Llama-8B MATH-500	128	+8.8	+11.2	7.3	+3.9	65%	59%

Table 15: The four restriction and budget cells of App. D, per tier: held-out gap top-p − min-p, with * marking a permutation $p < 0.05$. A: both full pools of 24. B: top-p's 24 against the 7 audit min-p draws with $T \leq 1.5$. C: a random 7-subset of top-p against those same 7. D: random 7-subsets of both full pools. C and D are the budget-matched pair; C is the cell to read beside the $K{=}24$ redraw of Table 13, with which it disagrees. All four cells share one bootstrap stream per tier, separate from Table 3's, so cell A reproduces Table 9's top-p-vs-min-p gaps only to Monte-Carlo error.

Tier	A 24/24	B 24/7	C $7/7^{T\leq 1.5}$	D 7/7
1.5B GSM8K	+.025 *	+.024	+.011	+.040
3B GSM8K	+.009	+.016	+.011	+.010
7B GSM8K	+.006	+.000	+.006	+.008
14B GSM8K	+.008	+.005	+.013	+.012
1.5B MATH-500	+.033 *	+.030	+.038	+.058
3B MATH-500	+.054 *	+.052 *	+.052 *	+.056
7B MATH-500	+.007	+.002	+.006	+.008
14B MATH-500	+.019 *	+.017	+.020	+.016
Llama-8B GSM8K	+.054 *	+.057	+.034	+.052
Llama-8B MATH-500	+.080 *	+.072 *	+.074 *	+.079

Table 16: **Arms added to the audit.** Held-out gaps at $K{=}24$ per arm, each on the items its arm and the audit's pool were both scored on (*items*); * marks a permutation $p < 0.05$ (configurations of the two arms pooled and re-split, selection re-run). *Chain order*: min-p − top-p with the audit's configurations and seeds under temperature-first (the audit) and truncation-first ordering. *Symmetric prior*: min-p's audit pool against top-p's own prior and against top-p redrawn over min-p's temperature range $U[0.7, 3.0]$; *mean of 4* averages min-p's gap to all four rivals so redrawn. *top-*$n\sigma$: the sixth family minus the audit's top-p pool.

		chain order			symmetric prior		
Tier	items	temp. 1st	trunc. 1st	top-p own	top-p, $T{\in}[0.7, 3]$	mean of 4	top-$n\sigma$
1.5B GSM8K	325	$-.041^*$	.004	$-.041^*$	$-.031$	$-.022$	$-.024$
3B GSM8K	325	$-.008$	$-.015^*$	–	–	–	$-.001$
7B GSM8K	325	$-.005$	$-.009^*$	–	–	–	$-.003$
1.5B MATH-500	325	$-.035^*$	$-.033^*$	–	–	–	$-.035^*$
3B MATH-500	325	$-.054^*$	.001	–	–	–	$-.041^*$

Table 17: **Independent draws of every prior** (1.5B GSM8K, 325 items). Held-out score of each family's selected configuration at $K{=}24$ under the audit draw (first build) and 4 fresh draws of the same priors (new RNG seeds, second build), with the SD over the complete draws (so it includes the build's drift) beside the item-sampling SD of one draw's estimate. Each fresh draw also drew a restricted min-p arm ($T \leq 1.5$) on the continuation of its stream; the audit draw's restricted counterpart is Table 13's arm, generated on the first build like the audit pool.

Family	audit	redraw 1	redraw 2	redraw 3	redraw 4	SD draws	SD items
temperature	.516	.519	.522	.528	.516	.005	.052
top-k	.498	.532	.500	.492	.498	.016	.053
top-p	.518	.510	.514	.512	.535	.010	.053
min-p	.478	.491	.490	.511	.517	.016	.054
typical	.522	.530	.523	.531	.527	.004	.052
min-p ($T \leq 1.5$)	.527	.519	.508	.534	.500	.014	
min-p − top-p	$-.040$	$-.019$	$-.024$	$-.000$	$-.018$	.014	.061
min-p ($T \leq 1.5$) − top-p	.009	.009	$-.006$	.022	$-.035$	.022	

Table 18: **Where dev selection lands.** Median over 600 item-first dev halves of the selected configuration's temperature and truncation parameter per family (T / parameter), with the percentage of halves whose selection sits at $T < 0.7$ in parentheses (typical's and min-p's priors start at $T{=}0.7$) and, for min-p, the percentage above $T{=}1.5$ in brackets. Median selected T is below 1 for temperature in 10 of 11 tiers (below 0.7 in 8) and for top-k in 10 of 10 (below 0.7 in 6). top-p is selected near $T{=}1$, where its truncation removes tokens, as is typical, whose prior is centred there; min-p is selected above $T{=}1.5$ in part of the halves on several tiers.

Tier	temperature	top-k	top-p	typical	min-p
1.5B GSM8K	0.43 (100)	0.58 / 43 (99)	0.71 / .80 (37)	0.85 / .52	0.78 / .012 [0]
3B GSM8K	0.59 (74)	0.74 / 140 (19)	1.02 / .84 (7)	0.98 / .60	1.64 / .175 [64]
7B GSM8K	1.09 (20)	0.86 / 25 (39)	0.98 / .83 (20)	0.98 / .61	1.64 / .077 [51]
14B GSM8K	0.81 (48)	0.91 / 66 (30)	0.98 / .83 (18)	1.12 / .60	1.45 / .063 [42]
1.5B MATH-500	0.52 (100)	0.57 / 24 (95)	0.84 / .82 (17)	0.96 / .49	0.78 / .012 [4]
3B MATH-500	0.47 (96)	0.61 / 24 (81)	1.00 / .72 (34)	0.91 / .57	0.94 / .015 [7]
7B MATH-500	0.52 (77)	0.75 / 24 (45)	1.02 / .80 (11)	0.98 / .61	1.45 / .063 [27]
14B MATH-500	0.59 (90)	0.64 / 43 (50)	0.98 / .80 (3)	1.12 / .76	1.45 / .063 [26]
4o-mini MATH-500	0.94 (49)	–	1.00 / .80 (3)	–	–
Llama-8B GSM8K	0.47 (93)	0.61 / 43 (84)	1.08 / .77 (12)	0.94 / .57	0.98 / .058 [4]
Llama-8B MATH-500	0.47 (94)	0.64 / 24 (62)	0.84 / .73 (34)	0.96 / .60	0.98 / .058 [11]

Table 19: **Where the generations are.** Record files and generations per kind, as released (the supplement's manifest lists every file with its record count and checksum); records are counted as released, including the duplicated cells of App. C; judge calls on the creative leg are not generations; the 14B GSM8K file also holds the 2,400 superseded replicate records of App. C, not counted here.

Kind	files	generations
audit panels and seed-replicate legs	11	511,982
14B seed-replicate leg	1	2,400
restricted min-p redraw, $T \leq 1.5$ (Table 13)	10	69,072
8-bit control	1	40,300
creative leg (generations)	5	16,403
added: symmetric temperature range	1	31,200
added: truncation-first order	5	78,000
added: top-$n\sigma$	5	39,000
added: redraws and second-build regeneration	9	227,500
added: fixed T=1	1	31,200
added: HuggingFace second stack	2	54,600
added: min-p with p paired to T	8	62,400
added: top-p regenerated on the second build	7	54,600
total	66	1,218,657

E CREATIVE-WRITING LEG AND JUDGE PROTOCOL

The same thirty story prompts serve each of five generators, Qwen2.5 1.5B–14B and GPT-4o-mini (Llama-3.1-8B did not run this leg; not a replication of the min-p evaluation's creative-writing setup); each configuration generates one story per prompt (seven 1.5B stories, five of them min-p, carry no score and are left out of those configurations' means); stories are scored 1–10 by GPT-4o-mini at temperature 0 with a fixed rubric (coherence and fluency, creativity and interest, adherence to the prompt and its 150–250-word length), and family scores aggregate as means of selected-configuration scores under the same EBET machinery (15 dev and 15 test prompts per split), MRSG computed identically; scores enter the $[0, 1]$ scale of §3.1 divided by 10 (so on $[0.1, 1]$) but are reported here on the 1–10 scale. GPT-4o-mini judges its own stories on the 4o-mini tier (a self-preference risk; Panickssery et al., 2024); since comparisons are within one generator, a constant self-preference shifts levels only, though it could interact with temperature. Selected-family means span 6.36–6.49 (1.5B), 7.50–7.61 (3B), 7.95–8.00 (7B), 7.98–8.05 (14B) and 8.44 (both 4o-mini families, temperature and top-p), and greedy scores $6.16/7.68/8.00/8.00/8.43$. The floors are the reason: per-tier MRSGs are $0.50/0.23/0.11/0.14/0.16$ on the 1–10 scale, which at every tier exceeds the entire observed spread between families. Thirty prompts, fifteen held out per split, cannot resolve these differences, and we do not claim they do. Judge self-consistency was measured by issuing 3,297 duplicate calls: 9.2% score differently (mean absolute difference 1.02 among those; single-call SD .221; $r = .990$ between passes). We report the creative leg as inconclusive: per-call judge noise on a given story is residual noise, which selection harvests and pairing does not cancel, and it adds to the prompt-sampling noise that, with 30 prompts, sets the resolution of any judged comparison at this scale.

F ARMS ADDED TO THE AUDIT

Every arm below was generated after the audit, on its items and after its results were seen, is exploratory (its p-values are not corrected across arms), and was drawn on its own RNG stream into its own file, so no configuration of the audit pool is touched, and each is read with the audit's own estimator: shared item-first splits, dev-argmax within each arm at $K{=}24$, the held-out gap, its MRSG, and a permutation test that pools the two arms' configurations and re-runs the selection (2,000 permutations), on the items both arms were scored on (Table 16). Where an arm ran on the second build and its comparison pool on the first, the permutation null also absorbs the build's drift ("Regeneration on a second build" below), so such a test is of build, prior and rule jointly.

Regeneration on a second build. The audit's 1.5B GSM8K pool was regenerated with identical configurations, seeds and items on a rebuilt llama.cpp (release b10240) on a fresh instance: 79% of item scores agree, against 88% for regeneration on the same build (App. C); the mean configuration's accuracy moves by $-.004$ (SD .025 over configurations); and no family's tuned held-out score moves by more than .011, inside every per-tier MRSG. Regenerating the audit's top-p configurations on the second build on eight tiers (1.5B GSM8K, 3B GSM8K, 7B GSM8K, 14B GSM8K, 1.5B MATH-500, 3B MATH-500, 7B MATH-500, 14B MATH-500) moves top-p's tuned held-out score by $-.015$ to .003. Gaps taken against an audit pool (the symmetric-prior rivals, top-$n\sigma$) carry that much drift, and so do level comparisons with the audit (the redraws' SD over draws, min-p's own score under truncation-first); gaps taken within one build (min-p $-$ top-p in the redraws and under truncation-first, η/ϵ against same-stack top-p) do not.

η and ϵ on a second stack. llama.cpp implements neither, so both run on HuggingFace transformers at 16-bit on a 325-item prefix of the 1.5B GSM8K tier, beside the audit's own top-p, min-p and typical configurations on the same stack. The model's shipped generation defaults are removed, so each family is its sampler alone, applied after temperature. Against same-stack top-p, the gaps are $-.047$ for η (MRSG .054, permutation $p = 0.005$), $-.050$ for ϵ (MRSG .052, $p = 0.002$) and $-.058$ for min-p (MRSG .059, $p = 0.0005$), none past its MRSG, against $-.042$ for min-p under llama.cpp on the same items: the sign of the min-p deficit replicates on a second implementation at a second precision, on one tier, though min-p's prior shares the $T{=}0.7$ floor discussed next, which this arm does not separate from its rule. All three priors start at $T{=}0.7$, and top-p's best configuration on this stack sits at $T{=}0.62$, below that bound. Searched over $T \sim U[0.3, 1.5]$ instead, the temperature family's range, η and ϵ tie same-stack top-p ($-.014$, $p = 0.968$; $.004$, $p = 0.973$), gaining .032 and .053 over their own priors with best configurations near $T{=}0.3$, which points to their temperature floor (the change also raised their temperature cap from 1.3 to 1.5, and both ranges are ours). Typical,

whose prior starts at the same T=0.7, ties top-p anyway (.009, $p = 0.599$; .003 under llama.cpp), so on this tier a range's cost differs by rule. The stack moves levels: tuned top-p scores .071 higher at 16-bit than under 4-bit llama.cpp, while the min-p deficit keeps its sign; temperature and top-k were not run on it.

Truncation-first ordering. llama.cpp's default chain truncates before tempering; the audit pins temperature first, the ordering of vLLM and current HuggingFace transformers. With the audit's own min-p and top-p configurations and seeds and the order reversed, on five tiers and on the second build (so the comparison with the audit's first-build order also carries build drift), min-p minus top-p pools to $-.010$ (negative in three, permutation $p < 0.05$ in three) against $-.029$ for the same configurations temperature-first on the same items (Table 16). Reversal does what min-p's high-temperature case predicts: no min-p configuration on a GSM8K tier falls below .231 accuracy truncation-first, against .003 temperature-first, so its hot draws stop being wasted budget. Tier by tier the deficit vanishes at 1.5B GSM8K and 3B MATH-500, persists at 1.5B MATH-500 ($-.033$, $p = 0.001$) and moves by less than the build drift at 3B GSM8K ($-.008$ to $-.015$) and 7B GSM8K ($-.005$ to $-.009$), and min-p's own tuned score rises by up to .041 (a comparison across builds); three tiers reject under truncation-first against three temperature-first. The five tiers include the three with the largest temperature-first deficits, so regression toward the pooled deficit accounts for part of the pooled fall. The ordering moves where the deficit sits more than whether it exists.

Symmetric temperature range. The audit's priors give min-p alone the range above T=1.5. This control gives it to every rival: each keeps its own truncation prior and draws $T \sim U[0.7, 3.0]$ on its own RNG stream (which also lifts the floors of temperature, from 0.3, and of top-k and top-p, from 0.5, removing the cool region where their dev selection lands on this tier, Table 18; floor and cap are not separated), against the audit's min-p pool on the same items (Table 16). On 1.5B GSM8K (325 items) min-p's mean deficit to the four rivals shrinks from $-.037$ under their own priors to $-.022$ under its range. The rivals respond differently (levels compared across builds, so within about .011 of drift and a draw's SD): temperature loses most (tuned .466, against min-p's .475), top-k less (.479), top-p keeps most of its score (.506), and typical gains (.539). Under the shared wide range the rivals separate from one another, in three of six pairs at $p < 0.05$, two past their MRSGs (largest gap .073, $p = 0.008$), against the chance rate under their own priors: rule and prior together separate them where temperature is high enough for truncation to matter, and dev selection under their own priors stays out of that region (Table 18). The rivals ran on the second build: against min-p's pool regenerated on that build, min-p trails top-p by .035 ($p = 0.060$) and typical by .068 ($p = 0.003$); against the audit's first-build pool, by .031 ($p = 0.095$) and .064 ($p = 0.011$).

Fixed temperature. The same question at a standard temperature: the four truncation families redrawn with T pinned at 1, each over its own truncation prior at K=24 (1.5B GSM8K, 325 items). Tuned held-out accuracies are .505 (top-p), .475 (top-k), .509 (typical) and .514 (min-p), against .516 for the audit's top-p with temperature searched (on this arm's bootstrap stream). The four rules separate only through top-k, which trails the other three by .030–.039 (three of six pairs at $p < 0.05$), while the other three lie within .009 of each other, with min-p nominally first, against a median MRSG of .052: on this tier, at a fixed moderate temperature, min-p's pool shows no handicap and top-k's does (a pool-level difference; every tuned gap is below the MRSG).

A sixth family: top-$n\sigma$. top-$n\sigma$ (Tang et al., 2025) keeps the tokens within n standard deviations of the top logit, a set temperature does not change, and is promoted, like min-p, for robustness at high temperature. Searched over min-p's own temperature range (App. B) against the audit's top-p pool on five tiers, it sits at $-.041$ to $-.001$ (pooled $-.021$). It trails top-p at permutation $p < 0.05$ at 1.5B and 3B MATH-500 (the latter where min-p trails most) and is not resolvably behind elsewhere. Against the audit's top-p configurations regenerated on the second build (1.5B GSM8K, 3B GSM8K, 7B GSM8K, 1.5B MATH-500, 3B MATH-500), the arm's own build, it pools to $-.014$ (per tier $-.038$ to .014; permutation $p < 0.05$ in two), against $-.021$ across builds on the same tiers. A temperature-invariant kept set does not stop sampling within it from degrading at T=3, so a wide range may cost this family too (no top-$n\sigma$ arm at $T \leq 1.5$ was run).

Independent redraws of every prior. Every conclusion above rests on one draw of each prior. On the 1.5B GSM8K tier the five priors were redrawn four times on fresh seeds, and each fresh draw also drew a restricted min-p arm ($T \leq 1.5$) on the continuation of its RNG stream (Table 17). Across the five draws a family's tuned held-out score moves by an SD of .004–.016, and a between-family gap by .004–.026, against a median item-sampling SD of .059 for one draw's gap (median ratio 0.24;

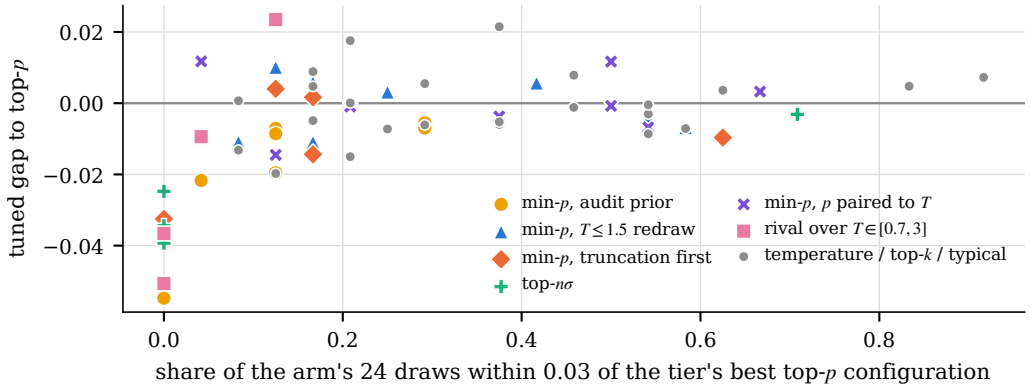

Figure 6: **Where the budget lands.** One point per (tier, arm): the share of the arm's 24 configurations whose accuracy on the shared items lies within 0.03 of the tier's best top-p configuration (itself a maximum of 24 noisy means), against the arm's tuned held-out gap to top-p at $K{=}24$ (llama.cpp arms of App. F and the audit's own pools, Qwen tiers; complete arms only). The truncation-first min-p arm is measured against truncation-first top-p on its own build; every other arm against the audit's top-p pool, so second-build arms carry that build's drift.

each SD has four degrees of freedom, so these are rough), which would widen the audit's intervals by a median 3%. Counting the audit draw (with its first-build restricted arm), the audit min-p prior trails top-p in four of five draws and ties in one (to within half a point) ($-.040$ to $-.0004$); the restricted prior in two of five ($-.035$ to .022, mean $-.000$). The audit draw, the only one generated on the first build, is the least favourable to min-p of the draws. Five of the six pairwise gaps among the other four families change sign between draws, as exchangeable pools allow. Every added arm's per-item accuracy profile correlates best with its own model's audit pool (own $r \geq 0.92$, any other model at most 0.92), the profile check of App. C.

min-p with p paired to T. The audit's min-p prior draws p independently of T, so most hot draws get a low p; this arm keeps $T \sim U[0.7, 3.0]$ and pairs p with T as the min-p guidelines do (App. B), 24 configurations on a fresh seed. On eight Qwen tiers (1.5B GSM8K, 3B GSM8K, 7B GSM8K, 14B GSM8K, 1.5B MATH-500, 3B MATH-500, 7B MATH-500, 14B MATH-500; 325 items or more), against the audit's unchanged top-p pool, it pools to .000 (SE .017; negative in four, permutation $p < 0.05$ in zero as a deficit and zero as a lead; per tier $-.015$ to .014), against $-.023$ for the audit arm and $-.001$ for the $T \leq 1.5$ redraw on the same tiers (Table 13); against those two arms directly it sits at .022 and .002. It ran on the second build, so its gaps to the audit's pools carry that build's drift (at most .011 per family on 1.5B GSM8K). Against the audit's top-p configurations regenerated on the second build (same seeds and items; 1.5B GSM8K, 3B GSM8K, 7B GSM8K, 14B GSM8K, 1.5B MATH-500, 3B MATH-500, 7B MATH-500, 14B MATH-500), the within-build comparison pools to .005 over eight tiers (negative in four, permutation $p < 0.05$ in zero), while regenerating top-p moves its tuned score by $-.015$ to .003.

Where the budget lands. One descriptive statistic tracks the arms: the share of a family's 24 draws whose pool truth (accuracy on the shared items) lies within 0.03 of the tier's best top-p configuration. Over the 62 (tier, arm) pairs of Fig. 6, a family's tuned gap to top-p tracks that share ($r = 0.46$; the points share tiers, and both axes are relative to top-p's pool): min-p's audit arm has among the lowest shares, and restricting its range raises the share in every tier (pairing p with T in all but one, where it ties). Every deficit larger than two points sits at a share below one in four, a threshold read off the figure rather than tested; the association is partly built in, since an arm with few draws near top-p's best can rarely tune past it.