SAMPLING AT INFINITE TEMPERATURE: KL* AND KL-BUDGET DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Applying temperature before truncation lets unlikely tokens into the candidate set, and a high temperature can then derail generation. We reverse the order. The model's original probabilities choose the candidates, and exploration happens only among them. At infinite temperature every candidate is equally likely, so the candidate set alone decides what is sampled. Global KL* picks the candidate set whose uniform draw is closest to the model in KL divergence. We also study local KL*, a greedy variant, and KL-budget, which flattens a min-p candidate set within a fixed KL allowance. On the complete GSM8K test set, both KL* rules match each model's recommended sampling settings without any tuning, and global KL* has no significant loss in pass@8 or majority vote against any standard default sampler. Across six tasks, the KL* rules gain two to three times as often as they lose against these defaults. On real next-token distributions they explore below the top token while rarely reaching the tail, and at high temperature truncating first prevents the collapse of high temperature sampling. The rules remain competitive in creative writing and with reasoning models. We release all generations and analysis code.

1 INTRODUCTION

Sampling lets a language model explore alternatives, such as several attempts at a math problem or several drafts of a story. Exploration helps only if the model keeps its confident predictions. Most samplers combine two steps. Truncation decides which tokens are allowed, and temperature decides how evenly they are weighted. Many pipelines apply temperature first, as in P-less (Tan et al., 2026). A higher temperature then also changes which tokens pass the cutoff, and a hot enough draw admits tokens the model considers implausible (Figure 1).

We reverse the order. The model's original probabilities choose the candidates, and temperature only reweights them. This follows the view of truncation as removing smoothing mass (Hewitt et al., 2022), the logit rules top-$n\sigma$ and min-k (Tang et al., 2025, Ding et al., 2026), and the default sampler chain of `llama.cpp` (llama.cpp contributors, 2026). A controlled thirteen-model study finds that truncation matters most when high temperature degrades a model (La Rosa, 2026). Infinite temperature is the extreme case. Every candidate becomes equally likely, so the candidate set alone decides what is sampled, and no temperature is left to tune.

Which candidates should an equal-weight draw use? Global KL* picks the set whose uniform draw is closest to the model in KL divergence. Adding a token makes the draw more diverse but gives an unlikely token an equal chance, and KL* balances the two. One sort and one scan find the exact answer, and nothing needs tuning. Local KL* stops at the first token that makes the divergence worse. KL-budget keeps a min-p candidate set and flattens its weights as far as a fixed KL allowance permits, using the solution of OD-Stega (Huang et al., 2026). Our contributions are:

- **A tuning-free rule for infinite-temperature sampling.** Global KL* finds the KL-closest uniform candidate set exactly. We prove that the cheaper local rule can be arbitrarily worse.
- **A default-settings comparison.** We compare seven samplers at their standard defaults, including each model's recommended settings. Against baselines tuned on development data in both sampling orders, tuning-free KL* stays within 3.3 points in pass@8 and Vote@8 (Appendix B).

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

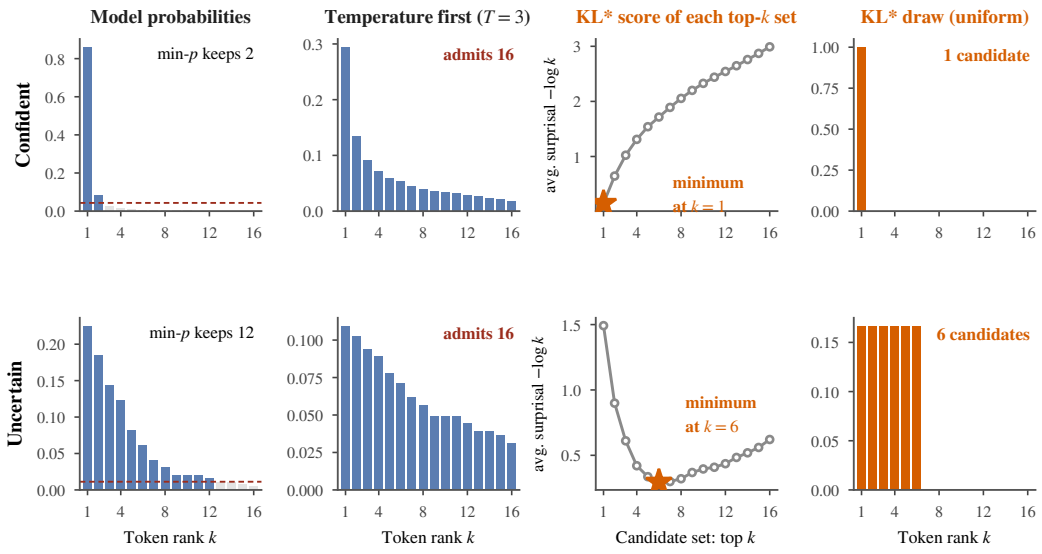


Figure 1: **How KL* chooses candidates.** Two example distributions, one confident and one uncertain. Blue bars are admitted tokens and gray bars are excluded. A min-p .05 cutoff on the model's probabilities keeps 2 and 12 tokens, but applying temperature 3 first admits all 16. KL* scores every top-k set by the average surprisal of its tokens minus $\log_e k$ and keeps the minimum: one token in the confident case and six in the uncertain case, each drawn with equal probability.

- **A complete test-set evaluation.** On all 1,319 GSM8K test questions for three checkpoints, we report all 189 pre-specified comparisons. Neither KL* rule differs significantly from any model's recommended settings, and global KL* has no significant pass@8 or Vote@8 loss against any default. Real next-token statistics show that KL* explores below the top token but rarely reaches the tail. Further studies cover five more tasks, checkpoints up to 123B, a reasoning model, and creative writing, and all generations are released.

Figure 2 works through all three rules on one distribution, and Figure 3 shows stories written with KL*, KL-budget, and the model's recommended settings.

2 Related Work

Truncation samplers. Top-k keeps a fixed number of tokens (Fan et al., 2018), and top-p keeps a probability-mass nucleus (Holtzman et al., 2020). Min-p scales its cutoff by the top token's probability (Nguyen et al., 2025). P-less sets its cutoff from the distribution itself and has no truncation parameter (Tan et al., 2026). These four are our baselines. Truncation can be read as removing smoothing mass (Hewitt et al., 2022), and sampling adapters trade recall for precision (Meister et al., 2023a). Locally typical sampling (Meister et al., 2023b) and Top-H (Baghaei Potraghloo et al., 2025) use entropy to shape a nonuniform draw, whereas KL* chooses the candidates of a uniform draw. Quality and diversity should be evaluated together (Zhang et al., 2021). La Rosa (2026) finds that truncation gains concentrate on models that degrade at higher temperature and that truncating first retains more accuracy on such a model.

Separating candidates from weights. Top-$n\sigma$ (Tang et al., 2025) and min-k (Ding et al., 2026) select tokens from logits in ways that temperature does not change (Appendix O). KL* shares this separation but selects candidates with a KL objective on the original probabilities. CUTS draws uniformly within thresholded top-k candidates during reinforcement learning (Liang et al., 2026). Sparse Bregman decoding optimizes candidates and weights together under a sparsity penalty (Noarov et al., 2025). Restricting to uniform draws gives a different objective, whose first local minimum need not be global.

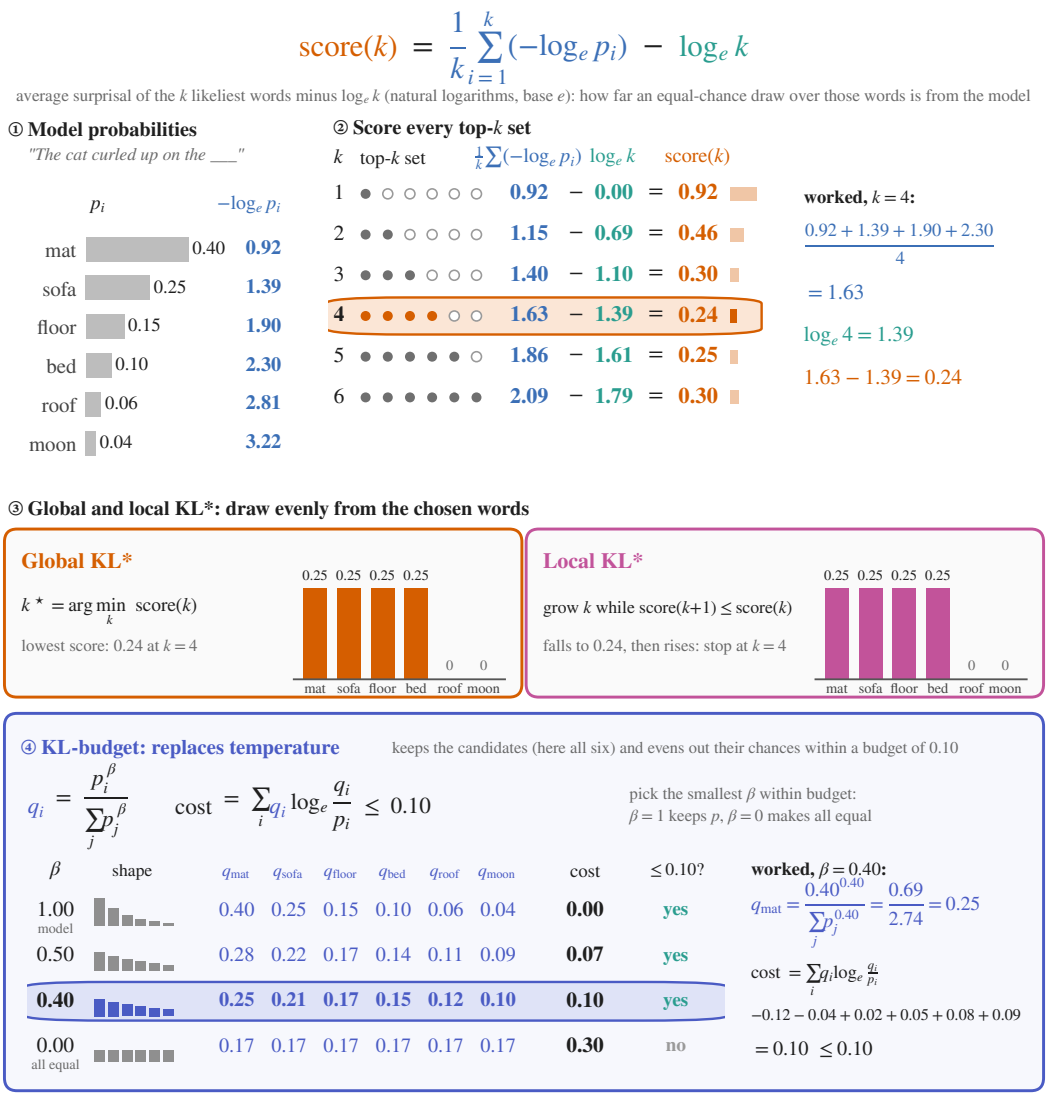


Figure 2: **KL*, local KL*, and KL-budget worked on one distribution.** Blue marks each word's surprisal $-\log_e p_i$ and its average over the top k words, teal marks $\log_e k$, and orange marks their difference, the KL* score. Global KL* keeps the k with the lowest score, and local KL* stops at the first rise. KL-budget replaces temperature. It raises each probability to a power β and tries values from 1, the model's own weights, down to 0, all equal. It keeps the flattest weights whose cost stays within a 0.10 budget, here $\beta = 0.40$.

Adaptive temperature and KL constraints. Maximum entropy is a standard principle for choosing distributions (Jaynes, 1957). Mirostat targets a perplexity (Basu et al., 2021), and KL-guided temperature has been used for source grounding (Chang et al., 2023). OD-Stega derives the most diverse reweighting under a KL constraint and analyzes truncation before it (Huang et al., 2026). KL-budget applies that solution within a min-p candidate set.

Multiple attempts. Chain-of-thought prompting elicits reasoning (Wei et al., 2022). Self-consistency votes over sampled paths (Wang et al., 2023), and pass@k asks whether any attempt is correct (Chen et al., 2021). Arithmetic sampling coordinates draws to reduce duplicates (Vilnis et al., 2023). We report individual-path accuracy, coverage, and voting separately.

3 KL* AND KL-BUDGET SAMPLING

3.1 TRUNCATE FIRST, THEN EXPLORE

With temperature first, a high temperature flattens the distribution until a fixed cutoff admits many unlikely tokens (Figure 1). With truncation first, the model's original probabilities decide which tokens survive, and temperature changes only their weights. As temperature grows, the surviving weights become equal. A single survivor gives a deterministic step, and several survivors leave room to explore. The remaining question is which tokens should survive.

3.2 GLOBAL KL*: THE CLOSEST UNIFORM DRAW

Global KL* ranks tokens by probability and scores each top-k set by the KL divergence of a uniform draw over it:

$$\text{score}(k) = \frac{1}{k}\sum_{i=1}^{k}(-\log_e p_i) - \log_e k, \tag{1}$$

the average surprisal of the top k tokens minus $\log_e k$, with natural logarithms throughout. It keeps the k with the lowest score and samples uniformly from those tokens (Figure 2 works through one example). Adding a token raises diversity through the $\log_e k$ term but also raises the average surprisal when the token is unlikely. A confident prediction therefore keeps one token, and an uncertain one keeps several. The best uniform set is always a top-k set, so one sort and one scan find the exact optimum (Appendix N). The search is also cheap, taking 0.10–0.14 ms per recorded 70B first-token CPU replay versus 0.09–0.11 ms for min-p (Appendix P.7).

3.3 LOCAL KL*: STOP AT THE FIRST INCREASE

Local KL* adds tokens in rank order and stops as soon as the score rises. In words, it admits the next token when its probability is at least 25–37% of the geometric mean of the tokens already admitted. The factor starts at one quarter and approaches $1/e$. Min-p instead admits every token above a fixed 5% of the top probability. Local KL* can stop too early. In Appendix Figure 16a the score rises at the second token and then falls, so local selection keeps one token while global selection keeps six. We prove that this gap can be arbitrarily large (Appendix N). A better score does not guarantee better answers, so we evaluate both rules.

3.4 KL-BUDGET: FLATTEN WITHIN A FIXED ALLOWANCE

KL-budget replaces temperature rather than truncation. Given a candidate set, it makes the weights as even as possible while staying within a KL allowance of the model's weights on that set. It can follow any candidate rule, including KL*. We evaluate it after min-p .05. OD-Stega gives the form of the solution (Huang et al., 2026). It amounts to a temperature chosen separately at each step (Figure 2, panel 4). The allowance needs tuning. For reasoning we choose it per checkpoint from .03, .1, and .3 nats on development data, and writing uses .1 nats. The allowance covers only reweighting, since truncation adds its own divergence (Appendix M). Our new contribution is the candidate-set analysis, and KL-budget reuses the prior solution.

4 EXPERIMENTS

4.1 SETUP

Models and tasks. The main evaluation uses all 1,319 GSM8K test questions (Cobbe et al., 2021) on Qwen2.5 1.5B, Mistral 7B, and Qwen2.5 7B (Yang et al., 2024, Mistral AI, 2024a). Each prompt has eight worked examples from the training split. Smaller-sample studies add five checkpoints from 27B to 123B (Qwen Team, 2026, Meta, 2024, Mistral AI, 2026, 2024b) and the MATH-500 and GPQA tasks (Hugging Face H4, 2024, Lightman et al., 2024, Rein et al., 2024) (Appendix Q).

Baselines. Every baseline runs at a standard default with no tuning. These are temperature one alone, top-p .95, top-k 40, min-p .05, and top-$n\sigma$ 1 (La Rosa, 2026), P-less (Tan et al., 2026), and

Prompt 1. *write about the strangest/scariest/saddest dream you've ever had in less than 200 words.*

Model card	Global KL*	KL-budget
In the dead of night, I found myself standing at the edge of a vast, endless forest. The air was thick with mist, and the trees loomed overhead like ancient sentinels. As I stepped forward, the path beneath my feet turned into a river of liquid darkness. I tried to turn back, but the river seemed to draw me in, pulling me deeper until I was submerged up to my …	In the quiet of a midnight, I found myself wandering through a forest shrouded in an eerie silence. The trees, tall and ancient, twisted and gnarled, loomed over me like sentinels of a forgotten era. The moon was a pale, distant crescent, casting an otherworldly glow that revealed a path of shimmering bones beneath my feet.	In the dead of night, I found myself in a vast, dark forest, the only sound the rustling leaves underfoot and the distant hooting of an owl. The moon, a sickly green sliver, barely pierced the canopy. As I walked, the trees began to shift, their gnarled roots reaching out like bony fingers.

Prompt 2. *You died. You were reborn. You remembered your previous life. But that's actually pretty useless - because you weren't reborn human.*

Model card	Global KL*	KL-budget
In the quiet of dawn, under a canopy of stars, Lila felt a strange sensation wash over her. She had always been a human, but now, as she lay in the soft grass, she realized something was different. Her body felt lighter, more fluid, and when she looked down, she saw shimmering wings sprouting from her back.	You died. You were reborn. You remembered your previous life. But that's actually pretty useless—because you weren't reborn human. In the bright, neon glow of the alien city, you found yourself, a shimmering figure with tentacles and iridescent scales. Your previous life as a human had been a distant memory, a blur of faces and places.	You died, and as you floated above your own body, the world blurred and faded, leaving only a distant whisper of voices and colors. You were reborn, and as consciousness returned, the first thing you remembered was the sensation of wings. Not the memory of them, but the actual feeling of them, beating against the air.

Figure 3: **Story openings from the same model under three samplers.** Qwen2.5 7B at default settings on the first two prompts of the WritingPrompts pool, seed 5100. The prompts were fixed before generation, and the samplers shown before any story was read. Appendix J lists every sampler.

each checkpoint's recommended model-card settings (Appendix C). Global and local KL* have nothing to choose, and KL-budget uses min-p .05 with a .1-nat allowance. Greedy decoding is a single-answer reference. A separate comparison tunes temperature, top-p, min-p, P-less, and KL-budget on development questions in both sampling orders (Appendix B).

Metrics. Every sampler produces eight seeded answers per question. *Individual-path accuracy* averages their correctness. *Pass@8* asks whether any of the eight is correct, as a verifier would. *Vote@8* scores the most common answer (Wang et al., 2023). Answers use flexible numerical extraction (EleutherAI, 2026). Unparseable outputs count as wrong, and outputs that reach the length limit are scored as returned.

Statistics. Questions are the unit of uncertainty. Intervals come from paired question bootstraps, and pass@8 uses Wilson intervals. Within each checkpoint we fixed a family of 63 comparisons in advance, three KL methods against seven defaults on three metrics, with Bonferroni-adjusted simultaneous 95% intervals (Appendix C). We call a comparison significant when its interval excludes zero. Of the 1,319 questions, 224 appeared in our smaller-sample studies, and we also report the other 1,095 separately.

Implementation. All decoding uses a pinned `llama.cpp` build, checksum-verified Q8_0 weights and caches for the three main checkpoints, and a 1,536-token output limit. Writing is judged blind by the instruction-tuned Gemma 4 31B in both presentation orders, and diversity uses BGE-M3 embeddings (Google DeepMind, 2026, Chen et al., 2024). A second writing study uses three 70B-class generators including R1-Distill-Llama (Guo et al., 2025).

4.2 GSM8K: THE COMPLETE TEST SPLIT

We evaluate every sampler on all 1,319 GSM8K test questions for three checkpoints, with eight answers per question. Figure 4 compares the tuning-free KL* rules with two references. One is the model-card setting that most users run. The other is the strongest default baseline for each checkpoint and metric.

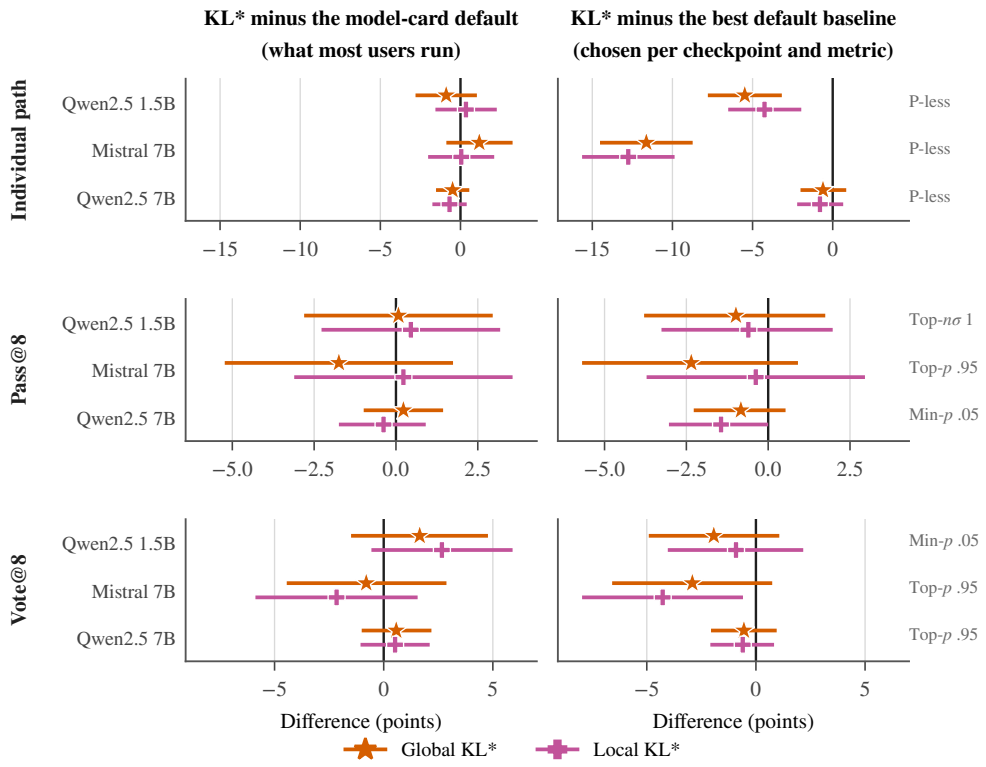


Figure 4: **Tuning-free KL* matches every model's recommended settings.** Differences on all 1,319 GSM8K test questions with Bonferroni-adjusted simultaneous 95% intervals. Left, KL* minus the model-card settings. Right, KL* minus the default baseline with the highest mean for each checkpoint and metric, named at the right. Each row shares one scale.

KL* matches the model card without tuning. All 18 intervals against the model-card settings include zero, for both rules, all three checkpoints, and all three metrics. For Qwen the model card combines a lower temperature, two truncation rules, and a repetition penalty. KL* reaches the same accuracy at infinite temperature with one rule and no parameter.

No coverage or voting loss for global KL*. Across the 42 pass@8 and Vote@8 comparisons with the seven defaults, global KL* has no significant loss and four significant gains. Both rules solve 2.4 to 9.7 points more questions than P-less, the other rule without a truncation parameter, on every checkpoint. Appendix Figure 12 shows the absolute values, with the KL rules inside the cluster of defaults and far above greedy decoding. On the 1,095 questions outside our smaller-sample studies, neither rule has a pass@8 or Vote@8 loss, and every model-card interval includes zero.

Exploration rescues the questions greedy decoding misses. A pre-registered subgroup audit splits the questions by whether greedy decoding of the same checkpoint solves them. On the questions greedy decoding gets wrong, where exploration can recover an answer, neither KL* rule has a significant loss on any checkpoint, and both solve 17 to 28 points more of them than P-less (Appendix K). Appendix A.1 lists every significant loss of the KL methods.

The candidate set decides how well uniform draws work. Uniform min-p draws uniformly from the min-p .05 set, exactly like KL* but with a fixed cutoff. In the development-tuned comparison, global KL* beats it in individual-path accuracy by 14.6, 11.9, and 2.6 points, and every simultaneous interval excludes zero (Appendix B).

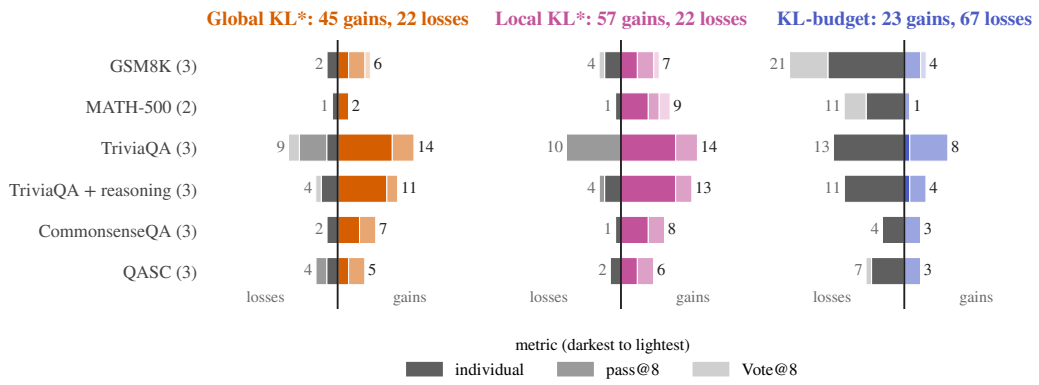


Figure 5: **Across six tasks, the KL* rules gain two to three times as often as they lose against default samplers.** Significant comparisons of each KL method with the seven defaults, with Bonferroni-adjusted simultaneous 95% intervals within each checkpoint and task. Gains are to the right and losses to the left, shaded by metric. Parentheses give the number of checkpoints. The appendix tables list every mean.

Accuracy and variation. The descriptive accuracy–variation frontier includes KL-budget on every checkpoint and local KL* at 1.5B, and in the smaller-sample studies up to 123B at least one KL method is on it in all 15 comparisons (Appendix Figures 13 and 22).

4.3 Other tasks

The same pre-registered design covers MATH-500 on the two Qwen checkpoints, TriviaQA with direct and brief-reasoning prompts, CommonsenseQA, and QASC (Hugging Face H4, 2024, Joshi et al., 2017, Talmor et al., 2019, Khot et al., 2020). Figure 5 counts the significant comparisons with the seven defaults, 357 per rule. Local KL* has 57 significant gains and 22 losses, and global KL* has 45 and 22. Most gains are in individual-path accuracy over temperature one, top-p, and top-k, and in pass@8 over P-less. On TriviaQA, local KL* has 20 significant gains in individual accuracy and 3 losses. On MATH-500 at 1.5B, local KL* has the highest pass@8 and Vote@8 of every sampler, and it also has the highest Vote@8 on CommonsenseQA at 1.5B. KL-budget has 20 pass@8 gains and no pass@8 loss (Appendices F and G).

4.4 Which tokens the rules sample

The distribution in Figure 2 is a constructed example. To see what the rules do on real distributions, we recorded the candidate-set size and the rank of every sampled token for 128 GSM8K questions and 32 writing prompts, with one generation per sampler on each checkpoint (Appendix I). Figure 6 shows how often each sampler picks a token other than the model's top choice, and how often it reaches past rank 10.

KL* explores, but stays near the top. Local KL* keeps a single candidate on 84 to 89% of GSM8K steps, where the model is confident, and opens up to a few candidates elsewhere. On GSM8K it picks a token below the top one more often than min-p and top-$n\sigma$, yet reaches past rank 10 less often than every sampler except P-less. On story writing it leaves the top token on 29% of steps, more than min-p, top-$n\sigma$, and the model card, and reaches past rank 10 on 0.9%, against 5.5% for top-p .95 and 11.9% for temperature one. KL-budget explores the most on GSM8K and the second most in writing, with a tail share close to that of global KL*. In both tasks every sampler off the frontier is dominated by a KL rule.

4.5 Robustness to high temperature

In a fixed-setting study on four larger checkpoints, each gate kept its standard setting at temperature 3 and only the order changed (Appendix Figure 30). With temperature first, accuracy collapses to 3%

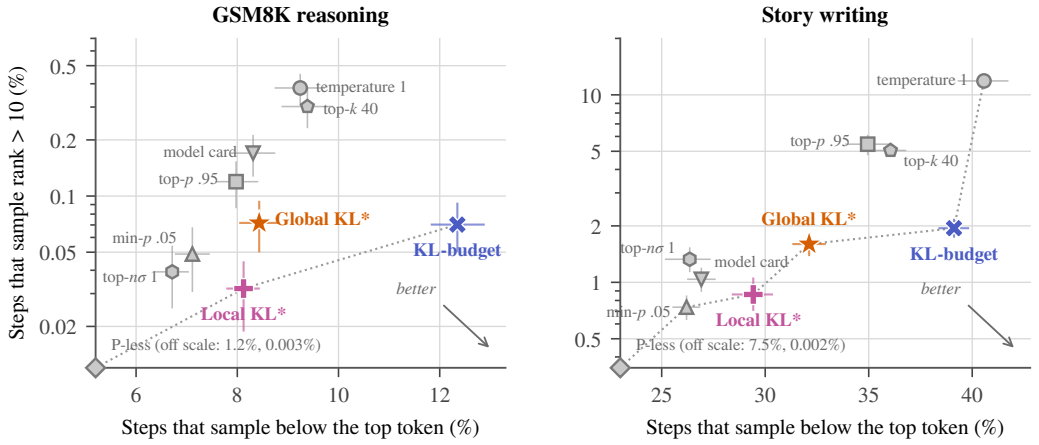


Figure 6: **KL rules explore below the top token but rarely reach the tail.** Real next-token statistics averaged over Qwen2.5 1.5B, Qwen2.5 7B, and Mistral 7B. Horizontal, steps whose sampled token is not the model's top token. Vertical (log scale), steps whose sampled token ranks 11th or lower. Bars are 95% prompt intervals, and the dotted line joins nondominated samplers. P-less stays so close to greedy decoding that it is pinned to the corner with its values.

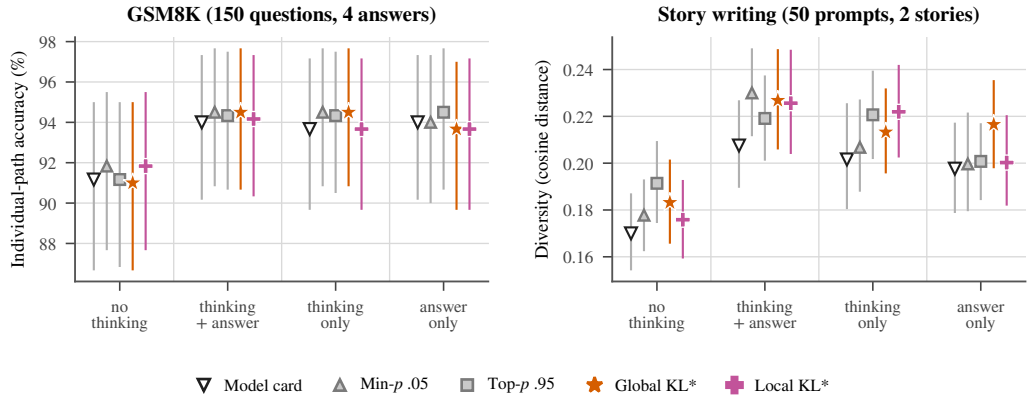


Figure 7: **With a reasoning model, where the sampler runs matters little.** Qwen3-8B at default settings. The scope says which part of the output the sampler generates, with greedy decoding for the rest. Bars are 95% question or prompt intervals (Appendix E).

or less in 8 of 12 model and gate pairs. With truncation first no pair falls below 68.8%, and global KL* reaches 81.2 to 100% with nothing to tune. This extends the fragility reported up to $T = 2$ (La Rosa, 2026) to complete collapse. These are fixed settings, not a tuned comparison.

4.6 Reasoning models

Qwen3-8B can think before answering or answer directly with the same weights (Yang et al., 2025). We ran five samplers in four scopes: no thinking, the sampler on both the thinking and the answer, on the thinking only with a greedy answer, and on the answer only after greedy thinking (Figure 7). Thinking adds about three GSM8K points for every sampler, and in every scope all samplers stay within about a point of each other, near the ceiling. None of the 24 pre-registered comparisons between a KL rule and the model card is significant. Applying KL* only to the thinking keeps its accuracy, and its stories are at least as varied as the model card's in every scope.

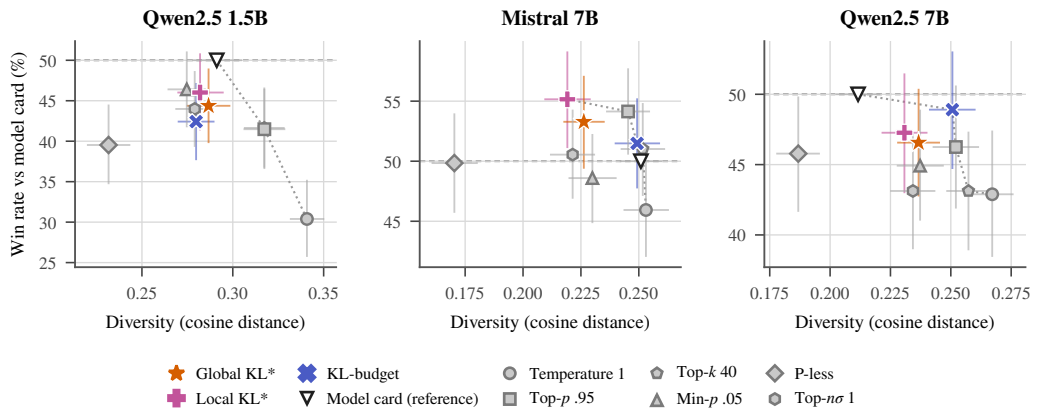


Figure 8: **The KL methods are among the closest challengers to each model's recommended settings in writing.** Win rate against the model-card story for the same prompt and seed, judged blind in both orders on 160 prompts and two seeds, against the diversity of each sampler's four stories per prompt. Bars are 95% prompt intervals, the dashed line marks a tie, and the dotted line joins nondominated samplers (Appendix D).

4.7 Creative writing

A pre-registered study asks each checkpoint for four stories on each of 160 WritingPrompts prompts. A blinded Gemma 4 31B judge compares each story with the model-card story for the same prompt and seed, in both presentation orders (Figure 8). Among the challengers, local KL* has the highest win rate on Mistral 7B, 55.2%, and the three KL methods are the three closest challengers on Qwen2.5 7B. Neither KL* rule loses a comparison on any checkpoint, and both beat temperature one significantly on two. KL-budget writes more varied stories than the KL* rules on Mistral 7B and Qwen2.5 7B, where it sits on the observed frontier. A second study with three 70B-class generators found KL-budget ahead of min-p on R1-Distill-Llama 70B in 68.8% of comparisons (Appendix S).

5 Conclusion

Infinite-temperature sampling works when the model chooses the candidates. Global KL* finds the closest uniform draw exactly and needs no tuning. At default settings both KL* rules match every model's recommended settings on GSM8K, gain two to three times as often as they lose across six tasks, explore below the top token while rarely reaching the tail, and stay within a few points of baselines tuned on development data.

Choosing a rule. Use global or local KL* when settings cannot be tuned and a verifier or a vote picks among attempts. When a single answer must be right, a conservative sampler such as P-less is more accurate per answer (Appendix A.1). KL-budget explores the most and has no significant pass@8 loss. The comparisons do not consistently favor global or local KL*.

Limitations

We evaluate quantized checkpoints in one inference engine and do not tune KL*. Writing uses an automated judge. Appendix A discusses all limitations.

Reproducibility statement

Appendix M defines every sampler and gives pseudocode for both KL* rules, and Appendix N proves the formal claims. Appendices C and P give the models, prompts, seeds, sampler settings, answer extraction, and statistical procedures, and Appendix L describes the timing setup. The release provides model and dataset revisions, weight checksums, all development and evaluation generations,

and scripts for every table and figure. Tests cover candidate selection, budget feasibility, exact uniform draws, sampler parity, answer extraction, and voting.

AI USE STATEMENT

Generative AI tools were used throughout the preparation of this work. Gemma 4 31B was used for the automated writing judgments described in Section 4.7. AI-assisted work was reviewed and validated by the authors, who take responsibility for the final work.

REFERENCES

Erfan Baghaei Potraghloo, Seyedarmin Azizi, Souvik Kundu, and Massoud Pedram. Top-H decoding: Adapting the creativity and coherence with bounded entropy in text generation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 28482–28513. Curran Associates, Inc., 2025. doi: 10.52202/085713-0958. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/294de0fa7149adcb88aa3119c239c63e-Paper-Conference.pdf.

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*, 2021. URL https://openreview.net/forum?id=W1G1JZEIy5_.

Chung-Ching Chang, David Reitter, Renat Aksitov, and Yun-Hsuan Sung. KL-divergence guided temperature sampling, 2023. URL https://arxiv.org/abs/2306.01286.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

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, 2021. URL https://arxiv.org/abs/2110.14168.

Yuanhao Ding, Meimingwei Li, Esteban Garces Arias, Matthias Aßenmacher, Christian Heumann, and Chongsheng Zhang. Min-k sampling: Decoupling truncation from temperature scaling via relative logit dynamics. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14932–14948, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.681. URL https://aclanthology.org/2026.acl-long.681/.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled AlpacaEval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

EleutherAI. Language model evaluation harness: GSM8K chain-of-thought evaluation, 2026. URL https://github.com/EleutherAI/lm-evaluation-harness/blob/ad8737ae7fad24cf64e50fc7fc31397bff586b9e/lm_eval/tasks/gsm8k/gsm8k-cot.yaml. Pinned task configuration and extraction rules.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082/.

Google DeepMind. Gemma 4 31B-it model card, 2026. URL https://huggingface.co/google/gemma-4-31B-it. Instruction-tuned checkpoint, model revision 842da3794eaa0b77d5f08bae87a17459d91ff475, run as the Q8_0 GGUF conversion unsloth/gemma-4-31B-it-GGUF revision c1ac76e99d5513b141e8adde7288b85c3f9c32ec, accessed September 13, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

John Hewitt, Christopher Manning, and Percy Liang. Truncation sampling as language model desmoothing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.249. URL https://aclanthology.org/2022.findings-emnlp.249/.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Yu-Shin Huang, Peter Just, Hanyun Yin, Krishna Narayanan, Ruihong Huang, and Chao Tian. ODStega: LLM-based relatively secure steganography via optimized distributions. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 827–851, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.36. URL https://aclanthology.org/2026.eacl-long.36/.

Hugging Face H4. MATH-500, 2024. URL https://huggingface.co/datasets/HuggingFaceH4/MATH-500. Revision 6e4ed1a2a79af7d8630a6b768ec859cb5af4d3be, accessed September 13, 2026.

E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, May 1957. ISSN 0031-899X. doi: 10.1103/PhysRev.106.620. URL https://doi.org/10.1103/PhysRev.106.620.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. QASC: A dataset for question answering via sentence composition. *Proceedings of the AAAI Conference on Artificial*

Intelligence, 34(05):8082–8090, April 2020. ISSN 2159-5399. doi: 10.1609/aaai.v34i05.6319. URL http://dx.doi.org/10.1609/aaai.v34i05.6319.

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, Koblenz, Germany, October 2023. Association for Computing Machinery. doi: 10.1145/3600006.3613165.

Francesco La Rosa. Temperature fragility and the conditional benefits of truncation sampling, 2026. URL https://arxiv.org/abs/2609.15476.

Zhenwen Liang, Yujun Zhou, Sidi Lu, Xiangliang Zhang, Haitao Mi, and Dong Yu. Too correct to learn: Reinforcement learning on saturated reasoning data. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 205–215, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-391-3. doi: 10.18653/v1/2026.acl-short.19. URL https://aclanthology.org/2026.acl-short.19/.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

llama.cpp contributors. Default sampler chain in llama.cpp, 2026. URL https://github.com/ggml-org/llama.cpp/blob/4d9176092d00586775af140581bb0b558ddc4389/common/common.h. Pinned commit 4d9176092d00586775af140581bb0b558ddc4389, common_params_sampling, accessed September 13, 2026.

Clara Meister, Tiago Pimentel, Luca Malagutti, Ethan Wilcox, and Ryan Cotterell. On the efficacy of sampling adapters. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1437–1455, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.80. URL https://aclanthology.org/2023.acl-long.80/.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023b. doi: 10.1162/tacl_a_00536. URL https://aclanthology.org/2023.tacl-1.7/.

Meta. Llama 3.3 70B Instruct model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

Mistral AI. Mistral-7B-Instruct-v0.3 model card, 2024a. URL https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. Specific checkpoint documentation, accessed September 2026.

Mistral AI. Mistral Large Instruct 2411 model card, 2024b. URL https://huggingface.co/mistralai/Mistral-Large-Instruct-2411.

Mistral AI. Mistral Small 4 119B 2603 model card, 2026. URL https://huggingface.co/mistralai/Mistral-Small-4-119B-2603/tree/a11f36bebf709121056b1dbcc943d1c6afbe494d.

Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 70333–70366, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/afa5f124e36bed5cc2125067005d43f5-Paper-Conference.pdf.

Georgy Noarov, Soham Mallick, Tao Wang, Sunay Joshi, Yan Sun, Yangxinyu Xie, Mengxin Yu, and Edgar Dobriban. Foundations of Top-k decoding for language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 116713–116756. Curran Associates, Inc., 2025. doi: 10.52202/085713-3890. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a96d4fda3017f1773b261a52a3efc8dd-Paper-Conference.pdf.

Qwen Team. Qwen3.8-27B model card, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B. Model revision 1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0, accessed September 13, 2026.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Runyan Tan, Shuang Wu, and Phillip Howard. p-less sampling: A robust hyperparameter-free approach for LLM decoding. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ItFuNJQGH4.

Chenxia Tang, Jianchun Liu, Hongli Xu, and Liusheng Huang. Top-$n\sigma$: Eliminating noise in logit space for robust token sampling of LLM. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10758–10774, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.528. URL https://aclanthology.org/2025.acl-long.528/.

Luke Vilnis, Yury Zemlyanskiy, Patrick Murray, Alexandre Tachard Passos, and Sumit Sanghai. Arithmetic sampling: Parallel diverse decoding for large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35120–35136. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/vilnis23a.html.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Shanli Xing, Zihao Ye, Bohan Hou, Luis Ceze, and Tianqi Chen. Sorting-free GPU kernels for LLM sampling. FlashInfer blog, March 2025. URL https://flashinfer.ai/2025/03/10/sampling.html. Accessed September 23, 2026.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin

Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zihao Ye, Lequn Chen, Ruihang Lai, Wuwei Lin, Yineng Zhang, Stephanie Wang, Tianqi Chen, Baris Kasikci, Vinod Grover, Arvind Krishnamurthy, and Luis Ceze. FlashInfer: Efficient and customizable attention engine for LLM inference serving. In M. Zaharia, G. Joshi, and Y. Lin (eds.), *Proceedings of Machine Learning and Systems*, volume 7. MLSys, 2025. URL https://proceedings.mlsys.org/paper_files/paper/2025/file/dbf02b21d77409a2db30e56866a8ab3a-Paper-Conference.pdf.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. Trading off diversity and quality in natural language generation. In Anya Belz, Shubham Agarwal, Yvette Graham, Ehud Reiter, and Anastasia Shimorina (eds.), *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pp. 25–33, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.humeval-1.3/.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. SGLang: Efficient execution of structured language model programs. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 62557–62583. Curran Associates, Inc., 2024. doi: 10.52202/079017-2000. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/724be4472168f31ba1c9ac630f15dec8-Paper-Conference.pdf.

A Limitations and Where the KL Methods Lose

A.1 Where the KL methods lose

Default settings. Against the seven defaults, global KL* has 2 significant losses in 63 comparisons, local KL* has 4, and KL-budget has 21. Both global losses are in individual-path accuracy against P-less, by 5.5 points at 1.5B and 11.6 points on Mistral 7B. Local KL* trails P-less by 4.3 and 12.8 points on the same two checkpoints, trails top-$n\sigma$ by 2.9 points on Mistral 7B, and trails top-p .95 by 4.3 Vote@8 points on Mistral 7B. Neither rule loses anything on Qwen2.5 7B. KL-budget at .1 nats explores more than any other sampler on GSM8K (Section 4.4). It has 14 losses in individual accuracy and 7 in Vote@8, 13 of them on Mistral 7B, and none in pass@8. The KL* rules have 6 and 7 significant gains, and KL-budget has 4.

Other tasks and writing. In the six-task census (Figure 5), most losses of the KL* rules are in individual accuracy against the two most conservative samplers, P-less and the model card, or in pass@8 against more exploratory ones. KL-budget has 55 losses in individual accuracy. In writing at default settings no sampler beats the model card significantly. On Qwen2.5 1.5B, whose model card adds a repetition penalty, every sampler trails it, and KL-budget trails it significantly (Appendix D).

Development-tuned comparison. Individual-path accuracy separates methods more often. The KL* rules beat temperature-only sampling by 5.8 to 8.4 points at 1.5B and Qwen 7B but trail it by 4.3 points on Mistral 7B. They trail temperature-first min-p by 2.5 to 5.3 points at 1.5B and Mistral 7B

and beat it by 3.1 to 3.3 points at Qwen 7B. Truncation-first P-less has the best individual accuracy of any sampler on all three checkpoints and the lowest pass@8. KL-budget has 20 of the 34 significant KL losses, 15 of them in individual accuracy and none in pass@8. On Mistral 7B, truncation-first top-p beats its Vote@8 by 6.9 points (adjusted interval 3.1 to 10.7). The 128-question subset study (Appendix Q.1) had ranked the KL* rules first in 1.5B pass@8, and the complete test split does not reproduce that order.

The full census (Figure 17) shows where these losses fall. Global KL* has 8 significant losses in 72 comparisons, local KL* has 6, and KL-budget has 20. Every loss of the two KL* rules is at 1.5B or Mistral 7B, and all but one are in individual-path accuracy. Tuned baselines keep the highest pass@8 mean on every checkpoint, leading global KL* by 3.2, 2.0, and 0.9 points. The descriptive accuracy–variation frontiers include both KL* rules at 1.5B, only global KL* at Qwen 7B, and neither rule at Mistral 7B.

A.2 Limitations

Models and implementation. All comparisons use quantized checkpoints and one inference engine. Training histories, weight precision, and the total and active parameter counts of Mistral Small 4 differ across checkpoints, so the model range does not isolate a scaling law. GPU batching can change logits under a fixed seed.

Baselines and tuning. The tuning grids do not exhaust either sampling order. Top-$n\sigma$ and min-k inform the motivation but are not evaluated baselines. The subset studies fix truncation before temperature, and only the complete GSM8K test split also tunes temperature-first baselines. Settings were chosen by pass@8 on 24 development questions, and the other metrics use the same settings. We do not tune KL*. Its generalized objective defines a family whose candidate sets grow with one parameter (Appendix N). A descriptive sweep of that parameter (Appendix H) suggests that smaller values raise individual accuracy on the weaker checkpoints, but it reuses test questions, so no tuned variant is evaluated.

Scoring. The GSM8K answer extractor changed after a pilot and before any confirmation question was scored (Appendix P). Output limits and extraction affect correctness, and every failure remains a measured outcome.

Writing. Writing preferences come from one automated judge. The 70B-class study is small and showed substantial presentation-order sensitivity, which averaging both orders mitigates. Neither writing study establishes human preference or creativity.

Scope of the theory. A better token-level divergence need not improve text. The KL-budget allowance excludes the divergence added by truncation. The comparison of global and local KL* concerns one next-token distribution, whereas generated paths visit different contexts. Frontier membership uses point estimates and is not a significance test.

B The Development-Tuned Comparison

We also compare the KL rules with temperature, top-p, min-p, and P-less, each tuned on development data. Each truncation baseline runs with truncation before temperature and, separately, after it, which includes the published P-less order. Each family and KL-budget choose one setting per checkpoint from three to five candidates by pass@8 on 24 development questions, frozen before testing (Appendix P). Global and local KL* have nothing to choose. Uniform min-p draws uniformly from the min-p .05 set and isolates the effect of the candidate set. The evaluation covers the same 1,319 GSM8K test questions and three checkpoints, 352,173 responses in total, with 72 planned comparisons per checkpoint. Figure 9 places the KL* rules between uniform min-p and the strongest tuned baseline.

The candidate set decides how well uniform draws work. Uniform min-p and the KL* rules use the same equal-weight draw and differ only in the candidates. A min-p cutoff of .05 gives a token with one-twentieth of the top probability the same chance as the top token. KL* admits such a token

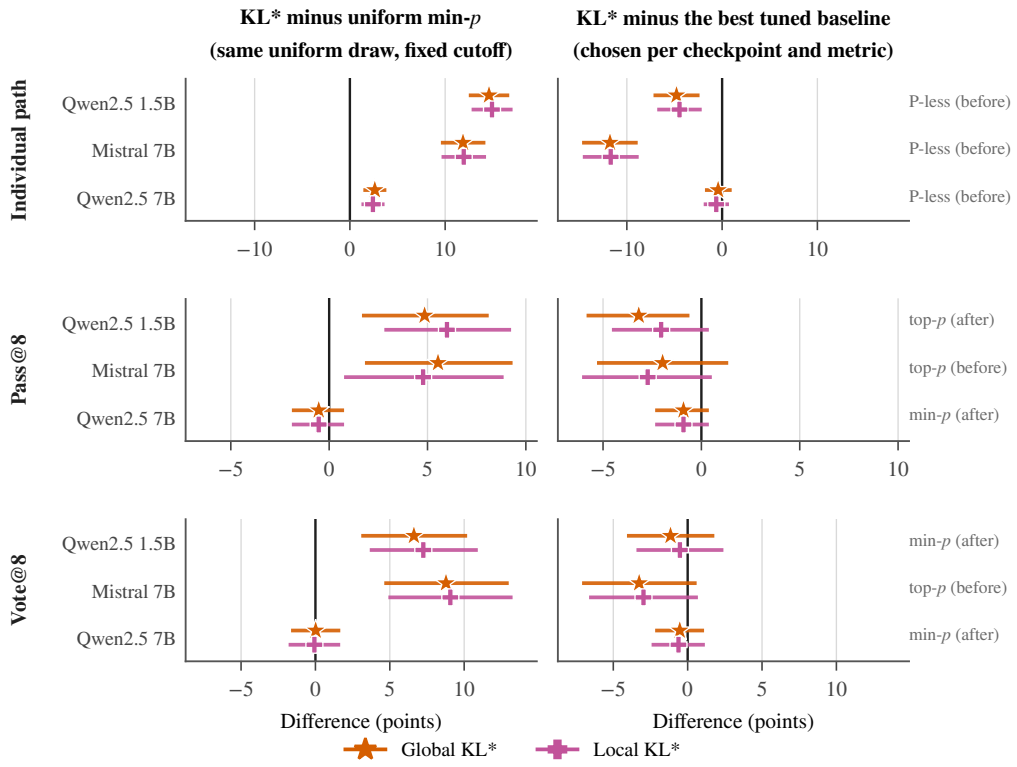


Figure 9: **Tuning-free KL* far outperforms a fixed-cutoff uniform draw and stays close to the best tuned sampler.** Differences on all 1,319 GSM8K test questions with Bonferroni-adjusted simultaneous 95% intervals. Left, KL* minus uniform min-p, which draws uniformly from the min-p .05 set. Right, KL* minus the tuned baseline with the highest mean for each checkpoint and metric, named at the right. Each row shares one scale.

only when it lowers the score. Global KL* improves individual-path accuracy by 14.6, 11.9, and 2.6 points, and every simultaneous interval excludes zero. On the two weaker checkpoints it also improves pass@8 by 4.9 and 5.5 points and Vote@8 by 6.6 and 8.8 points.

Without tuning, KL* stays close to the best tuned sampler. In pass@8 and Vote@8, the gap between KL* and the best tuned baseline is at most 3.3 points, and 11 of the 12 intervals include zero. Across all 48 pass@8 and Vote@8 comparisons with the eight baselines, local KL* has no significant loss, and global KL* has one. Both rules solve 3.1 to 9.1 points more questions than truncation-first P-less on every checkpoint. The same pattern holds on the 1,095 previously unused questions. Figure 10 shows the absolute values. Figure 11 plots individual-path accuracy against response variation at 1.5B, where both KL* rules are on the observed frontier. Global KL* is also on the frontier at Qwen 7B (Appendix P.1).

C Default-Settings Evaluation

Every default-settings protocol was fixed before any of its generations began. Protocol files, prompts, seeds, raw responses, and analysis code are released. All methods for a question ran on one server with two H100 SXM GPUs, using the pinned research build of `llama.cpp` from the tuned evaluation (identical source archive, native sampler tests passed before generation), Q8_0 weights, and Q8_0 caches.

Settings. Each baseline uses a published default rather than a development-tuned setting. Temperature one alone, top-p .95, top-k 40, min-p .05, and top-$n\sigma$ 1 at temperature one follow the standard

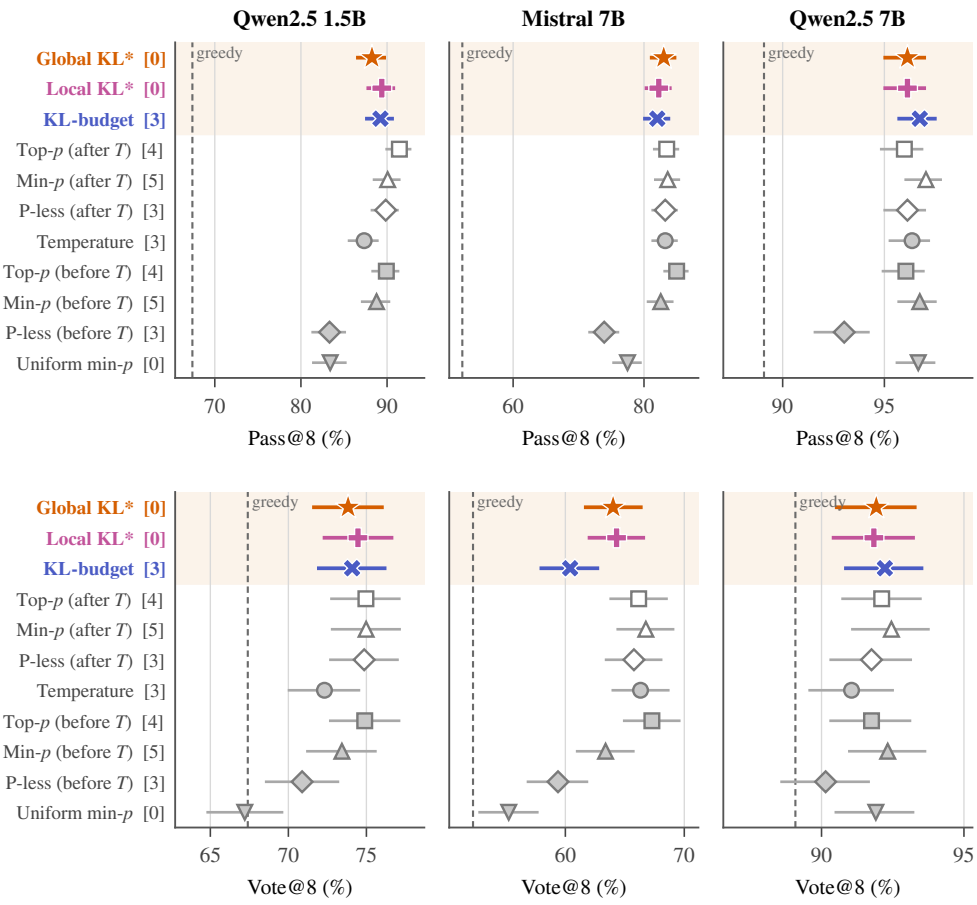


Figure 10: **In the tuned comparison, the KL rules sit inside the cluster of tuned samplers.** Pass@8 and Vote@8 on all 1,319 GSM8K test questions with 95% intervals. Brackets give the number of development settings searched. The dashed line is greedy decoding. Filled gray markers truncate before temperature and hollow markers after it.

settings used by La Rosa (2026). P-less has no truncation parameter and runs at temperature one. At temperature one the order of truncation and temperature is immaterial. The model-card setting is each checkpoint's official `generation_config.json`, applied in Transformers order: repetition penalty, temperature, top-k, top-p. Qwen2.5 1.5B uses $T = .7$, top-p .8, top-k 20, and repetition penalty 1.1. Qwen2.5 7B uses the same settings with penalty 1.05. The Mistral 7B configuration specifies no sampling fields, so the Transformers defaults with sampling enabled apply, $T = 1$ and top-k 50. Repetition penalties cover the whole context, as in Transformers. The KL* rules keep their canonical settings, and KL-budget uses its canonical allowance of .1 nats on min-p .05. Prompts, the eight seeds, and the 1,536-token limit are those of the tuned evaluation.

Statistics. Within each checkpoint, the planned family compares three KL methods with seven default baselines at three endpoints, 63 contrasts in total. Intervals are paired question bootstraps, with Bonferroni-adjusted simultaneous intervals over the family from one million joint question resamples, computed with the analysis code of the tuned evaluation unchanged. Every contrast is reported, and Table 1 gives every mean with its interval.

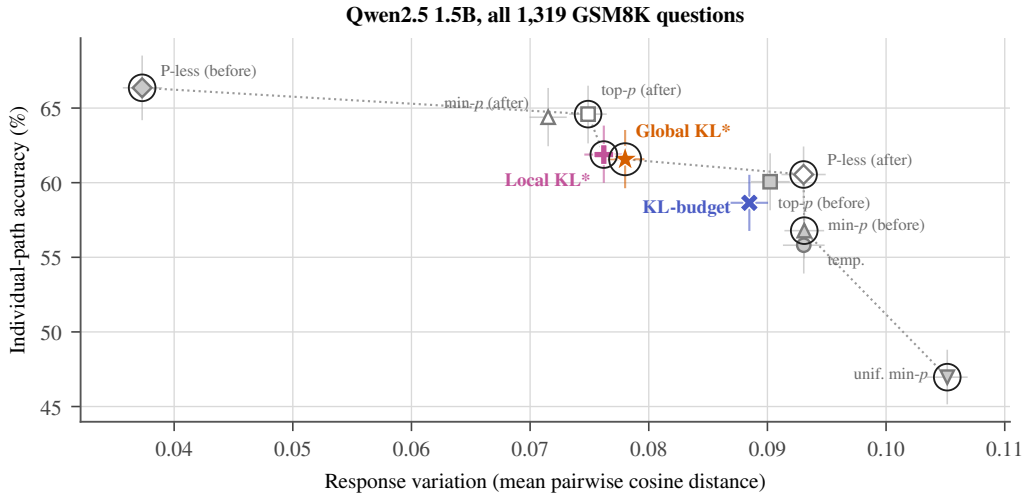


Figure 11: **Both KL* rules are on the 1.5B accuracy–variation frontier.** All eleven samplers on all 1,319 questions with eight responses each. Circles mark nondominated means, joined by the dotted frontier. Bars are 95% question intervals. Frontier membership is descriptive and is not a significance test.

Table 1: **Complete default-settings outcomes** on all 1,319 GSM8K test questions, with 95% question intervals. Pass@8 and greedy use Wilson intervals, and other means use question bootstraps. Bold marks the largest mean in each checkpoint and column, with ties all bold.

Method	Individual	Pass@8	Vote@8
Qwen2.5 1.5B ($n = 1319$)			
Greedy	**67.2** [64.7, 69.7]	67.2 [64.7, 69.7]	67.2 [64.7, 69.8]
Model card	62.1 [60.1, 64.0]	89.4 [87.6, 90.9]	72.1 [69.7, 74.5]
Temperature 1	56.5 [54.6, 58.3]	88.1 [86.2, 89.7]	73.1 [70.8, 75.3]
Top-p .95	59.6 [57.8, 61.5]	90.2 [88.5, 91.7]	75.5 [73.3, 77.7]
Top-k 40	55.4 [53.6, 57.3]	88.4 [86.6, 90.0]	73.5 [71.2, 75.8]
Min-p .05	61.4 [59.5, 63.3]	89.7 [87.9, 91.2]	**75.7** [73.4, 77.9]
P-less	66.7 [64.5, 68.8]	84.5 [82.5, 86.4]	71.3 [68.9, 73.6]
Top-$n\sigma$ 1	61.9 [60.0, 63.8]	**90.4** [88.7, 91.9]	75.1 [72.8, 77.3]
Global KL*	61.2 [59.2, 63.1]	89.5 [87.7, 91.0]	73.7 [71.4, 76.0]
Local KL*	62.4 [60.5, 64.3]	89.8 [88.1, 91.4]	74.7 [72.4, 77.0]
KL-budget	55.9 [54.0, 57.8]	89.4 [87.6, 90.9]	72.5 [70.2, 74.7]
Mistral 7B ($n = 1319$)			
Greedy	**51.3** [48.6, 54.0]	51.3 [48.6, 54.0]	51.3 [48.6, 54.1]
Model card	38.3 [36.8, 39.8]	83.3 [81.2, 85.2]	65.3 [62.9, 67.8]
Temperature 1	39.3 [37.8, 40.9]	83.8 [81.7, 85.7]	65.1 [62.6, 67.5]
Top-p .95	40.0 [38.5, 41.5]	**83.9** [81.8, 85.8]	**67.4** [65.0, 69.9]
Top-k 40	37.6 [36.2, 39.1]	83.8 [81.7, 85.7]	64.5 [62.0, 67.0]
Min-p .05	40.0 [38.4, 41.5]	82.4 [80.3, 84.4]	65.1 [62.6, 67.5]
P-less	51.1 [48.9, 53.3]	73.8 [71.4, 76.1]	59.9 [57.3, 62.4]
Top-$n\sigma$ 1	41.2 [39.6, 42.8]	83.2 [81.1, 85.1]	65.4 [62.9, 67.8]
Global KL*	39.5 [37.9, 41.0]	81.6 [79.4, 83.6]	64.5 [62.0, 67.0]
Local KL*	38.4 [36.8, 39.9]	83.5 [81.5, 85.5]	63.2 [60.7, 65.6]
KL-budget	33.0 [31.6, 34.3]	80.6 [78.4, 82.6]	60.6 [58.1, 63.0]
Qwen2.5 7B ($n = 1319$)			
Greedy	89.2 [87.4, 90.8]	89.2 [87.4, 90.8]	89.2 [87.6, 90.9]
Model card	89.4 [88.0, 90.7]	95.9 [94.7, 96.8]	91.4 [89.9, 92.8]
Temperature 1	88.4 [87.0, 89.7]	96.5 [95.4, 97.4]	92.1 [90.6, 93.4]
Top-p .95	88.4 [87.0, 89.7]	96.4 [95.3, 97.3]	92.5 [91.1, 93.8]
Top-k 40	88.1 [86.7, 89.5]	96.3 [95.1, 97.2]	91.9 [90.4, 93.4]

	Qwen2.5 1.5B		Mistral 7B		Qwen2.5 7B	
Sampler	Win (%)	Diversity	Win (%)	Diversity	Win (%)	Diversity
Global KL*	44.4 [40, 49]	0.287	53.3 [49, 57]	0.226	46.6 [43, 50]	0.237
Local KL*	46.0 [41, 51]	0.282	**55.2** [51, 59]	0.219	47.3 [43, 51]	0.231
KL-budget	42.4 [38, 47]	0.280	51.5 [48, 55]	0.249	**48.9** [45, 53]	0.251
Temperature 1	30.4 [26, 35]	**0.341**	45.9 [42, 50]	**0.253**	42.9 [38, 47]	**0.267**
Top-p .95	41.5 [37, 46]	0.317	54.1 [51, 58]	0.245	46.2 [42, 51]	0.252
Top-k 40	41.6 [37, 47]	0.317	51.0 [47, 55]	0.252	43.1 [39, 47]	0.257
Min-p .05	**46.4** [42, 51]	0.275	48.6 [45, 52]	0.230	44.9 [41, 49]	0.237
P-less	39.5 [35, 45]	0.232	49.8 [46, 54]	0.170	45.8 [42, 50]	0.187
Top-$n\sigma$ 1	44.0 [39, 49]	0.279	50.5 [47, 54]	0.221	43.1 [39, 47]	0.234
Model card (reference)	50	0.291	50	0.251	50	0.212

Table 2: **Writing at default settings.** Win rate against the model-card story for the same prompt and seed, with 95% prompt intervals, judged blind in both orders on 160 prompts and two seeds. Diversity is the mean cosine distance between a sampler's four stories per prompt. Bold marks the largest value among the challengers in each column.

Method	Individual	Pass@8	Vote@8
Min-p .05	88.9 [87.5, 90.2]	**97.0** [95.9, 97.8]	91.9 [90.5, 93.3]
P-less	**89.5** [87.9, 91.0]	93.1 [91.6, 94.3]	90.3 [88.7, 91.8]
Top-$n\sigma$ 1	89.3 [87.9, 90.6]	96.5 [95.4, 97.4]	92.1 [90.7, 93.5]
Global KL*	88.9 [87.5, 90.2]	96.1 [95.0, 97.0]	91.9 [90.5, 93.3]
Local KL*	88.7 [87.3, 90.0]	95.5 [94.3, 96.5]	91.9 [90.4, 93.3]
KL-budget	88.0 [86.7, 89.3]	96.9 [95.8, 97.7]	**92.7** [91.3, 94.0]

Reproducibility. Global and local KL* ran again at settings identical to the tuned evaluation. The generated texts differ because GPU batching changes logits under a fixed seed, but the aggregates agree within 1.4 points on every checkpoint and metric. On Qwen2.5 7B, global KL* scores 96.1% pass@8 and 91.9% Vote@8 in both runs.

Accuracy and variation. The rule for this secondary analysis was written before it was computed. It uses the variation measure of the tuned evaluation on all ten stochastic samplers and all 1,319 questions for every checkpoint. Figure 13 shows the result. Frontier membership is descriptive, and many samplers lie close to one trade-off line.

D Writing Study at Default Settings

The protocol was frozen before generation. Each checkpoint writes four stories, seeds 5100 to 5103, for each of the 160 prompts of the frozen WritingPrompts pool, with a 768-token limit and each sampler at its default setting. The judge compares the stories of seeds 5100 and 5101 with the model-card stories of the same prompt and seed, in both presentation orders, using the judging protocol of the 70B-class study (Appendix S) unchanged. On the same judge build it passed every held-out control, preferring the intact story in every degraded pair, tying every identical pair, and rating every empty story lowest, before any preference was accepted. A win counts one, a tie one half. Orders and seeds average within a prompt, and intervals resample prompts. The planned family compares each KL method with each default, 21 contrasts per checkpoint, with Bonferroni-adjusted simultaneous intervals. Table 2 gives every win rate and diversity value. The significant contrasts are the two KL* rules and KL-budget over temperature one at 1.5B, the two KL* rules over temperature one on Mistral 7B, and KL-budget below the model card at 1.5B. A secondary comparison against min-p .05 follows the same protocol.

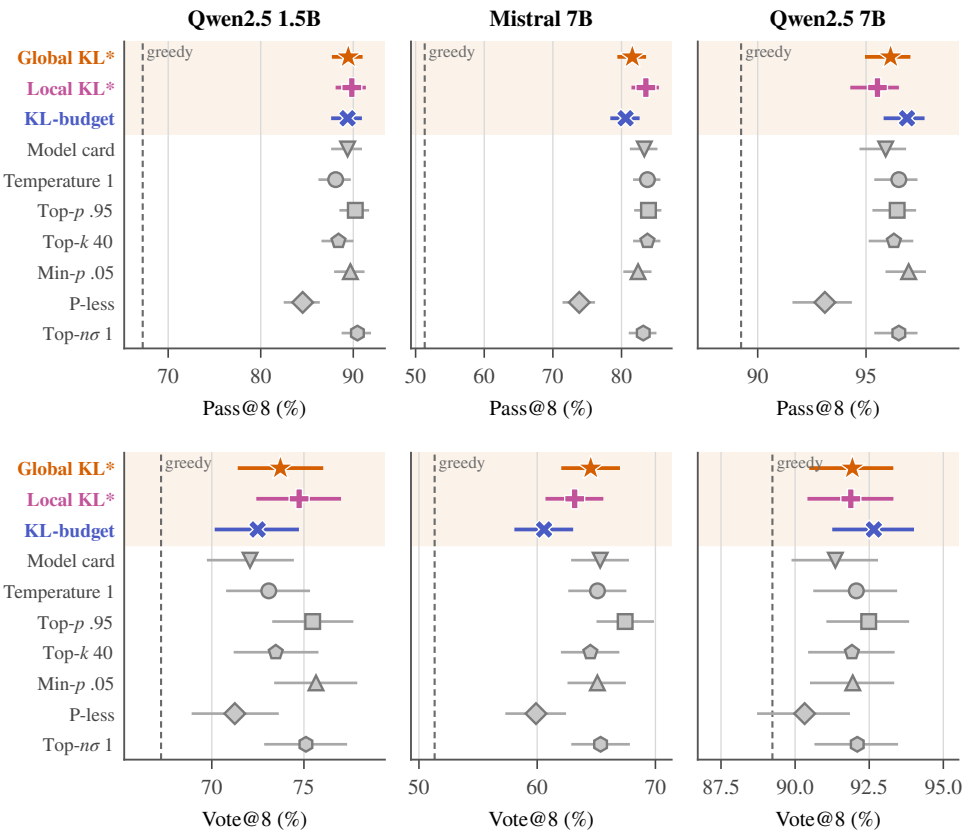


Figure 12: **The KL rules sit inside the cluster of default samplers.** Pass@8 and Vote@8 on all 1,319 GSM8K test questions with 95% intervals. The dashed line is greedy decoding. No setting was tuned.

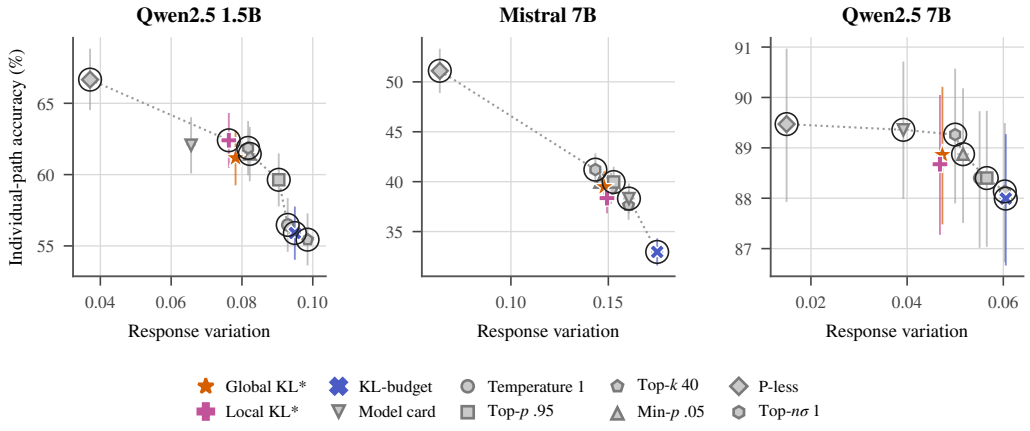


Figure 13: **Accuracy against response variation at default settings.** All ten stochastic samplers on all 1,319 GSM8K questions with eight responses each. Variation is the mean pairwise cosine distance between responses. Circles mark nondominated means, joined by the dotted frontier. Bars are 95% question intervals.

Sampler	Individual	Pass@4	Vote@4	Story diversity	Thinking tokens
No thinking					
Model card	91.2	93.3	92.0	0.170	0
Min-p .05	**91.8**	**95.3**	92.5	0.178	0
Top-p .95	91.2	94.0	92.0	**0.191**	0
Global KL*	91.0	94.0	91.5	0.183	0
Local KL*	**91.8**	94.0	**92.7**	0.176	0
Thinking + answer					
Model card	94.0	95.3	**95.0**	0.208	614
Min-p .05	**94.5**	**96.0**	94.7	**0.230**	646
Top-p .95	94.3	95.3	94.7	0.219	647
Global KL*	**94.5**	95.3	**95.0**	0.227	609
Local KL*	94.2	95.3	**95.0**	0.226	632
Thinking only					
Model card	93.7	94.7	94.0	0.202	617
Min-p .05	**94.5**	**95.3**	**94.7**	0.207	670
Top-p .95	94.3	**95.3**	**94.7**	0.221	667
Global KL*	**94.5**	**95.3**	**94.7**	0.213	615
Local KL*	93.7	94.7	94.0	**0.222**	623
Answer only					
Model card	94.0	**95.3**	94.7	0.198	638
Min-p .05	94.0	94.7	94.3	0.200	632
Top-p .95	**94.5**	**95.3**	**95.0**	0.201	631
Global KL*	93.7	94.7	94.7	**0.217**	634
Local KL*	93.7	**95.3**	93.7	0.200	620

Table 3: **Qwen3-8B with and without thinking.** GSM8K accuracy (%) on the first 150 test questions with four answers each, story diversity on the first 50 WritingPrompts prompts with two stories each, and mean GSM8K thinking length. The scope says which part of the output the sampler generates, with greedy decoding for the rest. Bold marks the largest value within each scope and column.

E Reasoning Ablation

The protocol was frozen before generation. Qwen3-8B runs the model card (`generation_config.json`: $T = .6$, top-k 20, top-p .95), min-p .05, top-p .95, and the two KL* rules on the first 150 GSM8K test questions with four seeds and on the first 50 WritingPrompts prompts with two seeds. Each thinking scope runs as two requests with the same prefix, with budgets of 3,584 thinking and 1,024 answer tokens for GSM8K and 2,048 and 1,024 for writing. A thinking budget that runs out is closed with `</think>` under one rule for every scope, which happened in 1.2 to 2.8% of GSM8K paths. GSM8K answers after the thinking use the primary extractor. The planned family compares each KL* rule with the model card in each scope on the three GSM8K metrics, 24 contrasts with Bonferroni adjustment, and none is significant. Table 3 gives every value.

F TriviaQA: Factual Recall With and Without Reasoning

A protocol frozen before generation draws 1,000 questions at random from the TriviaQA `rc.nocontext` validation split with seed 20260923 (Joshi et al., 2017). Each question is asked twice with each checkpoint's chat template, once for only a short answer (64 tokens) and once for brief reasoning ending in a marked final answer. Answers are scored by exact match against every released alias after the standard TriviaQA normalization. Each sampler draws eight answers per question with the seeds of the GSM8K evaluation, and each checkpoint and prompt has its own family of 63 comparisons with the same simultaneous intervals.

Tables 4 and 5 give every mean. Across the six checkpoint and prompt combinations, the KL* rules raise individual-path accuracy above most default samplers. Local KL* has 20 significant gains and 3 losses in individual accuracy, and global KL* has 19 gains and 5 losses. More exploratory samplers cover more questions. Local KL* has 7 gains and 11 losses in pass@8, most of the losses with direct answers, and global KL* has 6 and 5. KL-budget has 10 pass@8 gains and no pass@8 loss, but 24

	Qwen2.5 1.5B			Mistral 7B			Qwen2.5 7B		
Sampler	I	P	V	I	P	V	I	P	V
Greedy	**27.8**	27.8	27.8	**54.3**	54.3	54.3	**47.0**	47.0	47.0
Model card	25.1	35.4	26.9	50.2	**64.6**	**55.7**	46.0	52.8	47.5
Temperature 1	22.2	41.3	27.4	51.0	62.3	54.7	42.6	56.7	47.2
Top-p .95	21.4	**42.4**	27.5	50.9	63.5	55.6	41.9	56.7	47.3
Top-k 40	20.2	41.2	**28.0**	50.1	64.3	55.1	41.6	**58.5**	47.6
Min-p .05	24.1	40.3	27.6	51.7	61.5	54.4	44.5	56.3	**47.8**
P-less	26.7	32.2	27.1	53.9	55.2	53.8	46.7	49.6	47.4
Top-$n\sigma$ 1	23.8	40.7	27.7	52.2	60.2	54.3	44.9	55.8	**47.8**
Global KL*	24.6	38.5	27.1	52.4	60.1	53.9	44.6	54.3	45.4
Local KL*	24.8	38.3	26.8	52.5	59.6	54.2	44.8	53.6	45.9
KL-budget	22.2	41.3	26.7	49.4	63.9	54.4	42.5	56.8	46.3

Table 4: **TriviaQA with a direct answer.** 1,000 validation questions, eight answers each, default settings. I is individual-path accuracy, P is pass@8, and V is Vote@8, in percent. Bold marks the largest value in each column, with ties all bold.

	Qwen2.5 1.5B			Mistral 7B			Qwen2.5 7B		
Sampler	I	P	V	I	P	V	I	P	V
Greedy	31.3	31.3	31.3	**50.8**	50.8	50.8	**54.5**	54.5	54.5
Model card	29.8	51.0	**37.5**	41.8	68.6	56.9	53.6	68.9	57.6
Temperature 1	23.1	49.7	34.0	41.9	69.1	55.6	50.5	70.8	58.3
Top-p .95	24.2	49.7	35.0	43.5	69.0	56.8	51.3	69.8	58.5
Top-k 40	23.7	51.3	35.1	41.1	68.3	55.8	50.6	70.8	57.5
Min-p .05	27.7	**53.1**	35.7	44.4	**69.7**	56.3	52.1	71.4	58.5
P-less	**31.9**	47.2	34.9	50.4	59.8	53.9	53.9	65.9	55.9
Top-$n\sigma$ 1	26.1	50.7	34.4	45.9	68.8	**57.0**	52.7	**71.6**	58.8
Global KL*	27.1	50.6	33.8	45.7	66.9	56.3	52.1	70.6	58.2
Local KL*	28.1	51.5	35.4	45.9	66.6	56.4	52.3	70.8	57.8
KL-budget	24.5	52.2	34.3	40.7	68.8	55.9	50.7	71.3	**58.9**

Table 5: **TriviaQA with brief reasoning.** The same 1,000 questions, asked to reason briefly before answering. I is individual-path accuracy, P is pass@8, and V is Vote@8, in percent. Bold marks the largest value in each column, with ties all bold.

losses in individual accuracy. No KL method has a significant Vote@8 gain, and global KL* has 3 Vote@8 losses. Factual recall therefore places the KL* rules between the conservative samplers and the exploratory ones, as on GSM8K.

The protocol also compares the two prompts on identical questions for each sampler. Brief reasoning raises every sampler's means on Qwen2.5 1.5B and Qwen2.5 7B. On Mistral 7B it lowers every sampler's individual accuracy and raises every sampler's pass@8. On Qwen2.5 7B the KL* rules gain the most from reasoning, 17.2 pass@8 points for local KL* and 12.8 Vote@8 points for global KL*. On Mistral 7B they lose less individual accuracy than temperature one, top-k 40, and the model card, 6.5 to 6.7 points against 8.4 to 9.2. These per-sampler differences have 95% paired intervals that overlap across samplers, so we report them descriptively.

G CommonsenseQA, QASC, and MATH-500

A protocol frozen before generation runs every default sampler and KL method on all 1,221 CommonsenseQA and all 926 QASC validation questions with 512 output tokens, following the P-less evaluation (Tan et al., 2026, Talmor et al., 2019, Khot et al., 2020), and on all 500 MATH-500 problems with 2,048 tokens on the two Qwen checkpoints (Hugging Face H4, 2024). Each sampler draws eight answers per question with the GSM8K seeds. Multiple-choice answers take the last marked final letter, then the last standalone option letter, and unparseable answers count as wrong.

	Qwen2.5 1.5B			Mistral 7B			Qwen2.5 7B		
Sampler	I	P	V	I	P	V	I	P	V
Greedy	60.9	60.9	60.9	**71.4**	71.4	71.4	79.3	79.3	79.3
Model card	**64.2**	89.4	70.1	68.0	88.5	72.1	79.8	90.8	81.0
Temperature 1	56.8	**91.3**	68.5	68.7	89.1	73.2	79.5	91.2	81.4
Top-p .95	56.3	89.7	69.3	68.8	87.7	73.0	**80.0**	91.2	**81.7**
Top-k 40	55.1	90.1	70.1	68.2	87.9	72.8	79.6	**91.8**	81.3
Min-p .05	59.7	89.5	69.9	69.7	88.1	72.4	79.8	91.4	81.4
P-less	61.1	82.4	70.1	70.3	81.5	71.6	79.4	85.7	80.3
Top-$n\sigma$ 1	59.3	89.9	69.7	69.5	88.0	71.7	79.4	90.7	81.1
Global KL*	58.4	89.7	69.2	70.0	86.7	72.9	79.6	90.4	81.3
Local KL*	59.1	88.7	**70.7**	70.1	87.6	73.1	79.9	91.0	81.3
KL-budget	56.3	90.9	69.1	68.8	**89.3**	**73.4**	79.6	91.7	81.6

Table 6: **CommonsenseQA at default settings.** All 1,221 validation questions, eight answers each. I is individual-path accuracy, P is pass@8, and V is Vote@8, in percent. Bold marks the largest value in each column, with ties all bold.

	Qwen2.5 1.5B			Mistral 7B			Qwen2.5 7B		
Sampler	I	P	V	I	P	V	I	P	V
Greedy	54.4	54.4	54.4	**68.7**	68.7	68.7	79.0	79.0	79.0
Model card	54.8	84.1	65.5	66.8	84.8	**72.3**	78.8	89.0	80.8
Temperature 1	47.8	85.4	63.2	66.9	84.4	71.7	78.6	90.6	80.5
Top-p .95	48.6	85.5	63.4	67.0	84.4	70.8	78.4	89.7	80.2
Top-k 40	46.6	86.4	64.0	67.3	**86.2**	71.7	78.8	**90.8**	**81.5**
Min-p .05	51.8	**86.6**	64.2	67.7	85.1	70.8	78.4	90.2	80.5
P-less	**56.0**	76.9	64.7	68.4	78.4	69.8	**79.1**	85.6	80.3
Top-$n\sigma$ 1	50.6	85.2	**65.9**	68.5	84.2	71.3	78.7	89.0	81.0
Global KL*	50.7	85.4	63.7	67.4	82.0	71.0	78.9	89.2	80.8
Local KL*	51.5	85.6	**65.9**	67.7	83.8	70.9	78.9	89.8	80.1
KL-budget	46.1	84.6	62.0	66.4	85.5	71.3	78.6	90.5	81.3

Table 7: **QASC at default settings.** All 926 validation questions, eight answers each. I is individual-path accuracy, P is pass@8, and V is Vote@8, in percent. Bold marks the largest value in each column, with ties all bold.

MATH-500 uses the symbolic checker of Appendix P, in which a comparison that exceeds one second counts as unequal. This happened three times in about 44,000 comparisons. Each checkpoint and task has its own family of 63 comparisons. Tables 6, 7, and 8 give every mean, and Figure 5 counts the significant comparisons.

H THE KL* FAMILY

Weighting the entropy term by λ gives a one-parameter family whose candidate sets grow with λ (Appendix N). Canonical KL* uses $\lambda = 1$ and is the only setting reported elsewhere. A sweep over $\lambda \in \{.5, .75, 1.5, 2\}$ on the first 512 GSM8K test questions, frozen before generation, maps the family (Table 9). Smaller values admit fewer candidates. They raise individual accuracy on Qwen2.5 1.5B and Mistral 7B, by up to 3.8 and 7.6 points at $\lambda = .5$, and leave pass@8 within 4 points. Canonical KL* has the highest pass@8 on Qwen2.5 1.5B. On Qwen2.5 7B every value from .5 to 1.5 stays within a point of the canonical individual accuracy. At $\lambda = 2$ the candidate sets grow too large, and accuracy falls on all three checkpoints and collapses to 3.8% on Qwen2.5 1.5B. The sweep uses test questions, so it describes the family and does not select a setting.

	Qwen2.5 1.5B			Qwen2.5 7B		
Sampler	I	P	V	I	P	V
Greedy	**54.6**	54.6	54.6	75.2	75.2	75.2
Model card	44.3	75.2	57.7	74.4	89.4	79.3
Temperature 1	39.6	71.2	56.5	73.5	89.6	80.8
Top-p .95	44.7	74.2	60.2	73.5	90.2	79.9
Top-k 40	39.3	72.2	55.4	73.3	90.6	80.8
Min-p .05	45.8	75.6	61.0	74.5	89.8	81.2
P-less	53.8	75.8	61.0	**75.8**	86.6	78.1
Top-$n\sigma$ 1	46.2	75.0	60.1	74.5	90.8	80.1
Global KL*	47.0	76.2	60.3	74.2	89.8	80.7
Local KL*	48.6	**78.0**	**62.2**	74.2	89.0	**81.3**
KL-budget	36.2	70.6	52.2	73.2	**91.2**	**81.3**

Table 8: **MATH-500 at default settings.** All 500 problems, eight answers each, on the two Qwen checkpoints. I is individual-path accuracy, P is pass@8, and V is Vote@8, in percent. Bold marks the largest value in each column, with ties all bold.

	Qwen2.5 1.5B			Mistral 7B			Qwen2.5 7B		
λ	I	P	V	I	P	V	I	P	V
0.5	**65.0**	87.3	74.0	**47.4**	82.2	**64.6**	88.9	95.9	91.2
0.75	64.8	88.7	**76.2**	43.2	**84.0**	63.6	**89.6**	95.9	93.2
1 (KL*)	61.2	**90.8**	73.4	39.8	81.6	64.4	89.4	95.9	92.3
1.5	54.2	87.1	70.7	31.8	81.8	61.0	88.7	**96.9**	**93.4**
2	3.8	19.5	4.1	10.1	44.5	12.9	72.7	92.8	84.1

Table 9: **The** KL* **family on the first 512 GSM8K test questions.** Individual-path accuracy (I), pass@8 (P), and Vote@8 (V) in percent, eight answers per question, for global KL* with the entropy term weighted by λ. The $\lambda = 1$ row is canonical KL* from the default-settings runs. Bold marks the largest value in each column. The sweep is descriptive and selected no setting.

I Real Next-Token Statistics

For the first 128 GSM8K test questions and the first 32 WritingPrompts pool prompts, one seeded generation per sampler recorded the candidate-set size after the sampler chain and the rank of the sampled token at every step, as reported by the server. Every sampler ran natively, so no statistic is reconstructed counterfactually. The protocol was frozen before the run. Table 10 and Figure 14 summarize the steps. Qwen3-8B ran five samplers with thinking disabled. On its GSM8K steps no sampler goes past rank 10 more than 0.001% of the time. In writing, local KL* leaves the top token on 15.4% of steps, against 13.8% for min-p, 14.3% for top-p, and 8.1% for the model card, and it reaches past rank 10 on 0.008%, against 0.04% for min-p and 0.18% for top-p.

J Story Openings for Every Sampler

Figure 3 follows a rule fixed in the writing protocol before generation: the first two prompts of the WritingPrompts pool in manifest order, seed 5100. The generator (Qwen2.5 7B, the largest in the default-settings study), the three samplers in the figure, and the cut rule were fixed before any story was read. Each opening ends at the first sentence end after 50 words, or at 70 words. Every sampler is listed below.

Prompt 1. *write about the strangest/scariest/saddest dream you've ever had in less than 200 words.*

Model card. In the dead of night, I found myself standing at the edge of a vast, endless forest. The air was thick with mist, and the trees loomed overhead like ancient sentinels. As I stepped forward, the path beneath my feet turned into a river of liquid darkness. I tried to turn back, but the river seemed to draw me in, pulling me deeper until I was submerged up to my ...

Sampler	GSM8K reasoning			Story writing		
	One candidate	Below top	Past rank 10	One candidate	Below top	Past rank 10
Global KL*	85.8	8.4	0.072	53.1	**32.1**	**1.60**
Local KL*	86.2	**8.1**	**0.032**	56.1	**29.4**	**0.86**
KL-budget	72.0	**12.3**	**0.070**	28.9	**39.1**	**1.94**
Model card	57.9	8.3	0.17	30.9	26.9	1.04
Temperature 1	0.0	9.2	0.38	0.0	**40.6**	**11.87**
Top-p .95	72.1	8.0	0.12	22.0	35.0	5.46
Top-k 40	0.0	9.4	0.30	0.0	36.1	5.05
Min-p .05	74.7	7.1	0.049	31.1	**26.2**	**0.74**
P-less	97.7	**1.2**	**0.003**	84.5	**7.5**	**0.002**
Top-$n\sigma$ 1	77.3	6.7	0.039	33.7	26.4	1.33

Table 10: **Real next-token statistics.** Percentage of generated steps with a single candidate, with a sampled token other than the model's top token, and with a sampled token ranked 11th or lower. Each value averages Qwen2.5 1.5B, Qwen2.5 7B, and Mistral 7B, over 128 GSM8K questions and 32 WritingPrompts prompts with one seeded generation per sampler. Bold marks the samplers that no other sampler beats on both exploration and tail share, the dotted frontier of Figure 6. Temperature and top-k always keep more than one candidate.

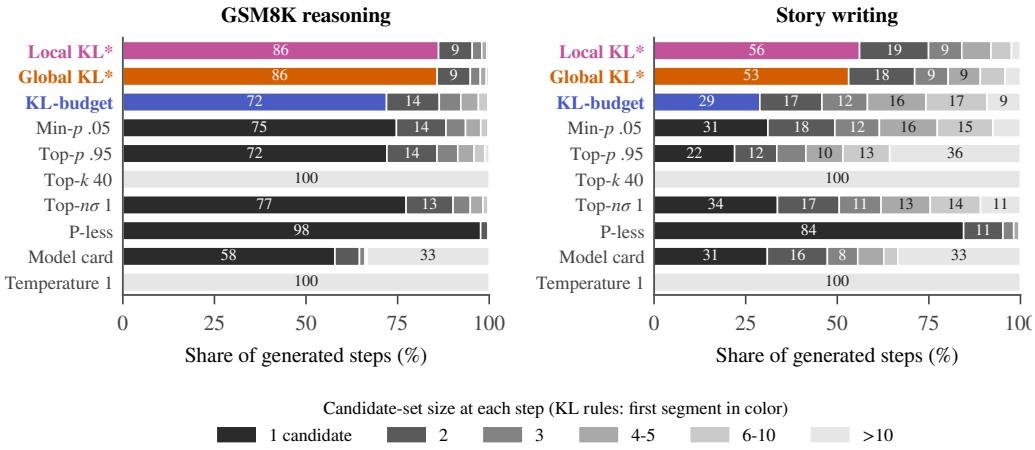


Figure 14: **How many candidates each sampler keeps.** Share of generated steps by candidate-set size, averaged with equal weight over Qwen2.5 1.5B, Qwen2.5 7B, and Mistral 7B. KL rules are colored in their single-candidate segment.

Temperature 1. In the dead of night, I found myself in a dream so strange it left me aghast. The landscape was surreal, a patchwork of colors that shifted with every breath. The sky was a deep shade of blood red, punctuated by a single, flickering star. I walked through a forest where the trees were twisted, their branches clawing at the sky like the fingers of the dead.

Top-p .95. In the quiet of a restless night, I found myself in a dream as strange and terrifying as any nightmare could be. The landscape was a surreal blend of the familiar and the unfamiliar, with towering, whispering trees that seemed to move and shift with every breath. The air was thick with an eerie silence, broken only by the distant, ethereal moans that lingered in the background.

Top-k 40. In the quiet of a cold winter night, I found myself in a place that felt both familiar and utterly alien—my dreamscape. The air was thick with a fog that felt like a tangible weight, pressing against my skin. The streets were empty, save for a lone figure shrouded in darkness, its eyes glowing like red coals.

Min-p .05. In the dead of night, I found myself in a dream so strange it left me in a cold sweat. The sky above was a canvas of swirling, dark hues, and the air was thick with the scent of decay. I stood before a dilapidated house, its windows broken and the door hanging on one hinge.

P-less. In the dead of night, I found myself in a desolate, endless forest, the trees towering and twisted like ancient sentinels. The moon was a sickly yellow, casting eerie shadows that danced and morphed into

faces. I walked, my heart pounding, until I stumbled upon a clearing where a figure stood, cloaked and silent.

Top-$n\sigma$ 1. In the dead of night, I found myself in a dream so strange it left me in a cold sweat. The sky above was a canvas of swirling, dark hues, and the air was thick with the scent of decay. I stood before a dilapidated house, its windows broken and the door hanging on one hinge.

Global KL*. In the quiet of a midnight, I found myself wandering through a forest shrouded in an eerie silence. The trees, tall and ancient, twisted and gnarled, loomed over me like sentinels of a forgotten era. The moon was a pale, distant crescent, casting an otherworldly glow that revealed a path of shimmering bones beneath my feet.

Local KL*. In the quiet of a moonlit night, I found myself wandering through a dense, foggy forest, the air thick with the scent of damp earth and something else—something indefinable and slightly off. The trees were not just dark; they seemed to pulse with an eerie, unnatural light, as if alive with a silent, malevolent force.

KL-budget. In the dead of night, I found myself in a vast, dark forest, the only sound the rustling leaves underfoot and the distant hooting of an owl. The moon, a sickly green sliver, barely pierced the canopy. As I walked, the trees began to shift, their gnarled roots reaching out like bony fingers.

Greedy. In the dead of night, I found myself in a dream so strange and terrifying that it lingered in my mind long after I awoke. The dream began in a familiar park, but as I walked, the trees grew taller, their branches twisting into grotesque shapes. The air was thick with an eerie silence, broken only by the distant, muffled cries of something I couldn't quite see.

Prompt 2. *You died. You were reborn. You remembered your previous life. But that's actually pretty useless - because you weren't reborn human.*

Model card. In the quiet of dawn, under a canopy of stars, Lila felt a strange sensation wash over her. She had always been a human, but now, as she lay in the soft grass, she realized something was different. Her body felt lighter, more fluid, and when she looked down, she saw shimmering wings sprouting from her back.

Temperature 1. Amara felt a strange sensation as she opened her eyes. The world around her was a blur of colors and shadows, eerily familiar yet completely alien. She remembered the life she had lived as a human—her friends, her job, the streets of the city she called home. But all of that was replaced by a feeling of weightlessness, like she was swimming in vast, unknown seas.

Top-p .95. Amara opened her eyes to a place that was both familiar and strange. The air was thick with the scent of incense and her body felt lighter, almost weightless. She sat up and looked around, seeing a room adorned with golden symbols and intricate carvings. It was then that she realized she wasn't in a human body; she had been reborn as an angel.

Top-k 40. Once a vibrant human, now reborn as a towering tree, I looked out over the world from my high canopy. Everything seemed strange and yet eerily familiar. I remembered—my life as a person, filled with laughter and tears, dreams and disappointments. But that was useless now, because I was not reborn as a human.

Min-p .05. You died, and in the blink of an eye, you were reborn. But this time, you weren't human. Instead, you were a falcon, soaring through the clear blue sky. The memory of your previous life as a human flickered in your mind, a ghostly echo of days long gone. You remembered the warmth of the sun on your skin, the sound of laughter, and the taste of a good cup ...

P-less. In the quiet of the night, under the vast, starlit sky, Lila found herself in a place she had never known before. The air was crisp, and the ground beneath her feet was soft and green. She looked down and saw that she was no longer a human, but a majestic eagle, with wings that stretched wide and strong.

Top-$n\sigma$ 1. You died, and in the blink of an eye, you were reborn. But this time, you weren't human. You were a bird, a beautiful falcon with piercing eyes and wings that glided gracefully through the sky. As you opened your beak, you remembered your previous life—your name was Alex, a writer with a passion for words and stories.

Global KL*. You died. You were reborn. You remembered your previous life. But that's actually pretty useless—because you weren't reborn human. In the bright, neon glow of the alien city, you found yourself, a shimmering figure with tentacles and iridescent scales. Your previous life as a human had been a distant memory, a blur of faces and places.

Local KL*. You died. You were reborn. You remembered your previous life. But that's actually pretty useless—because you weren't reborn human. The air was cool and crisp as you awoke, your eyes blinking to adjust to the dim light filtering through the dense foliage. You were in a vast forest, far from any known human settlement.

KL-budget. You died, and as you floated above your own body, the world blurred and faded, leaving only a distant whisper of voices and colors. You were reborn, and as consciousness returned, the first thing

	Qwen2.5 1.5B		Mistral 7B		Qwen2.5 7B	
	Greedy wrong pass@8	Greedy right individual	Greedy wrong pass@8	Greedy right individual	Greedy wrong pass@8	Greedy right individual
Questions	432	887	642	677	142	1177
Model card	70.6	77.7	69.2	52.6	62.7	96.3
Temperature 1	67.8	72.4	**70.1**	54.3	67.6	95.1
Top-p .95	72.2	75.8	**70.1**	54.5	66.9	95.2
Top-k 40	68.8	70.6	69.5	52.3	65.5	94.9
Min-p .05	71.1	78.0	67.1	55.8	**72.5**	95.5
P-less	53.2	**89.5**	47.5	**82.7**	37.3	**98.2**
Top-$n\sigma$ 1	**72.5**	78.3	68.8	58.3	67.6	95.8
Global KL*	70.4	77.8	67.4	54.9	64.8	95.6
Local KL*	71.3	78.9	69.6	53.6	60.6	95.4
KL-budget	72.2	70.9	66.4	45.6	71.8	94.2

Table 11: **Exploration rescues the questions greedy decoding misses.** Default settings on GSM8K, split by whether greedy decoding of the same checkpoint answers the question correctly. Pass@8 on the questions greedy gets wrong, and individual-path accuracy on the questions it gets right, in percent. Bold marks the largest value in each column.

you remembered was the sensation of wings. Not the memory of them, but the actual feeling of them, beating against the air.

Greedy. In the quiet of dawn, under the vast canopy of stars, I awoke to a world that was both familiar and alien. The memories of my past life flooded back, vivid and sharp—my family, my friends, the laughter and the tears. But as I stood there, the world around me shimmered, and I realized I was no longer human.

K Subgroup Audit

The subgroup protocol was fixed after the aggregate default-settings results for Qwen2.5 1.5B and Qwen2.5 7B were known and before any subgroup statistic was computed. Its primary split separates the questions that greedy decoding of the same checkpoint answers wrongly, where exploration can rescue an answer, from those it answers correctly, where exploration must keep it. Its secondary split uses terciles of question word count. Every KL method is compared with every default baseline in every subgroup on the three metrics, with Bonferroni adjustment over all 315 comparisons per checkpoint.

On the questions greedy decoding gets wrong, neither KL* rule has a significant loss on any checkpoint. Global KL* has 10 significant gains there and local KL* 11. Most are against P-less, which solves 17 to 28 points fewer of these questions in pass@8 than either rule, and the rest are individual-path gains over temperature one and top-k 40 at 1.5B. P-less stays close to greedy decoding and has the highest individual-path accuracy on the questions greedy already solves (Table 11), which is the source of its lead on the whole split. KL-budget has 8 gains and 7 losses on greedy-wrong questions, 6 of the losses on Mistral 7B, and 24 losses on greedy-right questions. Across the nine checkpoint and length-tercile subgroups, global KL* has 10 gains and 5 losses, local KL* 11 and 5, and KL-budget 3 and 35, so no length range reverses the whole-split pattern.

L Computational Cost

KL* adds no model computation. Its only cost is the sampler itself, which runs once per generated token on the CPU in `llama.cpp`. Table 12 replays recorded logits of four checkpoints from our smaller-sample studies. Local KL* is the fastest sampler on every checkpoint, and global KL* costs about as much as min-p, 0.10 to 0.14 ms. Both are faster than temperature-only sampling, because truncation shrinks the final normalization and draw from the whole vocabulary to a few tokens. P-less is slow in its pinned implementation, which sorts the vocabulary, and a direct linear threshold would make it comparable to temperature sampling.

Sampler Vocabulary	Llama 3.3 70B 128,256	R1-Distill 70B 128,256	Qwen2.5 72B 152,064	Qwen3.8 27B 248,320
Greedy	268.2	281.4	319.8	524.0
Temperature 1	244.9	256.1	293.9	480.9
Top-p .95	314.6	316.0	390.6	612.9
Min-p .05	90.0	90.6	107.6	176.3
P-less, pinned sorted	3972.2	4165.2	4904.6	8790.1
P-less, direct linear	297.3	297.3	353.8	581.4
Global KL*	102.7	138.7	137.8	129.5
Local KL*	**47.6**	**57.8**	**58.5**	**73.8**
KL-budget .1	90.0	90.4	107.8	176.4

Table 12: **Sampler time per token on real logits, in microseconds.** Recorded first-token logit vectors of 24 prompts per checkpoint, replayed through each native sampler chain, 31 timed rounds after five warmups, one core of an Apple M4 Pro. Times include admission, weighting, and the draw, and exclude model inference. Bold marks the fastest sampler per checkpoint. Replays of fixed reference histories give the same ordering.

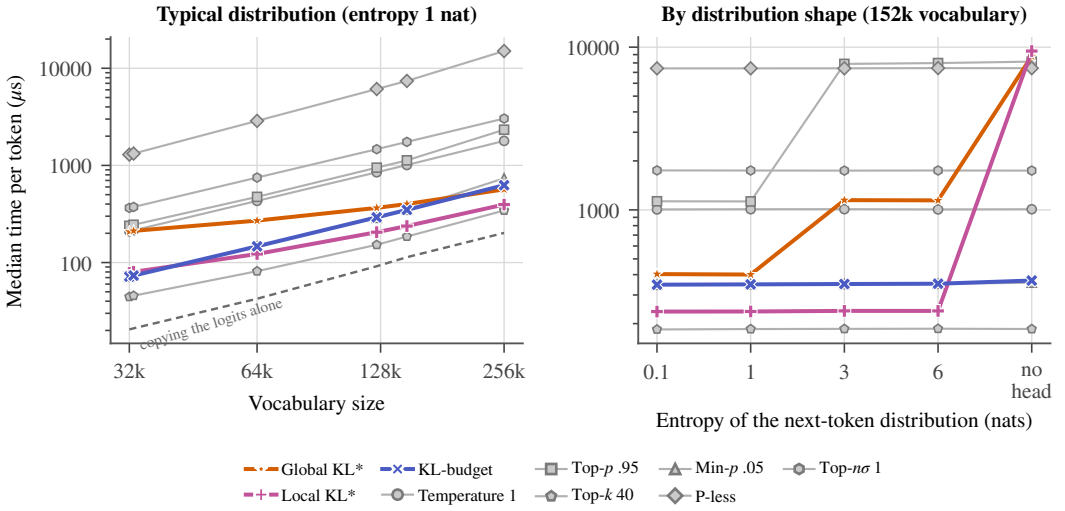


Figure 15: **Vocabulary size, not the choice of sampler, drives the cost.** Median CPU time per token of each default sampler chain on synthetic logits, including the copy of the logits and the final draw. Left, a distribution with entropy 1 nat across vocabulary sizes. Right, five distribution shapes at the Qwen vocabulary. KL-budget adds almost nothing to min-p, so their lines overlap.

Vocabulary size drives the cost. Figure 15 times every default sampler chain on synthetic logits for six vocabulary sizes from 32,000 to 262,144 and five distribution shapes, with 2,000 tokens per cell on one core of an Intel Xeon Platinum 8470 in a shared server. Copying the logits into the candidate array alone grows tenfold across this range, and every chain grows with it. In a typical distribution, the choice among the fast samplers changes the cost by less than a factor of three, whereas a 256k vocabulary costs roughly three to ten times as much as a 32k one. Global KL* sorts only a 512-token head, bounds the tail, and falls back to a full sort when the bound cannot certify the optimum. Local KL* scans an ordered head of 128 tokens with the same fallback. The fallback is needed only for very flat distributions, where it costs as much as top-p and P-less. In the real-token probe of Appendix I, 8.7% of story-writing steps and 0.3% of GSM8K steps have entropy above 3 nats, and 0.18% and none have entropy above 6 nats.

GPU inference engines. vLLM and SGLang sample on the GPU (Kwon et al., 2023, Zheng et al., 2024). On supported GPUs vLLM uses the sorting-free rejection kernels of FlashInfer (Ye et al., 2025) for top-k and top-p by default, and SGLang offers FlashInfer kernels for top-k, top-p, and

Algorithm 1 Global and local KL* for one next-token distribution

Require: probabilities p on V tokens and a rule, global or local
  sort tokens so that $p_1 \geq \cdots \geq p_V$ and set $C \leftarrow 0$
  for $k = 1, \ldots, V$ **do**
    $C \leftarrow C - \log p_k$ and $J(k) \leftarrow C/k - \log k$
    if rule is local, $k > 1$, and $J(k) > J(k-1)$ **then return** $\mathrm{Unif}\{1, \ldots, k-1\}$
  end for
  if rule is local **then return** $\mathrm{Unif}\{1, \ldots, V\}$
  return $\mathrm{Unif}\{1, \ldots, k^\star\}$, where $k^\star = \min \arg\min_k J(k)$

min-p. These kernels implement fixed threshold tests (Xing et al., 2025). KL* is not such a test. Its global objective need not be unimodal in the cutoff, so a single pivot search cannot certify the minimum. A GPU version would use a partial top-m selection and a prefix scan with a sort fallback, as the CPU version does, and local KL* needs only the scan. KL* is therefore close to free in engines that sample on the CPU, and in fused-kernel engines it needs a new kernel rather than a configuration change. vLLM's custom logits processors could host it as batched tensor operations, which we have not benchmarked.

M Formal Definitions

Throughout, $\log$ is the natural logarithm, so every divergence is in nats. Let $p_i = \mathrm{softmax}(z)_i > 0$ be the original next-token probabilities on a finite vocabulary. A gate chooses a nonempty support S. Subsequent temperature gives

$$q_{S,T}(i) = \frac{\mathbf{1}\{i \in S\} p_i^{1/T}}{\sum_{j \in S} p_j^{1/T}}, \qquad \lim_{T \to \infty} q_{S,T} = \mathrm{Unif}(S). \tag{2}$$

For min-p threshold α, the temperature-first gate instead satisfies

$$p_i(T) \geq \alpha p_{\max}(T) \quad \Longleftrightarrow \quad p_i \geq \alpha^T p_{\max}. \tag{3}$$

Retuning its threshold to $\alpha^{1/T}$ recovers the truncation-first support at positive finite temperature. Fixed-setting order comparisons quantify the effect of keeping the numerical threshold constant.

The retained baseline gates are compact. Top-p keeps the smallest ranked prefix whose original probability mass reaches its chosen threshold. Min-p keeps tokens with $p_i \geq \alpha \max_j p_j$. Order-two P-less keeps tokens with $p_i \geq \sum_j p_j^2$, the collision-probability threshold (Tan et al., 2026). We use its standard unnormalized-threshold variant, with `p_less_norm=false`. Temperature-only sampling keeps the full vocabulary. Every baseline's admitted probabilities are then reweighted as in Equation 2. Uniform min-p uses the exact limiting draw.

Sort $p_1 \geq \cdots \geq p_V$ and let U_k be uniform on the first k tokens. Global KL* uses the smallest globally minimizing prefix:

$$k^\star = \min \arg \min_{1 \leq k \leq V} J(k), \qquad J(k) = D_{\mathrm{KL}}(U_k \| p) = -\log k - \frac{1}{k} \sum_{i=1}^{k} \log p_i. \tag{4}$$

With $A(k) = -k^{-1} \sum_{i \leq k} \log p_i$, local KL* accepts adjacent extensions while

$$J(k+1) - J(k) = \frac{-\log p_{k+1} - A(k)}{k+1} - \log(1 + 1/k) \leq 0, \tag{5}$$

then stops before the first strict increase. For $p = (.5, .1, .1, .1, .1, .1)$, local selection chooses one token while global selection chooses six. Algorithm 1 gives both rules.

For a separately chosen support, write $p_S(i) = p_i / P(S)$. KL-budget solves

$$\max_{q \in \Delta(S)} H(q) \quad \text{subject to} \quad D_{\mathrm{KL}}(q \| p_S) \leq \epsilon. \tag{6}$$

Token	Original	Top-p	Min-p	P-less	Global KL*	Local KL*	Budget
Lower confidence: $p_{\max} = 0.564$ *Prefix:* It was a typical Friday evening at Fact Restaurant, with the usual bustle of diners and the sound of clinking glasses. Two events were listed on the							
␣reservation	56.35	**36.50**	**36.50**	**100.00**	**50.00**	**50.00**	**41.34**
␣calendar	19.22	**25.50**	**25.50**	0.00	**50.00**	**50.00**	**25.16**
␣board	9.25	**19.98**	**19.98**	0.00	0.00	0.00	**17.95**
␣evening	6.78	**18.02**	**18.02**	0.00	0.00	0.00	**15.55**
␣reservations	2.71	0.00	0.00	0.00	0.00	0.00	0.00
␣restaurant	2.38	0.00	0.00	0.00	0.00	0.00	0.00
Other tokens	3.30	0.00	0.00	0.00	0.00	0.00	0.00
Support size	V	4	4	1	2	2	4
Retained mass (%)	100.0	91.6	91.6	56.4	75.6	75.6	91.6
Higher confidence: $p_{\max} = 1.000$ *Prefix:* In the land of Azura, where knights were noble and chivalry was a way of life, a sinister force lurked in the shadows. Prince Edward							
’	100.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
␣’	0.00	0.00	0.00	0.00	0.00	0.00	0.00
’\n	0.00	0.00	0.00	0.00	0.00	0.00	0.00
.	0.00	0.00	0.00	0.00	0.00	0.00	0.00
␣was	0.00	0.00	0.00	0.00	0.00	0.00	0.00
␣(	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Other tokens	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Support size	V	1	1	1	1	1	1
Retained mass (%)	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 13: **Admission and weighting respond differently to model confidence.** Actual Llama 3.3 70B next-token probabilities (%). Top-p .9, min-p .05, and P-less truncate first and draw at $T = 3$. Global/local KL* draw uniformly. KL-budget uses min-p .05 and $\epsilon = .1$. The two cases are the smallest and largest original $p_{\max}$ among the twelve prescribed WritingPrompts reference-history probes, selected without quality scores. Prefix text ends immediately before the predicted token. ␣ denotes a space in a token. The top six tokens and all remaining probability mass are shown. Bold marks the tokens each rule keeps. Displayed zeroes can be nonzero below rounding precision. These are float64 distribution calculations on saved logits.

Its solution is $q_\beta(i) \propto p_S(i)^\beta$, with $\beta \in [0, 1]$, as derived in OD-Stega's Theorem 1 (Huang et al., 2026). Uniform weighting is chosen when feasible. Otherwise the constraint is tight. An allowance of zero returns the conditional model distribution. For every supported draw,

$$D_{\mathrm{KL}}(q\|p) = -\log P(S) + D_{\mathrm{KL}}(q\|p_S). \tag{7}$$

Thus truncation and subsequent reweighting spend distinct amounts of divergence. OD-Stega's Section 4.2 and Theorem 3 give this two-stage accounting for its optimized draw. The identity above holds for any q supported on S.

M.1 Recorded next-token examples

Table 13 records actual Llama 3.3 70B distributions. In its less certain context, the two KL* rules admit two candidates and KL-budget retains unequal weights on four. In the confident context, all gates retain the same single candidate.

N Proofs and Exact Scope

N.1 Optimal uniform supports are ranked prefixes

Proposition 1 (Exact tuning-free selection)**.** *For every positive next-token distribution p, a ranked prefix minimizes $D_{\mathrm{KL}}(U_S\|p)$ over all nonempty candidate sets S. Global* KL* *therefore returns an optimal uniform candidate set after one sort and one linear scan.*

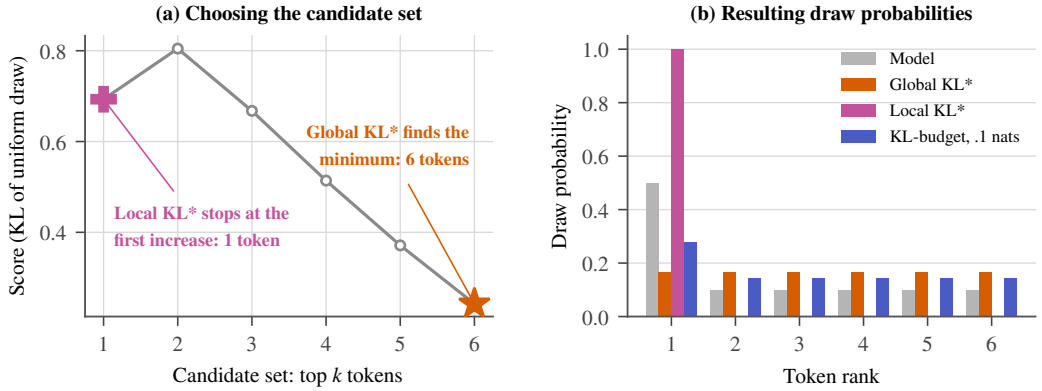


Figure 16: **Global and local KL* on one distribution.** (a) The score of Equation 1 for $p = (.5, .1, .1, .1, .1, .1)$. Local KL* stops at the first increase and keeps one token. Global KL* keeps all six. (b) The resulting draw probabilities. KL-budget with a .1-nat allowance keeps unequal weights on all six tokens.

Proof. For a support S of size k,

$$D_{\mathrm{KL}}(\mathrm{Unif}(S)\|p) = -\log k - \frac{1}{k}\sum_{i\in S}\log p_i.$$

The first term is fixed. Replacing an admitted token with an excluded token of larger probability decreases the second term. Thus a top-k support is optimal at every size, and minimizing Equation 4 over sizes gives a global optimum. Nonuniqueness at a fixed size requires a probability tie at the boundary. Distinct sizes may also tie. Our smallest-size convention resolves the latter. □

Generalized objective and reference temperature. For interpretation, define $J_\lambda(k) = A(k) - \lambda \log k$, where $A(k) = -k^{-1}\sum_{i\le k}\log p_i$. Canonical KL* uses $\lambda = 1$. For $\lambda_1 < \lambda_2$, adding the two global optimality inequalities gives $(\lambda_2 - \lambda_1)(\log k_2 - \log k_1) \ge 0$, so increasing λ cannot shrink the chosen prefix. If the gate instead reads $p(\tau_0) \propto p^{1/\tau_0}$, then

$$J_\lambda(k, p(\tau_0)) = \tau_0^{-1} J_{\lambda\tau_0}(k, p) + c(\tau_0),$$

with $c(\tau_0)$ independent of k. This identity applies only to admission. The draw temperature and conditional budget remain separate parameters. The main comparison keeps $\lambda = 1$ throughout.

N.2 Local stopping does not approximate global KL in general

Let $G_k = (\prod_{i\le k} p_i)^{1/k}$. Equation 5 accepts a new candidate exactly when

$$p_{k+1} \ge G_k \left(\frac{k}{k+1}\right)^{k+1}.$$

This only compares adjacent prefixes. A later prefix can improve the objective after an initial increase, as the six-token example demonstrates.

Proposition 2 (No constant regret for local stopping)**.** *For every $c > 0$ there is a finite positive distribution on which local* KL* *exceeds the global minimum of $D_{\mathrm{KL}}(U_S\|p)$ by more than c nats.*

Proof. For integer $m \ge 4$, take $V = m^2$ and $p_1 = 1/m$, $p_i = 1/[m(m+1)]$ for $2 \le i \le m^2$. The probabilities sum to one. Because $p_2/p_1 = 1/(m+1) < 1/4$, local selection stops at one token, with divergence $\log m$. The feasible full-vocabulary uniform draw has

$$D_{\mathrm{KL}}(U_{m^2}\|p) = \log(1 + 1/m) - \frac{\log(m+1)}{m^2} < 1/m.$$

The global optimum is no worse, so local regret exceeds $\log m - 1/m$, which is unbounded in m. □

The proposition characterizes support selection under the KL objective. Output quality is assessed empirically.

N.3 THE SCOPE OF THE ENTROPY–SURPRISAL FRONTIER

For fixed contexts x drawn under a common measure μ, write $c_x(k) = -k^{-1} \sum_{i \le k} \log p_i^{(x)}$ and $h(k) = \log k$. Pointwise optimality of $k_x^\star$ for $c_x(k) - \lambda h(k)$ implies

$$E_\mu[c_x(k_x^\star)] \le E_\mu[c_x(k_x)] - \lambda\{E_\mu[h(k_x)] - E_\mu[h(k_x^\star)]\}.$$

Thus global selection lies on the supported frontier within the uniform-support class at its attained entropy. Discrete supports need not realize every entropy target. If arbitrary nonuniform q are allowed, the minimizer of $E_q[-\log p] - \lambda H(q)$ is instead $q_i \propto p_i^{1/\lambda}$. Ordinary temperature therefore describes the unrestricted frontier. The comparison must use the same contexts. Different decoders visit different generated histories.

N.4 KL-BUDGET SOLUTION

For completeness, introduce a multiplier $\eta \ge 0$ for Equation 6. Stationarity of $H(q) - \eta[D_{\mathrm{KL}}(q\|p_S) - \epsilon]$ gives $q_i \propto p_S(i)^{\eta/(1+\eta)}$ (Huang et al., 2026). Put $\beta = \eta/(1+\eta)$, $\ell_i = \log p_S(i)$, and $F(\beta) = \log \sum_i e^{\beta \ell_i}$. Then

$$H(q_\beta) = F(\beta) - \beta F'(\beta), \qquad D(\beta) = D_{\mathrm{KL}}(q_\beta\|p_S) = (\beta - 1)F'(\beta) - F(\beta).$$

Since $F''(\beta) = \mathrm{Var}_{q_\beta}(\ell)$,

$$H'(\beta) = -\beta F''(\beta) \le 0, \qquad D'(\beta) = (\beta - 1)F''(\beta) \le 0.$$

Uniform weights have maximal entropy if feasible. Otherwise the constraint is tight and, for a nonuniform positive reference, its boundary root is unique. A uniform or singleton reference is unchanged. Near zero budget,

$$D(\beta) = \tfrac{1}{2} \mathrm{Var}_{p_S}(\log p_S)(1 - \beta)^2 + O((1 - \beta)^3).$$

A fixed budget therefore induces a context-dependent temperature even close to the original conditional distribution. Substituting $p_i = P(S)p_S(i)$ in the KL sum gives Equation 7. A full-reference budget B would leave only $B + \log P(S)$ for reweighting and would be infeasible on S if this were negative.

Finite-horizon accounting. For a fixed horizon n, let Q be the generated sequence law and P the original model law. At history h_t, write $m_t = P_t(S_t \mid h_t)$ for retained mass. The KL chain rule and Equation 7 give

$$D_{\mathrm{KL}}(Q\|P) = \sum_{t=1}^{n} \mathbb{E}_Q\left[-\log m_t + D_{\mathrm{KL}}(q_t\|p_{t,S_t})\right] \le \sum_{t=1}^{n} \mathbb{E}_Q[-\log m_t + \epsilon_t].$$

Against the reference process that uses p_{t,S_t} at the same histories, the bound is simply $\sum_t \epsilon_t$ for fixed per-step allowances. Against the original model, truncation costs remain. Maximizing entropy at each next-token distribution also does not establish a global optimum for sequence entropy, because draws change future histories.

O SAMPLING ORDER AND LIMITS

For a fixed finite support, $p_i^{1/T} \to 1$, proving the uniform limit in Equation 2. This result does not authorize substituting an exactly uniform input into an arbitrary discontinuous admission rule.

When does a gate commute with temperature? Let $G(z)$ return token indices from logits z, with a fixed tie convention. If $G(z/T) = G(z)$ for every $T > 0$, gating before or after that scaling gives the same support and the same normalized draw. Top-k satisfies this because positive scaling preserves ranks. Top-$n\sigma$ satisfies it because $z_i \ge z_{\max} - n\ \mathrm{sd}(z)$ is unchanged when both sides

are divided by T (Tang et al., 2025). Min-k's core score $(z_i - z_{i+1})/[i(z_1 - z_V)]$ has the same property when $z_1 > z_V$ (Ding et al., 2026). Its complete algorithm also uses a fallback proportional to $1/(z_1 - z_V)$, which changes if recomputed after scaling. Algorithm 1 of that paper computes the fallback from raw logits before temperature. We therefore cite its pre-temperature selection and invariant core, without assuming that every fallback implementation commutes with scaling. KL* likewise keeps the original reference fixed: changing its reference temperature changes its support objective as shown above.

Min-p. For a finite positive vector and $0 < \alpha < 1$, every token eventually satisfies $p_i(T)/p_{\max}(T) \geq \alpha$ as T increases. At the boundary $\alpha = 1$, only tied maxima survive at any finite temperature. Equation 3 gives the exact finite-temperature threshold equivalence.

Top-p. For mass threshold $\rho \in (0, 1)$, the ranked cumulative mass at size k converges to k/V. Generically the eventual size is $\lceil \rho V \rceil$. Token ranks still depend on the model: a size primarily determined by vocabulary size does not erase ranking information. At exactly uniform input, the implementation's tie convention matters.

P-less. Order-two P-less admits $p_i \geq \sum_j p_j^2$. Write $\bar{z} = V^{-1} \sum_j z_j$. At high finite temperature,

$$p_i(T) = \frac{1}{V} + \frac{z_i - \bar{z}}{VT} + O(T^{-2}), \qquad \sum_j p_j(T)^2 = \frac{1}{V} + O(T^{-2}).$$

If no logit equals $\bar{z}$, the eventual set is $\{i : z_i > \bar{z}\}$. In contrast, an exactly uniform vector satisfies the weak inequality for every token. We therefore keep the following experiments distinct: finite-temperature-first P-less, P-less applied to the original model followed by uniform draws, and a gate evaluated at an exactly uniform input. Only the first two occur in the generation comparison.

P Experimental Protocol and Reproduction

P.1 Complete GSM8K tests: three small checkpoints

This evaluation covers all 1,319 official test questions on Qwen2.5 1.5B, Mistral 7B, and Qwen2.5 7B, using the development-selected settings of Table 22. Each stochastic method generates eight outputs. Greedy generates once. The nine truncation-first methods and three separately development-tuned temperature-first controls therefore require $1,319(11 \cdot 8 + 1) = 117,391$ responses per checkpoint, or 352,173 in total. The prompts, demonstrations, answer rules, seeds 5100–5107, and 1,536-token output allowance are shared with the subset studies. No official test question is excluded. The release separates the 224 previously used questions from the 1,095 previously unused questions and retains every invalid or capped output.

All methods for a question run on the same hardware. Table 14 gives the assignment: Qwen1.5B uses 960 H100 and 359 RTX 5090 questions, Mistral 7B uses 896 and 423, respectively, and Qwen 7B uses the H100s for all 1,319 questions. Both backends use checksum-verified weights and the pinned sampler source. The Windows CUDA build passes native sampler and parallel-execution checks before receiving complete question blocks. Per-response build and execution identities are retained. Cross-device bitwise agreement is not assumed. Interrupted requests resume with the same identities. Committed responses are never replaced.

The complete split is primary. Descriptive intervals use 10,000 paired question resamples, with Wilson intervals for binary coverage. Within each checkpoint, the planned family contains three KL methods against eight stochastic baselines at three endpoints: individual-path accuracy, pass@8, and Vote@8. Its 72 simultaneous intervals use Bonferroni-adjusted tails of one million joint question resamples. This family was fixed before any complete full-test result was scored, when the subset results were already known. Exact coverage checks use Holm adjustment over the 24 binary contrasts within each checkpoint. For global KL* versus temperature-first top-p at 1.5B, the discordant counts are 32 versus 74 questions and adjusted $p = .0010$. These families do not adjust across checkpoints or the separate exposure diagnostics. An independent raw-response recount checks all identities, settings, seeds, means, paired differences, and discordant counts.

Table 14: Execution composition by question and eight-path stage. All methods share a backend at each question and seed. Counts describe execution assignment, not causal hardware effects. Greedy retains its single original draw.

Checkpoint	Task	Paths	First eight	Next eight	Questions
Qwen2.5 1.5B	GSM8K	8	H100	–	960
Qwen2.5 1.5B	GSM8K	8	RTX 5090	–	359
Mistral 7B	GSM8K	8	H100	–	896
Mistral 7B	GSM8K	8	RTX 5090	–	423
Qwen2.5 7B	GSM8K	8	H100	–	1319

Table 15: Complete-split outcomes with 95% question intervals. Pass@8 and greedy individual accuracy use Wilson intervals. Other means use question bootstraps. Gate labels specify whether truncation is applied before or after temperature T. Bold marks the largest mean in each checkpoint and column, with ties all bold.

Method	Individual	Pass@8	Vote@8
Qwen2.5 1.5B / GSM8K ($n = 1319$)			
Greedy	**67.4** [64.8, 69.9]	67.4 [64.8, 69.9]	67.4 [64.9, 70.0]
Temperature	55.8 [53.9, 57.7]	87.3 [85.4, 89.0]	72.3 [70.0, 74.6]
Top-p (before T)	60.1 [58.2, 62.0]	89.9 [88.2, 91.4]	74.9 [72.6, 77.2]
Min-p (before T)	56.8 [54.9, 58.6]	88.8 [87.0, 90.4]	73.4 [71.1, 75.7]
P-less (before T)	66.4 [64.2, 68.5]	83.3 [81.2, 85.2]	70.9 [68.5, 73.3]
Uniform min-p	47.0 [45.1, 48.8]	83.4 [81.3, 85.3]	67.2 [64.7, 69.7]
Top-p (after T)	64.6 [62.6, 66.5]	**91.4** [89.8, 92.8]	**75.0** [72.7, 77.2]
Min-p (after T)	64.4 [62.4, 66.3]	90.1 [88.3, 91.6]	**75.0** [72.7, 77.2]
P-less (after T)	60.5 [58.6, 62.4]	89.8 [88.1, 91.4]	74.9 [72.6, 77.1]
Global KL*	61.6 [59.6, 63.5]	88.2 [86.4, 89.9]	73.8 [71.5, 76.1]
Local KL*	61.9 [60.0, 63.8]	89.4 [87.6, 90.9]	74.5 [72.2, 76.7]
KL-budget	58.7 [56.8, 60.5]	89.2 [87.4, 90.8]	74.1 [71.8, 76.3]
Mistral 7B / GSM8K ($n = 1319$)			
Greedy	**52.2** [49.5, 54.9]	52.2 [49.5, 54.9]	52.2 [49.5, 55.0]
Temperature	43.6 [41.9, 45.3]	83.2 [81.1, 85.2]	66.3 [63.9, 68.8]
Top-p (before T)	40.1 [38.6, 41.7]	**85.0** [83.0, 86.8]	**67.3** [64.8, 69.7]
Min-p (before T)	36.2 [34.7, 37.6]	82.6 [80.4, 84.5]	63.4 [60.9, 65.8]
P-less (before T)	51.0 [48.8, 53.2]	73.9 [71.5, 76.2]	59.4 [56.7, 61.9]
Uniform min-p	27.4 [26.2, 28.6]	77.5 [75.2, 79.7]	55.2 [52.7, 57.8]
Top-p (after T)	39.7 [38.2, 41.2]	83.5 [81.4, 85.4]	66.2 [63.7, 68.6]
Min-p (after T)	44.6 [42.9, 46.3]	83.6 [81.5, 85.5]	66.8 [64.3, 69.2]
P-less (after T)	41.3 [39.7, 42.9]	83.2 [81.1, 85.2]	65.8 [63.3, 68.2]
Global KL*	39.2 [37.7, 40.8]	83.0 [80.9, 84.9]	64.0 [61.6, 66.5]
Local KL*	39.3 [37.7, 40.9]	82.3 [80.1, 84.2]	64.3 [61.9, 66.7]
KL-budget	32.8 [31.5, 34.2]	82.0 [79.9, 84.0]	60.4 [57.8, 62.8]
Qwen2.5 7B / GSM8K ($n = 1319$)			
Greedy	89.1 [87.3, 90.7]	89.1 [87.3, 90.7]	89.1 [87.3, 90.8]
Temperature	80.6 [79.1, 82.1]	96.4 [95.2, 97.2]	91.1 [89.5, 92.5]
Top-p (before T)	88.2 [86.8, 89.5]	96.1 [94.9, 97.0]	91.8 [90.3, 93.2]
Min-p (before T)	88.0 [86.7, 89.4]	96.7 [95.6, 97.6]	92.3 [90.9, 93.7]
P-less (before T)	**89.5** [87.9, 91.0]	93.0 [91.5, 94.3]	90.1 [88.6, 91.7]
Uniform min-p	86.4 [85.0, 87.8]	96.7 [95.6, 97.5]	91.9 [90.5, 93.3]
Top-p (after T)	89.1 [87.7, 90.4]	96.0 [94.8, 96.9]	92.1 [90.7, 93.5]
Min-p (after T)	85.8 [84.4, 87.1]	**97.0** [96.0, 97.8]	**92.5** [91.0, 93.8]
P-less (after T)	89.1 [87.7, 90.5]	96.1 [95.0, 97.0]	91.8 [90.3, 93.2]
Global KL*	89.0 [87.6, 90.4]	96.1 [95.0, 97.0]	91.9 [90.5, 93.3]
Local KL*	88.8 [87.4, 90.2]	96.1 [95.0, 97.0]	91.8 [90.4, 93.3]
KL-budget	87.0 [85.7, 88.4]	96.7 [95.6, 97.6]	92.2 [90.8, 93.6]

Method	Qwen2.5 1.5B $n = 1095$	Mistral 7B $n = 1095$	Qwen2.5 7B $n = 1095$
Greedy	67.1	52.3	88.7
Temperature	86.5	83.1	96.3
Top-p (before T)	89.9	**84.6**	95.8
Min-p (before T)	88.7	83.0	96.5
P-less (before T)	83.1	73.8	93.0
Uniform min-p	83.3	76.8	96.5
Top-p (after T)	**91.4**	83.3	95.8
Min-p (after T)	89.8	83.0	**96.9**
P-less (after T)	89.9	83.0	96.2
Global KL*	88.0	82.2	96.1
Local KL*	89.2	81.9	96.0
KL-budget	88.9	81.9	96.3

Table 16: **Pass@8 on GSM8K.** Percentages on the previously unused questions from the official evaluation split, with development settings frozen. Bold marks every largest unrounded mean in each column, not statistical significance. Gate labels specify whether truncation is applied before or after temperature T. All stochastic methods use eight attempts. Greedy uses one. All invalid and capped responses remain outcomes.

Table 17: Completion and extraction on complete evaluation splits. Capped and parsed values are percentages. Tokens are per generated response. Values are descriptive means. Question-level intervals are in the evidence. Parsing uses the primary extractor for each benchmark. An extracted answer need not be correct. All capped and unparseable outputs remain outcomes. Gate labels specify order relative to temperature T. Bold marks the lowest capped share in each block, with ties all bold.

Method	Capped	Parsed	Mean tokens
Qwen2.5 1.5B / GSM8K ($n = 1319$, **8 paths**)			
Greedy	0.2	100.0	158.5
Temperature	**0.0**	100.0	169.5
Top-p (before T)	**0.0**	100.0	164.3
Min-p (before T)	**0.0**	100.0	170.1
P-less (before T)	0.3	100.0	162.2
Uniform min-p	0.2	100.0	190.0
Top-p (after T)	0.2	100.0	162.8
Min-p (after T)	0.1	100.0	161.2
P-less (after T)	0.2	100.0	172.8
Global KL*	0.1	100.0	164.5
Local KL*	0.1	100.0	164.0
KL-budget	**0.0**	100.0	167.2
Mistral 7B / GSM8K ($n = 1319$, **8 paths**)			
Greedy	6.3	100.0	283.7
Temperature	7.9	100.0	350.4
Top-p (before T)	5.0	100.0	353.2
Min-p (before T)	5.7	100.0	380.2
P-less (before T)	7.3	100.0	294.8
Uniform min-p	**2.7**	100.0	402.2
Top-p (after T)	5.7	100.0	359.0
Min-p (after T)	9.2	100.0	355.6
P-less (after T)	6.2	100.0	361.5
Global KL*	8.4	100.0	384.4
Local KL*	9.7	100.0	398.3
KL-budget	3.0	100.0	372.3
Qwen2.5 7B / GSM8K ($n = 1319$, **8 paths**)			
Greedy	**0.0**	100.0	163.8
Temperature	3.8	100.0	249.2
Top-p (before T)	**0.0**	100.0	166.2

Method	Capped	Parsed	Mean tokens
Min-p (before T)	**0.0**	100.0	168.7
P-less (before T)	**0.0**	100.0	165.3
Uniform min-p	**0.0**	100.0	172.3
Top-p (after T)	**0.0**	100.0	165.8
Min-p (after T)	**0.0**	100.0	174.3
P-less (after T)	**0.0**	100.0	165.5
Global KL*	**0.0**	100.0	166.0
Local KL*	**0.0**	100.0	166.2
KL-budget	**0.0**	100.0	170.2

Table 18: GSM8K extraction sensitivity on complete splits, at unchanged sampler settings. Primary and sensitivity scoring use identical outputs and questions. No retuning or confirmatory significance claim is made for this diagnostic. Values are percentages. Descriptive intervals and all question-level outcomes are in the evidence. Greedy uses one draw. Bold marks the largest value in each checkpoint and column, with ties all bold.

Method	Primary pass	Requested marker pass	Primary vote	Requested marker vote
Qwen2.5 1.5B ($n = 1319$**, 8 paths)**				
Greedy	67.4	51.7	67.4	51.7
Temperature	87.3	79.9	72.3	68.4
Top-p (before T)	89.9	**82.9**	74.9	71.9
Min-p (before T)	88.8	79.9	73.4	67.1
P-less (before T)	83.3	71.8	70.9	63.6
Uniform min-p	83.4	73.3	67.2	60.7
Top-p (after T)	**91.4**	82.5	**75.0**	71.2
Min-p (after T)	90.1	82.3	**75.0**	**72.2**
P-less (after T)	89.8	82.0	74.9	70.9
Global KL*	88.2	78.9	73.8	69.2
Local KL*	89.4	79.9	74.5	69.2
KL-budget	89.2	81.0	74.1	69.2
Mistral 7B ($n = 1319$**, 8 paths)**				
Greedy	52.2	51.7	52.2	51.7
Temperature	83.2	82.6	66.3	66.8
Top-p (before T)	**85.0**	**84.4**	**67.3**	**67.8**
Min-p (before T)	82.6	81.1	63.4	62.6
P-less (before T)	73.9	73.8	59.4	60.0
Uniform min-p	77.5	74.3	55.2	54.5
Top-p (after T)	83.5	82.3	66.2	66.5
Min-p (after T)	83.6	83.1	66.8	67.4
P-less (after T)	83.2	81.5	65.8	66.2
Global KL*	83.0	81.7	64.0	64.7
Local KL*	82.3	81.2	64.3	64.4
KL-budget	82.0	80.6	60.4	60.5
Qwen2.5 7B ($n = 1319$**, 8 paths)**				
Greedy	89.1	87.9	89.1	87.9
Temperature	96.4	96.1	91.1	92.0
Top-p (before T)	96.1	96.1	91.8	92.2
Min-p (before T)	96.7	96.9	92.3	**93.0**
P-less (before T)	93.0	92.9	90.1	90.6
Uniform min-p	96.7	96.7	91.9	92.3
Top-p (after T)	96.0	96.1	92.1	92.6
Min-p (after T)	**97.0**	**97.0**	**92.5**	92.7
P-less (after T)	96.1	96.4	91.8	92.2
Global KL*	96.1	96.2	91.9	92.3
Local KL*	96.1	96.3	91.8	92.3
KL-budget	96.7	96.9	92.2	**93.0**

Checkpoint	Weights	GSM8K	MATH-500	GPQA
Qwen2.5 1.5B	Q8_0	128 × 8	96 × 8	n/a
Mistral 7B	Q8_0	128 × 8	n/a	n/a
Qwen2.5 7B	Q8_0	128 × 8	253 × 8	n/a
Qwen3.8 27B	Q8_0	64 × 8	32 × 8	n/a
Llama 3.3 70B	Q8_0	64 × 8	32 × 8	n/a
Qwen2.5 72B	Q8_0	64 × 8	32 × 8	n/a
Mistral Small 4 119B†	Q4_K_M	64 × 8	32 × 8	32 × 8
Mistral Large 123B	Q4_K_M	64 × 8	32 × 8	n/a

Table 20: Subset-study coverage: questions × paths per stochastic method. Greedy uses one draw. Counts are whole balanced blocks. Missing or partial blocks have no reported accuracy. †Mistral Small 4 has 119B total and 6.5B active parameters. Quantization and training differ across checkpoints. Comparisons of samplers are within a checkpoint. The release pins full model revisions and weight checksums. The complete test-split evaluation is in Figure 10.

Complete variation census. Table 19 includes all completed checkpoints and all eleven stochastic methods in both views. On the full split, global KL* appears in four of the six checkpoint–view comparisons and local KL* in two. Neither appears on Mistral 7B. These are point-estimate memberships, not significant dominance claims. The shared-correct view retains 783, 480, and 1,180 questions at Qwen1.5B, Mistral 7B, and Qwen 7B, respectively. On previously unused questions, global remains on both Qwen frontiers, whereas local appears only in the 1.5B all-response view. Neither appears on either Mistral view. Every view and competitor is retained in the evidence.

Complete contrast census. Figure 17 displays all 216 planned full-test contrasts, three KL methods against eight baselines at three endpoints on each checkpoint, and Figure 9 summarizes them. Resolution uses the same simultaneous intervals as the rest of this section. Every resolved coverage and voting gain is against truncation-first P-less or uniform min-p. Resolved losses concentrate in individual-path accuracy and in KL-budget on Mistral 7B. On the 1,095 previously unused questions, with the same 72-contrast adjustment per checkpoint, the totals of resolved gains / unresolved / resolved losses are 17 / 47 / 8 for global KL*, 16 / 50 / 6 for local KL*, and 13 / 39 / 20 for KL-budget. The only resolved coverage or voting loss of either KL* rule is again temperature-first top-p coverage at 1.5B, 3.4 points (adjusted interval [0.5, 6.4]). The census is generated from the verified result files, and its source checksums are recorded with the output.

Table 19: Complete census of observed accuracy–response-variation frontiers. Every stochastic method is included. Membership uses point estimates and is not a significance test. A dash means the method is dominated, and n/a means no shared question set satisfies that view. Bold marks frontier membership.

Checkpoint / task	Response view	n	Global	Local	Budget
Qwen2.5 1.5B / GSM8K	All responses	1319	**Yes**	**Yes**	–
Qwen2.5 1.5B / GSM8K	Shared correct	783	**Yes**	**Yes**	–
Mistral 7B / GSM8K	All responses	1319	–	–	**Yes**
Mistral 7B / GSM8K	Shared correct	480	–	–	**Yes**
Qwen2.5 7B / GSM8K	All responses	1319	**Yes**	–	**Yes**
Qwen2.5 7B / GSM8K	Shared correct	1180	**Yes**	–	**Yes**

The following protocol sections describe the smaller-sample subset studies unless explicitly stated otherwise.

P.2 Models, questions, and the development grid

The data manifest pins repository revisions, file checksums, original row identifiers, option order, and complete prompts. A permutation with seed 20260914 fixes the question order. The eight GSM8K worked examples are training rows 612, 1922, 2223, 2294, 2476, 5386, 6141, and 6923.

Qwen2.5 1.5B ($n = 1{,}319$): 20 favor KL, 11 favor a baseline, 41 not separated

		Temp.	Top-p before T	Min-p before T	P-less before T	Uniform min-p	Top-p after T	Min-p after T	P-less after T
Individual	Global KL*	**+5.8**	+1.5	**+4.8**	**-4.8**	**+14.6**	**-3.0**	**-2.8**	+1.0
	Local KL*	**+6.1**	**+1.8**	**+5.1**	**-4.5**	**+14.9**	**-2.7**	**-2.5**	+1.3
	KL-budget	**+2.8**	-1.4	**+1.9**	**-7.7**	**+11.7**	**-5.9**	**-5.7**	**-1.9**
Pass@8	Global KL*	+0.9	-1.7	-0.5	**+4.9**	**+4.9**	**-3.2**	-1.8	-1.6
	Local KL*	+2.0	-0.5	+0.6	**+6.1**	**+6.0**	-2.0	-0.7	-0.5
	KL-budget	+1.9	-0.7	+0.5	**+5.9**	**+5.8**	-2.2	-0.8	-0.6
Vote@8	Global KL*	+1.5	-1.1	+0.4	+2.9	**+6.6**	-1.1	-1.2	-1.0
	Local KL*	+2.1	-0.4	+1.0	**+3.6**	**+7.3**	-0.5	-0.5	-0.4
	KL-budget	+1.8	-0.8	+0.7	+3.2	**+6.9**	-0.9	-0.9	-0.8

Mistral 7B ($n = 1{,}319$): 16 favor KL, 19 favor a baseline, 37 not separated

		Temp.	Top-p before T	Min-p before T	P-less before T	Uniform min-p	Top-p after T	Min-p after T	P-less after T
Individual	Global KL*	**-4.3**	-0.9	**+3.1**	**-11.8**	**+11.9**	-0.5	**-5.3**	**-2.0**
	Local KL*	**-4.3**	-0.8	**+3.2**	**-11.7**	**+12.0**	-0.4	**-5.3**	-2.0
	KL-budget	**-10.7**	**-7.2**	**-3.3**	**-18.2**	**+5.5**	**-6.8**	**-11.7**	**-8.4**
Pass@8	Global KL*	-0.2	-2.0	+0.5	**+9.1**	**+5.5**	-0.5	-0.6	-0.2
	Local KL*	-1.0	-2.7	-0.3	**+8.3**	**+4.8**	-1.2	-1.4	-1.0
	KL-budget	-1.2	-3.0	-0.5	**+8.1**	**+4.5**	-1.4	-1.6	-1.2
Vote@8	Global KL*	-2.3	-3.3	+0.6	**+4.6**	**+8.8**	-2.1	-2.8	-1.8
	Local KL*	-2.0	-3.0	+0.9	**+4.9**	**+9.1**	-1.9	-2.5	-1.5
	KL-budget	**-5.9**	**-6.9**	-3.0	+1.0	**+5.2**	**-5.8**	**-6.4**	**-5.4**

Qwen2.5 7B ($n = 1{,}319$): 14 favor KL, 4 favor a baseline, 54 not separated

		Temp.	Top-p before T	Min-p before T	P-less before T	Uniform min-p	Top-p after T	Min-p after T	P-less after T
Individual	Global KL*	**+8.4**	+0.9	+1.0	-0.4	**+2.6**	-0.0	**+3.3**	-0.1
	Local KL*	**+8.2**	+0.7	+0.8	-0.6	**+2.4**	-0.2	**+3.1**	-0.3
	KL-budget	**+6.4**	-1.1	**-1.0**	**-2.4**	+0.6	**-2.0**	**+1.3**	**-2.1**
Pass@8	Global KL*	-0.2	+0.1	-0.6	**+3.1**	-0.5	+0.2	-0.9	+0.0
	Local KL*	-0.2	+0.1	-0.6	**+3.1**	-0.5	+0.2	-0.9	+0.0
	KL-budget	+0.4	+0.7	+0.0	**+3.7**	+0.1	+0.8	-0.3	+0.6
Vote@8	Global KL*	+0.9	+0.2	-0.4	**+1.8**	+0.0	-0.2	-0.5	+0.2
	Local KL*	+0.8	+0.1	-0.5	**+1.7**	-0.1	-0.3	-0.6	+0.1
	KL-budget	+1.2	+0.5	-0.1	**+2.1**	+0.3	+0.1	-0.2	+0.5

Simultaneous 95% interval above zero (KL method higher)
Simultaneous 95% interval below zero (baseline higher)
Interval includes zero

Figure 17: **Every planned full-test contrast.** Cells give the KL method minus the baseline in percentage points on all 1,319 GSM8K test questions. Blue and red cells have simultaneous 95% intervals, Bonferroni-adjusted over the 72 contrasts of their checkpoint, entirely above or below zero. Gray cells are unresolved and do not establish equivalence. Individual is mean path accuracy. Pass@8 and Vote@8 use eight attempts. The KL* rules have no development-selected parameter. Uniform min-p fixes $\alpha = .05$. The other baselines use their frozen development selections.

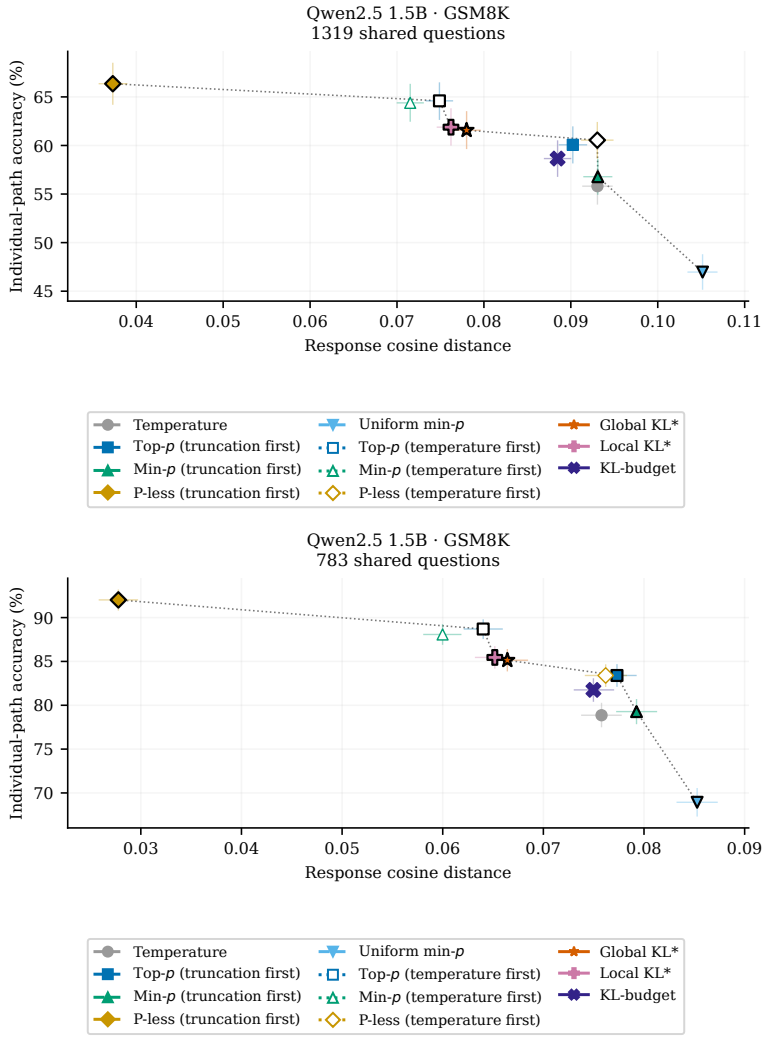


Figure 18: **Complete Qwen2.5 1.5B variation diagnostics.** The upper panel uses all eight responses on all 1,319 questions. The lower panel uses the 783 questions where every method has at least two correct responses: only distance excludes incorrect paths, while accuracy still scores all eight. Both axes use the same questions, and all eleven stochastic methods are displayed. Both global and local KL* lie on both observed frontiers. Black outlines mark every nondominated point estimate. Bars are descriptive 95% question intervals. Distance measures response variation, not distinct verified reasoning strategies, and membership is not a significance test.

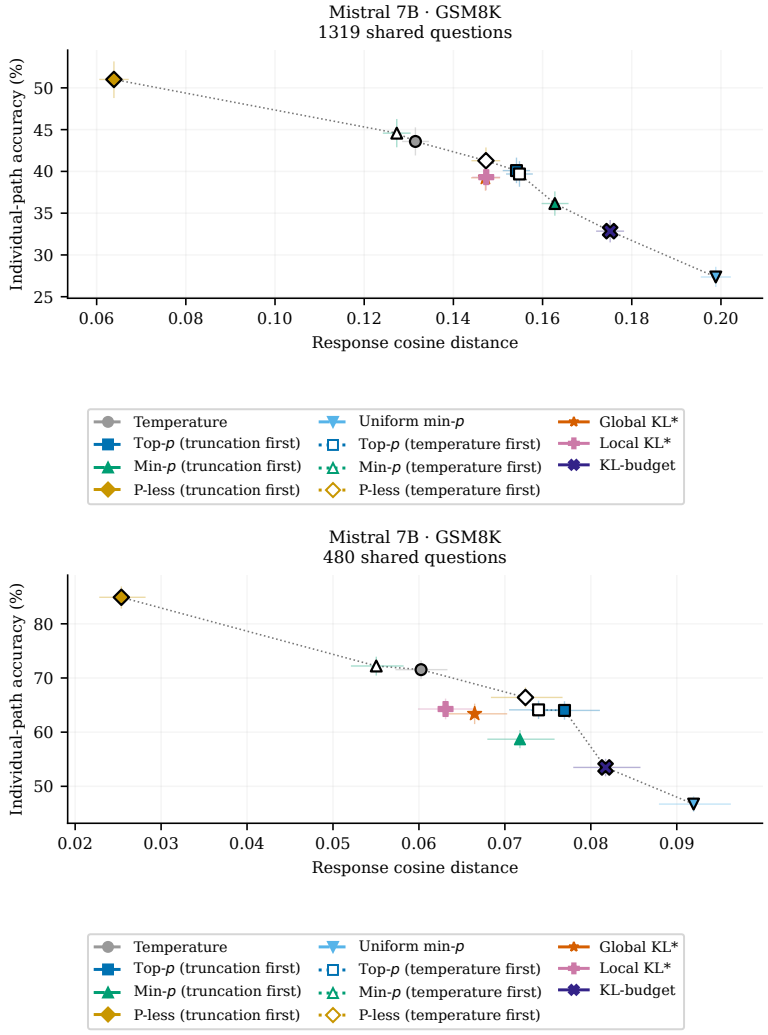


Figure 19: **Complete Mistral 7B variation diagnostics.** The upper panel uses all eight responses on all 1,319 questions. The lower panel uses the 480 questions where every method has at least two correct responses: only distance excludes incorrect paths, while accuracy still scores all eight. Both axes use the same questions, and all eleven stochastic methods are displayed. Neither global nor local KL* lies on either observed frontier. Black outlines mark every nondominated point estimate. Bars are descriptive 95% question intervals. Distance measures response variation, not distinct verified reasoning strategies, and membership is not a significance test.

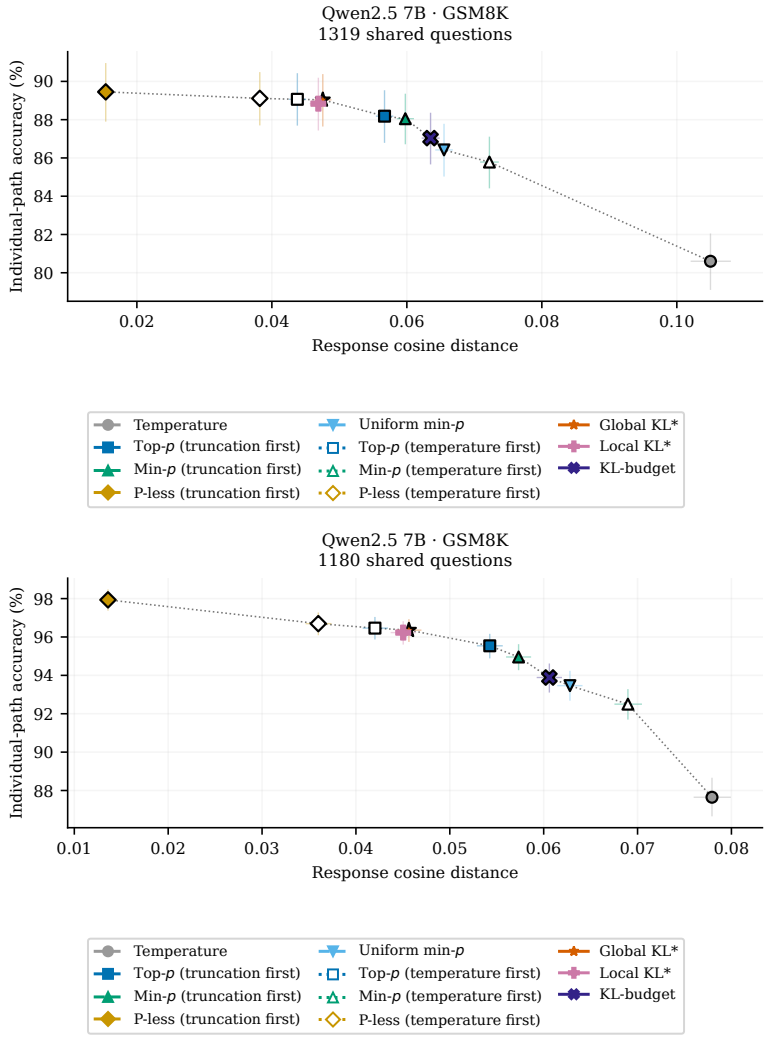


Figure 20: **Complete Qwen2.5 7B variation diagnostics.** The upper panel uses all eight responses on all 1,319 questions. The lower panel uses the 1,180 questions where every method has at least two correct responses: only distance excludes incorrect paths, while accuracy still scores all eight. Both axes use the same questions, and all eleven stochastic methods are displayed. Global KL* lies on both observed frontiers. Local KL* lies on neither. Black outlines mark every nondominated point estimate. Bars are descriptive 95% question intervals. Distance measures response variation, not distinct verified reasoning strategies, and membership is not a significance test.

Family	Development candidates	Count
Temperature	$T = .7, 1, 1.5$	3
Top-p	$(p, T) = (.9, .7), (.9, 1), (.9, 2), (.95, 1)$	4
Min-p	$(\alpha, T) = (.05, .7), (.05, 1), (.05, 2), (.1, 1), (.1, 2)$	5
P-less	Order two. $T = .7, 1, 2$	3
KL-budget	Min-p .05. $\epsilon = .03, .1, .3$ nats	3

Table 21: The complete 18-setting development grid. Each setting receives the same 24 questions and eight paths. One setting per family is selected by pass@8, then Vote@8, single-answer accuracy, and a fixed lexicographic identifier. Global/local KL* and the uniform min-p control do not enter this selection.

Model	Temperature	Top-p	Min-p	P-less	KL-budget
Qwen2.5 1.5B	$T = 1$	$T = 1, p = 0.95$	$T = 2, \alpha = 0.05$	$T = 2$	$\epsilon = 0.03$
Mistral 7B	$T = 0.7$	$T = 1, p = 0.95$	$T = 2, \alpha = 0.1$	$T = 2$	$\epsilon = 0.1$
Qwen2.5 7B	$T = 1.5$	$T = 1, p = 0.95$	$T = 2, \alpha = 0.05$	$T = 1$	$\epsilon = 0.3$
Qwen3.8 27B	$T = 0.7$	$T = 1, p = 0.95$	$T = 1, \alpha = 0.05$	$T = 0.7$	$\epsilon = 0.03$
Llama 3.3 70B	$T = 0.7$	$T = 1, p = 0.95$	$T = 0.7, \alpha = 0.05$	$T = 0.7$	$\epsilon = 0.03$
Qwen2.5 72B	$T = 0.7$	$T = 0.7, p = 0.9$	$T = 0.7, \alpha = 0.05$	$T = 0.7$	$\epsilon = 0.03$
Mistral Small 4 119B	$T = 0.7$	$T = 2, p = 0.9$	$T = 1, \alpha = 0.05$	$T = 2$	$\epsilon = 0.03$
Mistral Large 123B	$T = 0.7$	$T = 1, p = 0.95$	$T = 1, \alpha = 0.05$	$T = 2$	$\epsilon = 0.03$

Table 22: Settings selected on 24 development questions with eight paths each, then frozen for all confirmation tasks and the sixteen-path extension. Global/local KL* have no development-selected parameter. Uniform min-p fixes $\alpha = .05$. This table records settings, not held-out performance rankings.

The 24 development questions exclude these examples. The test permutation excludes 64 GSM8K pilot questions, and another 32 GSM8K and 32 MATH-500 pilot questions are excluded for every checkpoint. Repeated requests use seeds 4100–4107 on development and 5100–5115 on evaluation. Within a question, all methods see the same prompt and option order. The complete manifests accompany the release.

The different family grid sizes reflect their available parameters: min-p and top-p vary both a threshold and a temperature, whereas P-less has only a temperature and KL-budget has only an allowance after its fixed gate. The selection objective is eight-attempt coverage. A table's Single and Vote columns evaluate that same coverage-selected setting. All three endpoints reuse one setting per family. The development questions and all settings' outputs remain available.

P.3 Answer scoring

A pilot required the requested final-answer marker. Its 1.5B GSM8K outputs had 54.7–74.6% marker coverage across stochastic methods despite few output-limit hits, because completed answers often appeared in prose. Before scoring any confirmation question, we therefore switched to the evaluation harness's flexible numerical endpoint and excluded the pilot questions.

The numerical extractor follows the pinned harness source at commit `ad8737ae7fad24cf64e50fc7fc31397bff586b9e` (EleutherAI, 2026). It takes the last regex-matched number and applies the harness's string normalization, including removing commas and terminal periods. It can accept a last intermediate number and does not equate every numerically equivalent string. This behavior is frozen for every method. The original requested-marker score is also computed for every new GSM8K output. The released min-p results use a strict harness endpoint and different demonstrations. Our comparison follows their tasks and experimental questions, with these explicit protocol differences. Table 27 compares both extraction rules without retuning. They need not give nested correctness sets: a response can place another number after a correct marked answer, making marker scoring higher than the last-number endpoint. Conversely, a correct answer in prose can fail the requested-marker rule.

MATH-500 requires an explicit boxed final answer and uses `math-verify` 0.8.0, `latex2sympy2-extended` 1.10.2, `antlr4-python3-runtime` 4.13.2, and `sympy` 1.14.0.

Method	1	2	3	4	5	6	7	8	Correct
Greedy	3	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0/1
Temperature	3	3	3	3	3	12	3	3	0/8
Top-p	3	3	3	6	3	3	3	3	0/8
Min-p	3	3	4	4	3	3	4	3	**3/8**
P-less	3	3	3	3	3	3	3	3	0/8
Uniform min-p	7	3	4	3	3	3	4	3	2/8
Global KL*	4	4	3	3	3	3	3	3	2/8
Local KL*	3	3	3	3	4	3	3	3	1/8
KL-budget	3	3	4	3	3	3	4	4	**3/8**

Table 23: **A sampled alternative can recover the correct counting operation.** Qwen2.5 1.5B answers on the fixed illustrative GSM8K question 770. Columns follow the first eight seeds. Greedy has one draw. Blue marks the correct final answer, and bold marks both largest correct-answer fractions: min-p and KL-budget. This selected example illustrates answer coverage on one question.

The release checks every reference answer against itself and tests nontrivial equivalent and inequivalent expressions. MATH votes group symbolically equivalent final answers without consulting the gold answer, using deterministic first-seed representatives and symmetric equivalence checks. Multiple-choice tasks use the fixed final-answer extractor and the common option permutation. An invalid extraction is incorrect and abstains from a vote. A group with no valid votes is incorrect. Token-limited responses remain in every denominator.

P.4 An illustrative set of eight answers

We select, post hoc, the first local-KL*/top-p coverage disagreement in the frozen 128-question order. Four vanilla varieties sell five bottles each. Eight fruity varieties sell two each. The required difference is $4 \cdot 5 - 8 \cdot 2 = 4$. Table 23 shows every method's answers. Top-p never finds four, whereas KL-budget finds it three times, global KL* twice, and local KL* once. The recorded local-KL* answer explicitly computes the two totals and their difference. Min-p also finds the correct answer three times. Every method's majority answer is wrong: the example separates useful coverage from successful answer selection.

A source-linked diagnostic reads the first correct output on each of the twelve local-KL*/top-p coverage disagreements, including both directions. All twelve inspected outputs end normally and state the reference final answer. Some contain flawed intermediate bookkeeping. This qualitative check changes no scores and does not estimate the frequency of reasoning errors. Our endpoint evaluates final answers, and the semantic-variation analysis does not certify entire derivations.

P.5 Multiple-path estimators and uncertainty

Single-answer accuracy averages the correctness indicators of all generated paths. For Pass@k, the primary endpoint checks whether the first k fixed evaluation seeds contain a correct answer. Vote@k uses the most frequent normalized answer among those paths. It assigns fractional credit if the gold answer is tied with other answers. The same first eight paths define the eight-path endpoint in a sixteen-path extension. Curves use these fixed prefixes, including the first path at $k = 1$. Their first point can differ from Single, which averages all paths. Greedy contributes one actual draw at every displayed path count.

For a question with c correct answers among n independent sampled paths, we also release the unbiased estimator $1 - \binom{n-c}{k} / \binom{n}{k}$ of pass@k for $k \leq n$. This is the estimator used by Chen et al. (2021). It is distinct from the fixed-prefix endpoint and does not determine the main rankings. All uncertainty treats a question, with its full collection of paths, as one unit. Binary question accuracy has a Wilson interval. Fractional single-answer and vote outcomes use 4,000 question-bootstrap resamples. An all-zero or all-one bounded question score instead receives a conservative boundary interval based on the number of independent questions. The degenerate bootstrap interval is retained in the raw analysis. This prevents 256 correct paths on 32 questions from being reported as a population accuracy known without uncertainty. For n observed unit scores the interval is $[0.025^{1/n}, 1]$, with

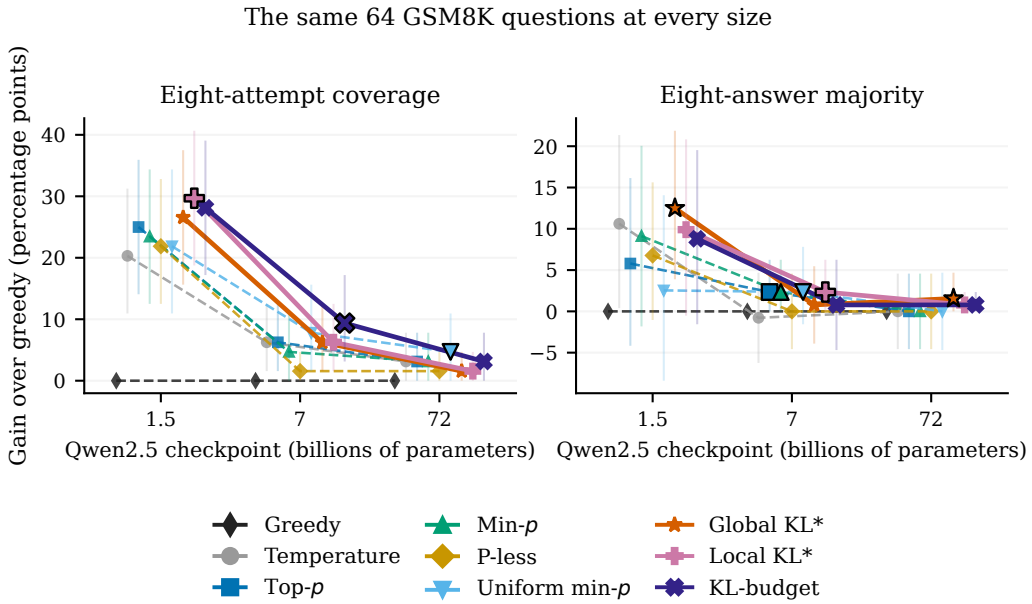


Figure 21: **Eight paths create more opportunity when the first answer is weaker.** Every completed Qwen2.5 checkpoint uses the same 64 GSM8K questions and eight paths. Local KL* has the largest coverage gain at 1.5B, KL-budget at 7B, and uniform min-p at 72B. Majority rankings vary by size. Black outlines mark every tied maximum. Methods are offset within each checkpoint for visibility. Bars give pointwise 95% paired intervals, with finite uncertainty after zero observed changes. This matched subset is distinct from the broader 128-question 1.5B estimate.

the reflected bound for zeros: for a score $Y \in [0, 1]$, $\Pr(Y = 1) \leq \mathbb{E}[Y]$. The same interval code supplies accuracy axes in the exploratory diversity plots. Paired comparisons resample the same questions for both methods. Bonferroni-adjusted intervals cover five baseline contrasts for one KL variant, endpoint, checkpoint, and task. They do not cover a search across every model and endpoint. Complete paired results accompany the plotted means.

A post-hoc sensitivity adds two-sided exact McNemar tests for every binary Pass@8 and Pass@16 comparison. The conditional binomial test uses the questions covered by only one method. All five baselines remain in its Bonferroni adjustment. We retain both discordant counts and the original bootstrap intervals. With few discordant questions, a percentile bootstrap can exclude zero while the exact test remains inconclusive. On the 128-question 1.5B GSM8K block, local KL* has eleven wins and one loss against temperature, and thirteen wins and one loss against uniform min-p. Their adjusted values are .0317 and .0092. These families are specific to one KL method, endpoint, model, and task. Fractional Single and Vote scores do not enter this binary test.

The model-size comparison uses the shared question intersection across every completed Qwen2.5 checkpoint. It plots each method's paired change from the same recorded greedy answers. When every sampled question has zero change, we retain the zero bootstrap interval in the evidence and display the finite bound $\pm(1 - 0.05^{1/n})$ for n independent questions. A question's change lies in $[-1, 1]$, so its absolute population mean cannot exceed the probability of a nonzero change. Only the identical greedy reference is known to have exactly zero change.

P.6 Inference, completion, and exact sampler execution

The subset studies use two H100 NVL GPUs, each serving one checkpoint at a time. Requests use Q8_0 key/value caches, no context shifting, and a prompt-length preflight. The per-cohort slot context must fit the full rendered prompt and complete output allowance. GSM8K uses a 4,096-token context and allows 1,536 output tokens. The core MATH-500 and GPQA runs allow 2,048. Output allowances are identical across methods within a comparison. Thinking is disabled for the Qwen3.8

template, Mistral Small 4 uses `reasoning_effort=none`, and the Llama 3.3 chat template uses a fixed system date. Every model load must confirm that all layers fit on the GPU. Exact commands, template arguments, model revisions, and weight checksums are retained.

A request identity covers the rendered prompt, data and model revisions, method, seed, and output allowance. Infrastructure failures are separate from valid generations that give wrong or malformed answers. Results require a complete balanced block of all nine methods on the declared questions. Compatible whole blocks can be pooled, counting each question–seed identity once, and partial blocks are never reported. Some MATH-500 and GPQA comparisons use balanced 16-question halves, and tables always report the actual question count.

The native backend starts from `llama.cpp` commit `4d917609` and the checksum-pinned research patches. Global selection uses certified bounds to skip prefixes that cannot beat the best candidate. Local selection implements its own stopping rule. The budget solver combines bracketed Newton steps with bisection and returns the feasible side of the bracket. Exact uniform draws set admitted finite logits to zero and preserve excluded logits, giving exact uniform categorical draws. Native acknowledgements and returned generation settings are checked for every request.

Independent candidate arrays and random-number states allow sampling across batch slots to run in parallel. Sampling runs on eight or sixteen threads, set at model load. For each count, a 768-decision full-vocabulary test checks bitwise agreement with serial execution for admitted candidates, logits, probabilities, and token draws on identical input logits. Grammar, speculative decoding, and reasoning-budget cases retain the serial path. This checks sampler parity. Different GPU batches can still produce slightly different model logits. Each model load records its binary checksum and thread count.

Certificate for global search. Let $g_i = z_1 - z_i$ be ranked logit gaps and $C_h = \sum_{i \le h} g_i$. Up to a constant independent of k, the canonical objective is $f(k) = C_k/k - \log k$. After sorting the leading $h = 512$ candidates, every later prefix satisfies

$$f(k) \ge L_h(k) = g_h + \frac{C_h - h g_h}{k} - \log k, \qquad k > h.$$

Since $C_h - h g_h \le 0$, the derivative of L_h has at most one zero, a maximum. Its minimum over $[h+1, V]$ is therefore at an endpoint. If both endpoint bounds strictly exceed the best score in the sorted head, that head contains a global optimum. A tighter certificate rounds the remaining gaps downward into 256 bins. Each bin gives an affine lower bound on cumulative gaps, and hence a score of the same form whose endpoints suffice. These bins only certify that a tail cannot win. They do not approximate the probabilities returned by the sampler. Floating-point margins reject borderline certificates and trigger the full global search.

P.7 Recorded candidate sets and sampler cost

The recorded next-token examples in Appendix M show the mechanism directly. In a confident Llama continuation, the KL rules keep one candidate, so an infinite-temperature draw is deterministic. In a less certain continuation, they admit several alternatives. Global and local KL* agree on all 98 recorded within-response contexts across the three 70B-class checkpoints, with median support size one. Their different free-generation outcomes and the formal counterexample nevertheless justify treating them as separate algorithms.

The exact global search need not sort every tail token in typical contexts. Appendix L reports its cost on recorded logits and across vocabulary sizes. The release includes the native parity checks, raw timings, and generated-token costs, which keep sampler overhead distinct from the cost of generating more reasoning paths.

P.8 Exploratory diversity of reasoning text

After the first accuracy results and before any embedding scores, we specified an exploratory semantic-variation analysis. It reuses the writing study's BGE-M3 revision, float32 normalized dense embeddings, and verified cache implementation (Chen et al., 2024). The full visible response is encoded, without shortening it. Every pair among the first eight paths contributes to the all-path

Checkpoint	Task	Shared	Global	Local	Budget	Controls
Qwen2.5 1.5B	GSM8K	80/128	–	**Yes**	**Yes**	4
Qwen2.5 1.5B	MATH-500	41/96	**Yes**	–	**Yes**	5
Mistral 7B	GSM8K	47/128	**Yes**	–	**Yes**	4
Qwen2.5 7B	GSM8K	112/128	**Yes**	–	**Yes**	4
Qwen2.5 7B	MATH-500	188/253	**Yes**	–	–	5
Qwen3.8 27B	GSM8K	61/64	**Yes**	–	–	3
Qwen3.8 27B	MATH-500	27/32	–	–	**Yes**	5
Llama 3.3 70B	GSM8K	61/64	–	–	**Yes**	3
Llama 3.3 70B	MATH-500	24/32	–	–	**Yes**	2
Qwen2.5 72B	GSM8K	59/64	**Yes**	–	**Yes**	3
Qwen2.5 72B	MATH-500	27/32	–	**Yes**	**Yes**	1
Mistral Small 4 119B	GSM8K	55/64	–	–	**Yes**	5
Mistral Small 4 119B	MATH-500	18/32	**Yes**	**Yes**	–	2
Mistral Large 123B	GSM8K	60/64	**Yes**	–	–	3
Mistral Large 123B	MATH-500	20/32	–	–	**Yes**	2

Table 24: **Complete census of observed accuracy–variation frontiers.** Yes means nondominated in both shared-subset individual accuracy and correct-response distance among all eight stochastic methods. Controls counts the nondominated temperature, top-p, min-p, P-less, and uniform min-p methods. Global KL* appears in 8/15 comparisons and local KL* in 3/15. At least one appears in 10/15, and both appear in 1/15. Including KL-budget gives at least one KL method in 15/15. These are point-estimate memberships, without a claim of statistically established dominance. All 15 completed comparisons from the eight-method semantic study are included. Bold marks frontier membership.

cosine-distance mean. Question means receive equal weight. A pair containing an empty response contributes zero. Encoder overflow remains flagged and makes the required semantic endpoint incomplete. Responses remain intact.

For the answer-correct view, we first find the shared intersection of questions where each of the eight stochastic methods has at least two correct final answers. On those same questions, each method's distance averages only its answer-correct pairs. Single-answer accuracy still scores all eight paths on the shared subset. This avoids comparing methods' diversity on different selected questions, while explicitly conditioning the result on the easier shared subset. Each figure reports the shared sample size alongside the full question count. Greedy has one observed response and does not enter pairwise-diversity comparisons. Intervals use 4,000 paired question resamples. Black outlines identify all nondominated point estimates. The intervals describe each point's coordinates. Raw vectors, text identities, eligible question IDs, and every complete comparison accompany the release. Correct final answers do not certify the intermediate reasoning, and semantic distance does not establish human creativity or distinct solution strategies.

A post-hoc paired analysis, specified after the first five complete embedding comparisons, evaluates all three KL methods against all five stochastic baselines. It uses 200,000 question-bootstrap resamples and adjusts over both distance endpoints and every completed checkpoint/task comparison. The complete contrasts, including question-level differences, are released. Equal observed accuracy does not establish population equivalence.

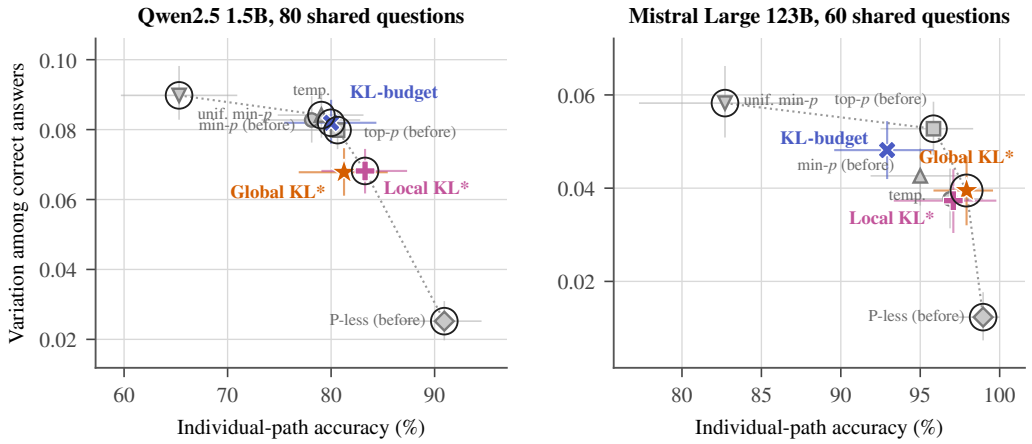


Figure 22: **KL methods are on the observed frontier from 1.5B to 123B.** Smaller-sample subset studies of the eight truncation-first methods. Accuracy uses all eight paths on the shared questions, and variation uses responses with correct answers. Circles mark nondominated means. Appendix Table 24 lists all 15 completed comparisons.

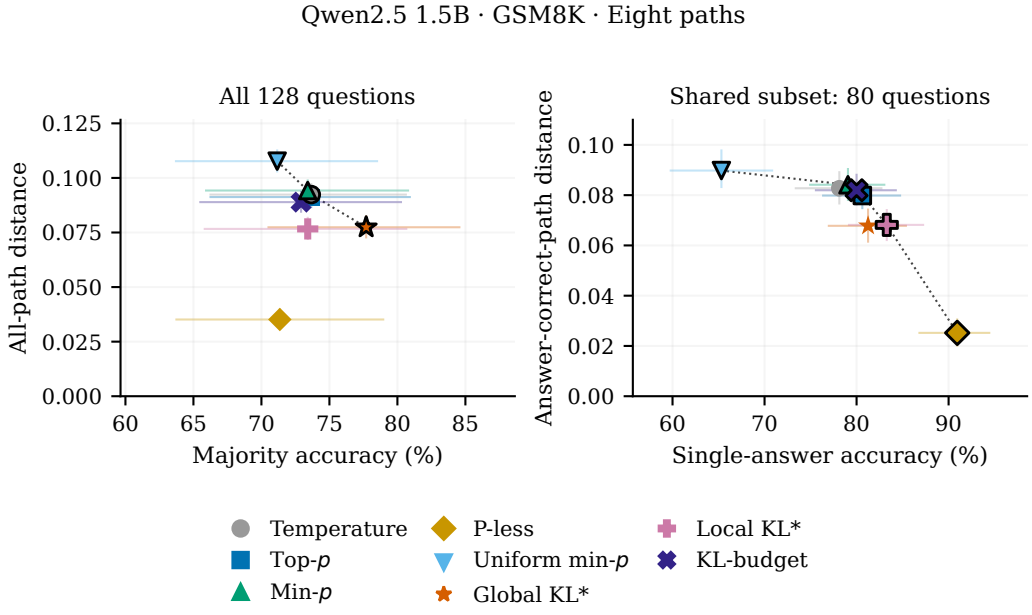


Figure 23: **The full-set accuracy–diversity comparison complements the shared subset.** Left: global KL* has the highest observed majority accuracy across all 128 questions. Right: local KL* and KL-budget lie on the observed frontier for the shared 80-question subset. Distance on the left includes all paths. Distance on the right includes only paths with correct final answers. Accuracy always uses every path on the displayed question set. Bars give 95% question intervals. Outlines mark every nondominated mean.

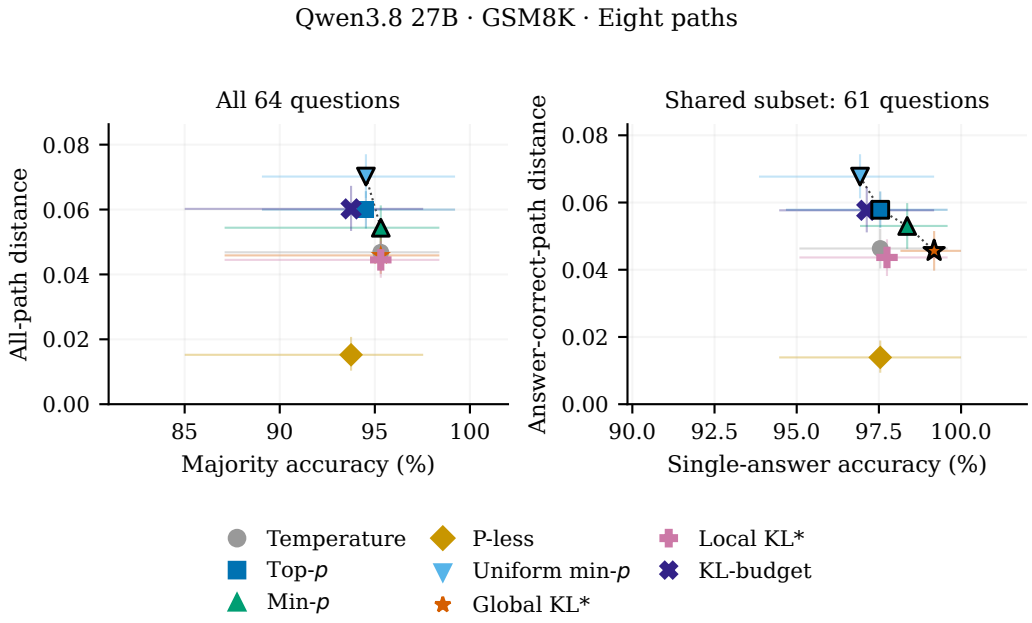


Figure 24: **Global KL* retains more correct-response variation than P-less at higher observed accuracy on the larger 27B sample.** Of 64 Qwen3.8 27B GSM8K questions, 61 satisfy the shared-subset requirement. On this subset, global KL* has 99.2% individual accuracy and correct-response distance 0.0456, versus 97.5% and 0.0139 for P-less. Local KL* has distance 0.0437. Both KL* distance increases over P-less remain positive after the adjustment across all completed semantic comparisons. The leading accuracy differences remain unresolved. Left: full-set majority accuracy and all-path distance. Right: accuracy uses all eight paths on the shared questions, while distance uses only pairs with correct final answers. Bars show 95% question intervals. Black outlines mark all nondominated means.

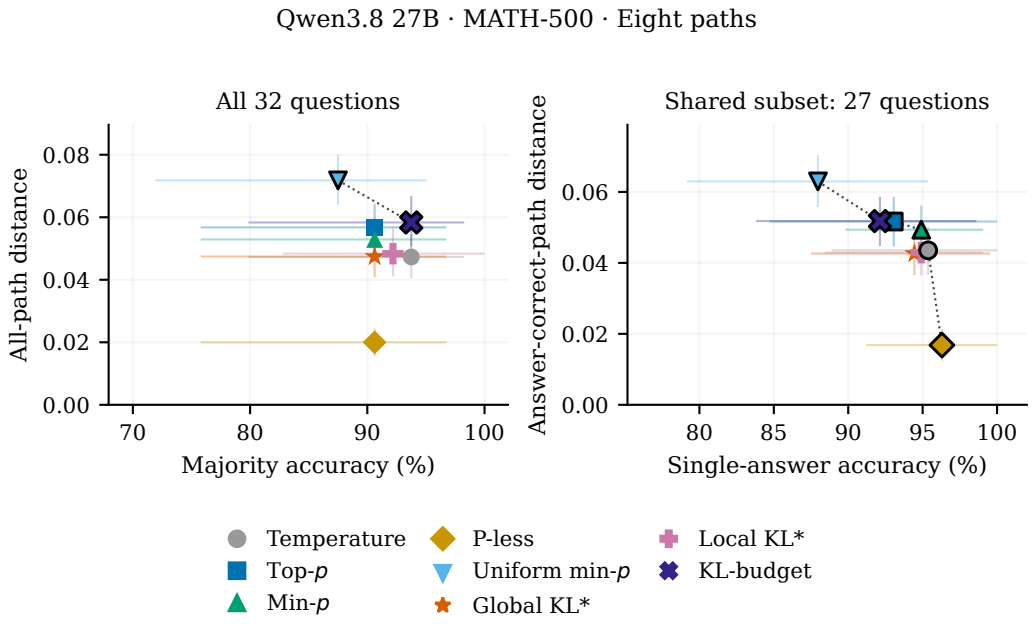


Figure 25: **KL-budget combines strong observed voting with response variation on the expanded 27B mathematics slice.** Across 32 Qwen3.8 27B MATH-500 questions, KL-budget ties temperature for the highest Vote@8, 93.8%, with all-path distance 0.0583 versus 0.0473. On the shared 27 questions, global KL*, local KL*, and KL-budget have correct-response distances 0.0426, 0.0429, and 0.0518, versus 0.0168 for P-less. All three distance contrasts remain positive after the adjustment across all completed semantic comparisons. P-less has the highest shared-subset individual accuracy among stochastic methods, 96.3%. The common output allowance is 2,048 tokens. Bars show 95% question intervals. Black outlines mark all nondominated means.

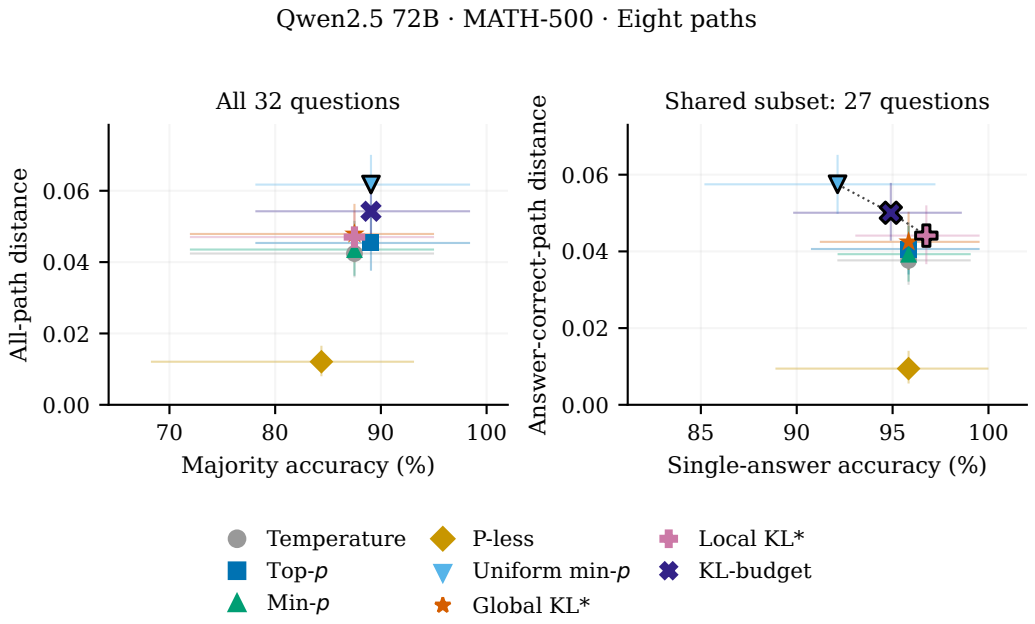


Figure 26: **Uniform KL* draws retain variation on 72B competition mathematics.** Left: KL-budget ties the highest observed majority accuracy across all 32 questions. Right: local KL* has the highest stochastic individual accuracy on the shared 27-question subset and lies on the observed frontier with KL-budget and uniform min-p. Correct-response distances are 0.0425 and 0.0441 for global/local KL*, versus 0.0094 for P-less. Both differences remain positive after the adjustment across all completed semantic comparisons. The shared subset requires two correct final answers from every stochastic method. Bars show 95% question intervals. All nondominated means are outlined.

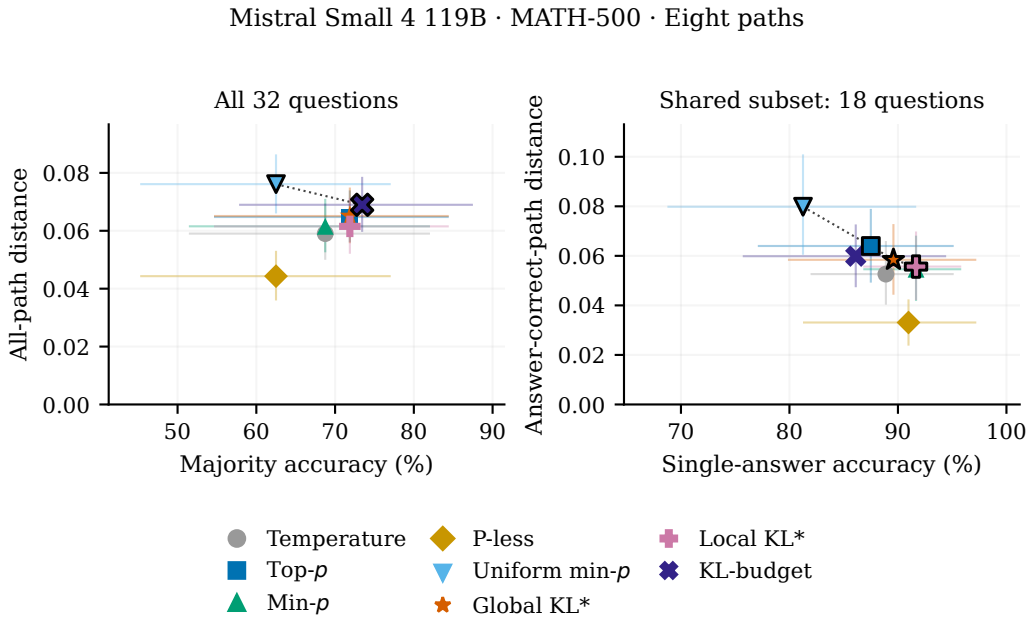


Figure 27: **KL-budget leads observed majority accuracy on 2026 Mistral Small 4 mathematics under a fixed output allowance.** Left: all 32 questions at the common 2,048-token cap. Right: global and local KL* lie on the observed frontier on the shared 18-question subset. Local KL* ties min-p for the highest stochastic individual accuracy. The right panel measures distance only among responses with correct final answers. Bars show 95% question intervals. Outlines mark every nondominated mean. The full panel retains capped responses and does not measure uncapped mathematical ability. The model has 119B total and 6.5B active parameters.

Q SUBSET STUDIES ACROSS MODEL SIZES

These studies use smaller question samples but extend the comparison to checkpoints up to 123B and to MATH-500 and GPQA. For Qwen2.5 1.5B, Mistral 7B, and Qwen2.5 7B, the complete GSM8K test split of Section 4.2 supersedes their GSM8K results. Table 25 summarizes them across all eight checkpoints. The sixteen-path follow-up adds attempts on the same 128 questions: global KL* leads coverage there, whereas temperature leads majority accuracy (Figure 28). The transfer studies likewise favor different rules at different endpoints. On 253 Qwen 7B MATH-500 questions, local KL* has the highest coverage and top-p the highest voting accuracy. The leading coverage difference includes zero. GPQA highlights a limit of coverage itself: eight independent uniform guesses on a four-choice question would include the correct answer about 90% of the time, so individual accuracy and voting are necessary alongside coverage (Appendix Q.3).

Q.1 RESULTS ACROSS MODEL SIZES

Table 25 reports the truncation-first comparison. On its 128 GSM8K questions with Qwen2.5 1.5B, local KL* covers 120 (93.8%), global KL* covers 118 (92.2%), and KL-budget covers 116 (90.6%). Top-p, the strongest baseline, covers 114 (89.1%). Greedy solves 86 (67.2%). All three KL methods recover more solutions with the same eight attempts. Local KL* gains 7.8 points over temperature, with adjusted exact paired $p = .032$ across its complete baseline contrasts (Appendix P). Global and local KL* average 165 and 166 generated tokens per path, respectively, compared with 167 for top-p. Their greater coverage uses similar response lengths as well as the same number of attempts.

Global KL* also leads observed Vote@8 at 77.7%, versus 73.7% for temperature and top-p. Its individual accuracy is 62.6%. Recovering a correct alternative does not always make it the majority answer. Table 25 reports the development objective, pass@8, alongside individual-path accuracy. Appendix R gives every endpoint and paired uncertainty. The paired intervals for the leading majority-vote contrasts include zero, so the observed ranking does not establish that global KL* selects the correct answer more often than the strongest baseline.

The predefined final-answer-marker check also puts both KL* variants first for coverage (87.5%, versus 82.8% for the strongest baseline) and global KL* first for Vote@8 (77.3%). Table 27 reports both endpoints at the same settings.

The larger Qwen2.5 7B GSM8K slice separates methods that tied on 32 questions. On 128 questions, KL-budget covers 127 (99.2%), versus 125 (97.7%) for temperature, the strongest finite-temperature baseline. Global KL* has the highest stochastic individual-path accuracy, 88.7%, while min-p leads Vote@8 at 94.1%. The leading paired coverage difference remains unresolved. These endpoints distinguish finding a correct alternative from repeatedly producing or selecting one. On 128 Mistral 7B questions, local KL* ties top-p for the highest pass@8, 85.9%. Its 13.3-point coverage gain over P-less has adjusted exact $p = .012$. Temperature leads Vote@8 at 66.8%, while P-less leads individual accuracy at 49.3%. The preferred sampler depends on how its answers are used.

On 64 Qwen3.8 27B GSM8K questions, KL-budget ties min-p and top-p for the highest pass@8, 96.9%. Global KL* leads individual accuracy at 94.5%. On 64 Qwen2.5 72B questions, global KL* leads observed Vote@8 at 93.8%, versus 92.2% for temperature, top-p, min-p, and P-less. On 64 Mistral Small 4 questions, local KL* ties min-p for pass@8 (96.9%) and temperature for Vote@8 (92.2%). On 64 Mistral Large questions, all three KL methods tie temperature, top-p, and min-p for pass@8 at 96.9%, versus 93.8% for P-less. Min-p leads Vote@8 at 95.3%. Leading paired coverage differences remain unresolved. The model-size diagnostic uses a shared question intersection (Appendix Figure 21), with sampler comparisons within each checkpoint.

Q.2 FROM EIGHT TO SIXTEEN REASONING PATHS

The sixteen-path experiment revisits the same 128 GSM8K questions and adds eight fresh seeds per stochastic method, keeping the original eight and the frozen settings. Global KL* has the highest observed pass@16, 123/128 (96.1%). Local KL* and KL-budget each cover 121, versus 120 for the strongest baseline, min-p. Local KL* already covers 120 questions with eight paths, matching the best baseline's sixteen-path count. Majority voting behaves differently: temperature leads Vote@16 at 79.0%, versus 78.1% for local KL* and 78.0% for global KL*. The leading differences remain

Method	Qwen2.5 1.5B		Mistral 7B		Qwen2.5 7B	
	GSM	MATH	GSM	MATH	GSM	MATH
Greedy	67.2 / 67.2	55.2 / 55.2	46.9 / 46.9	n/a	89.8 / **89.8**	79.1 / **79.1**
Temperature	85.9 / 57.4	76.0 / 44.5	82.8 / 42.9	n/a	97.7 / 79.3	88.9 / 66.3
Top-p	89.1 / 59.9	76.0 / 49.9	**85.9** / 39.9	n/a	96.9 / 88.4	92.1 / 77.1
Min-p	87.5 / 57.7	78.1 / 40.4	80.5 / 35.5	n/a	96.9 / 87.9	92.9 / 76.3
P-less	86.7 / **67.7**	**80.2** / **58.2**	72.7 / **49.3**	n/a	93.8 / 88.5	88.9 / 78.0
Global KL*	92.2 / 62.6	77.1 / 52.2	82.0 / 38.0	n/a	96.9 / 88.7	91.7 / 77.4
Local KL*	**93.8** / 62.8	**80.2** / 50.0	**85.9** / 38.2	n/a	96.9 / 87.6	**93.7** / 77.4
KL-budget	90.6 / 59.7	79.2 / 45.4	79.7 / 33.1	n/a	**99.2** / 87.8	92.1 / 72.9
Questions	128	96	128	n/a	128	253

Method	Qwen3.8 27B		Llama 3.3 70B		Qwen2.5 72B	
	GSM	MATH	GSM	MATH	GSM	MATH
Greedy	93.8 / 93.8	90.6 / **90.6**	95.3 / 95.3	75.0 / 75.0	92.2 / 92.2	84.4 / **84.4**
Temperature	95.3 / 93.0	**93.8** / 86.7	**96.9** / 95.3	87.5 / 76.6	**95.3** / 92.2	90.6 / 83.2
Top-p	**96.9** / 93.6	**93.8** / 81.6	**96.9** / **95.7**	**90.6** / **79.3**	**95.3** / **92.6**	**93.8** / 83.6
Min-p	**96.9** / 93.9	**93.8** / 83.2	**96.9** / 95.3	87.5 / 76.6	**95.3** / 92.4	90.6 / 83.2
P-less	95.3 / 93.0	90.6 / 85.2	95.3 / 94.9	81.2 / 70.7	93.8 / 92.4	87.5 / 84.0
Global KL*	95.3 / **94.5**	**93.8** / 83.2	**96.9** / 95.1	87.5 / **79.3**	93.8 / 92.4	**93.8** / 84.0
Local KL*	95.3 / 93.2	**93.8** / 84.0	**96.9** / 94.1	87.5 / 78.1	93.8 / 92.0	**93.8** / 84.0
KL-budget	**96.9** / 92.8	**93.8** / 81.2	**96.9** / 94.7	87.5 / 77.3	**95.3** / 92.0	90.6 / 83.2
Questions	64	32	64	32	64	32

Method	Mistral S4 119B		Mistral L. 123B	
	GSM	MATH	GSM	MATH
Greedy	90.6 / **90.6**	53.1 / 53.1	92.2 / 92.2	71.9 / **71.9**
Temperature	95.3 / 86.1	68.8 / 52.0	**96.9** / 91.4	**90.6** / 66.8
Top-p	95.3 / 82.2	71.9 / 52.3	**96.9** / 91.4	84.4 / 62.1
Min-p	**96.9** / 85.2	68.8 / 55.1	**96.9** / 90.4	84.4 / 68.0
P-less	95.3 / 88.5	62.5 / 54.7	93.8 / **92.8**	81.2 / 65.6
Global KL*	95.3 / 85.2	71.9 / 55.1	**96.9** / 92.6	81.2 / 63.3
Local KL*	**96.9** / 86.7	71.9 / **56.6**	**96.9** / 92.2	87.5 / 67.2
KL-budget	95.3 / 83.6	**78.1** / 52.0	**96.9** / 88.1	**90.6** / 63.3
Questions	64	32	64	32

Table 25: **Answer coverage and reliability across model sizes.** Each cell gives Pass@8 / individual-path accuracy (%). GSM is GSM8K and MATH is MATH-500. The second value distinguishes reliability when eight-attempt coverage approaches its ceiling. Methods share questions and the first eight fixed seeds. Greedy generates once. Development settings transfer unchanged to MATH-500. Bold marks every largest unrounded mean among the displayed methods, separately for each endpoint. Question counts are explicit. Complete controls, majority votes, precision, and uncertainty appear in Appendix R.

uncertain. More paths narrow the majority gap while global KL* achieves the highest observed answer coverage at sixteen paths.

Q.3 Transfer to mathematics and science

On 96 Qwen2.5 1.5B MATH-500 questions, local KL* ties P-less for the highest observed pass@8, 80.2%, and leads Vote@8 at 65.2%. P-less has the highest individual-path accuracy, 58.2%.

On 253 Qwen2.5 7B mathematics questions, local KL* covers 237 (93.7%), versus 235 (92.9%) for min-p, the strongest baseline. Its 4.7-point coverage gains over temperature and P-less have adjusted

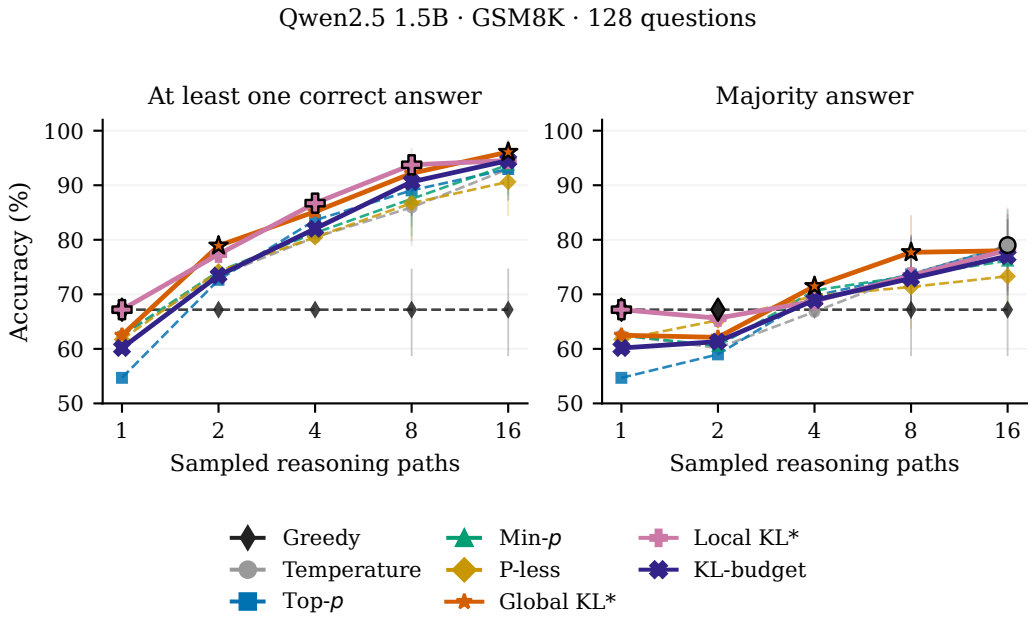


Figure 28: **More paths improve answer coverage and majority selection in different ways.** On all 128 questions, local KL* has the highest coverage at eight paths and global KL* at sixteen. Global KL* leads Vote@8. Temperature leads Vote@16. Black outlines mark every best observed value at each path count. Vertical lines give 95% question intervals at eight and sixteen paths. The restricted vertical range makes the differences visible while retaining every point and interval. Each curve uses the first k fixed seeds. Greedy uses one draw. Settings and questions are unchanged as paths are added.

exact $p = .021$ and $p = .009$, respectively. The smaller gap from min-p remains unresolved. Top-p and local KL* reach 84.0% and 83.5% Vote@8, averaging 572 and 570 output tokens, respectively. Local and global KL* also improve Vote@8 over temperature by 5.2 and 3.6 points, with positive adjusted paired bootstrap intervals.

The large-model mathematics comparisons in Table 25 use a common 2,048-token allowance. On 32 Mistral Large questions, KL-budget ties temperature for pass@8 at 90.6%, while min-p leads Vote@8 at 75.0%. The leading coverage gaps remain unresolved. Output-limit hits are especially frequent for Mistral Small 4, so its results measure completion reliability within that allowance. Appendix R reports every endpoint, cap rates, and paired uncertainty.

On 32 Mistral Small 4 GPQA questions, KL-budget ties top-p for pass@8 at 81.2%. Its Vote@8 is 50.0%, versus 46.9% for temperature and 40.1% for top-p. The leading differences remain uncertain. Eight-draw coverage has limited discriminatory power on these four-choice questions: eight independent uniform guesses would include the answer about 90% of the time. We therefore interpret coverage together with voting and individual-answer accuracy.

Q.4 Accuracy and variation among reasoning paths

Figure 22 examines variation among responses with correct final answers at small and large scales. Within each checkpoint, the comparison uses the shared questions where every stochastic method has at least two correct answers. On the shared 80-question 1.5B subset, local KL* and KL-budget lie on the observed accuracy–response-variation frontier. Here variation means BGE-M3 embedding distance between response texts, not a count of distinct mathematical solution strategies or a verification that intermediate reasoning is correct.

At 123B, 60 of 64 questions are eligible. Global and local KL* have mean correct-response distances of 0.0395 and 0.0373, compared with 0.0123 for P-less. Their shared-subset individual accuracies

are 97.9% and 97.1%, versus 99.0% for P-less. Global KL* remains on the observed accuracy–response-variation frontier. Across all 64 questions, global KL* and P-less average 172 and 171 generated tokens, respectively. Both KL* diversity differences remain positive in the post-hoc paired adjustment across all completed semantic comparisons (Appendix P). On the expanded Qwen3.8 27B comparison, global KL* has the highest observed individual accuracy. Its correct-response distance on the shared 61-question subset is 0.0456, while local KL* reaches 0.0437 and P-less 0.0139. Both KL* distance contrasts remain positive after the adjustment (Appendix Figure 24). The pattern also appears on Qwen2.5 72B MATH-500: both KL* variants match P-less's 84.0% individual accuracy, with all-path distances of 0.0479 and 0.0470 versus 0.0121. Both correct-response contrasts on the shared 27 questions remain positive under the same adjustment (Appendix Figure 26). The expanded 27B mathematics slice also has positive correct-response distance contrasts for all three KL methods against P-less (Appendix Figure 25). Shared-subset results are conditional. Correct final answers do not verify every intermediate step. Appendix Figure 23 gives the full 1.5B view. The complete frontier census includes every stochastic control, including uniform min-p (Table 24). Global KL* is nondominated in 8 of the 15 semantic comparisons, local KL* in 3, and at least one of the two in 10. Both occur together in one comparison. Including KL-budget gives a KL method on all 15 observed frontiers. These counts characterize the measured trade-offs, without implying that both uniform rules dominate downstream accuracy and variation.

R Complete Subset-Study Outcomes

The following compact tables retain every completed model/task comparison. Intervals are descriptive unless a paired adjusted contrast is explicitly cited. Bold marks all largest unrounded means, including ties. The release also contains the development grid, exact question-level outcomes, and sixteen-path and pass-estimator variants.

Table 26: Eight-path confirmation outcomes with question-level uncertainty. Binary outcomes use Wilson intervals. Fractional votes use a question bootstrap. All-zero or all-one bounded means retain finite-question uncertainty. Bold marks all largest accuracy means within a model/task block. Cap is the fraction reaching the output limit. Mark is requested-marker coverage for GSM8K and multiple choice, and boxed-answer coverage for MATH-500. These values diagnose completion reliability. Exact settings are in Table 22.

Method	Single [95% CI]	Pass@8 [95% CI]	Vote@8 [95% CI]	Cap	Mark
Qwen2.5 1.5B, GSM8K, 128 questions, 8 paths, 1,536 output tokens					
Greedy	67.2 [58.7, 74.7]	67.2 [58.7, 74.7]	67.2 [58.7, 74.7]	0.0	75.8
Temperature	57.4 [51.3, 63.4]	85.9 [78.9, 90.9]	73.7 [66.2, 80.9]	0.0	58.8
Top-p	59.9 [53.9, 65.4]	89.1 [82.5, 93.4]	73.7 [66.2, 80.9]	0.1	60.3
Min-p	57.7 [51.9, 63.5]	87.5 [80.7, 92.2]	73.4 [65.8, 80.6]	0.0	60.4
P-less	**67.7** [60.8, 74.1]	86.7 [79.8, 91.5]	71.4 [63.7, 78.5]	0.2	72.1
Uniform min-p	47.1 [41.3, 52.6]	84.4 [77.1, 89.7]	71.2 [63.3, 78.6]	0.2	55.7
Global KL*	62.6 [56.9, 68.2]	92.2 [86.2, 95.7]	**77.7** [70.1, 84.5]	0.0	60.7
Local KL*	62.8 [56.8, 68.8]	**93.8** [88.2, 96.8]	73.4 [65.6, 81.0]	0.2	66.0
KL-budget	59.7 [53.7, 65.3]	90.6 [84.3, 94.6]	72.9 [65.5, 80.4]	0.0	61.2
Mistral 7B, GSM8K, 128 questions, 8 paths, 1,536 output tokens					
Greedy	46.9 [38.4, 55.5]	46.9 [38.4, 55.5]	46.9 [38.4, 55.5]	3.9	93.0
Temperature	42.9 [37.4, 48.3]	82.8 [75.3, 88.4]	**66.8** [58.7, 74.6]	6.6	89.9
Top-p	39.9 [35.1, 44.9]	**85.9** [78.9, 90.9]	64.8 [56.8, 72.6]	4.7	85.8
Min-p	35.5 [31.0, 40.4]	80.5 [72.8, 86.4]	64.6 [56.5, 72.2]	5.2	82.9
P-less	**49.3** [42.3, 56.5]	72.7 [64.4, 79.6]	59.0 [50.9, 67.3]	6.4	94.9
Uniform min-p	27.3 [23.4, 31.2]	77.3 [69.4, 83.7]	57.2 [48.9, 65.4]	3.0	73.0
Global KL*	38.0 [32.9, 43.5]	82.0 [74.5, 87.7]	59.9 [51.7, 68.4]	9.7	87.3
Local KL*	38.2 [33.6, 43.2]	**85.9** [78.9, 90.9]	61.6 [53.4, 69.8]	10.1	87.9
KL-budget	33.1 [28.9, 37.6]	79.7 [71.9, 85.7]	62.2 [54.4, 70.4]	2.6	78.9

Table 26 (continued)

Method	Single [95% CI]	Pass@8 [95% CI]	Vote@8 [95% CI]	Cap	Mark
Qwen2.5 7B, GSM8K, 128 questions, 8 paths, 1,536 output tokens					
Greedy	**89.8** [83.4, 94.0]	89.8 [83.4, 94.0]	89.8 [83.4, 94.0]	0.0	97.7
Temperature	79.3 [74.3, 84.1]	97.7 [93.3, 99.2]	89.5 [84.0, 94.1]	3.7	91.6
Top-p	88.4 [83.8, 92.3]	96.9 [92.2, 98.8]	91.8 [86.7, 96.1]	0.0	97.5
Min-p	87.9 [83.3, 91.9]	96.9 [92.2, 98.8]	**94.1** [89.8, 97.7]	0.0	98.2
P-less	88.5 [83.1, 93.3]	93.8 [88.2, 96.8]	89.8 [83.4, 94.0]	0.0	97.8
Uniform min-p	86.8 [82.2, 90.9]	98.4 [94.5, 99.6]	91.0 [85.5, 95.7]	0.0	97.8
Global KL*	88.7 [83.9, 92.9]	96.9 [92.2, 98.8]	91.4 [86.3, 95.7]	0.0	98.0
Local KL*	87.6 [82.8, 91.8]	96.9 [92.2, 98.8]	91.4 [86.3, 95.7]	0.1	97.9
KL-budget	87.8 [83.4, 91.8]	**99.2** [95.7, 99.9]	92.2 [87.5, 96.5]	0.0	98.5
Qwen3.8 27B, GSM8K, 64 questions, 8 paths, 1,536 output tokens					
Greedy	93.8 [85.0, 97.5]	93.8 [85.0, 97.5]	93.8 [85.0, 97.5]	0.0	100.0
Temperature	93.0 [87.1, 97.9]	95.3 [87.1, 98.4]	**95.3** [87.1, 98.4]	1.8	98.2
Top-p	93.6 [88.1, 98.0]	**96.9** [89.3, 99.1]	94.5 [89.0, 99.2]	1.4	98.6
Min-p	93.9 [88.3, 98.4]	**96.9** [89.3, 99.1]	**95.3** [87.1, 98.4]	1.0	99.0
P-less	93.0 [86.9, 98.2]	95.3 [87.1, 98.4]	93.8 [85.0, 97.5]	1.2	98.8
Uniform min-p	92.4 [86.3, 97.5]	95.3 [87.1, 98.4]	94.5 [88.3, 99.2]	2.7	97.7
Global KL*	**94.5** [88.7, 99.2]	95.3 [87.1, 98.4]	**95.3** [87.1, 98.4]	0.6	99.2
Local KL*	93.2 [87.1, 98.0]	95.3 [87.1, 98.4]	**95.3** [87.1, 98.4]	1.2	98.8
KL-budget	92.8 [87.1, 97.7]	**96.9** [89.3, 99.1]	93.8 [85.0, 97.5]	1.8	98.0
Llama 3.3 70B, GSM8K, 64 questions, 8 paths, 1,536 output tokens					
Greedy	95.3 [87.1, 98.4]	95.3 [87.1, 98.4]	95.3 [87.1, 98.4]	0.0	100.0
Temperature	95.3 [90.0, 99.4]	**96.9** [89.3, 99.1]	95.3 [87.1, 98.4]	0.0	100.0
Top-p	**95.7** [90.8, 99.4]	**96.9** [89.3, 99.1]	**96.9** [89.3, 99.1]	0.0	100.0
Min-p	95.3 [90.0, 99.4]	**96.9** [89.3, 99.1]	95.3 [87.1, 98.4]	0.4	99.6
P-less	94.9 [88.9, 99.6]	95.3 [87.1, 98.4]	95.3 [87.1, 98.4]	0.0	100.0
Uniform min-p	93.2 [87.7, 97.5]	**96.9** [89.3, 99.1]	95.3 [87.1, 98.4]	0.8	99.2
Global KL*	95.1 [90.0, 98.8]	**96.9** [89.3, 99.1]	**96.9** [89.3, 99.1]	0.0	100.0
Local KL*	94.1 [89.1, 98.0]	**96.9** [89.3, 99.1]	**96.9** [89.3, 99.1]	0.0	100.0
KL-budget	94.7 [89.5, 98.6]	**96.9** [89.3, 99.1]	96.1 [91.4, 100.0]	0.0	100.0
Qwen2.5 72B, GSM8K, 64 questions, 8 paths, 1,536 output tokens					
Greedy	92.2 [83.0, 96.6]	92.2 [83.0, 96.6]	92.2 [83.0, 96.6]	0.0	100.0
Temperature	92.2 [85.4, 97.7]	95.3 [87.1, 98.4]	92.2 [83.0, 96.6]	0.0	100.0
Top-p	**92.6** [85.4, 98.4]	95.3 [87.1, 98.4]	92.2 [83.0, 96.6]	0.0	100.0
Min-p	92.4 [85.7, 98.0]	95.3 [87.1, 98.4]	92.2 [83.0, 96.6]	0.0	100.0
P-less	92.4 [85.0, 98.4]	93.8 [85.0, 97.5]	92.2 [83.0, 96.6]	0.0	100.0
Uniform min-p	92.0 [85.5, 97.3]	**96.9** [89.3, 99.1]	92.2 [83.0, 96.6]	0.0	100.0
Global KL*	92.4 [85.5, 97.9]	93.8 [85.0, 97.5]	**93.8** [85.0, 97.5]	0.0	100.0
Local KL*	92.0 [85.0, 97.7]	93.8 [85.0, 97.5]	93.0 [85.9, 98.4]	0.0	100.0
KL-budget	92.0 [85.2, 97.7]	95.3 [87.1, 98.4]	93.0 [85.9, 98.4]	0.0	100.0
Mistral Small 4 119B, GSM8K, 64 questions, 8 paths, 1,536 output tokens					
Greedy	**90.6** [81.0, 95.6]	90.6 [81.0, 95.6]	90.6 [81.0, 95.6]	0.0	98.4
Temperature	86.1 [78.9, 92.4]	95.3 [87.1, 98.4]	**92.2** [83.0, 96.6]	0.0	99.2
Top-p	82.2 [74.4, 89.1]	95.3 [87.1, 98.4]	89.8 [82.0, 96.9]	0.0	96.7
Min-p	85.2 [78.1, 91.2]	**96.9** [89.3, 99.1]	89.6 [81.8, 96.4]	0.0	98.6
P-less	88.5 [81.4, 94.5]	95.3 [87.1, 98.4]	89.8 [82.0, 96.1]	0.0	99.4
Uniform min-p	73.0 [65.6, 79.9]	95.3 [87.1, 98.4]	86.7 [78.1, 94.5]	0.8	93.9
Global KL*	85.2 [78.3, 91.2]	95.3 [87.1, 98.4]	91.1 [84.1, 96.9]	0.0	98.8
Local KL*	86.7 [80.3, 92.6]	**96.9** [89.3, 99.1]	**92.2** [85.2, 97.7]	0.0	99.6
KL-budget	83.6 [76.4, 89.8]	95.3 [87.1, 98.4]	89.8 [82.0, 96.9]	0.0	97.1

Table 26 (continued)

Method	Single [95% CI]	Pass@8 [95% CI]	Vote@8 [95% CI]	Cap	Mark
Mistral Large 123B, GSM8K, 64 questions, 8 paths, 1,536 output tokens					
Greedy	92.2 [83.0, 96.6]	92.2 [83.0, 96.6]	92.2 [83.0, 96.6]	0.0	100.0
Temperature	91.4 [85.4, 96.5]	**96.9** [89.3, 99.1]	93.8 [85.0, 97.5]	0.0	99.6
Top-p	91.4 [85.7, 96.1]	**96.9** [89.3, 99.1]	93.8 [85.0, 97.5]	0.0	100.0
Min-p	90.4 [84.8, 95.3]	**96.9** [89.3, 99.1]	**95.3** [87.1, 98.4]	0.0	99.6
P-less	**92.8** [86.1, 98.0]	93.8 [85.0, 97.5]	93.8 [85.0, 97.5]	0.0	100.0
Uniform min-p	77.5 [70.3, 84.4]	93.8 [85.0, 97.5]	92.2 [85.2, 97.7]	0.0	99.4
Global KL*	92.6 [86.7, 97.5]	**96.9** [89.3, 99.1]	93.8 [85.0, 97.5]	0.0	99.8
Local KL*	92.2 [85.9, 97.3]	**96.9** [89.3, 99.1]	93.0 [86.7, 98.4]	0.0	99.6
KL-budget	88.1 [82.2, 93.4]	**96.9** [89.3, 99.1]	93.8 [87.5, 98.4]	0.0	99.6
Qwen2.5 1.5B, MATH-500, 96 questions, 8 paths, 2,048 output tokens					
Greedy	55.2 [45.3, 64.8]	55.2 [45.3, 64.8]	55.2 [45.3, 64.8]	5.2	92.7
Temperature	44.5 [37.4, 51.6]	76.0 [66.6, 83.5]	64.8 [55.5, 73.6]	0.1	94.1
Top-p	49.9 [42.2, 57.6]	76.0 [66.6, 83.5]	60.9 [51.0, 70.3]	0.7	95.7
Min-p	40.4 [33.9, 46.9]	78.1 [68.9, 85.2]	58.6 [49.2, 67.9]	0.3	96.4
P-less	**58.2** [49.7, 66.5]	**80.2** [71.1, 86.9]	62.5 [52.8, 71.6]	4.7	93.6
Uniform min-p	27.6 [21.2, 34.0]	62.5 [52.5, 71.5]	45.6 [36.0, 55.2]	0.3	95.1
Global KL*	52.2 [44.1, 60.0]	77.1 [67.7, 84.4]	63.4 [53.5, 72.6]	1.3	95.4
Local KL*	50.0 [42.3, 57.3]	**80.2** [71.1, 86.9]	**65.2** [55.8, 73.9]	1.4	95.8
KL-budget	45.4 [38.1, 52.7]	79.2 [70.0, 86.1]	59.2 [49.7, 68.1]	0.5	96.1
Qwen2.5 7B, MATH-500, 253 questions, 8 paths, 2,048 output tokens					
Greedy	**79.1** [73.6, 83.6]	79.1 [73.6, 83.6]	79.1 [73.6, 83.6]	0.8	99.2
Temperature	66.3 [61.7, 70.7]	88.9 [84.5, 92.2]	78.3 [73.3, 83.0]	9.6	85.5
Top-p	77.1 [72.9, 81.2]	92.1 [88.1, 94.8]	**84.0** [79.4, 88.3]	0.4	99.5
Min-p	76.3 [71.9, 80.4]	92.9 [89.0, 95.5]	83.2 [78.7, 87.3]	0.3	99.7
P-less	78.0 [73.4, 82.3]	88.9 [84.5, 92.2]	79.8 [75.0, 84.4]	0.9	99.1
Uniform min-p	73.0 [68.5, 77.3]	89.7 [85.4, 92.9]	80.7 [75.7, 85.2]	0.0	100.0
Global KL*	77.4 [73.0, 81.4]	91.7 [87.6, 94.5]	81.9 [77.1, 86.3]	0.6	99.4
Local KL*	77.4 [73.2, 81.2]	**93.7** [90.0, 96.1]	83.5 [78.8, 87.7]	0.4	99.6
KL-budget	72.9 [68.6, 77.1]	92.1 [88.1, 94.8]	81.4 [76.5, 85.8]	0.0	100.0
Qwen3.8 27B, MATH-500, 32 questions, 8 paths, 2,048 output tokens					
Greedy	**90.6** [75.8, 96.8]	90.6 [75.8, 96.8]	90.6 [75.8, 96.8]	6.2	93.8
Temperature	86.7 [75.8, 95.3]	**93.8** [79.9, 98.3]	**93.8** [79.9, 98.3]	12.1	87.9
Top-p	81.6 [69.1, 92.2]	**93.8** [79.9, 98.3]	90.6 [75.8, 96.8]	16.0	84.0
Min-p	83.2 [71.5, 92.6]	**93.8** [79.9, 98.3]	90.6 [75.8, 96.8]	14.1	85.9
P-less	85.2 [72.7, 94.9]	90.6 [75.8, 96.8]	90.6 [75.8, 96.8]	11.7	88.3
Uniform min-p	76.2 [62.9, 87.5]	87.5 [71.9, 95.0]	87.5 [71.9, 95.0]	21.9	78.1
Global KL*	83.2 [71.1, 93.0]	**93.8** [79.9, 98.3]	90.6 [75.8, 96.8]	14.1	85.9
Local KL*	84.0 [72.3, 93.4]	**93.8** [79.9, 98.3]	92.2 [82.8, 100.0]	14.5	85.5
KL-budget	81.2 [68.8, 91.4]	**93.8** [79.9, 98.3]	**93.8** [79.9, 98.3]	16.8	83.2
Llama 3.3 70B, MATH-500, 32 questions, 8 paths, 2,048 output tokens					
Greedy	75.0 [57.9, 86.7]	75.0 [57.9, 86.7]	75.0 [57.9, 86.7]	3.1	96.9
Temperature	76.6 [63.3, 87.5]	87.5 [71.9, 95.0]	84.4 [68.2, 93.1]	1.6	98.4
Top-p	**79.3** [66.8, 89.8]	**90.6** [75.8, 96.8]	82.8 [68.8, 93.8]	1.6	98.4
Min-p	76.6 [63.3, 87.9]	87.5 [71.9, 95.0]	81.2 [64.7, 91.1]	1.6	98.4
P-less	70.7 [55.5, 84.0]	81.2 [64.7, 91.1]	73.4 [57.8, 87.5]	7.8	92.6
Uniform min-p	75.4 [63.3, 85.9]	**90.6** [75.8, 96.8]	**87.5** [71.9, 95.0]	0.8	99.2
Global KL*	**79.3** [66.4, 89.8]	87.5 [71.9, 95.0]	**87.5** [71.9, 95.0]	2.7	97.3
Local KL*	78.1 [65.6, 88.3]	87.5 [71.9, 95.0]	**87.5** [71.9, 95.0]	1.6	98.4
KL-budget	77.3 [64.8, 87.9]	87.5 [71.9, 95.0]	85.9 [73.4, 96.9]	0.8	99.2

Table 26 (continued)

Method	Single [95% CI]	Pass@8 [95% CI]	Vote@8 [95% CI]	Cap	Mark
Qwen2.5 72B, MATH-500, 32 questions, 8 paths, 2,048 output tokens					
Greedy	**84.4** [68.2, 93.1]	84.4 [68.2, 93.1]	84.4 [68.2, 93.1]	0.0	100.0
Temperature	83.2 [71.5, 93.4]	90.6 [75.8, 96.8]	87.5 [71.9, 95.0]	0.8	99.2
Top-p	83.6 [71.9, 93.8]	93.8 [79.9, 98.3]	**89.1** [78.1, 98.4]	1.2	98.8
Min-p	83.2 [71.1, 93.0]	90.6 [75.8, 96.8]	87.5 [71.9, 95.0]	1.2	98.8
P-less	84.0 [70.3, 94.5]	87.5 [71.9, 95.0]	84.4 [68.2, 93.1]	1.2	98.8
Uniform min-p	81.6 [70.7, 91.4]	**96.9** [84.3, 99.4]	**89.1** [78.1, 98.4]	0.0	100.0
Global KL*	84.0 [72.7, 93.8]	93.8 [79.9, 98.3]	87.5 [71.9, 95.0]	0.4	99.6
Local KL*	84.0 [72.3, 93.8]	93.8 [79.9, 98.3]	87.5 [71.9, 95.0]	0.4	99.6
KL-budget	83.2 [71.9, 93.0]	90.6 [75.8, 96.8]	**89.1** [78.1, 98.4]	0.8	99.2
Mistral Small 4 119B, MATH-500, 32 questions, 8 paths, 2,048 output tokens					
Greedy	53.1 [36.4, 69.1]	53.1 [36.4, 69.1]	53.1 [36.4, 69.1]	43.8	56.2
Temperature	52.0 [36.7, 66.8]	68.8 [51.4, 82.0]	68.8 [51.4, 82.0]	43.0	56.6
Top-p	52.3 [37.1, 66.8]	71.9 [54.6, 84.4]	71.9 [54.6, 84.4]	44.9	55.5
Min-p	55.1 [39.5, 69.5]	68.8 [51.4, 82.0]	68.8 [51.4, 82.0]	42.2	59.8
P-less	54.7 [38.7, 69.9]	62.5 [45.3, 77.1]	62.5 [45.3, 77.1]	40.6	58.6
Uniform min-p	47.7 [32.8, 62.1]	65.6 [48.3, 79.6]	62.5 [45.3, 77.1]	49.6	51.2
Global KL*	55.1 [39.8, 69.5]	71.9 [54.6, 84.4]	71.9 [54.6, 84.4]	40.6	59.0
Local KL*	**56.6** [41.4, 70.7]	71.9 [54.6, 84.4]	71.9 [54.6, 84.4]	38.7	61.7
KL-budget	52.0 [37.5, 66.0]	**78.1** [61.2, 89.0]	**73.4** [57.8, 87.5]	44.5	55.5
Mistral Large 123B, MATH-500, 32 questions, 8 paths, 2,048 output tokens					
Greedy	**71.9** [54.6, 84.4]	71.9 [54.6, 84.4]	71.9 [54.6, 84.4]	0.0	100.0
Temperature	66.8 [53.5, 78.9]	**90.6** [75.8, 96.8]	74.9 [59.4, 88.3]	0.0	99.6
Top-p	62.1 [48.4, 74.6]	84.4 [68.2, 93.1]	74.5 [59.4, 87.5]	0.4	98.4
Min-p	68.0 [53.9, 80.1]	84.4 [68.2, 93.1]	**75.0** [57.9, 86.7]	0.4	98.4
P-less	65.6 [50.8, 79.3]	81.2 [64.7, 91.1]	68.8 [53.1, 84.4]	0.0	99.2
Uniform min-p	59.8 [44.5, 73.8]	75.0 [57.9, 86.7]	68.8 [51.4, 82.0]	0.0	98.4
Global KL*	63.3 [48.8, 77.0]	81.2 [64.7, 91.1]	72.3 [56.2, 85.9]	0.8	98.4
Local KL*	67.2 [52.7, 80.1]	87.5 [71.9, 95.0]	73.3 [57.4, 87.5]	0.8	98.8
KL-budget	63.3 [48.8, 76.6]	**90.6** [75.8, 96.8]	69.6 [53.1, 84.8]	0.0	99.2
Mistral Small 4 119B, GPQA, 32 questions, 8 paths, 2,048 output tokens					
Greedy	34.4 [20.4, 51.7]	34.4 [20.4, 51.7]	34.4 [20.4, 51.7]	6.2	90.6
Temperature	36.3 [25.8, 47.7]	75.0 [57.9, 86.7]	46.9 [30.9, 63.6]	5.1	89.8
Top-p	32.0 [23.0, 41.4]	**81.2** [64.7, 91.1]	40.1 [24.0, 57.3]	8.6	85.9
Min-p	37.5 [26.6, 48.8]	78.1 [61.2, 89.0]	42.2 [25.0, 59.4]	7.8	88.3
P-less	**38.3** [25.8, 51.2]	71.9 [54.6, 84.4]	48.4 [31.2, 65.6]	2.0	92.6
Uniform min-p	37.9 [27.7, 47.7]	78.1 [61.2, 89.0]	**51.6** [35.9, 67.2]	13.3	84.8
Global KL*	**38.3** [27.3, 49.6]	78.1 [61.2, 89.0]	42.2 [25.0, 59.4]	3.5	91.8
Local KL*	33.2 [23.4, 43.4]	78.1 [61.2, 89.0]	42.2 [25.0, 59.4]	7.4	89.5
KL-budget	37.1 [27.0, 47.3]	**81.2** [64.7, 91.1]	50.0 [32.8, 65.6]	7.0	90.6

R.1 Sensitivity to the requested final-answer marker

Table 27: GSM8K answer-extraction sensitivity at unchanged sampler settings. Flexible is the primary harness numerical endpoint. Marker requires the requested final-answer tag. Bold marks all largest unrounded means in each model/column. Every method uses the same first eight paths, with no sensitivity-specific retuning. Greedy uses one draw. All values are percentages.

	Pass@8		Vote@8	
Method	Flexible	Marker	Flexible	Marker
Qwen2.5 1.5B, 128 questions				
Greedy	67.2	53.1	67.2	53.1
Temperature	85.9	80.5	73.7	69.3
Top-p	89.1	82.8	73.7	74.4
Min-p	87.5	81.2	73.4	69.9
P-less	86.7	75.0	71.4	64.2
Uniform min-p	84.4	75.0	71.2	59.4
Global KL*	92.2	**87.5**	**77.7**	**77.3**
Local KL*	**93.8**	**87.5**	73.4	75.5
KL-budget	90.6	79.7	72.9	69.7
Mistral 7B, 128 questions				
Greedy	46.9	46.9	46.9	46.9
Temperature	82.8	**83.6**	**66.8**	**69.0**
Top-p	**85.9**	82.8	64.8	64.5
Min-p	80.5	77.3	64.6	63.3
P-less	72.7	69.5	59.0	58.3
Uniform min-p	77.3	73.4	57.2	59.0
Global KL*	82.0	82.0	59.9	62.4
Local KL*	**85.9**	**83.6**	61.6	63.4
KL-budget	79.7	77.3	62.2	63.1
Qwen2.5 7B, 128 questions				
Greedy	89.8	89.1	89.8	89.1
Temperature	97.7	97.7	89.5	91.6
Top-p	96.9	96.9	91.8	93.0
Min-p	96.9	96.9	**94.1**	**94.5**
P-less	93.8	93.8	89.8	90.6
Uniform min-p	98.4	98.4	91.0	91.3
Global KL*	96.9	96.9	91.4	92.6
Local KL*	96.9	96.9	91.4	92.6
KL-budget	**99.2**	**99.2**	92.2	92.2
Qwen3.8 27B, 64 questions				
Greedy	93.8	93.8	93.8	93.8
Temperature	95.3	95.3	**95.3**	**95.3**
Top-p	**96.9**	**96.9**	94.5	**95.3**
Min-p	**96.9**	**96.9**	**95.3**	**95.3**
P-less	95.3	95.3	93.8	93.8
Uniform min-p	95.3	95.3	94.5	94.5
Global KL*	95.3	95.3	**95.3**	**95.3**
Local KL*	95.3	95.3	**95.3**	**95.3**
KL-budget	**96.9**	95.3	93.8	93.8
Llama 3.3 70B, 64 questions				
Greedy	95.3	**96.9**	95.3	**96.9**
Temperature	**96.9**	**96.9**	95.3	**96.9**
Top-p	**96.9**	**96.9**	**96.9**	**96.9**
Min-p	**96.9**	**96.9**	95.3	**96.9**
P-less	95.3	**96.9**	95.3	**96.9**
Uniform min-p	**96.9**	**96.9**	95.3	**96.9**
Global KL*	**96.9**	**96.9**	**96.9**	**96.9**
Local KL*	**96.9**	**96.9**	**96.9**	**96.9**
KL-budget	**96.9**	**96.9**	96.1	**96.9**

	Top-p .9		Min-p .05		P-less		
Checkpoint (n)	Temp. first	Trunc. first	Temp. first	Trunc. first	Temp. first	Trunc. first	Global KL*
Qwen3.8 27B (32)	0.0	**100.0**	3.1	**100.0**	3.1	96.9	**100.0**
Llama 3.3 70B (64)	0.0	**95.3**	93.8	**95.3**	**95.3**	**95.3**	**95.3**
Qwen2.5 72B (64)	0.0	93.8	89.1	**95.3**	93.8	**95.3**	93.8
R1-Distill 70B (32)	0.0	68.8	0.0	75.0	0.0	**90.6**	81.2

Table 28: **Fixed-setting order study at $T = 3$.** GSM8K accuracy (%) on the same questions per checkpoint in the subset study, with one seed and a 768-token limit (2,048 for R1-Distill). Only the order changes within each pair of columns. Global KL* uses an exact uniform draw on the same questions. These are fixed settings, separate from the development-tuned comparison. Bold marks the highest accuracy for each checkpoint.

Table 27 (continued)

	Pass@8		Vote@8	
Method	Flexible	Marker	Flexible	Marker
Qwen2.5 72B, 64 questions				
Greedy	92.2	93.8	92.2	93.8
Temperature	95.3	95.3	92.2	93.8
Top-p	95.3	95.3	92.2	93.8
Min-p	95.3	95.3	92.2	93.8
P-less	93.8	95.3	92.2	93.8
Uniform min-p	**96.9**	**96.9**	92.2	**95.3**
Global KL*	93.8	95.3	**93.8**	**95.3**
Local KL*	93.8	95.3	93.0	94.5
KL-budget	95.3	95.3	93.0	94.5
Mistral Small 4 119B, 64 questions				
Greedy	90.6	89.1	90.6	89.1
Temperature	95.3	95.3	**92.2**	**92.2**
Top-p	95.3	95.3	89.8	89.1
Min-p	**96.9**	**96.9**	89.6	90.1
P-less	95.3	95.3	89.8	89.8
Uniform min-p	95.3	95.3	86.7	86.7
Global KL*	95.3	95.3	91.1	91.1
Local KL*	**96.9**	**96.9**	**92.2**	**92.2**
KL-budget	95.3	95.3	89.8	89.8
Mistral Large 123B, 64 questions				
Greedy	92.2	92.2	92.2	92.2
Temperature	**96.9**	**96.9**	93.8	93.8
Top-p	**96.9**	**96.9**	93.8	93.8
Min-p	**96.9**	**96.9**	**95.3**	**95.3**
P-less	93.8	93.8	93.8	93.8
Uniform min-p	93.8	95.3	92.2	93.0
Global KL*	**96.9**	**96.9**	93.8	93.8
Local KL*	**96.9**	**96.9**	93.0	93.0
KL-budget	**96.9**	**96.9**	93.8	94.5

S Writing and Fixed-Setting Order Studies

These studies use Llama 3.3 70B, Qwen2.5 72B, and R1-Distill-Llama 70B, with two seeded stories per prompt. Llama and Qwen have 32 writing prompts each. R1 has 16. Llama and Qwen have a 768-token generation allowance. R1 has 2,048. An independent Q8_0 Gemma 4 31B model judges full visible responses in both presentation orders. It sees the instruction and responses without method names, rates coherence, instruction fulfillment, and originality, gives a rationale, then selects A, B, or a tie. Greedy judging uses a 384-token cap and disabled thinking. Scores average orders and seeds

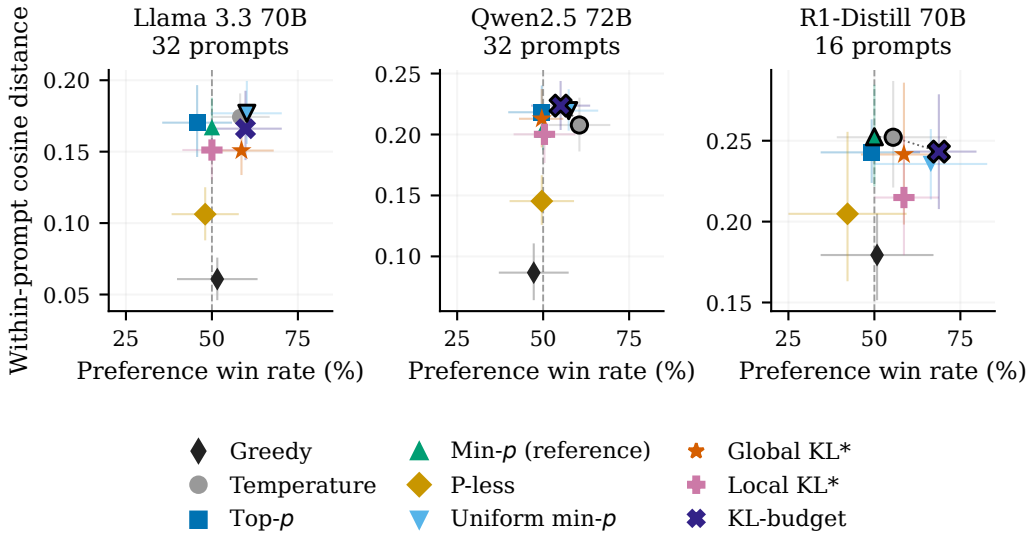


Figure 29: **Writing preference and response variation.** Automated preference against min-p .05 at temperature one and within-prompt response variation, with 95% prompt intervals. Black outlines and dotted lines mark nondominated means, including baselines. KL-budget lies on the Qwen and R1 frontiers.

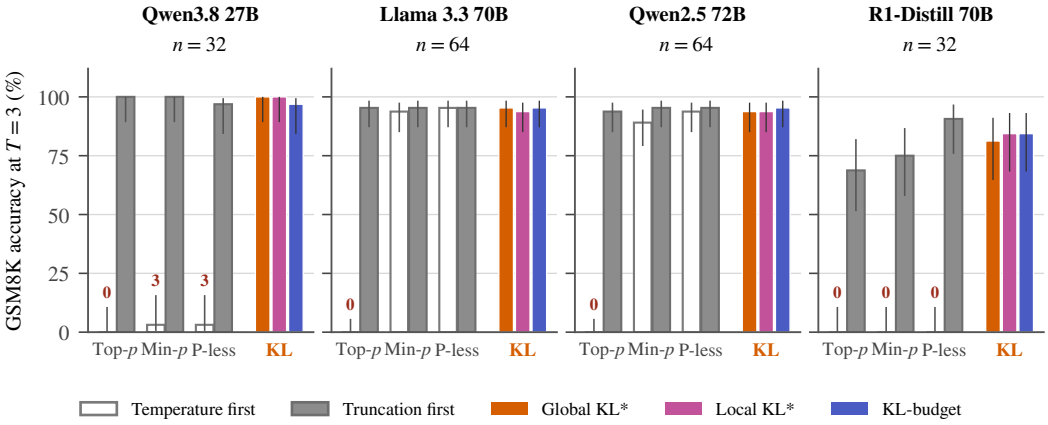


Figure 30: **Truncating first prevents collapse at high temperature.** GSM8K accuracy at $T = 3$ for top-p .9, min-p .05, and P-less in both orders, on the same questions per checkpoint from the subset study. The three KL rules run at their canonical settings. Bars are 95% Wilson intervals. These fixed settings diagnose a failure mode and are not a tuned comparison (Appendix Table 28).

within a prompt before uncertainty is computed. The min-p .05, $T = 1$ reference has a defined 50% win rate.

The judge format passed 80 development controls and 104 held-out controls: all 168 degraded pairs favored the coherent response, all 16 identical pairs tied, and all 24 empty responses received minimum ratings. All control records are retained. Swapping story presentation order nevertheless changes 25.0–51.6% of KL methods' win/tie/loss outcomes across the generators. Averaging both orders mitigates this sensitivity. The study does not include new human annotations.

Length and diversity. The supplementary LC estimate fits a local logistic mean model with a generator–candidate intercept, a shared log word-count-ratio coefficient, a shared presentation-position coefficient, and an ℓ_2 penalty of .01. It evaluates equal word counts and averages the two

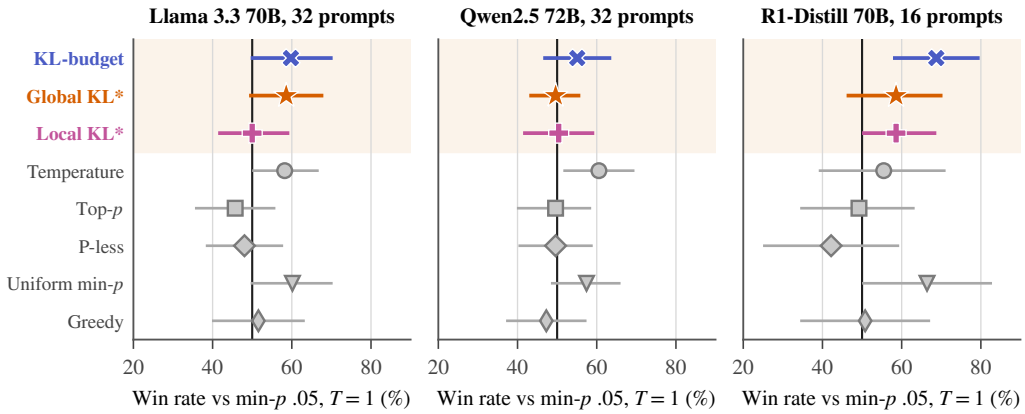


Figure 31: **KL-budget has the highest writing preference on R1-Distill 70B.** Blinded automated win rate against min-p .05 at temperature one, with 95% prompt intervals. The vertical line marks a tie with that reference. Figure 29 pairs preference with response variation.

Method	Llama 3.3 70B Win	Llama 3.3 70B LC	Qwen2.5 72B Win	Qwen2.5 72B LC	R1-Distill 70B Win	R1-Distill 70B LC
Greedy	51.6	51.5	47.3	47.3	50.8	51.0
Top-p	45.7	45.7	49.6	49.6	49.2	49.1
Min-p (reference)	50.0	50.0	50.0	50.0	50.0	50.0
P-less	48.0	48.0	49.6	49.6	42.2	42.4
Global KL*	58.6	58.6	49.6	49.6	58.6	58.5
Local KL*	50.0	50.0	50.4	50.4	58.6	58.4
KL-budget	**59.8**	**59.7**	**55.1**	**55.1**	**68.8**	**68.5**
Prompts × seeds	32 × 2		32 × 2		16 × 2	

Table 29: **KL-budget has the highest observed R1 writing preference.** Blinded win rate (%) against min-p .05, $T = 1$ outputs from the same generator. Standard truncation baselines use $T = 1$. KL* draws uniformly. KL-budget uses min-p .05 and $\epsilon = .1$. LC is a local length adjustment. Bold marks all highest means among the displayed methods. Appendix S retains the complete comparison, including temperature-only sampling, uncertainty, and the judge protocol.

positions. Intervals use 2,000 prompt-cluster bootstrap refits. This local associational sensitivity uses its own estimator and differs from official AlpacaEval scoring (Dubois et al., 2024). BGE-M3 embeddings give the mean cosine distance between responses to the same prompt (Chen et al., 2024). Empty pairs contribute zero, and inputs exceeding encoder capacity are flagged. Neither length adjustment nor semantic distance establishes human creativity independently of quality.

Generator	Method	Versus min-p, $T = 1$	Versus uniform min-p
Llama 3.3 70B	Global KL*	+8.6 [-5.5, +22.7]	-1.6 [-19.5, +16.4]
Llama 3.3 70B	Local KL*	+0.0 [-13.3, +14.1]	-10.2 [-27.7, +7.4]
Llama 3.3 70B	KL-budget	+9.8 [-6.2, +25.0]	-0.4 [-19.9, +18.0]
Qwen2.5 72B	Global KL*	-0.4 [-10.5, +9.4]	-7.8 [-23.0, +5.9]
Qwen2.5 72B	Local KL*	+0.4 [-13.0, +13.7]	-7.0 [-22.7, +9.4]
Qwen2.5 72B	KL-budget	+5.1 [-8.2, +18.4]	-2.3 [-17.2, +12.5]
R1-Distill 70B	Global KL*	+8.6 [-10.9, +26.6]	-7.8 [-37.5, +20.3]
R1-Distill 70B	Local KL*	+8.6 [-3.9, +24.2]	-7.8 [-35.9, +25.0]
R1-Distill 70B	KL-budget	**+18.8 [+1.6, +35.9]**	+2.3 [-21.9, +30.5]

Table 30: Writing preference differences in percentage points with Bonferroni-adjusted prompt-bootstrap intervals for all 18 displayed contrasts (three KL methods, three generators, two controls). This post-hoc check uses 200,000 paired prompt resamples. Orders and seeds remain grouped within each prompt. Bold identifies an interval entirely above zero. A difference versus uniform min-p compares their respective win rates against the common $T = 1$ reference. Each method is judged against that reference.