RETHINKING POWER SAMPLING FOR LLM REASONING: FROM SHARPENING–LOSS THEORY TO MARGINAL-AWARE RESCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

During decoding of large language models, temperature governs the trade-off between exploration and precision, yet locally optimal preferences do not necessarily yield globally better reasoning trajectories. Power Sampling addresses this by sharpening global sequence likelihood—favoring tokens that lead to confident continuations rather than merely locally confident ones—and nearly matches or even outperforms RL post-training on several reasoning benchmarks. However, its MCMC inference is costly, its joint-likelihood objective favors shorter outputs, and the preferred length-normalized mean likelihood cannot be sharpened directly. Best-of-N (BoN) is a natural alternative but has struggled to match this performance. We develop a unified sharpening-loss theory that explains this gap: Power Sampling and BoN share the same sequence-level sharpening mechanism, governed by sharpening direction and sharpening strength. A wrong direction concentrates probability on inferior candidates, causing *preference loss*, while insufficient strength leaves probability dispersed among distractors, causing *dispersion loss*. With length-normalized mean log-likelihood scored at an elevated temperature, BoN achieves a better sharpening direction, but its core bottleneck is insufficient strength: the probability of selecting high-scoring trajectories grows too slowly with the sampling budget. Building on this theory, we propose Marginal-Aware Rescaling (MAR), a training-free method that computes expected logits to identify the boundary of tokens worth exploring and thereby adaptively controls both direction and strength. MAR lowers the expected sampling budget needed to find the optimal candidate, and across diverse reasoning tasks and model families it matches or surpasses Power Sampling and RL post-training at substantially lower sampling cost.

1 INTRODUCTION

Many reasoning problems admit a way to check whether a candidate solution is good: a proof checks out, code passes tests, or the answer matches. A large language model (LLM) asked to solve such problems by greedy decoding leaves this verification unused: with a single deterministic pass there is nothing to select among, and accuracy cannot grow with inference compute. Sampling removes this ceiling by turning the model's output distribution into a proposal — diverse candidates allow sequence-level selection scale with the sampling budget.

At every decoding step the model is in one of two states. In a *lock*, the top-ranked continuation is already correct, and sampling needs precision against the distractor tail; in a *fork*, a correct continuation exists but is ranked below the top choice — a top-1 preference error — and sampling needs exploration. A single reasoning sequence interleaves the two states, and they demand opposite temperatures: cooling protects locks but starves forks, and heating does the opposite. No single global temperature serves both (Zhao et al., 2024); the conflict has so far been resolved only at training time (Zhang et al., 2026); temperature nonetheless dominates decoding because it is cheap — a single pass with no search. Figure 1 previews the argument that follows.

Behind the conflict sits a simple mechanism: *temperature is nothing but sharpening the model's own preference*, and whether sharpening helps is determined entirely by the preference it amplifies.

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.

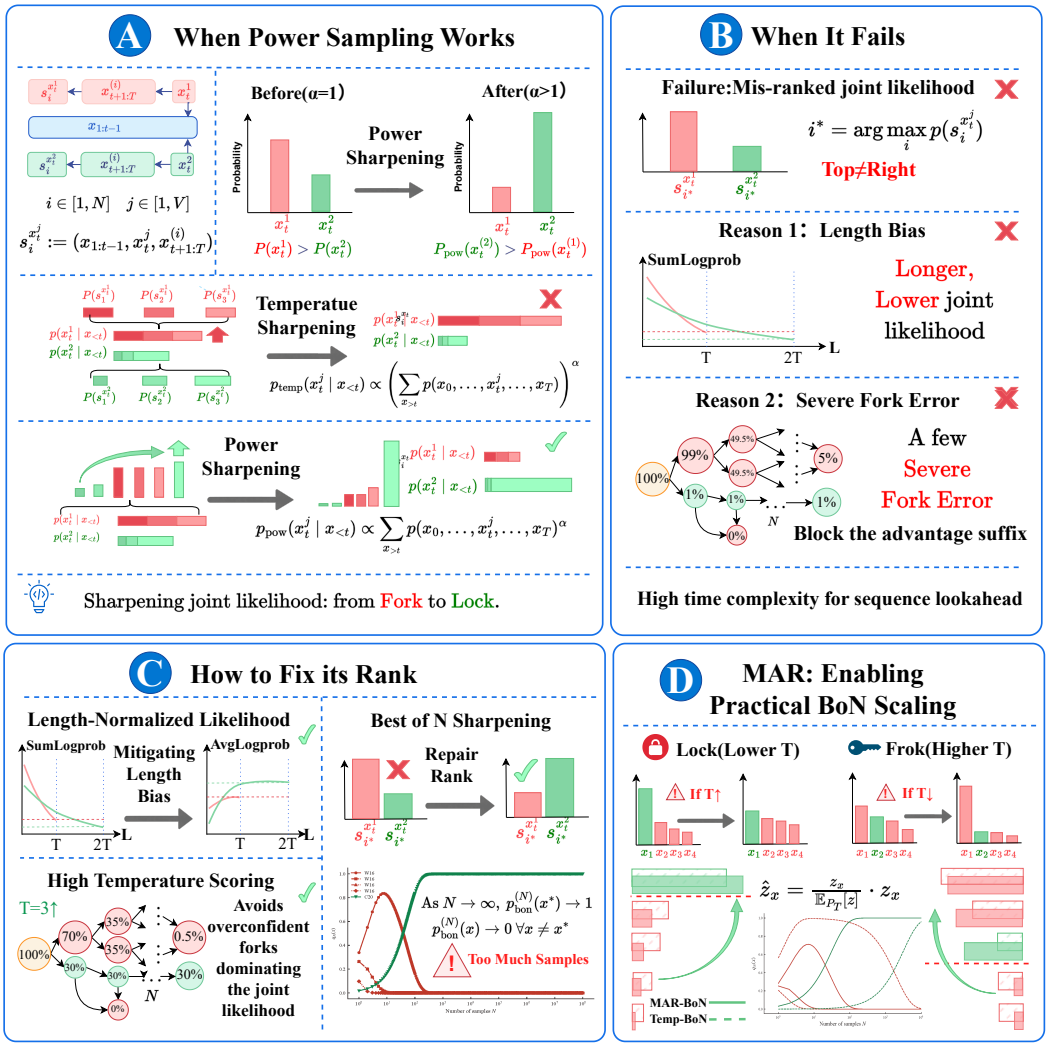


Figure 1: **Overview of the paper.** (**A**) Power sampling sharpens the joint likelihood, overturning wrong token-level preferences: forks become locks. (**B**) But the joint likelihood is mis-ranked—length bias and severe forks—and expensive to sample from. (**C**) Average log-likelihood and elevated-temperature scoring repair the rank; repairing the proposal runs into the lock–fork temperature tension. (**D**) MAR rescales each logit by its depth below the expected logit, $\lambda_x = z_x / \mathbb{E}_{P_{T_b}}[z]$, serving locks and forks under one temperature and bringing Best-of-N scaling into practical budgets.

If the model's top choice is correct, sharpening is unconditionally good. If the preference is only partially correct, sharpening helps only within a degree. If the preference is completely wrong, sharpening is unambiguously harmful: it concentrates mass on the wrong peak. These three regimes are precisely the lock, the mild fork, and the severe fork, each with a distinct loss profile: a wrong direction causes *preference loss*, insufficient strength causes *dispersion loss* (Appendix B). In short, the ranking of the preference determines the *direction* of sharpening, and its quality determines the *strength* sharpening can safely reach. Preference at the sequence level aggregates evidence across positions: a correct sequence can remain the model's global favorite even when several of its steps are locally misranked. Sharpening in this better direction is what power sampling (Karan & Du, 2026) achieves: marginalizing a power-tilted target over all continuations, it draws each token in proportion to the tilted mass of the complete sequences it can lead to, $p_{\mathrm{pow}}(x_t^j \mid x_{<t}) \propto \sum_{x_{>t}} p(x_0, \ldots, x_t^j, \ldots, x_T)^\alpha$ (Figure 1, panel A). Tokens heading toward sequences the model

is more confident in are boosted, many forks are repaired into locks, and sharpening beats low-temperature sampling at high strength.

Power sampling shows that repairing the direction works, but it leaves two deficiencies. Its chains need sequence-level lookahead and return a single trajectory each; and the target itself is mis-ranked (Figure 1, panel B): joint likelihood shrinks with length and is dominated by severe forks, while what we want from a reasoning trajectory is confidence in the *whole* sequence. Both defects are repairable at the level of the score. Average log-probability removes the length bias exactly. Elevated-temperature scoring de-emphasizes severe forks: heating can flip near-tie rankings in favor of globally confident continuations—e.g. it moves $(0.5, 0.4, 0.05, 0.05)$ toward $(0.3, 0.35, 0.25, 0.1)$, where the correct second choice loses mass but becomes top-ranked—and since the scoring temperature never enters the sampler, heating the score is free improvement for selection even though heating the proposal would harm locks (panel C). This reliably recovers about a point of accuracy, and what remains is a residue no rescoring can repair. The repaired object, however, is a *score*, not a probability distribution, and power sharpening has no purchase on a score. Selection needs only a ranking, which points to Best-of-N (BoN); yet BoN is widely regarded as mere selection that cannot sharpen, and in practice it has not matched power sampling even at large budgets.

We propose a *sharpening-loss theory* that says otherwise (Appendix B). Any score, together with a proposal q, induces a family of sharpened distributions with two degrees of freedom: a *direction*—which candidates receive mass, fixed by the score—and a *strength*—how far mass concentrates at a fixed parameter; as the parameter grows, strength accumulates at a *rate* that differs from mechanism to mechanism. If we instead model the BoN output distribution directly—the probability that a sequence is returned after drawing N candidates and keeping the highest-scoring one—we see that any suboptimal candidate is increasingly likely to be beaten by a higher-scoring draw, so its return probability decays with N; in the limit the output concentrates on the score's top-scoring region. Thus BoN *is* a sharpening mechanism, whose direction is supplied by the score, and at the limit it shares power sampling's ceiling. The practical gap is therefore not one of direction—BoN over the repaired score sharpens in a better direction than power sampling over raw likelihood—but one of rate: the strength reachable within a budget N is throttled by the proposal's per-sample hit probability, since a candidate must be sampled before it can be selected. And the proposal cannot be repaired by any global temperature, because the lock–fork conflict reappears inside it: a correct trajectory is a thin spike whose mass is diverted at every fork into a large family of plausible wrong continuations—cooling shields its path at every fork, heating destroys the locks. On fork-heavy problems, BoN's scaling begins only at budgets far beyond the practical (Figure 1, panel D).

In this paper, we present a surprising result: with the repaired score and a proposal shaped by *marginal-aware rescaling* (MAR), plain Best-of-N matches or surpasses power sampling at a fraction of its sampling budget, and matches or surpasses RL post-training without any training, with the largest gains on MATH.

MAR computes, at each step, the expected logit $\mathbb{E}_{P_T}[z]$ at a boundary temperature T_b and rescales every logit by its own depth below this boundary: the distractor tail at locks is pushed down, the viable competitors at forks are pulled toward equal footing, the model's ordering is preserved, and no scalar temperature can reproduce the transform (Figure 1, panel D). MAR is *verifier-agnostic*: any signal that ranks candidates can serve as the score. For a direct comparison with power sampling, our main experiments use the model's own repaired likelihood, requiring no external verifier, and the gains persist when the score is replaced by external verifiers (Appendix D).

2 Preliminaries

2.1 Setup and notation

Let $\mathcal{V}$ denote a vocabulary. For an autoregressive model, at decoding step t with prefix $h_t = (x_1, \ldots, x_{t-1})$, let $z_x(h_t)$ be the logit assigned to token $x \in \mathcal{V}$; its base distribution is the temperature-one next-token distribution $p(x \mid h_t) \propto \exp z_x(h_t)$, and at temperature $T > 0$, $p_T(x \mid h_t) \propto \exp\big(z_x(h_t)/T\big)$. A complete sequence $X = (x_1, \ldots, x_L)$ has joint probability $p(X) = \prod_{t=1}^{L} p(x_t \mid h_t)$ and log-likelihood $\ell(X) = \log p(X)$. Throughout, T_g denotes the temperature at which candidates are generated and T_s the temperature at which they are scored; the two are decoupled. When comparing sequences of different lengths we use the length-normalized

average log-likelihood $S_T^{\text{avg}}(X) = \ell_T(X)/L(X)$, which removes the length dependence exactly: $\ell_T(X) = L(X)\, S_T^{\text{avg}}(X)$ (Appendix B.1, Eq. (20)). A *score* S is any function that ranks complete sequences; unless otherwise specified, $S = S_{T_s}^{\text{avg}}$.

2.2 Locks and forks

Reasoning sequences interleave two types of steps: the lock and the fork (Definition B.1 of Appendix B.1). Fix a target token x_t^* (e.g., the token taken by a reference solution at step t); write $M_t = \max_x z_x(h_t)$ and $m_t = \arg\max_x z_x(h_t)$. A step is a *lock* if the model's top choice is already the target, $m_t = x_t^*$, and a *fork* otherwise. The *preference gap* $\Delta_t^* = M_t - z_{x_t^*}(h_t) \geq 0$ measures how far the target is ranked below the top choice: forks with small Δ_t^* are *mild*, while forks whose wrong peak concentrates most of the mass (e.g., $p(m_t \mid h_t) \approx 0.9$) are *severe*. For a candidate token x whose correctness is not assumed, $\Delta_t(x) = M_t - z_x(h_t)$ is the gap that would be incurred if x were chosen.

2.3 Sharpening operators

The mechanisms below are all *sharpening operators*: they reweight, or select from, the base distribution so that mass concentrates on candidates the model itself prefers more strongly. They differ in *where they look* when deciding which candidates to promote (Appendix B.1).

Temperature: local sharpening. Temperature is the cheapest operator: it rescales the logits, $p_T(x \mid h_t) \propto \exp\big(z_x(h_t)/T\big)$, which is exactly a power of the base distribution taken *one step at a time*, $p_T(\cdot \mid h_t) \propto p(\cdot \mid h_t)^{1/T}$, each step renormalized independently of every other. Sharpening ($T < 1$) amplifies whatever the model prefers *locally*—but no step's mass is informed by the quality of the continuations it leads to, so a step the model is locally confident in may still head toward globally worse conclusions.

Power sampling: global sharpening. Power sampling removes this locality. Instead of sharpening each step separately, it raises the probability of the *complete sequence* to a power, $p_\alpha(X) = p(X)^\alpha / Z_\alpha$ with $\alpha > 1$ (Karan & Du, 2026). Marginalizing over continuations, a single step then draws each token in proportion to the tilted mass of the complete sequences it can lead to, $p_{\text{pow}}(x_t^j \mid x_{<t}) \propto \sum_{x_{>t}} p(x_1, \ldots, x_t^j, \ldots, x_L)^\alpha$: a token that is locally dispreferred but heads toward confident continuations is promoted, and wrong token-level preferences are overturned. The price is *global* normalization: exact sampling requires lookahead over continuations, so in practice p_α is approximated by Metropolis–Hastings with block resampling, at significant per-trajectory cost.

Best-of-N: sharpening by selection. Sharpening, however, need not produce a normalized distribution at all—selection needs only a ranking. Best-of-N (BoN) draws $X_1, \ldots, X_N$ i.i.d. from a *proposal* q—by default the model's own distribution $q = p_{T_g}$—and returns the champion $\hat{X} = \arg\max_i S(X_i)$. For a candidate X with proposal mass $q_X = q(X)$, the *winning mass* $W_X = \sum_{X' :\, S(X') > S(X)} q(X')$ is the probability that a single draw outscores it. The output distribution then has the closed form

$$p_{\text{bon}}^{(N)}(X) = (1 - W_X)^N - (1 - W_X - q_X)^N. \tag{1}$$

Eq. (1) separates two views: W_X sees only the score's *ranking*—the sharpening *direction*—while q_X sees only the proposal's *sampling mass*—the raw material any selection mechanism has to work with (Appendix B.1).

3 Sharpening-Loss Theory and Marginal-Aware Rescaling

All operators of Section 2.3 reweight the same base distribution, so they can be compared on one loss. The comparison separates inference failure into a *direction* component and a *rate* component; we repair the two independently—a repaired score fixes the direction (Section 3.2), marginal-aware rescaling fixes the rate (Section 3.3). Proofs are in Appendix B.

3.1 SHARPENING-LOSS THEORY

One loss, three decompositions. Under the *one-hot assumption*—a single best token x_t^* per step and a single best sequence X^*—a sharpened distribution r is measured by $\mathcal{L} = -\log r(x_t^*)$ (step) or $-\log r(X)$ (sequence). Instantiating r with the three operators gives the same decomposition each time, a *preference loss* (distance to the operator's favorite) plus a *dispersion loss* (mass spread over non-favorites; Appendix B.2):

$$-\log p_T(x_t^* \mid h_t) = \underbrace{\frac{M_t - z_{x_t^*}(h_t)}{T}}_{\text{preference}} + \underbrace{\varepsilon_{T,t}}_{\text{dispersion}}, \tag{2}$$

$$-\log p_\alpha(X^*) = \underbrace{\alpha\big(-\log p(X^*)\big)}_{\text{preference}} + \underbrace{\log Z_\alpha}_{\text{dispersion}}, \tag{3}$$

$$-\log p_{\text{bon}}^{(N)}(X^*) = \underbrace{\big(-N\log(1-W_{X^*})\big)}_{\text{preference}} + \underbrace{\Big(-\log\big(1-(1-\tilde{q}_{X^*})^N\big)\Big)}_{\text{dispersion}}, \tag{4}$$

where $\varepsilon_{T,t} = \log\big(1 + \sum_{x \neq m_t} \exp\frac{z_x(h_t) - M_t}{T}\big)$ is the softmax remainder, $Z_\alpha = \sum_X p(X)^\alpha$, and $\tilde{q}_{X^*} := q_{X^*}/(1 - W_{X^*})$ is the target's *conditional hit probability*. The operators differ only in the favorite: the local top choice for temperature, the largest *continuation mass* for power, the score's top region for BoN.

One trade-off, three parameters. Raising the strength ($1/T$, α, or N) amplifies the preference term but suppresses dispersion, with diminishing relief as the output concentrates: the loss is strictly convex in the strength, minimized at infinite strength at a lock and at a finite, misranking-dependent strength at a fork (Appendix B.3; Proposition B.5 for BoN).

Direction sets the ceiling; rate is BoN's constraint. The preference term is fixed by the favorite, hence by the direction: as strength grows, every operator collapses onto its favorite, and if that favorite is not the target the loss diverges—for BoN the preference term grows at least linearly, $-N\log(1 - W_{X^*}) \geq NW_{X^*}$ (Eq. (33) of Appendix B.2). Budget never repairs a wrong direction; direction sets the ceiling, and strength only decides how much of it is realized. The binding question is therefore the *rate*: BoN's dispersion relief grows only like $\log N$ and saturates at $1/\tilde{q}_{X^*}$, so the optimal budget scales as $N^* \approx \tilde{q}_{X^*}^{-1} \log(1 + \tilde{q}_{X^*}/W_{X^*})$ (Proposition B.5). Yet $\tilde{q}_{X^*}$ is tiny on hard problems: a correct trajectory is a thin spike whose mass decays multiplicatively along the forks it threads, and no global temperature repairs this without breaking locks (Section 2.2). The defect is one of rate, not ceiling—as $N \to \infty$ BoN concentrates on the score's top region, power sampling's limit as well (Remark B.10).

3.2 REPAIRING THE DIRECTION: AVERAGE LIKELIHOOD AT AN ELEVATED TEMPERATURE

We score candidates by the average log-likelihood at a scoring temperature T_s decoupled from generation,

$$S(X) := S_{T_s}^{\text{avg}}(X) = \frac{\ell_{T_s}(X)}{L(X)}, \tag{5}$$

which removes the length bias exactly, $\ell_T(X) = L(X)\, S_T^{\text{avg}}(X)$ (Appendix B.1, Eq. (20)). Warming the score further repairs the ranking: for a correct competitor X^+ and an incorrect one X^-, the margin $R(T) = S_T^{\text{avg}}(X^+) - S_T^{\text{avg}}(X^-) = -A/T - D(T)$—with $A > 0$ the excess preference penalty the correct path accumulates and $D(T)$ the dispersion difference—rises from low temperature to a unique interior optimum then declines: warming dissolves the penalty for locally dispreferred steps until dispersion dominates (Appendix B.5, Proposition B.11). Since the score never enters the sampler, raising T_s is free; a genuinely wrong preference remains unrepaired.

3.3 REPAIRING THE RATE: MARGINAL-AWARE RESCALING

Per-step optimal strengths are incompatible (locks demand ∞, severe forks 0), and one scalar can only average over them (Eq. (45) of Appendix B.3). The fix must act per step, in both directions.

Differentiating Eq. (2),

$$\frac{d}{dT}\big(-\log p_T(x \mid h_t)\big) = \frac{z_x(h_t) - \mathbb{E}_t(T)}{T^2}, \qquad \mathbb{E}_t(T) = \textstyle\sum_x p_T(x \mid h_t)\, z_x(h_t),$$

so the expected logit $\mathbb{E}_t(T)$—strictly below the peak on non-degenerate steps—separates contenders (above; harmed by heating) from diluters (below). Computing it at a *boundary temperature* T_b and rescaling each max-shifted logit ($z_x \le 0$) by its own depth below the boundary gives *marginal-aware rescaling* (MAR):

$$\lambda_x = \frac{z_x}{\mathbb{E}_t(T_b)} \ge 0, \qquad \tilde{z}_x = \lambda_x\, z_x = \frac{z_x^2}{\mathbb{E}_t(T_b)}. \tag{6}$$

Every pairwise gap is rescaled by the depth of its pair, $\hat{\Delta}_{ij} = \kappa_{ij}\,\Delta_{ij}$, $\kappa_{ij} = (z_i + z_j)/\mathbb{E}_t(T_b)$ (Proposition B.6): deep-tail pairs at locks get $\kappa \gg 1$ and are pushed down (mass decays as $\exp(-\Delta^2/T_g)$ instead of $\exp(-\Delta/T_g)$), near-peak pairs at forks get $\kappa \ll 1$ and are leveled. The transform preserves the model's ordering and is not reproducible by any single temperature (Appendix B.4): locks are protected and forks activated under one global T_g. We recommend $T_b = 3$; $[2, 3]$ is indistinguishable in practice.

The method. Best-of-N with the repaired score and the MAR-shaped proposal: (i) generate N candidates from the rescaled logits at T_g; (ii) score by $S = S^{\mathrm{avg}}_{T_s}$; (iii) return the champion. No training, search, or external signal is needed beyond the score, and MAR adds $O(V)$ overhead per step. It does not raise BoN's ceiling (Remark B.10); it raises $\tilde{q}_{X^*}$, so the optimal coverage is reached at a smaller budget.

4 Experiments

4.1 Experimental Setup

Benchmarks. For a fair comparison with power sampling we adopt the benchmark suite of Karan & Du (2026): **MATH500** (competition mathematics), **HumanEval** (program synthesis), and **GPQA Diamond** (graduate-level science QA). Dataset descriptions and grading protocols are given in Appendix C.

Models and baselines. We experiment on four open base models, Qwen2.5-Math-7B, Qwen2.5-7B, Qwen3-4B, and Llama3-8B-Instruct (Yang et al., 2024b;a; 2025; Grattafiori et al., 2024). On these models we compare MAR against three inference-time baselines, standard autoregressive sampling at $T{=}1$ (**Base**), its low-temperature variant at $T{=}0.25$ (**Low-temperature**), and MCMC **Power Sampling** ($\alpha{=}4$) of Karan & Du (2026), together with one training baseline, **GRPO(MATH)**, obtained by GRPO post-training on the MATH training split. Checkpoint sources and baseline configurations are given in Appendix C.

Experiment settings. All methods share the same benchmarks, models, and answer protocol; full details are given in Appendix C. For MAR we use $N{=}8$ candidates, a scoring temperature $T_s{=}3$, and a boundary temperature $T_b{=}3$. The scoring temperature is fixed at 3 because accuracy saturates there (Section 3.2), and since the score never enters the sampler, warming it is free; the boundary temperature is set to the same value, as performance is insensitive across $[2, 3]$. The generation temperature T_g is set to the stronger of the temperature-sampling baselines on each benchmark, and no other tuning is performed. Hyperparameters for all methods are summarized in Table 1.

4.2 Main Results

Table 2 presents the main results across four model families and three benchmarks. On MATH500, MAR comprehensively surpasses both power sampling and GRPO(MATH); on GPQA Diamond and HumanEval it performs on par with the strongest baselines. The single large deficit is Qwen3-4B's MATH500, where GRPO(MATH) remains far ahead: this model is small and its global preference is itself unreliable, so RL post-training, which reshapes the model rather than selecting from it, outperforms any inference-time scaling—yet the same RL training severely degrades generalization, and GRPO(MATH) falls behind MAR on the other two benchmarks.

Table 1: Main hyperparameter settings for each method.

Method	Hyperparameters
Temperature sampling	temperature $\in \{1, 0.25\}$ (per model & dataset)
Standard BoN	$N = 8$; T_g same as MAR; selection: joint log-likelihood
MCMC power sampling	$\alpha = 4$, $B = 192$, $N_{\text{MCMC}} = 10$, p_{prop}: base temperature $1/\alpha$
GRPO (MATH)	Standard GRPO setup (learning rate, batch size, KL penalty, etc.)
MAR (ours)	$N = 8$; $T_g \in \{\text{base}, 0.25\}$ (per model & dataset); $T_s = 3$; $T_b = 3$

Table 2: Accuracy on MATH500, HumanEval, and GPQA Diamond (experimental details in Appendix C). **Bold**: best inference-time method per model; GRPO rows are excluded as they spend a training rather than a sampling budget. Colored deltas are relative to the GRPO(MATH) row of the same model.

Method	MATH500	HumanEval	GPQA
Qwen2.5-Math-7B			
Base(T=1)	0.496	0.329	0.278
Low-temperature(T=0.25)	0.690	0.512	0.353
Standard BoN	0.734	0.549	0.263
Power Sampling(MCMC)	0.748 (-0.037)	0.573 (+0.036)	0.389 (-0.010)
MAR-BoN(ours)	**0.806** (+0.021)	**0.610** (+0.073)	**0.404** (+0.005)
GRPO (MATH)	0.785	0.537	0.399
Qwen2.5-7B			
Base(T=1)	0.498	0.329	0.278
Low-temperature(T=0.25)	0.628	0.524	0.303
Standard BoN	0.686	0.518	0.313
Power Sampling(MCMC)	0.706 (-0.034)	**0.622** (+0.061)	0.318 (-0.036)
MAR-BoN(ours)	**0.744** (+0.004)	0.610 (+0.049)	**0.338** (-0.016)
GRPO (MATH)	0.740	0.561	0.354
Qwen3-4B			
Base(T=1)	0.428	0.659	0.414
Low-temperature(T=0.25)	0.374	0.720	0.429
Standard BoN	0.450	0.726	0.414
Power Sampling(MCMC)	0.480 (-0.294)	**0.744** (+0.086)	0.475 (+0.122)
MAR-BoN(ours)	**0.538** (-0.236)	0.732 (+0.074)	**0.505** (+0.152)
GRPO (MATH)	0.774	0.658	0.353
Llama3-8B-Instruct			
Base(T=1)	0.334	0.549	0.293
Low-temperature(T=0.25)	0.474	0.592	0.323
Standard BoN	0.524	0.579	0.318
Power Sampling(MCMC)	0.512 (+0.064)	**0.621** (+0.060)	0.323 (-0.020)
MAR-BoN(ours)	**0.546** (+0.082)	0.609 (+0.048)	**0.328** (-0.015)
GRPO (MATH)	0.464	0.561	0.343

Elevated-temperature scoring repairs the ranking. Figure 2 sweeps T_s against the sampling temperature. At every sampling temperature, accuracy climbs with T_s and reaches a stable peak around $T_s = 3$: e.g., at $T_{\text{sample}} = 0.25$ it rises from 0.740 ($T_s = 0.25$) to 0.760 at $T_s = 3$, with essentially no gain beyond. The hot scorer with a cool sampler is again the best configuration, as the theory predicts—the score never enters the sampler, so warming it is free (Section 3.2)—and the peak-then-plateau shape is the two-phase trade-off of Proposition B.11: warming dissolves the preference penalty until dispersion growth cancels it. Since accuracy saturates at $T_s = 3$, we use $T = 3$ for MAR's boundary temperature as well.

MAR restores Best-of-N scaling. Figure 3(a) plots accuracy against the candidate budget N on Qwen2.5-Math-7B/MATH500, with MCMC power sampling (0.748) as a fixed reference. At $N = 1$ the two proposals are close (0.690 vs. 0.702), so MAR's advantage is not a one-shot effect but a scaling effect. The fixed-temperature proposal saturates after $N = 4$ (0.742–0.760 through $N = 20$), whereas MAR reaches 0.806 at $N = 8$, peaks at 0.810, and stays stable at 0.804 through

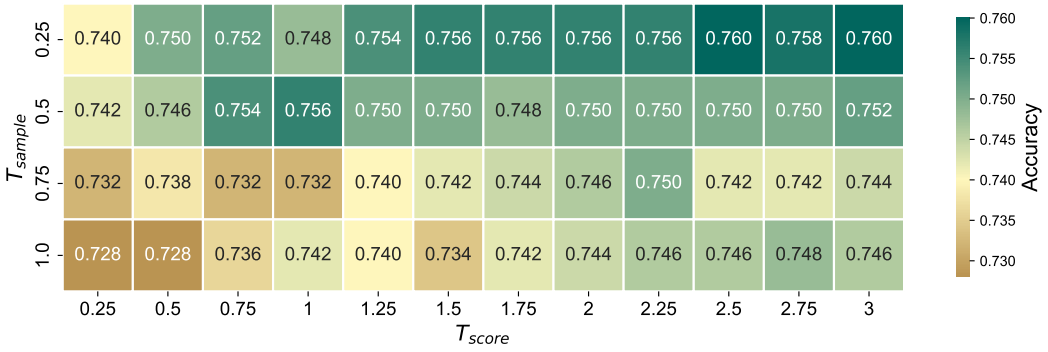
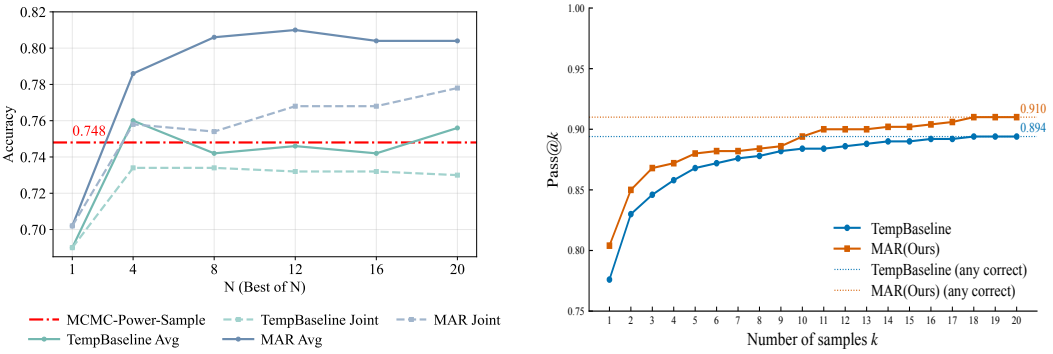


Figure 2: Accuracy of Best-of-N (Qwen2.5-Math-7B, MATH500) over $T_{\text{sample}} \times T_s$. Accuracy rises approximately monotonically with $T_s \in [0, 3]$ and stabilizes at its peak around $T_s = 3$.



(a) Best-of-N scaling

(b) Pass@k

Figure 3: MAR restores Best-of-N scaling on Qwen2.5-Math-7B/MATH500: accuracy grows with N while the baseline saturates, and pass@k confirms better coverage.

$N = 20$, above the MCMC reference. MAR protects locks by suppressing the distractor tail and activates forks by compressing near-peak gaps, improving precision and exploration under one proposal. Since the verifier and selection rule are identical, the contrast isolates the proposal: BoN's bottleneck is neither the verifier nor the selection mechanism, but the poor proposal, which prevents sharpening from manifesting at practical budgets. Figure 3(b) shows that this comes without any loss of diversity: MAR's pass@k curve lies above the baseline at every k, ending at 0.910 vs. 0.894 at $k = 20$. A higher pass@k means the candidate pool contains the correct answer more often, so MAR also raises the ceiling available to an oracle verifier.

Mean likelihood is the better selection score. Figure 3(a) also separates the score's form from the proposal's shape. Across every configuration and budget, mean-scored Best-of-N consistently and substantially outperforms sum-scored Best-of-N, under both the MAR and temperature proposals. This confirms that average log-likelihood is the better selection score: the summed log-likelihood conflates quality with brevity—short wrong candidates systematically outrank long correct ones—whereas the average score removes this length penalty exactly.

Ablation: both components are necessary. Table 3 ablates the two repairs on Qwen2.5-Math-7B. Removing either component degrades every number, and their contributions split along the theory's two axes: elevated-temperature scoring adds $+1.6/+1.2/+1.1$ points on MATH500/HumanEval/GPQA on top of the MAR proposal (fixing *selection*), while the MAR proposal adds $+4.6/+0.6/+4.1$ on top of the repaired score (fixing *coverage*)—coverage is the binding constraint on MATH500 and GPQA, whereas on HumanEval neither repair moves accuracy much, echoing the one-point gaps of Section 4.2. Standard BoN, with neither repair, is the weakest variant overall ($0.740/0.598/0.379$). The one non-monotonic cell is instructive rather than embarrassing: with the raw temperature proposal, $T_s = 3$ is slightly *worse* than $T_s = 1$ on GPQA (0.363 vs. 0.379)—rescoring cannot help when the pool rarely contains the correct candidate—which is precisely the coverage-first view of Section 3.1.

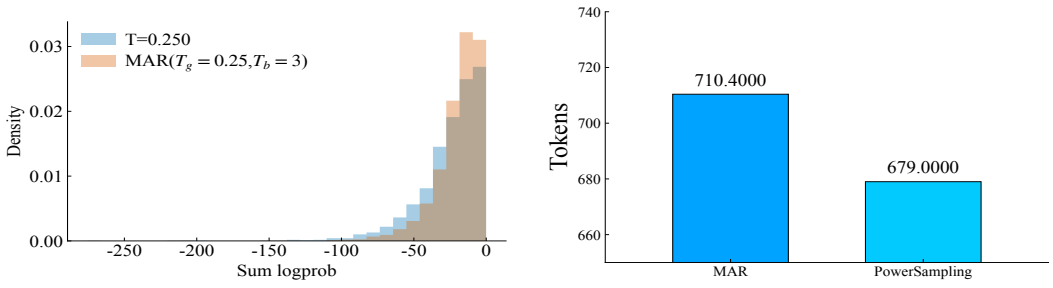
Table 3: Ablation on Qwen2.5-Math-7B. Removing either the elevated-temperature score (T_s) or the MAR proposal degrades accuracy.

Variant	MATH500	HumanEval	GPQA
MAR (full)	0.806	0.610	0.404
w/o elevated-T_s scoring ($T_s = 1$)	0.790	0.598	0.393
w/o MAR (temperature proposal)	0.760	0.604	0.363
w/o both (BoN-AvgLoglikelyhood)	0.740	0.598	0.379

Table 4: Cost per delivered answer: newly generated tokens counted at max output length Tmax = 3072; wall-clock is the average time per output over actual dataset runs; all multiples are relative to a single generation (single RTX 5090; both methods served by vLLM).

Method	New tokens	Serial wall-clock	Parallel wall-clock
Single generation (T=1)	3,072 (1×)	5.26 s (1.00×)	5.26 s (1.00×)
Power sampling (MCMC)	130,560 (42.5×)	242.64 s (46.1×)	242.64 s (46.1×)
MAR (N=8)	24,576 (8×)	42.12 s (8.01×)	5.69 s (1.08×)

Inference Efficiency Table 4 compares the cost per delivered answer at $T_{\max} = 3072$. MAR is $O(NT)$ serially—$8T$ tokens, $8.01\times$ baseline wall-clock—and $O(T)$ in parallel, since its N candidates batch in a single decode ($1.08\times$ baseline); power sampling is $O(N_{\mathrm{MCMC}}T^2/B)$ serially with no batched-parallel form, paying $42.5\times$ the tokens and $46.1\times$ the wall-clock. MAR is thus cheaper on both counts—$5.3\times$ fewer tokens and $42.6\times$ faster than MCMC in the parallel regime—and the rescaling itself adds almost no overhead: a single MAR sample takes 5.49 s against the 5.26 s baseline ($1.04\times$; full serial comparison in Appendix E).



(a) Correct-sequence likelihood

(b) Output length

Figure 4: (a) Sum log-probability of correct sequences under temperature sampling and MAR. (b) Average output length of MAR-BoN vs. power sampling.

MAR improves the proposal's likelihood and mitigates brevity bias. Figure 4a shows that MAR shifts the sum log-probability of correct sequences toward higher values and concentrates more mass near zero, raising their sampling probability—the rate repair predicted by the theory. Figure 4b shows that MAR-BoN produces longer outputs than power sampling (710.4 vs. 679.0 tokens on average), mitigating the brevity bias of joint-likelihood sharpening. Together, MAR improves the proposal in both likelihood and length.

5 Conclusion

We unified power sampling and Best-of-N as one mechanism—sharpening with a direction set by the amplified preference and a strength that accumulates at a mechanism-specific rate—and located BoN's gap to power sampling in the rate, not the direction. Two independent, training-free repairs follow: a length-normalized, elevated-temperature average-likelihood score fixes the direction at zero sampling cost, and marginal-aware rescaling fixes the rate, protecting locks and activating forks under one global temperature; together they let plain Best-of-N match or surpass both power sampling and RL post-training. The theory also marks the training-free boundary—MAR raises the rate toward the ceiling, never the ceiling itself, which is the model's own global preference—and points to scores with execution or verification signals as the natural next step.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure: assisting in the writing of proofs and providing critical ingredients for proving mathematical claims (e.g., assistance with formula derivations and proof steps), and implementing methods (specifically, generating parts of the code implementation based on our algorithm design). We have not used generative AI tools for the other tasks with required disclosure, and the remaining required disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools for the following tasks with recommended disclosure: editing a research paper to improve readability (language polishing) and creating or editing software code (as part of the code implementation described above). We have reviewed all AI-assisted work. The AI-assisted text was manually checked and revised by the authors; the AI-assisted formulas and proof steps were verified by the authors; and the AI-assisted code was reviewed, tested, and validated by the authors for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Seyedarmin Azizi, Erfan Baghaei Potraghloo, Minoo Ahmadi, Souvik Kundu, and Massoud Pedram. POWER-SMC: Low-latency sequence-level power sampling for training-free LLM reasoning, 2026. URL https://arxiv.org/abs/2602.10273.

Eric Bigelow, Ari Holtzman, Hidenori Tanaka, and Tomer Ullman. Forking paths in neural text generation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 17134–17167, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/2b6a8ca3d06ffa2e044bff8c723863ff-Paper-Conference.pdf.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Shehzaad Dhuliawala, Ilia Kulikov, Ping Yu, Asli Celikyilmaz, Jason Weston, Sainbayar Sukhbaatar, and Jack Lanchantin. Adaptive decoding via latent preference optimization, 2024. URL https://arxiv.org/abs/2411.09661.

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, 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/.

Weihua Du, Yiming Yang, and Sean Welleck. Optimizing temperature for language models with multi-sample inference. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/du25f.html.

Bryan Eikema and Wilker Aziz. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4506–4520, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.398. URL https://aclanthology.org/2020.coling-main.398/.

Gonçalo R. A. Faria, Sweta Agrawal, António Farinhas, Ricardo Rei, José G. C. de Souza, and André F.T. Martins. Quest: Quality-aware metropolis-hastings sampling for machine translation. 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. 89042–89068. Curran Associates, Inc., 2024. doi: 10.52202/079017-2825. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a221d22ff6a33599142c8299c7ed06bb-Paper-Conference.pdf.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Andre Wang He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting GRPO beyond distribution sharpening. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25548–25560, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1298. URL https://aclanthology.org/2025.emnlp-main.1298/.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks*, 2021.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J. Foster. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/huang25c.html.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty, 2025. URL https://arxiv.org/abs/2502.18581.

Aayush Karan and Yilun Du. Reasoning with sampling: Your base model is smarter than you think. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Vsgq2ldr4K&nesting=2&sort=date-desc. Oral.

Marvin Li, Aayush Karan, and Sitan Chen. Blink of an eye: a simple theory for feature localization in generative models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/li25at.html.

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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. Critical tokens matter: Token-level contrastive estimation enhances llm's reasoning capability. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second*

International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/lin25j.html.

Clara Meister, Ryan Cotterell, and Tim Vieira. If beam search is the answer, what was the question? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2173–2185, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.170. URL https://aclanthology.org/2020.emnlp-main.170/.

Nguyen Nhat Minh, Andrew Baker, Clement Neo, Allen G. Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=FBkpCyujtS.

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.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Felix Stahlberg and Bill Byrne. On NMT search errors and model errors: Cat got your tongue? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3356–3362, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1331. URL https://aclanthology.org/D19-1331/.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize from human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

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

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. Reasoning aware self-consistency: Leveraging reasoning paths for efficient LLM sampling. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3613–3635, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.184. URL https://aclanthology.org/2025.naacl-long.184/.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. In 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. 115452–115486. Curran Associates, Inc., 2025a. doi: 10.52202/085713-3850. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a797c2d2e0c1fdabf4d1ab8cd0b465c6-Paper-Conference.pdf.

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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=1PL1NIMMrw.

Yizhou Wang, Lingzhi Zhang, Yue Bai, Mang Tik Chiu, Zhengmian Hu, Mingyuan Zhang, Qihua Dong, Yu Yin, Sohrab Amirghodsi, and Yun Fu. Cautious next token prediction. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25685–25697, Vienna, Austria, 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1318. URL https://aclanthology.org/2025.findings-acl.1318/.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report, 2024a. URL https://arxiv.org/abs/2412.15115.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024b. URL https://arxiv.org/abs/2409.12122.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengren Huang, Chenxu Lv, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yunhao Yang, Yuexin Bian, Yunjie Tian, Di Fu, Tianjin Huang, Yuanyuan Shi, Ziang Xiao, Nuno Vasconcelos, and Yijiang Li. Co-rl: Unsupervised reasoning emerges from diverse cohort in multi-agent rl, 2026. URL https://arxiv.org/abs/2608.17253.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? 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. 57654–57689. Curran Associates, Inc., 2025. doi: 10.52202/085713-1933. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/537d5aa768c2d534016a4d06f87bc8fb-Paper-Conference.pdf.

Ruixiang Zhang, Richard He Bai, Huangjie Zheng, Navdeep Jaitly, Ronan Collobert, and Yizhe Zhang. Embarrassingly simple self-distillation improves code generation, 2026. URL https://arxiv.org/abs/2604.01193.

Stephen Zhao, Rob Brekelmans, Alireza Makhzani, and Roger Baker Grosse. Probabilistic inference in language models via twisted sequential monte carlo. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=frA0NNBS1n.

Yuqi Zhu, Jia Li, Ge Li, Yunfei Zhao, Jia Li, Zhi Jin, and Hong Mei. Hot or cold? adaptive temperature sampling for code generation with large language models. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 437–445. AAAI Press, 2024. doi: 10.1609/AAAI.V38I1.27798. URL https://doi.org/10.1609/aaai.v38i1.27798.

A Related Work

Sharpening and power sampling. A growing body of work interprets RLVR as *sharpening* rather than capability acquisition: RL-trained reasoners improve pass@1 by concentrating mass on high-likelihood, already-correct trajectories, while the base model retains comparable coverage at large budgets (Yue et al., 2025; Guo et al., 2025; Shao et al., 2024); He et al. (2025) make the sharpening mechanism explicit and lift GRPO beyond it, and Huang et al. (2025) show that Best-of-N over the base distribution can in principle match RL alignment. Power sampling takes this premise to the training-free limit, drawing from the base distribution raised to a power α via MCMC (Karan & Du, 2026), with follow-ups reducing its cost through scalable or SMC-based implementations (Azizi et al., 2026; Zhao et al., 2024; Faria et al., 2024); a parallel tradition selects instead of reweighting, from self-consistency and verifier-based Best-of-N to verifier-free selection (Wang et al., 2023; Wan et al., 2025; Stiennon et al., 2020; Kang et al., 2025). At the sequence level, however, all of these methods still target the *joint* likelihood—and that maximizing model likelihood need not maximize quality is a classical lesson of neural machine translation: wider beams find higher-likelihood but worse translations (Stahlberg & Byrne, 2019), and the mode itself can be an inadequate target (Eikema & Aziz, 2020; Meister et al., 2020). Our departure from this line is twofold. First, we do not seek the sequence of maximal joint likelihood; we sharpen selection toward sequences of higher *mean* likelihood, whose length normalization removes the brevity bias that spoiled joint likelihood, and whose ranking we further repair by elevated-temperature scoring. Second, at this repaired argmax limit the *estimation* paradigm these methods operate in—where fidelity to a normalized target sets the sample complexity—collapses into a far cheaper *coverage* problem: the remaining gap is not estimation quality but the proposal's per-sample hit probability.

The limits of global decoding. Practical decoding shapes the next-token distribution with truncation and a single scalar temperature (Holtzman et al., 2020; Minh et al., 2025; Tan et al., 2026; Ding et al., 2026), with variants adapting the shaping per step or per sample (Dhuliawala et al., 2024; Zhu et al., 2024; Wang et al., 2025b; Du et al., 2025). Such global control is fundamentally limited by the token-level structure of generation: *fork* positions, where several continuations genuinely diverge, interleave with *lock* positions, where a single continuation is essentially determined (Bigelow et al., 2025), and a few high-entropy or critical tokens drive most downstream behavior (Wang et al., 2025a; Lin et al., 2025; Li et al., 2025). Zhang et al. (2026) formalize the resulting precision–exploration conflict—lowering temperature secures locks but starves forks, raising it does the opposite—and resolve it by self-distillation, reshaping the model at training time to sharpen locks while protecting forks. MAR is driven by the same mechanism but realizes it entirely at inference time: its order-preserving transform compresses near-peak competition and deepens tail suppression at each position, sharpening locks while protecting forks without any training. A proposal shaped this way remains diverse where exploration matters and precise where commitment matters—precisely the property Best-of-N needs for its accuracy to keep growing with the candidate pool.

B SHARPENING-LOSS THEORY

B.1 SETUP AND NOTATION

The base distribution. Let $\mathcal{V}$ be a vocabulary of size V. An autoregressive model assigns, at each step t with prefix $h_t = (x_1, \ldots, x_{t-1})$, a logit $z_{x_t^j}(h_t)$ to every candidate token x_t^j, $j = 1, \ldots, V$. The *base distribution* of the model is its temperature-one next-token distribution,

$$p(x_t^j \mid h_t) = \frac{\exp z_{x_t^j}(h_t)}{\sum_{j'=1}^{V} \exp z_{x_t^{j'}}(h_t)}. \tag{7}$$

A complete sequence $X = (x_1, \ldots, x_L)$, where each x_t is the token realized at step t, receives probability

$$p(X) = \prod_{t=1}^{L} p(x_t \mid h_t), \qquad \log p(X) = \sum_{t=1}^{L} \log p(x_t \mid h_t). \tag{8}$$

All distributions in this paper are constructed from equation 7–equation 8 by a *sharpening operator*: a transformation that reweights the base distribution so that mass concentrates on candidates the model prefers. The sections below introduce three such operators—temperature, power, and selection—which differ in *where* they look when deciding which candidates to promote.

Locks and forks. Reasoning sequences are not uniform: a small number of steps decide which conclusion the sequence reaches, while most steps are forced. The following structure definition makes this distinction precise.

Definition B.1 (Locks, forks, and preference gaps)**.** Fix a target token x_t^* (e.g., the token taken by a reference solution at step t). A step is a *lock* if the model's top choice is already the target,

$$m_t = \arg\max_{1 \le j \le V} z_{x_t^j}(h_t) = x_t^*,$$

and a *fork* otherwise. The *preference gap*

$$\Delta_t^* = M_t - z_{x_t^*}(h_t) \ge 0, \qquad M_t = \max_{1 \le j \le V} z_{x_t^j}(h_t), \tag{9}$$

measures how far the target is ranked below the model's top choice. Forks with small Δ_t^* are *mild*; forks whose wrong peak concentrates most of the mass (e.g., $p(m_t \mid h_t) \approx 0.9$) are *severe*. For a candidate token x_t^j whose correctness is not assumed, $\Delta_t(x_t^j) = M_t - z_{x_t^j}(h_t)$ denotes the preference gap that the conditional distribution $p(\cdot \mid h_t)$ would exhibit if x_t^j were the target.

Loss under the one-hot assumption. To speak about which sharpened distribution is "better," we need a loss. We adopt the one-hot assumption: at each step there is a single best candidate x_t^*, so the target distribution is the point mass δ_t defined by

$$\delta_t(x_t^j) = \begin{cases} 1, & x_t^j = x_t^*, \\ 0, & x_t^j \ne x_t^*, \end{cases} \tag{10}$$

and, at the sequence level, the point mass δ_{X^*} on the best complete sequence X^*. For any candidate distribution r over the same candidates, the loss is the Kullback–Leibler divergence from the target to the candidate,

$$\mathcal{L}(r) = D_{\mathrm{KL}}(\delta_t \| r) = \sum_{j=1}^{V} \delta_t(x_t^j) \log \frac{\delta_t(x_t^j)}{r(x_t^j)}. \tag{11}$$

Under the one-hot target equation 10 this reduces to the familiar form

$$\mathcal{L}(r) = -\log r(x_t^*), \qquad \mathcal{L}(X) = -\log r(X) = \sum_{t=1}^{L} \big(-\log r(x_t \mid h_t) \big),$$

i.e., the negative log-likelihood of the target. We keep the general form equation 11 for now and instantiate r with each sharpening operator below; how this loss responds to sharpening, and how it decomposes, is the subject of the following sections.

Temperature: local sharpening. The temperature-T distribution rescales the logits:

$$p_T(x_t^j \mid h_t) = \frac{\exp\left(z_{x_t^j}(h_t)/T\right)}{\sum_{j'=1}^{V} \exp\left(z_{x_t^{j'}}(h_t)/T\right)}. \tag{12}$$

Since the base probability satisfies $p(x_t^j \mid h_t) \propto \exp z_{x_t^j}(h_t)$, temperature is exactly a power of the base distribution, taken *one step at a time*:

$$p_T(x_t^j \mid h_t) = \frac{p(x_t^j \mid h_t)^{1/T}}{\sum_{j'=1}^{V} p(x_t^{j'} \mid h_t)^{1/T}}, \qquad \text{i.e.} \qquad p_T(\cdot \mid h_t) \propto p(\cdot \mid h_t)^{1/T}. \tag{13}$$

Equation equation 13 makes precise what "local" means: for $T < 1$, each single-step distribution is raised to the power $1/T > 1$ and renormalized *independently of every other step*. The sequence-level distribution inherits this stepwise construction, $p_T(X) = \prod_t p_T(x_t \mid h_t) \propto \prod_t p(x_t \mid h_t)^{1/T}$: sharpening happens at every position separately, and no step's mass is informed by the quality of the continuations it leads to.

Power: global sharpening. The power distribution instead raises the *complete sequence* probability to a power $\alpha > 1$:

$$p_\alpha(X) = \frac{p(X)^\alpha}{Z_\alpha}, \qquad Z_\alpha = \sum_X p(X)^\alpha. \tag{14}$$

Marginalizing equation 14 over all continuations shows what a single step of this sampler amounts to. The exact conditional distribution is

$$p_{\mathrm{pow}}(x_t^j \mid x_{<t}) = \frac{\sum_{x_{>t}} p\big(x_1, \ldots, x_t^j, \ldots, x_L\big)^\alpha}{\sum_{j'=1}^{V} \sum_{x_{>t}} p\big(x_1, \ldots, x_t^{j'}, \ldots, x_L\big)^\alpha} \;\propto\; \sum_{x_{>t}} p\big(x_1, \ldots, x_t^j, \ldots, x_L\big)^\alpha, \tag{15}$$

i.e., each token is promoted in proportion to the total α-power mass of the *complete* sequences it can lead to. Sharpening is now *global*: a token that is locally dispreferred but heads toward sequences the model is confident in receives more mass than a locally preferred token that heads toward sequences the model is unsure of. This is the mechanism by which power sampling overturns wrong token-level preferences and repairs forks into locks—at the cost that exact sampling from equation 14 requires lookahead over continuations.

Selection: the Best-of-N distribution. The third operator does not reweight the base distribution directly; it *selects* from it. Let S be any score ranking complete sequences (a likelihood-based score, a verifier, or any other signal), and let q be the *proposal*: any distribution over complete sequences that the model induces, written $q := p_{\mathrm{LLM}}$. Best-of-N draws $X_1, \ldots, X_N \sim q$ i.i.d. and returns the champion

$$\hat{X} = \arg \max_{i \in \{1,\ldots,N\}} S(X_i).$$

For a candidate sequence X with proposal mass $q_X = q(X)$, let

$$W_X = \sum_{X' :\, S(X') > S(X)} q(X') \tag{16}$$

be the *winning mass* against X: the probability that a single draw outscores it. The output distribution of Best-of-N is then

$$p_{\mathrm{bon}}^{(N)}(X) = (1 - W_X)^N - (1 - W_X - q_X)^N = \sum_{k=1}^{N} \binom{N}{k} q_X^k \, (1 - W_X - q_X)^{N-k}. \tag{17}$$

Proposition B.2 (Best-of-N output distribution)**.** *Equation equation 17 holds for every candidate* X.

Proof. The event "X is returned" is the intersection of "no draw lands in $\{X' : S(X') > S(X)\}$," of probability $(1 - W_X)^N$, with "X appears at least once." Conditioning on the number k of draws equal to X gives probability $\binom{N}{k} q_X^k (1 - W_X - q_X)^{N-k}$; summing over $k = 1, \ldots, N$ yields the third form of equation 17, and the binomial theorem collapses it to the first. □

Equation equation 17 already displays the structure that the rest of the theory exploits: W_X sees only the score's *ranking* (the sharpening *direction*), while q_X sees only the proposal's *sampling mass* (the raw material selection has to work with).

Joint versus average log-likelihood. The joint log-likelihood of a sequence X is the sum accumulated in equation 8,

$$\ell(X) = \log p(X) = \sum_{t=1}^{L} \log p(x_t \mid h_t), \tag{18}$$

and its length-normalized *average log-likelihood* is

$$S^{\text{avg}}(X) = \frac{\ell(X)}{L(X)}. \tag{19}$$

The two are related by the exact identity

$$\ell(X) = L(X) \cdot S^{\text{avg}}(X), \tag{20}$$

which exposes the structural role of the length L: the joint log-likelihood compounds per-step confidence additively along the sequence, so two sequences with identical per-token confidence receive joint scores differing by a factor proportional to their lengths, and longer sequences are systematically penalized regardless of how confident each of their steps is. Dividing by L removes exactly this multiplicative length factor, leaving a score that reflects confidence per step rather than the number of steps.

Marginal-aware rescaling (concept). MAR reshapes the proposal q at inference time, before selection. Writing the logits in max-shifted form $z_{x_t^j} \leftarrow z_{x_t^j} - M_t \leq 0$, MAR computes the expected logit at a boundary temperature T_b,

$$\mathbb{E}_t(T_b) = \mathbb{E}_{p_{T_b}(\cdot \mid h_t)}[z] = \sum_{j=1}^{V} p_{T_b}(x_t^j \mid h_t)\, z_{x_t^j}, \tag{21}$$

and rescales every logit by a factor equal to its own depth below this expectation:

$$\tilde{z}_{x_t^j} = \lambda_{x_t^j}\, z_{x_t^j}, \qquad \lambda_{x_t^j} = \frac{z_{x_t^j}}{\mathbb{E}_t(T_b)}. \tag{22}$$

Conceptually, the expected logit acts as a per-step *boundary* between tokens worth exploring and tokens that only dilute the budget: each token's rescaling factor is proportional to how far it sits below this boundary, so logits near the peak are compressed toward equal footing while the deep tail is pushed down further. The proposal is then the tempered distribution applied to the rescaled logits, $q = p_{T_g;\, T_b}$. Why the expectation is taken at $T_b = 3$, and why the depth-proportional factor takes exactly the form equation 22, is developed in Appendix B.4.

The operators introduced above are thus comparable on a single axis—how, and how fast, they move mass toward the candidates a criterion prefers—and the following sections analyze this common structure.

B.2 The preference-dispersion loss decomposition

All losses in this paper derive from the one-hot loss of Appendix B.1: for a target x_t^* and a candidate distribution r, $\mathcal{L}(r) = -\log r(x_t^*)$ at the step level and $\mathcal{L}(X) = -\log r(X)$ at the sequence level. This section instantiates r with each sharpening operator of Appendix B.1 and shows that every one of them admits a decomposition of the same form,

$$\mathcal{L} = \underbrace{(\text{preference})}_{\text{how far the target sits below the operator's favorite candidate}} + \underbrace{(\text{dispersion})}_{\text{how much mass the operator spreads over non-favorites}},$$

which differ across operators only in *what counts as favorite* and *how the spread is measured*. These decompositions are the workhorse of the theory: the lock–fork tension, the rate bottleneck of Best-of-N, and the design of MAR are all statements about which term each operator can and cannot control.

The base model. Instantiating the step-level loss with $r = p$ (Eq. equation 7) and adding and subtracting the step maximum M_t in the log gives

$$\mathcal{L}_t = -\log p(x_t^* \mid h_t) = \underbrace{\big(M_t - z_{x_t^*}(h_t)\big)}_{\text{preference loss: } \Delta_t^*} + \underbrace{\log \sum_{j=1}^{V} \exp\big(z_{x_t^j}(h_t) - M_t\big)}_{\text{dispersion loss: } \varepsilon_t}, \tag{23}$$

where the two terms are exactly the preference gap Δ_t^* of Definition B.1 and the log-sum-exp remainder $\varepsilon_t \geq 0$ around the peak. The decomposition equation 23 is exact and interpretable term by term: the preference loss is zero precisely at a lock ($x_t^* = m_t$) and grows with the gap at a fork; the dispersion loss is zero only for a point-mass distribution and grows with the spread of the conditional distribution, reaching its maximum $\log V$ at the uniform distribution. Summed over a target path,

$$\mathcal{L}(X^*) = \sum_{t=1}^{L} \big(\Delta_t^* + \varepsilon_t\big). \tag{24}$$

Temperature. Instantiating with $r = p_T$ (Eq. equation 12) gives the same decomposition with both terms temperature-dependent:

$$\mathcal{L}_t(T) = -\log p_T(x_t^* \mid h_t) = \underbrace{\frac{\Delta_t^*}{T}}_{\text{preference loss}} + \underbrace{\log\Big(1 + \sum_{j:\, x_t^j \neq m_t} \exp \frac{z_{x_t^j}(h_t) - M_t}{T}\Big)}_{\text{dispersion loss: } \varepsilon_{T,t}}. \tag{25}$$

The two terms respond to temperature in opposite directions:

$$\frac{d}{dT}\,\frac{\Delta_t^*}{T} = -\frac{\Delta_t^*}{T^2} \leq 0, \qquad \frac{d\,\varepsilon_{T,t}}{dT} = \frac{M_t - \mathbb{E}_t(T)}{T^2} \geq 0, \tag{26}$$

where $\mathbb{E}_t(T) = \sum_j p_T(x_t^j \mid h_t)\, z_{x_t^j}(h_t)$ is the expected logit of Eq. equation 21, which lies strictly below the maximum M_t on non-degenerate steps. Qualitatively: **lowering T suppresses dispersion but amplifies the preference penalty; raising T dissolves the preference penalty but lets dispersion grow toward** $\log V$. The asymmetry between the two regimes is the lock–fork tension in loss form: at a *lock* ($\Delta_t^* = 0$) cooling is unconditionally good—the preference term vanishes and only dispersion is suppressed—while at a *fork* ($\Delta_t^* > 0$) cooling concentrates mass on the wrong peak and makes the target exponentially less likely. No single temperature is optimal at both position types; Section 3.3 of the main text resolves this tension not by tuning T but by reshaping the logits themselves.

Power. At the sequence level, instantiating with $r = p_\alpha$ (Eq. equation 14) gives

$$\mathcal{L}(p_\alpha) = -\log p_\alpha(X^*) = \alpha\,\big(-\log p(X^*)\big) + \log Z_\alpha = \alpha \sum_{t=1}^{L} \big(\Delta_t^* + \varepsilon_t\big) + \log Z_\alpha, \tag{27}$$

using equation 24. Two observations put power and temperature on the same footing. First, at the level of sequence probabilities the two operators are literally the same family: from Eq. equation 13,

$$p_T(X) = \prod_t p_T(x_t \mid h_t) \propto \prod_t p(x_t \mid h_t)^{1/T} = p(X)^{1/T}, \tag{28}$$

so temperature *is* the power distribution with $\alpha = 1/T$; what distinguishes them is not the exponent but the *scope of normalization*—temperature normalizes each step separately (the product of per-step Z_t in Eq. equation 13), while power normalizes once globally through Z_α. Second, this difference shows up precisely in the dispersion term: power scales the *entire* base loss by α (both

Δ_t^* and ε_t), whereas temperature scales the preference term by exactly $1/T$ but replaces ε_t by the renormalized $\varepsilon_{T,t} \neq \varepsilon_t/T$—per-step renormalization suppresses dispersion more aggressively than a global power of the same exponent.

At the token level the one-hot loss under power sampling is $-\log p_{\mathrm{pow}}(x_t^* \mid x_{<t})$ with p_{pow} given by Eq. equation 15. To decompose it, define the *continuation masses*

$$M_j \;=\; \sum_{x_{>t}} p\big(x_1, \ldots, x_t^j, \ldots, x_L\big)^\alpha, \qquad \tilde{m}_t = \arg\max_{1 \le j \le V} M_j, \tag{29}$$

i.e. M_j is the total α-power mass that the complete sequences passing through candidate x_t^j contribute to the tilted distribution. Adding and subtracting $\log M_{\tilde{m}_t}$ gives the decomposition.

Proposition B.3 (Token-level power decomposition)**.** *The one-hot loss of power sampling at step t decomposes exactly as*

$$-\log p_{\mathrm{pow}}(x_t^* \mid x_{<t}) = \underbrace{\log \frac{M_{\tilde{m}_t}}{M_*}}_{\text{preference loss: } \tilde{\Delta}_t^*} + \underbrace{\log \frac{\sum_{j=1}^V M_j}{M_{\tilde{m}_t}}}_{\text{dispersion loss: } \tilde{\varepsilon}_t}, \tag{30}$$

where M_* is the continuation mass of the target x_t^*, $\tilde{\Delta}_t^* \ge 0$, and $\tilde{\varepsilon}_t \ge 0$.

Proof. From Eq. equation 15,

$$-\log p_{\mathrm{pow}}(x_t^* \mid x_{<t}) = \log \sum_{j=1}^V M_j - \log M_* = \Big(\log \sum_{j=1}^V M_j - \log M_{\tilde{m}_t}\Big) + \Big(\log M_{\tilde{m}_t} - \log M_*\Big),$$

which is equation 30; non-negativity of both terms follows from $M_{\tilde{m}_t} = \max_j M_j$. □

The analogy with equation 25 is exact in structure: the preference loss is the log-gap from the target to the operator's favorite candidate, except that "favorite" is now decided by continuation mass (the total confidence of the futures a token leads to) rather than by the local logit; the dispersion loss is again the log-sum-exp remainder, now measured over continuation masses. Two consistency checks close the loop. If step t has no continuation ($t = L$), then $M_j = p(x_t^j \mid h_t)^\alpha$ and equation 30 collapses to $\alpha\,(\Delta_t^* + \varepsilon_t)$—the temperature loss equation 25 at $T = 1/\alpha$, as required by equation 28. And if the target has the largest continuation mass ($M_* = M_{\tilde{m}_t}$, a *global lock*), the preference term vanishes and only dispersion remains.

Best-of-N. Selection acts on sequences, so the relevant loss is the sequence-level one against the proposal q. Unlike the previous operators, the one-hot target X^* is *not* in general the operator's favorite: let $\bar{X}$ be the score's maximizer. The ranking term and the sampling term must therefore be kept apart from the start. Starting from Eq. equation 17 and factoring out $(1 - W_{X^*})^N$ gives the exact identity

$$p_{\mathrm{bon}}^{(N)}(X^*) = (1 - W_{X^*})^N \cdot \Big[1 - \big(1 - \tilde{q}_{X^*}\big)^N\Big], \qquad \tilde{q}_{X^*} := \frac{q_{X^*}}{1 - W_{X^*}}, \tag{31}$$

where $\tilde{q}_{X^*}$ is the *conditional hit probability*: the share of proposal mass that the target holds among all sequences the score does *not* rank above it. Applying the geometric-series factorization $1 - (1 - \tilde{q})^N = \tilde{q} \cdot A_N(\tilde{q})$ with

$$A_N(\tilde{q}) = \sum_{k=0}^{N-1} (1 - \tilde{q})^k = \frac{1 - (1 - \tilde{q})^N}{\tilde{q}}$$

to the bracket of equation 31 yields the decomposition.

Proposition B.4 (Best-of-N loss decomposition)**.** *For any target X^* with $W_{X^*} < 1$,*

$$\mathcal{L}_{\mathrm{bon}}(X^*) = \underbrace{\big(-N \log(1 - W_{X^*})\big)}_{\text{preference loss}} + \underbrace{\Big(-\log\big(1 - (1 - \tilde{q}_{X^*})^N\big)\Big)}_{\text{dispersion loss}}. \tag{32}$$

The decomposition has exactly the shape of equation 23–equation 30: the preference loss vanishes precisely when the target *is* the operator's favorite ($W_{X^*} = 0$, nothing ranked above it), and the dispersion loss is nonnegative and nonincreasing in the budget N, tending to zero as $N \to \infty$—more draws concentrate the output on the score's favorite, whatever that favorite is. For interpretation, the dispersion term further splits as

$$-\log\big(1-(1-\tilde{q}_{X^*})^N\big) = \underbrace{\big(-\log \tilde{q}_{X^*}\big)}_{\text{proposal loss}} + \underbrace{\big(-\log A_N(\tilde{q}_{X^*})\big)}_{\text{selection gain}}, \qquad \tilde{q}_{X^*}\cdot A_N(\tilde{q}_{X^*}) = 1-(1-\tilde{q}_{X^*})^N,$$

i.e. the proposal loss enters intact—selection does not reweight the proposal, it only draws from it—diminished by a nonpositive gain $-\log A_N \le 0$, $A_N \in [1, N]$, in the same structural role that $\log Z_\alpha$ plays for power (Eq. equation 27).

The preference loss behaves unlike any preference term seen so far. Since $-\log(1-W) \ge W$,

$$-N\log(1-W_{X^*}) \;\ge\; NW_{X^*}, \qquad -N\log(1-W_{X^*}) \xrightarrow[N\to\infty]{} \infty \quad (W_{X^*} > 0), \tag{33}$$

i.e. it grows linearly in N and diverges whenever *any* proposal mass sits on a sequence the score ranks above the target. This is the quantitative form of a qualitative fact: BoN's output collapses onto $\bar{X}$, so if $\bar{X} \neq X^*$ more sampling budget only pushes the output further from the target. Selection is only as good as the score's alignment with the one-hot target, and N amplifies the misalignment—sharpening direction is the score's responsibility, not the budget's.

The dispersion loss has the two classical regimes of coverage:

$$-\log A_N(\tilde{q}) \approx -\log N \quad (N\tilde{q} \ll 1), \qquad -\log A_N(\tilde{q}) \to -\log \tfrac{1}{\tilde{q}} \;(N \to \infty),$$

i.e. while the target is unlikely to be drawn at all, each draw contributes its full $\log N$ of budget gain; once $N\tilde{q} \gtrsim 1$ the gain saturates at $1/\tilde{q}$, exactly cancelling the proposal loss, as it must since the dispersion loss itself tends to zero. Two consistency checks: if the score ranks the target on top ($W_{X^*} = 0$, e.g. the score is the joint log-likelihood and the target is its maximizer), the preference loss vanishes and equation 32 reduces to the pure dispersion form $\mathcal{L}_{\text{bon}} = -\log\big(1-(1-q_{X^*})^N\big)$, i.e. more budget monotonically improves the loss; and expanding A_N to first order in $\tilde{q}$ recovers the classical $N\tilde{q}\,(1-\tilde{q})^{N-1}$ shape up to the exact duplicate-draw correction.

A token-level marginal of BoN exists in the same sense as for power:

$$p^{(N)}_{\text{bon}}\big(x^j_t \mid x_{<t}\big) \;:=\; \sum_{x_{>t}} p^{(N)}_{\text{bon}}\big(x_1, \ldots, x^j_t, \ldots, x_L\big), \tag{34}$$

summing the output sequence distribution over continuations through x^j_t. Unlike power, however, $p^{(N)}_{\text{bon}}(X)$ is not a product of per-step factors—the score couples every position of a sequence—so equation 34 has no closed form in general and no token-level decomposition analogous to Proposition B.3: the preference–dispersion structure of BoN is intrinsically sequence-level. This is exactly the sense in which BoN sharpens *globally* like power rather than locally like temperature.

Marginal-aware rescaling. MAR is not a sixth distribution to decompose from scratch: it is a reshaping of the proposal q that selection then draws from, so its loss has the temperature form equation 25 computed on the rescaled logits $\tilde{z}_{x^j_t} = z^2_{x^j_t}/\mathbb{E}_t(T_b)$ of Eq. equation 22:

$$\mathcal{L}^{\text{MAR}}_t = -\log q(x^*_t \mid h_t) = \underbrace{\frac{\hat{\Delta}^*_t}{T_g}}_{\text{preference loss}} + \underbrace{\hat{\varepsilon}_{T_g,t}}_{\text{dispersion loss}}, \qquad \hat{\Delta}^*_t = \frac{(\Delta^*_t)^2}{-\mathbb{E}_t(T_b)}, \tag{35}$$

where the second equality uses max-shifted logits ($M_t = 0$, $z_{x^*_t} = -\Delta^*_t$) and $-\mathbb{E}_t(T_b) > 0$. The contrast with temperature is now quantifiable. Temperature divides *both* terms by a single global factor $1/T$; MAR multiplies the preference gap by the per-pair factor

$$\kappa^*_t = \frac{\hat{\Delta}^*_t}{\Delta^*_t} = \frac{\Delta^*_t}{-\mathbb{E}_t(T_b)}, \qquad \kappa^*_t \to 0 \;(\Delta^*_t \ll |\mathbb{E}_t(T_b)|), \quad \kappa^*_t \gg 1 \;(\Delta^*_t \gg |\mathbb{E}_t(T_b)|), \tag{36}$$

so the same boundary $\mathbb{E}_t(T_b)$ that diagnoses the temperature response in equation 26 now *acts* on it: preference gaps of mild forks, which sit near the boundary, are compressed toward zero

($\kappa_t^* \ll 1$), healing the very penalty that cooling would amplify; preference gaps deep in the tail, which correspond to distractors at locks, are expanded ($\kappa_t^* \gg 1$), suppressing dispersion-adding tail mass that heating would release.

The comparison carries over to selection through Eq. equation 32: BoN under a temperature proposal and BoN under a MAR proposal share the identical two-term loss

$$\mathcal{L}_{\text{bon}}(X^*) = -N \log(1 - W_{X^*}) \; - \; \log\big(1 - (1 - \tilde{q}_{X^*})^N\big),$$

and the two terms separate cleanly by what MAR can and cannot touch. The *preference* loss is governed by the score: it vanishes only if no proposal mass sits on a sequence ranked above the target, and it grows with N otherwise; a proposal reshaping cannot repair a misaligned score—this term is the score's responsibility alone. The *dispersion* loss is where the proposal acts. Under a temperature proposal, each mild fork on the path contributes Δ_t^*/T_g to $-\log q_{X^*}$ and the hit probability decays multiplicatively with the number of forks,

$$q_{X^*}^{\text{temp}} = \exp\Big(-\sum_t \frac{\Delta_t^*}{T_g}\Big) \cdot \prod_t p_{T_g}(m_t \mid h_t),$$

exactly, since $\varepsilon_{T_g,t} = -\log p_{T_g}(m_t \mid h_t)$, so the dispersion loss sits in its unsaturated regime $-\log A_N \approx -\log N$, where the budget gain is the only lever and the target path is never actually covered—the rate bottleneck, quantified in the next section. Under a MAR proposal, the near-boundary compression ($\hat{\Delta}_t^* \ll \Delta_t^*$ at mild forks) raises the per-step target mass and with it both q_{X^*} and $\tilde{q}_{X^*}$: the proposal component of the dispersion loss shrinks, and the selection gain $-\log A_N$ saturates at its $1/\tilde{q}_{X^*}$ ceiling at a smaller budget, driving the dispersion loss down faster in N. If the compression also moves mass off sequences the score ranks above the target, W_{X^*} falls as a side effect, but the decomposition makes the division of labor explicit: MAR attacks the dispersion term—the rate—without any claim on the preference term—the direction.

B.3 Sharpening strength versus loss: one trade-off, three parameters

Each operator of Appendix B.1 is governed by a single scalar *sharpening strength*: the inverse temperature $\beta := 1/T$, the power exponent α, or the sampling budget N. This section differentiates the decompositions of Appendix B.2 with respect to that parameter. The outcome is uniform across operators. The derivative of the loss is always a sum of two terms with opposite signs,

$$\frac{d\mathcal{L}}{d\theta} = \underbrace{(\text{preference derivative})}_{\text{nonnegative, constant in } \theta} + \underbrace{(\text{dispersion derivative})}_{\text{nonpositive, increasing in } \theta},$$

the dispersion relief per unit of sharpening being largest when the distribution is diffuse and vanishing as the output collapses onto the operator's favorite. Two consequences follow in all three cases: the loss is strictly convex in the sharpening strength, so the trade-off has at most one interior resolution; and that resolution is pushed to infinity at a *lock* (where the preference derivative vanishes or is absorbed by the dispersion floor) while it is pinned to a finite value at a *fork* (where continued sharpening past θ^* makes the loss worse, not better).

Temperature: the optimal strength solves an expectation identity. Write the step loss equation 25 in terms of the sharpening strength $\beta = 1/T$:

$$\mathcal{L}_t(\beta) = \beta\, \Delta_t^* + \log \sum_{j=1}^{V} \exp\big(\beta\, (z_{x_t^j}(h_t) - M_t)\big). \tag{37}$$

Differentiating,

$$\frac{d\,\mathcal{L}_t(\beta)}{d\beta} = \underbrace{\Delta_t^*}_{\text{preference derivative} \geq 0} + \underbrace{\Big(\mathbb{E}_t(\beta) - M_t\Big)}_{\text{dispersion derivative} \leq 0}, \qquad \frac{d^2\,\mathcal{L}_t(\beta)}{d\beta^2} = \text{Var}_t(\beta) > 0, \tag{38}$$

where $\mathbb{E}_t(\beta) = \sum_j p_{1/\beta}(x_t^j \mid h_t)\, z_{x_t^j}(h_t)$ is the expected logit under the sharpened distribution—the same object as Eq. equation 21, evaluated at $T_b = 1/\beta$—and $\text{Var}_t(\beta)$ is the corresponding logit variance. The step loss is strictly convex in β; the dispersion derivative increases strictly from $\bar{z}_t - M_t < 0$ (at $\beta \to 0$, with $\bar{z}_t$ the mean logit) to 0 (at $\beta \to \infty$), while the preference derivative stays constant at Δ_t^*. Three regimes:

- *Lock* ($\Delta_t^* = 0$): the derivative is strictly negative for every finite β and tends to 0 only as $\beta \to \infty$. The optimal strength is $\beta^* = \infty$ ($T \to 0$): sharpening never stops helping.
- *Mild fork* ($\bar{z}_t < z_{x_t^*} < M_t$): the derivative changes sign exactly once, and the unique minimizer β_t^* is the solution of
$$\mathbb{E}_t(\beta_t^*) = z_{x_t^*}(h_t) : \tag{39}$$
sharpen until the sharpened distribution's expected logit is pulled down to the target's logit. Comparative statics: implicit differentiation of equation 39 gives $d\beta_t^*/d\Delta_t^* = -1/\operatorname{Var}_t(\beta_t^*) < 0$— the deeper the misranking, the *less* sharpening the step tolerates.
- *Severe fork* ($z_{x_t^*} \le \bar{z}_t$): the derivative is nonnegative for every β and the optimum is $\beta^* = 0$: the loss-minimizing response to a badly misranked step is not to sharpen at all.

Power: the tilted entropy against the target's log-deficiency. At the sequence level, differentiating equation 27 gives

$$\frac{d\,\mathcal{L}(p_\alpha)}{d\alpha} = \underbrace{-\ell(X^*)}_{\text{preference derivative} \ge 0} + \underbrace{\mathbb{E}_{p_\alpha}[\log p(X)]}_{\text{dispersion derivative} \le 0}, \qquad \frac{d^2\,\mathcal{L}(p_\alpha)}{d\alpha^2} = \operatorname{Var}_{p_\alpha}[\log p(X)] > 0, \tag{40}$$

since $d \log Z_\alpha / d\alpha = \mathbb{E}_{p_\alpha}[\log p(X)]$. The dispersion derivative increases from $\frac{1}{K}\sum_X \log p(X) = \log K - \mathcal{H}(p)$ (at $\alpha \to 0$, K the number of sequences with positive base probability) to $-\ell_{\min}$ (at $\alpha \to \infty$, where $\ell_{\min} = \min_X \ell(X)$ is the log-deficiency of the base distribution's mode), and the loss is strictly convex in α. The regimes mirror temperature:

- *Lock* (X^* is the base mode, $\ell(X^*) = \ell_{\min}$): the derivative equals $\ell_{\min} + \mathbb{E}_{p_\alpha}[\log p(X)] < 0$ for every finite α, since $\mathbb{E}_{p_\alpha}[\log p(X)] < \log p_{\max} = -\ell_{\min}$; it tends to 0 as $\alpha \to \infty$, so $\alpha^* = \infty$.
- *Mild fork* ($\ell_{\min} < \ell(X^*) < \log K - \mathcal{H}(p)$): the derivative crosses zero once, and the unique minimizer α^* solves
$$\mathbb{E}_{p_{\alpha^*}}[\log p(X)] = -\ell(X^*), \tag{41}$$
i.e. sharpen until the tilted distribution's expected log-deficiency equals the target's own. The threshold $\log K - \mathcal{H}(p)$ is the log-deficiency of the *average* sequence (drawn uniformly from the support): an interior optimum exists exactly when the target is worse than the mode but better than average.
- *Severe fork* ($\ell(X^*) \ge \log K - \mathcal{H}(p)$): the optimum is $\alpha^* = 0$—global sharpening is harmful.

Best-of-N: the optimal budget is finite. Treating the budget as a continuous parameter and differentiating equation 32,

$$\frac{d\,\mathcal{L}_{\text{bon}}(X^*)}{dN} = \underbrace{\big(-\log(1 - W_{X^*})\big)}_{\text{preference derivative} \ge 0} + \underbrace{\frac{(1-\tilde{q}_{X^*})^N \log(1-\tilde{q}_{X^*})}{1-(1-\tilde{q}_{X^*})^N}}_{\text{dispersion derivative} \le 0}, \qquad \frac{d^2\,\mathcal{L}_{\text{bon}}}{dN^2} > 0, \tag{42}$$

the dispersion derivative increasing from $-\infty$ (at $N \to 0$) to 0 (at $N \to \infty$). The preference derivative is zero exactly at a lock ($W_{X^*} = 0$) and strictly positive otherwise; convexity then gives the complete picture.

Proposition B.5 (Optimal budget of Best-of-N)**.**
(i) If $W_{X^*} = 0$ (lock), $\mathcal{L}_{\text{bon}}(X^*)$ is strictly decreasing in N and the optimum is $N^* = \infty$.
(ii) If $W_{X^*} > 0$ (fork), there is a unique finite optimal budget N^* characterized by

$$-\log(1 - W_{X^*}) = \frac{(1-\tilde{q}_{X^*})^{N^*}\,\big[-\log(1-\tilde{q}_{X^*})\big]}{1-(1-\tilde{q}_{X^*})^{N^*}}, \tag{43}$$

at which the marginal dispersion relief equals the marginal preference damage; N^* is strictly decreasing in W_{X^*}.
(iii) For $\tilde{q}_{X^*} \ll 1$,

$$N^* \;\approx\; \frac{1}{\tilde{q}_{X^*}} \log\Big(1 + \frac{\tilde{q}_{X^*}}{W_{X^*}}\Big), \qquad \textit{so} \qquad N^* \tilde{q}_{X^*} \approx \log \frac{\tilde{q}_{X^*}}{W_{X^*}} \;\; (W_{X^*} \ll \tilde{q}_{X^*}) : \tag{44}$$

the optimal coverage grows only logarithmically with the inverse winning mass.

Proof. (i) and the existence and uniqueness in (ii) follow from equation 42: the total derivative increases from $-\infty$ to $-\log(1 - W_{X^*})$, crossing zero exactly once when $W_{X^*} > 0$. Implicit differentiation of equation 43 gives $dN^*/dW_{X^*} = -\left[1/(1 - W_{X^*})\right]/\left[d^2\mathcal{L}_{\text{bon}}/dN^2\right]_{N^*} < 0$. For (iii), put $u = N\tilde{q}_{X^*}$: with $-\log(1 - \tilde{q}_{X^*}) \approx \tilde{q}_{X^*}$ and $(1 - \tilde{q}_{X^*})^N \approx e^{-u}$, the right-hand side of equation 43 is $\tilde{q}_{X^*}e^{-u}/(1 - e^{-u})$, and $-\log(1 - W_{X^*}) \approx W_{X^*}$; solving for u gives $e^u = 1 + \tilde{q}_{X^*}/W_{X^*}$. □

One trade-off, and why a scalar cannot win it. The three operators, their parameters, and their optima are one and the same statement at three granularities: the loss is strictly convex in the sharpening strength; the preference derivative is constant and the dispersion derivative increases toward its floor, so the two reconcile at most once; the deeper the misranking, the smaller the optimal strength; and at a lock the optimum is not finite. The practical obstacle is that these optima are declared *per step*, and a reasoning sequence is heterogeneous by construction: along a single target path, locks demand $\beta_t^* = \infty$, mild forks demand moderate finite β_t^* solving equation 39, and severe forks demand $\beta_t^* = 0$. A single scalar β shared by all steps must minimize the path-averaged loss, whose first-order condition,

$$\sum_{t=1}^{L} \mathbb{E}_t(\beta) = \sum_{t=1}^{L} z_{x_t^*}(h_t), \tag{45}$$

forces the sharpened average expectation to equal the average target logit: every lock is under-sharpened, every fork is over-sharpened, and the violation of each step's own equation 39 grows with the heterogeneity of the path. Power and Best-of-N face the same restriction at sequence level—one global α, one global N—with the same consequence: a scalar parameter distributes one compromise over steps whose requirements are incompatible. Appendix B.4 removes the scalar restriction, reshaping the logits step by step instead of tuning a single strength.

B.4 Why MAR protects locks and activates forks

Appendix B.3 ended with a structural impossibility: a single scalar sharpening strength must compromise among per-step optima that are mutually incompatible—locks demand infinite sharpening, mild forks a moderate finite strength, and badly misranked steps no sharpening at all. This section develops the operator that removes the scalar restriction, and explains why it acts in *opposite* directions at the two position types: it *protects locks* and *activates forks*.

Two design requirements. A reshaping of the logits $\{z_x\} \mapsto \{\tilde{z}_x\}$ removes the compromise only if it meets both of the following, which no scalar parameter can meet simultaneously (Appendix B.3):

- *Protect locks* ($\Delta_t^* = 0$): suppress dispersion without limit. The deep-tail distractors must be pushed down arbitrarily far as their gaps grow, while the peak—the target itself—stays fixed; and no preference gap may be amplified (there is none to amplify).
- *Activate forks* ($\Delta_t^* > 0$, mild): shrink, not amplify, the preference gap of viable competitors, pulling near-peak candidates toward equal footing. This is the exact opposite of what temperature does: cooling multiplies *every* gap by $1/T \geq 1$, crushing precisely the candidates a fork needs.

Requirement one asks for sharpening that intensifies with depth; requirement two asks for compression that vanishes at the peak. Both are met by one and the same rule: rescale each logit in proportion to its own depth below a per-step boundary. With max-shifted logits $z_x \leq 0$ and the boundary expectation $\mathbb{E}_t(T_b) < 0$ of Eq. equation 22, the transform is

$$\tilde{z}_x = \lambda_x\, z_x, \qquad \lambda_x = \frac{z_x}{\mathbb{E}_t(T_b)}, \qquad \text{i.e.} \qquad \tilde{z}_x = \frac{z_x^2}{\mathbb{E}_t(T_b)},$$

and the proposal is the tempered distribution $q = p_{T_g;\,T_b}$ applied to the rescaled logits. The rest of this section derives what the rule does and why it must take this form.

The pairwise gap law. The action of the transform on gaps, as opposed to logits, is what the loss theory consumes.

Proposition B.6 (Pairwise gap law)**.** *For any pair of tokens with $z_j \le z_i \le 0$ and gap $\Delta_{ij} = z_i - z_j \ge 0$, the transformed gap $\hat{\Delta}_{ij} = \tilde{z}_i - \tilde{z}_j$ satisfies*

$$\hat{\Delta}_{ij} = \kappa_{ij}\,\Delta_{ij}, \qquad \kappa_{ij} = \frac{z_i + z_j}{\mathbb{E}_t(T_b)} \ge 0. \tag{46}$$

In particular, for the target gap Δ_t^* against the peak ($z = 0$),

$$\hat{\Delta}_t^* = \kappa_t^*\,\Delta_t^*, \qquad \kappa_t^* = \frac{\Delta_t^*}{-\mathbb{E}_t(T_b)}. \tag{47}$$

Proof. Both statements are the same algebra: $\tilde{z}_i - \tilde{z}_j = (z_i^2 - z_j^2)/\mathbb{E}_t(T_b) = (z_i + z_j)(z_i - z_j)/\mathbb{E}_t(T_b)$, with $\mathbb{E}_t(T_b) < 0$ and $z_i + z_j \le 0$ so that $\kappa_{ij} \ge 0$; for the target pair, $z_i + z_j = -\Delta_t^*$. □

Three structural properties.

Proposition B.7 (Order preservation)**.** *$\tilde{z}_x$ is strictly increasing in z_x on $(-\infty, 0]$: MAR preserves the model's ranking exactly, and therefore never flips a preference—it reweights the* strength *of preferences only.*

Proof. On $(-\infty, 0]$, $z \mapsto z^2$ is strictly decreasing, and division by the negative constant $\mathbb{E}_t(T_b)$ reverses the order once more. □

Proposition B.8 (The boundary separates mild from severe)**.** *For the target gap, $\kappa_t^* < 1$ if and only if the target logit sits* above *the boundary, $z_{x_t^*} = -\Delta_t^* > \mathbb{E}_t(T_b)$; $\kappa_t^* > 1$ if and only if it sits below. MAR therefore operationalizes the mild/severe distinction of Definition B.1 with a single per-step quantity: a fork is mild exactly when its target lies above the boundary, and only mild forks are activated.*

Proof. Immediate from equation 47: $\kappa_t^* < 1 \iff \Delta_t^* < -\mathbb{E}_t(T_b) \iff -\Delta_t^* > \mathbb{E}_t(T_b)$. □

Proposition B.9 (MAR is not a temperature)**.** *Except for degenerate configurations (all shifted logits equal), no single temperature T' reproduces the MAR-transformed distribution.*

Proof. A temperature rescales all logits by one factor $1/T'$, hence preserves all pairwise gap *ratios*: $\Delta_{ij}/\Delta_{jk} = \Delta'_{ij}/\Delta'_{jk}$ for every triple. Under MAR, $\hat{\Delta}_{ij}/\hat{\Delta}_{jk} = \big[\kappa_{ij}/\kappa_{jk}\big]\,\big[\Delta_{ij}/\Delta_{jk}\big]$ by Proposition B.6. Equality of the ratios for every triple requires $\kappa_{ij} = \kappa_{jk}$ for every triple, i.e. $(z_i + z_j) = (z_j + z_k)$, i.e. $z_i = z_k$ for all i, k: a degenerate configuration. □

Propositions B.7- B.9 are the structural content of "one transform, two directions": at lock positions the distractor tail lies deep below the boundary and is expanded ($\kappa \gg 1$); at fork positions the viable competitors sit above the boundary and are compressed ($\kappa \ll 1$). The gap law and the three properties are structural; their effect on the proposal itself is quantified next.

What protection and activation mean in probability. The two directions admit a quantitative reading. At a lock, a distractor at gap Δ has $\kappa = \Delta/(-\mathbb{E}_t(T_b))$ against the peak, hence transformed gap

$$\hat{\Delta} = \frac{\Delta^2}{-\mathbb{E}_t(T_b)} \qquad (\Delta \gg -\mathbb{E}_t(T_b)), \tag{48}$$

so its mass under the proposal decays as $\exp\big(-\Delta^2/((-\mathbb{E}_t(T_b))\,T_g)\big)$ instead of temperature's $\exp(-\Delta/T_g)$: tail suppression is upgraded from exponential-in-gap to exponential-in-gap-squared, while the peak token sits at gap 0 and is untouched. At a mild fork, the target's proposal mass is raised by

$$\frac{q(x_t^*)}{p_{T_g}(x_t^*)} \;\approx\; \exp\Big(\frac{\Delta_t^* - \hat{\Delta}_t^*}{T_g}\Big) \;=\; \exp\Big(\frac{\Delta_t^*}{T_g}\,\big(1 - \kappa_t^*\big)\Big) \;\ge\; 1, \tag{49}$$

exact for the pair (peak, target) up to the renormalization of the remaining candidates, which the compression only reinforces. Locks are protected by super-linear tail suppression; forks are activated by a preference penalty that shrinks with the gap itself.

Why an expectation at temperature T_b. The boundary must sit where the contender set lives, and its job is to *price exploration*; it should therefore be measured under exploration conditions rather than under the base distribution's more greedy mass allocation. The expectation $\mathbb{E}_t(T_b) = \mathbb{E}_{p_{T_b}}[z]$ is the logit of the token the model would typically draw when forced to explore at T_b: tokens above it carry the softened distribution's mass and form the plausible contender set; tokens below it are tail diluters. Differentiating Eq. equation 21,

$$\frac{d\,\mathbb{E}_t(T_b)}{dT_b} = \frac{\mathrm{Var}_{p_{T_b}}[z]}{T_b^2} \geq 0, \tag{50}$$

so raising T_b lifts the boundary toward the peak, shrinking the compression region and pushing borderline tokens into the expanded regime; lowering T_b deepens the boundary and widens activation. The operator degenerates as $T_b \to 0^+$ (the expectation collapses onto the peak, all factors λ_x blow up, and the rescaled logits lose contrast), so the boundary temperature must stay bounded away from 0. Theory thus fixes the functional form and the *direction* of the T_b effect; it does not derive a specific constant. Empirically the operator is insensitive across $T_b \in [2, 3]$ and we recommend $T_b = 3$ throughout (Section 4 of the main text); the expectation is a single softmax average and adds $O(V)$ cost per decoding step.

Rate, not ceiling.

Remark B.10 (Shared asymptotic ceiling). If two proposals both assign positive mass to the score's top region, then as $N \to \infty$ their Best-of-N outputs concentrate on the same region: a proposal improvement such as MAR raises the *rate* at which the ceiling is approached, never the ceiling itself.

Indeed, for any x outside the top region, the winning mass W_x is positive under any such proposal, and Eq. equation 31 gives $p_{\mathrm{bon}}^{(N)}(x) \leq (1 - W_x)^N \to 0$. Remark B.10 delimits MAR's contract precisely: combined with Proposition B.5(iii), MAR is a repair of the *rate* at finite budgets—it raises $\tilde{q}_{X^*}$, so the optimal coverage $N^*\tilde{q}_{X^*}$ is reached with a smaller budget—while the asymptotic limit of BoN is left to the score.

Summary. The theory separates inference failure into a direction component (the preference loss, governed by the score's ranking) and a rate component (the dispersion loss of the proposal, governed by how quickly budget converts into coverage). The two repairs are correspondingly decoupled: the repaired average-likelihood score fixes the direction at zero sampling cost, since the score never enters the sampler; MAR fixes the proposal geometry that throttles Best-of-N's rate, at inference time and without touching the model's ordering (Proposition B.7). What MAR does not do is raise the ceiling (Remark B.10).

B.5 Why warming the score helps: two-phase scoring repair

Candidates are ranked by the repaired score $S(Y) = S_{T_s}^{\mathrm{avg}}(Y) = \ell_{T_s}(Y)/L(Y)$—the average log-likelihood evaluated at a scoring temperature T_s decoupled from generation. The length bias is already removed exactly, since $\ell(Y) = L(Y)\, S^{\mathrm{avg}}(Y)$ (Eq. equation 20 of Appendix B.1). What remains is how the *ranking* between two candidates responds to the scoring temperature T_s. This section shows that the response is two-phase—warming helps up to a unique interior optimum, then hurts—and identifies precisely which candidates the warming recovers.

The ranking margin. Take a correct competitor Y^+ and an incorrect one Y^-. Averaging the step decomposition equation 25 along a sequence Y gives

$$S_T(Y) = -\frac{\bar{\Delta}(Y)}{T} - \bar{\varepsilon}_T(Y), \qquad \bar{\Delta}(Y) = \frac{1}{L(Y)} \sum_{t=1}^{L(Y)} \Delta_t(y_t), \qquad \bar{\varepsilon}_T(Y) = \frac{1}{L(Y)} \sum_{t=1}^{L(Y)} \varepsilon_{T,t},$$

where $\Delta_t(y_t)$ is the preference gap of the token Y actually takes at step t (Definition B.1) and $\varepsilon_{T,t}$ is the step dispersion. The ranking margin of the correct candidate is

$$R(T) = S_T(Y^+) - S_T(Y^-) = -\frac{A}{T} - D(T), \qquad A = \bar{\Delta}(Y^+) - \bar{\Delta}(Y^-), \quad D(T) = \bar{\varepsilon}_T(Y^+) - \bar{\varepsilon}_T(Y^-). \tag{51}$$

Warming helps exactly where $R'(T) > 0$. Differentiating equation 51 and using $d\varepsilon_{T,t}/dT = (M_t - \mathbb{E}_t(T))/T^2$ (Eq. equation 26) sequence-averaged along each candidate's own path,

$$\frac{dR}{dT} = \frac{A - B(T)}{T^2}, \qquad B(T) = (\overline{M - \mathbb{E}})(Y^+; T) - (\overline{M - \mathbb{E}})(Y^-; T), \tag{52}$$

where $(\overline{M - \mathbb{E}})(Y; T) = \frac{1}{L(Y)} \sum_t \big(M_t - \mathbb{E}_t(T)\big)$. The structure is exactly that of Appendix B.3: a constant preference term A against an increasing dispersion term $B(T)$—the same trade-off, relocated from sampling to selection.

Proposition B.11 (Two-phase scoring repair)**.** *Let A and $B(T)$ be as in equation 51–equation 52, and assume (i) $A > 0$ (Y^+ carries the larger accumulated preference gaps: it is penalized at low temperature), and (ii) B crosses A at a first point T^* (sufficient: B is nondecreasing and $\sup_T B(T) > A$). Then R is strictly increasing on $(0, T^*)$ and strictly decreasing on (T^*, ∞), and T^* is the unique maximizer of the margin. The correct candidate wins at the optimum, $R(T^*) > 0$, if and only if*

$$\bar{\varepsilon}_{T^*}(Y^-) - \bar{\varepsilon}_{T^*}(Y^+) > \frac{A}{T^*}. \tag{53}$$

Moreover $dT^*/dA = 1/B'(T^*) > 0$: the more the correct candidate was penalized at low temperature, the higher the optimal scoring temperature.

Proof. Equation equation 52 gives $\operatorname{sign}(R') = \operatorname{sign}(A - B(T))$. By (i), $B(0^+) = 0 < A$ (each step satisfies $M_t - \mathbb{E}_t(T) \to 0$ as $T \to 0$), so R increases near 0—from $R(0^+) = -\infty$ when $A > 0$—and by (ii) the first crossing flips the sign from $+$ to $-$, giving the unique maximum. Evaluating equation 51 at T^* gives equation 53. Implicit differentiation of $B(T^*) = A$ yields $dT^*/dA = 1/B'(T^*) > 0$, since $B'(T^*) > 0$ at a first crossing. □

Where the gain comes from. Three remarks make the proposition operational. First, *which* candidates are recovered: assumption (i) is the mild-fork signature—a correct solution that somewhere took a locally dispreferred step accumulates $A > 0$ and is under-ranked at $T = 1$, while its incorrect competitor, having followed the model's local peaks, accumulates the smaller gaps. Warming from 1 to $T < T^*$ moves the margin by

$$R(T) - R(1) = A\Big(1 - \frac{1}{T}\Big) - \big(D(T) - D(1)\big) \;\approx\; A\Big(1 - \frac{1}{T}\Big),$$

to first order in the dispersion difference: the gain is the dissolved preference penalty, positive and growing in T until the dispersion term catches up at T^*. That the optimum is interior rather than $T_s \to \infty$ is the dispersion check of Appendix B.3 reappearing at selection level.

Second, *why the gain is free*: the score never enters the sampler, so T_s and T_g are fully decoupled. Warming the *proposal* would be punished at locks—locks demand $\beta^* = \infty$ (Appendix B.3), and heating the proposal destroys them—while warming the *score* only re-ranks the candidate pool. The direction of selection is repaired at zero sampling cost.

Third, *what warming cannot repair*: if $A \leq 0$ the low- temperature ranking already favors Y^+, or the preference is genuinely wrong, and no rescoring fixes it—the residual case left untouched in the main text. Nor does rescoring touch the rate: the score acts on Best-of-N only through the winning mass W (the direction term of Eq. equation 32), which is exactly the division of labor of Appendix B.4.

C Experimental Details

C.1 Benchmarks and grading

MATH500. MATH (Hendrycks et al., 2021) is a dataset of 12,500 competition-level mathematics problems with full step-by-step solutions. MATH500 (Lightman et al., 2024) is a 500-problem subset of the MATH test set that has become a standard test bed for reasoning; we follow the evaluation code released by Karan & Du (2026) for grading.

HumanEval. HumanEval (Chen et al., 2021) consists of 164 hand-written Python programming problems, each specified by a function signature and docstring. Generated programs are scored by program execution; we report pass@1.

GPQA Diamond. GPQA (Rein et al., 2024) is a multiple-choice benchmark of 448 questions written by domain experts in biology, physics, and chemistry. The questions are "Google-proof": PhD-level experts reach 65% accuracy while highly skilled non-expert validators reach only 34% with unrestricted web access. GPQA Diamond is the highest-quality subset retained after expert filtering; responses are scored by exact match after answer extraction.

Answer protocol. Every method delivers a single final response per problem: inference-time methods such as Best-of-N may sample several candidates internally, but only one candidate—the selected answer—is returne.

C.2 Models and RL checkpoints

We experiment on four open base models: Qwen2.5-Math-7B and Qwen2.5-7B (Yang et al., 2024b;a), Qwen3-4B (Yang et al., 2025), and Llama3-8B-Instruct (Grattafiori et al., 2024).

For RL baselines we use GRPO models post-trained on the MATH training split, reported as GRPO(MATH) in Table 2. For Qwen2.5-Math-7B and Qwen2.5-7B we use the publicly released GRPO checkpoints of Shao et al. (2024). For Qwen3-4B we use the publicly available community checkpoint **Qwen3-4B-Base-GRPO**,[1] obtained from full-parameter GRPO training on the MATH training split with a rule-based math reward (100 optimization steps, 25,600 training prompts, 8 rollouts per prompt, implemented with verl). For Llama3-8B-Instruct we use the publicly released checkpoint of Yang et al. (2026), [2] obtained from multi-agent GRPO training on math data.

C.3 Baseline configurations

Inference-time baselines. **Base**: ancestral sampling at $T = 1$ with the default generation configuration. **Low-temperature**: ancestral sampling at $T = 0.25$. **Power sampling**: MCMC with $\alpha = 4$, $B = 192$ resampling blocks, $N_{\mathrm{MCMC}} = 10$ chains, maximum output length $T_{\max} = 3072$, and proposal temperature $1/\alpha$, following Karan & Du (2026).

Number attribution. Numbers for Qwen2.5-7B and Qwen2.5-Math-7B under the Base, Low-temperature, and Power sampling settings are as reported by Karan & Du (2026); all remaining numbers, including all GRPO(MATH) and MAR results, are evaluated by us under the same protocol.

[1] https://huggingface.co/SeongryongJung/Qwen-4b-base-GRPO

[2] https://huggingface.co/logan7000/cogrpo-homo-llama31-8b-math345-groupB-llama31-8b-b-end

D VERIFIER-AGNOSTIC SELECTION WITH ALTERNATIVE SIGNALS

Motivation. The main text argues that MAR improves the *proposal* rather than exploiting a particular scoring function. To test this claim, we repeat the proposal comparison on Qwen2.5-Math-7B while replacing the repaired average-likelihood score with two alternative selection signals: answer-level consistency voting and joint likelihood selection. In all cases the proposal is the only variable: the baseline draws $N = 8$ candidates from standard temperature sampling at $T = 0.25$, while MAR draws $N = 8$ candidates from the rescaled logits with $T_g = 0.25$ and boundary temperature $T_b = 3$. The same selector is then applied to both candidate sets.

Selection signals. For *consistency voting*, we extract the final answer from each of the N candidates and return the majority answer, following the same answer-extraction protocol as the main experiments. For *joint-likelihood selection*, we return the candidate with the largest joint log-likelihood $\ell(X) = \log p(X)$, without length normalization or elevated-temperature rescoring. The latter is deliberately the unrepaired likelihood signal that the main text shows to be length-biased; it therefore provides a conservative test of whether MAR helps beyond the specific repaired score used in the main results.

Table 5: Alternative selection signals on Qwen2.5-Math-7B. All runs use $N = 8$ candidates. The temperature proposal uses $T = 0.25$; MAR uses $T_g = 0.25$ and $T_b = 3$. Within each block the selection signal is identical for both proposals; only the proposal distribution differs. Bold marks the better result in each pair.

Selection signal	Benchmark	Temperature proposal	MAR proposal	Δ
Consistency voting	MATH500	0.782	**0.824**	+0.042
Consistency voting	GPQA Diamond	0.349	**0.399**	+0.050
Joint likelihood	MATH500	0.734	**0.754**	+0.020
Joint likelihood	GPQA Diamond	0.263	**0.313**	+0.050

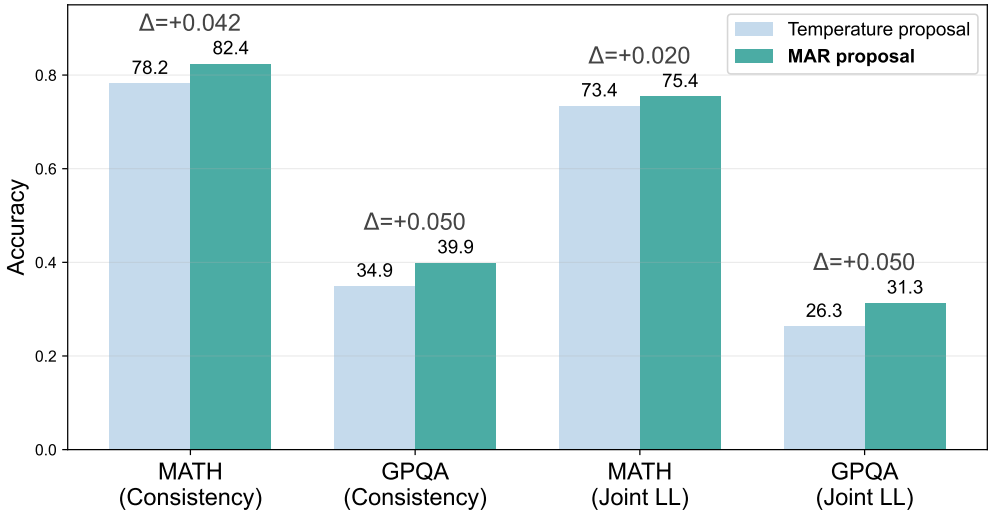


Figure 5: Verifier-agnostic behavior of MAR on Qwen2.5-Math-7B. Bars compare the temperature proposal and the MAR-shaped proposal under two different selection signals. The score is fixed within each pair, so the gap isolates the effect of the proposal. MAR improves every setting, including joint-likelihood selection, despite the latter being length-biased and unrepaired.

Results. Table 5 and Figure 5 summarize the results. Consistency voting improves from 0.782 to 0.824 on MATH500 and from 0.349 to 0.399 on GPQA Diamond. Joint-likelihood selection improves from 0.734 to 0.754 on MATH500 and from 0.263 to 0.313 on GPQA Diamond. The

absolute levels differ substantially across selectors, as expected, but the direction is uniform: in all four comparisons the MAR-shaped proposal delivers a better champion under the same budget and the same selector.

Takeaway. These results support the verifier-agnostic claim. MAR does not require the repaired average-likelihood score to be useful; it improves the candidate pool seen by any downstream selector. This is consistent with the sharpening-loss theory: MAR raises the per-sample hit probability of high-quality trajectories, so it attacks the rate bottleneck of Best-of-N rather than replacing the score.

E COST DETAILS

Table 6: Serial wall-clock and time complexity per delivered answer at $T_{\max} = 3072$. Complexity is asymptotic in T, instantiated at $T = 3072$ as a multiple of a single generation (1). Wall-clock is relative to the 5.26 s single-generation baseline (single RTX 5090, vLLM). MCMC is an inherently sequential chain with no batched-parallel form. All complexity and wall-clock entries are dimensionless multiples ($\times$); "Serial"/"Parallel" refer to the complexity of the serial and batched-parallel schedules, respectively.

	Complexity		**Wall-clock**
Method	**Serial**	**Parallel**	**Serial**
Single generation (T=1)	$O(T)$ (1)	$O(T)$ (1)	5.26 s (1.00)
MAR, single (N=1)	$O(T)$ (1)	$O(T)$ (1)	5.49 s (1.04)
Power sampling (MCMC)	$O(N_{\mathrm{MCMC}}T^2/B)$ (42.5)	$O(N_{\mathrm{MCMC}}T^2/B)$ (42.5)	242.64 s (46.1)
MAR (N=8)	$O(NT)$ (8)	$O(T)$ (1)	42.12 s (8.01)

Table 6 reports the serial wall-clock and serial/parallel complexity behind Table 4. Per decoding step, temperature sampling applies one softmax to the V logits, $O(V)$; MAR computes the boundary expectation $\mathbb{E}_t(T_b) = \sum_x p_{T_b}(x \mid h_t)\, z_x$—one further softmax—and rescales each logit as $\tilde{z}_x = z_x^2/\mathbb{E}_t(T_b)$, one elementwise pass. Its per-step complexity is therefore $O(V)$, identical to temperature sampling up to a small constant factor that vanishes against the dominant forward pass; measured accordingly, a single MAR sample takes 5.49 s against the 5.26 s baseline ($1.04\times$). With N candidates, BoN-MAR costs NT tokens: $O(NT)$ serial, $O(T)$ parallel, as the candidates share one batched decode. MCMC power sampling costs in expectation

$$\mathbb{E}[\text{tokens}] = N_{\mathrm{MCMC}} \sum_{k=1}^{M} \frac{kB}{2} = \frac{N_{\mathrm{MCMC}} BM(M+1)}{4} \approx \frac{N_{\mathrm{MCMC}} T^2}{4B}$$

(Karan & Du, 2026) ($M = \lceil T/B \rceil$), i.e., $O(N_{\mathrm{MCMC}}T^2/B)$, quadratic in T because each Metropolis proposal resamples a suffix whose expected length grows with block index; it admits no batched-parallel form since the chain is sequential. Serialized, BoN-MAR's eight candidates cost $8.01\times$ baseline wall-clock, versus $46.1\times$ for MCMC.