The Optimal Decoder is Tilted Temperature

Nikolaos Papagiannis
Purdue University
npapagia@purdue.edu

Ananth Grama
Purdue University
ayg@purdue.edu

Wojciech Szpankowski
Purdue University
szpan@purdue.edu

Abstract

It is known that larger sample budgets benefit from higher sampling temperatures in LLMs. However, we lack a formal understanding of the objective that temperature tuning optimizes and how it accomplishes this. We seek to answer these questions: we pose decoding as pass@k maximization subject to two constraints that encode critical decoding failure modes: an entropy floor preventing diversity collapse, and a cross-entropy ceiling preventing implausibility. We prove that this maximization takes the form of a single closed-form conditional sampling rule, the *tilted temperature* $q^\star \propto \left(pe^{\eta V}\right)^{1/T}$. Here, the tilt comes from a verifier potential V that quantifies how likely a partial generation is to eventually pass. Ordinary temperature sampling, verifier-greedy decoding, and reward-tilted policies represent special cases. We further show that, under a natural rule for scaling the diversity budget with sample size k, the resulting optimal temperature increases with k, turning empirical choice into a theorem. The optimum still needs the unknown verifier potential, which we estimate using two design approaches. In the online setting, our J-EOS algorithm jointly tracks the two constraint multipliers and learns the verifier potential from binary pass/ fail feedback alone. In the offline setting, we show that when the model is calibrated, its log-probability tracks correctness up to an affine map. Under this condition, the tilt collapses into the exponent and the optimum reduces to ordinary temperature sampling. Calibrated Temperature Decoding (CTD) then fits one scalar offline and samples with zero additional inference-time overhead relative to temperature sampling. Building on these decision rules, we design a practical decoding algorithm for code generation and evaluate it alongside CTD on benchmarks including HumanEval and MBPP, showing improvements over standard and adaptive decoding baselines.

1 Introduction

LLMs are often used and evaluated under a metric that bears little resemblance to the one that shaped them. Training maximizes the likelihood of a single next token; however, deployment in code generation, math, and reasoning pipelines is judged by $\mathrm{pass@}k$, the probability that at least one of k independently drawn sequences solves the task. Chain-of-thought traces with lower likelihood routinely outperform their higher-likelihood alternatives on reasoning benchmarks; best-of-k and self-consistency pipelines exploit precisely this gap; and increasing k yields accuracy at a rate that scales predictably with compute. The decoding hyper-parameter that mediates this trade-off, temperature, together with its truncated variants top-k and nucleus, is empirically tuned. Practitioners have converged on the hypothesis that small k warrants cold sampling and large k prefers hot. However, there is no principled account of which temperature is right, or even of which objective is being optimized when one is chosen.

Formulating these questions as an optimization problem and solving it is hard for two reasons. First, the optimum of an unconstrained $\mathrm{pass@}k$ problem is degenerate: the highest-$\mathrm{pass@}k$ sampler is supported on whichever sequences score best under the success criterion, leading to a Dirac mass on the single best sequence. This does not generalize, collapses sample diversity to zero, and exploits an imperfect verifier. We therefore need an *entropy floor* that forces the sampling distribution to retain spread. Second, the dual failure mode is just as bad: maximizing entropy alone produces uniform sampling, supported on enormous amounts of fictitious data, discarding the information about coherence that the base model distribution p encodes at considerable training cost. We therefore need a second constraint on the sampling distribution q, which is a plausible re-weighting of p, to enforce a *ceiling* on the average per-token cross-entropy of q against p. The resulting program is, for some parameters β and γ,

$$\max_q \ \mathrm{pass@}k(q) \ \text{ s.t. } \quad \tfrac{1}{n}\sum_t \mathbb{E}_{w^{t-1}\sim q} H\big(q_t(\cdot|w^{t-1})\big) \geq \beta, -\tfrac{1}{n}\sum_t \mathbb{E}_{w^{t-1}\sim q}\left[\sum_w q_t \log p_t(w \mid w^{t-1})\right] \leq \gamma. \tag{PE}$$

The two constraints encode the two failure modes that any decoder must avoid: entropy collapse and implausibility. What makes (PE) difficult is the structure of the objective. $\mathrm{pass@}k$ is a sequence-level quantity that depends on the whole generated sequence through a verifier observed only after the last token, while the decoder operates token by token. The bridge is the verifier potential $V_t(w \mid w^{t-1})$, a backward continuation free-energy associated with committing to token w after prefix w^{t-1}. It combines the terminal verifier reward with the downstream likelihood weighting induced by the two dual prices, and satisfies a backward Bellman recursion (Lemma 1). This yields an exact token-level dynamic-programming characterization of the sequence-level optimum. The resulting optimum has conditional form $q_t^\star(w \mid w^{t-1}) \propto \big(p_t(w \mid w^{t-1}) \exp(\eta V_t(w \mid w^{t-1}))\big)^{1/T}$, with $T = 1/\lambda^\star$ and $\eta = \mu^\star/\lambda^\star$. We call this the *tilted temperature* distribution. The family interpolates between likelihood-driven and increasingly verifier-driven decoding. At zero verifier price, its sequence law reduces to tempering of the base sequence distribution; ordinary per-token temperature sampling closely approximates this law when the remaining-horizon correction has limited variation across candidate tokens, and is exact in the zero-variation case. Since $\mathrm{pass@}k(q)$ is strictly increasing in the single-sample success, for fixed budgets (β, γ) the optimizer itself is independent of k; the k-dependence arises through the budget-selection rule for β and is discussed in Appendix F.

The form of the optimum is established by Theorem 1; what is not given in practice is the verifier potential V. We handle this unknown quantity in two complementary ways:

1. *Online analytical construction: jointly track the prices and estimate V.* Our J-EOS construction (Section 3.2) uses binary verdicts from completed rollouts to update a backward Monte-Carlo estimate of the verifier potential while causally tracking the two dual prices (λ, μ) associated with the active constraints. For a fixed potential estimate, we prove that prefix-wise joint price tracking remains close to the offline optimum in parameter space and consequently in sequence law and $\mathrm{pass@}k$ (Theorems 2 and 3). J-EOS is used as an analytical construction establishing how the two constrained prices can be tracked online; it is not our recommended deployment algorithm.

2. *Calibration-based, offline solution: making the verifier tilt redundant.* Rather than estimating V_t directly, we exploit the sequence-level form of the optimum. We formalize (s, ε)*-calibration* by requiring the terminal verifier to be approximately affine in the mean sequence log-probability, $U(w^n) = s\,\bar{\ell}(w^n) + \bar{c} + \bar{r}(w^n), \qquad |\bar{r}(w^n)| \le \varepsilon$. Under perfect calibration, maximizing success is equivalent to minimizing cross-entropy over the feasible set. The optimum therefore has sequence law $q^\star(w^n) \propto p(w^n)^{a_\beta}$, where the entropy floor pins a_β and the cross-entropy ceiling acts only as a feasibility check. When the remaining-horizon correction has limited variation, ordinary per-token temperature sampling closely approximates this sequence optimum. This yields *Calibrated Temperature Decoding (CTD)*: the regression estimates whether positive likelihood–correctness alignment holds and how large the residual is (Theorem 7).

For deployment beyond the calibrated regime, we use the theory as a design principle rather than run J-EOS directly. Our practical decoder, TMAD-GATED (Algorithm 3), intervenes only at uncertain tokens, uses short lookahead to construct a continuation-success proxy, and selects among tilted-temperature regimes subject to plausibility and diversity. TMAD-Gated achieves the best pass@k in 16 of 18 settings across three models and two code-generation benchmarks, outperforming both fixed-temperature and adaptive temperature sampling in most settings.

Contributions. Our main contributions can be summarized as follows: (i) *The decoding problem under entropy and plausibility constraints (Section 2).* We formulate (PE), in which the entropy floor enforces diversity and the cross-entropy ceiling enforces plausibility; the constraints correspond to the two distinct failure modes any deployable decoder must avoid. (ii) *Tilted temperature as the closed-form optimum (Section 3).* We prove the optimum is $q^\star \propto (p\, e^{\eta V})^{1/T}$ (Theorem 1); under the budget-selection rule of Proposition 3, the induced temperature is non-decreasing in k in the high-coverage regime and strictly increasing at interior points where the inequality is strict. (iii) *A sequence-level verifier admits an exact token-level potential characterization.* Although success is observed only after a complete generation, we show that the optimal decoder can be expressed through a backward verifier potential V_t satisfying a Bellman-style recursion. This converts the sequence-level $\mathrm{pass@}k$ objective into a token-level tilted decoding rule through an acyclic Bellman recursion. (iv) *Joint online price tracking as an analytical construction (Section 3.2).* We introduce J-EOS, which couples online potential estimation with causal tracking of both dual prices. Holding the potential estimate fixed, we prove $O(1/t)$ prefix-price tracking and corresponding sequence-law and $\mathrm{pass@}k$ stability bounds (Theorems 2 and 3). We use this construction to analyze the online constrained problem rather than as the practical decoder in our experiments. (v) *Calibration-based offline decoding: CTD (Section 3.3).* We show that sequence-level $(s, 0)$-calibration with $s > 0$ collapses the optimum to the tempered sequence law $q^\star(w^n) \propto p(w^n)^{a_\beta}$, where the entropy floor pins a_β and the cross-entropy ceiling checks feasibility. Ordinary per-token temperature sampling closely approximates this optimum when the remaining-horizon correction has limited variation, with prefix independence as the zero-variation special case. CTD estimates the likelihood–correctness alignment offline by OLS, uses it as a certificate rather than

as the temperature parameter, and admits a suboptimality bound controlled by calibration residual and continuation variation (Theorems 5–7). (vi) *Theory-guided practical decoding and code-generation experiments (Section 4).* We instantiate the tilted-temperature structure in TMAD-GATED (*Tilted Model-Adaptive Decoding*), a confidence-gated decoder that combines short-horizon continuation scoring with constrained selection of (T, η) regimes. On HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) we evaluate CTD and TMAD-GATED against fixed-temperature and adaptive decoding baselines, testing when calibration alone is sufficient and when explicit continuation information provides additional value.

Related work. A recent line of work treats the number of decoding samples as a resource to be scaled: coverage grows log-linearly in k over orders of magnitude (Brown et al., 2024), compute-optimal frontiers trade model size against sample count (Wu et al., 2024), and allocation can be adapted per prompt (Snell et al., 2025). These studies establish pass@k as the operational objective but hold the sampling distribution fixed at a hand-tuned temperature. We ask instead, which distribution maximizes pass@k under an explicit budget, and show that the answer is the tilted temperature (3). A complementary thread improves *selection* over a fixed pool—outcome verifiers (Cobbe et al., 2021), process reward models (Lightman et al., 2024), best-of-n reranking (Nakano et al., 2021), and self-consistency (Wang et al., 2023), whereas our potential V_t enters the *process* through Bellman recursion, and becomes redundant under calibration (Theorem 5). Heuristic reshaping rules —nucleus sampling (Holtzman et al., 2020), truncation-as-desmoothing (Hewitt et al., 2022), are justified empirically. Under perfect sequence-level calibration, the optimum becomes a pure power of the base sequence law, and ordinary per-token temperature sampling closely approximates it when the remaining-horizon correction has limited variation across candidate tokens; prefix independence is the zero-variation special case. The tilted form itself is the Gibbs family of maximum-entropy inference (Jaynes, 1957; Cover and Thomas, 2006) and KL-regularized control (Ziebart et al., 2008; Levine, 2018; Korbak et al., 2022; Rafailov et al., 2023). We differ in operating at inference time under a cross-entropy budget, which is what pins T to the budget rather than to a training-time penalty. Our calibration assumption builds on Platt scaling (Platt, 1999; Guo et al., 2017) and evidence that pretrained models partly track their own correctness (Desai and Durrett, 2020; Kadavath et al., 2022).

2 Problem Formulation

Let $\mathcal{W}$ denote the finite token alphabet and $\mathcal{W}^n$ the set of length-n sequences. Let $p_t(w) = \Pr(w \mid w^{t-1}) \in \Delta(\mathcal{W})$ be the model's conditional distribution at step t. A *policy* (sampling distribution) is $q = (q_1, \ldots, q_n)$ with q_t depending only on $\{p_i\}_{i \le t}$), inducing $q(w^n) = \prod_t q_t(w_t \mid w^{t-1})$. We assume $p_t(w) \ge \rho > 0$ for all t, w. Let $U : \mathcal{W}^n \to [0, 1]$ be the verifier; often binary, $U(w^n) \in \{0, 1\}$ (e.g. code: unit tests pass; math: answer matches). The single-sample success probability of q is $u_q := \mathbb{E}_{w^n \sim q}[U(w^n)]$, and drawing k i.i.d. samples gives $\mathrm{pass@}k(q) := 1 - (1 - u_q)^k$ leading to

$$\max_{q_t \in \Delta(\mathcal{W})} \; u_q \;\; \text{s.t.} \;\; \frac{1}{n} \sum_t \mathbb{E}_{w^{t-1} \sim q} \Big[\mathrm{H}\big(q_t(\cdot \mid w^{t-1})\big) \Big] \;\ge\; \beta, \;\; -\frac{1}{n} \sum_t \mathbb{E}_{w^{t-1} \sim q} \Big[\sum_w q_t \log p_t \Big] \;\le\; \gamma, \tag{1}$$

with $\mathrm{H}(q_t) = -\sum_w q_t(w \mid w^{t-1}) \log q_t(w \mid w^{t-1})$ the entropy of q. To analyze token-level optimization, we summarize a committed prefix–token pair by a value function. Let $\lambda \ge 0$ and $\mu_E \ge 0$ denote the KKT multipliers of the cross-entropy ceiling and entropy floor, respectively. Whenever $\mu_E > 0$, set $\lambda^\star := n\lambda/\mu_E \ge 0$ and $\mu^\star := n/\mu_E > 0$. If $\lambda^\star > 0$, we equivalently write $T := 1/\lambda^\star$ and $\eta := \mu^\star/\lambda^\star$, so that T is the temperature and η the verifier tilt. For $\mu^\star > 0$ the verifier potential at step t is the free energy of the tilted continuation,

$$V_t(w \mid w^{t-1}) \;:=\; \frac{1}{\mu^\star} \log \sum_{w_{t+1:n} \in \mathcal{W}^{n-t}} \prod_{s=t+1}^{n} p_s(w_s \mid w^{s-1})^{\lambda^\star} \, e^{\mu^\star U(w^n)}, \;\; w^n = (w^{t-1}, w, w_{t+1:n}), \tag{2}$$

with the convention that at $t = n$ the continuation set $\mathcal{W}^0$ is the single empty tuple and the product is empty, so that the terminal condition reads $V_n(w \mid w^{n-1}) = U(w^{n-1}, w)$. It scores the tilted continuation value of a prefix–token pair, combining terminal verifier reward with the downstream likelihood weighting induced by the dual prices. It depends on $(p, U, \lambda^\star, \mu^\star)$ but not on any policy. This is the object the optimum of Theorem 1 tilts on, and we denote it by V_t throughout the paper. The next lemma follows by peeling the innermost sum. The proof is in Appendix B.

Lemma 1 (Bellman consistency)**.** *For any $\lambda^\star \ge 0$, $\mu^\star > 0$, step $t \in \{1, \ldots, n-1\}$, prefix w^{t-1}, and token w,*

$$V_t(w \mid w^{t-1}) \;=\; \frac{1}{\mu^\star} \log \sum_{w' \in \mathcal{W}} p_{t+1}(w' \mid w^{t-1}w)^{\lambda^\star} \, \exp\big(\mu^\star V_{t+1}(w' \mid w^{t-1}w)\big),$$

with terminal condition $V_n(w \mid w^{n-1}) = U(w^{n-1}, w)$*. When* $\lambda^\star > 0$*, equivalently* $V_t = \frac{T}{\eta} \log Z_{t+1}(w^{t-1}w)$ *with*

$$Z_{t+1}(w^{t-1}w) \;=\; \sum_{w' \in \mathcal{W}} \Big(p_{t+1}(w' \mid w^{t-1}w)\, e^{\eta V_{t+1}(w' \mid w^{t-1}w)} \Big)^{1/T}.$$

The recursion involves $(p, U, \lambda^\star, \mu^\star)$*, so* V *exists, is unique, and is obtained by one backward sweep over prefix states.*

Remark 1. Program (PE) is written for a fixed prompt x: $p_t(w \mid w^{t-1})$ already conditions on x, so the optimum $q^\star$ is a per-prompt object. In Appendix D, we make the prompt dependence explicit through the natural prompt-level objectives (see Theorem 8). In Appendix E, we study adaptive (PE) where (β, γ) may take D different values in $[0, n]$.

3 Main Results

In Section 3.1, we derive the optimal tilted-temperature sampling distribution. Section 3.2 introduces J-EOS as an analytical online construction for joint price tracking and potential estimation, while Section 3.3 develops CTD, an offline calibration-based decoder that removes the need to explicitly estimate the potential under additional structure.

3.1 Optimal Distribution

Program (PE) is a finite-horizon entropy-regularized control problem: a completed sequence is rewarded by U, and each step by a mixture of the base log-likelihood and the policy entropy. Its optimum is the Gibbs policy on the verifier potential of Lemma 1. We use the following:

Lemma 2 (Gibbs variational principle)**.** *For* $c \in \mathbb{R}^{\mathcal{W}}$ *and* $\tau > 0$*,* $\max_{q \in \Delta(\mathcal{W})} \{\langle q, c\rangle + \tau H(q)\} = \tau \log \sum_w e^{c(w)/\tau}$*, uniquely attained at* $q(w) = e^{c(w)/\tau} / \sum_{w'} e^{c(w')/\tau}$*, where* $\langle q, c\rangle$ *is the scalar product of* q *and* c *over* w*.*

Proof. $\langle q, c\rangle + \tau H(q) = -\tau\, \mathrm{KL}(q \,\|\, g) + \tau \log \sum_w e^{c(w)/\tau}$ with $g(w) \propto e^{c(w)/\tau}$; with equality iff $q = g$. □

Theorem 1 (Tilted temperature optimum: Proof in Appendix C)**.** *Let* $\mu_E \geq 0$ *and* $\lambda \geq 0$ *be the multipliers of the entropy floor and cross-entropy ceiling of* (PE)*. For* $\mu_E > 0$*, set* $\lambda^\star := n\lambda/\mu_E$ *and* $\mu^\star := n/\mu_E$*. The optimum of* (PE) *is*

$$q_t^\star(w \mid w^{t-1}) = \frac{p_t(w)^{\lambda^\star} \exp(\mu^\star V_t(w \mid w^{t-1}))}{Z_t(w^{t-1}; \lambda^\star, \mu^\star)}, \qquad Z_t = \sum_w p_t(w)^{\lambda^\star} \exp(\mu^\star V_t(w \mid w^{t-1})), \tag{3}$$

with V_t *the verifier potential of Lemma 1. If* $\lambda^\star > 0$*, this is equivalent, with* $T = 1/\lambda^\star$ *and* $\eta = \mu^\star/\lambda^\star$*, to*

$$q_t^\star(w \mid w^{t-1}) \propto \big(p_t(w \mid w^{t-1}) \exp(\eta\, V_t(w \mid w^{t-1})) \big)^{1/T}. \tag{4}$$

Remark 2. We show in Appendix F that, under a natural rule for scaling the diversity budget with k, the induced optimal temperature is non-decreasing in the high-coverage regime (Proposition 3). A HumanEval temperature sweep with Qwen2.5-Coder-1.5B shows the same trend; see Figure 2.

3.2 ONLINE TRACKING OF THE JOINT DUAL PRICES

Theorem 1 pins the optimal decoder *given* the true dual prices $\theta^\star = (\lambda^\star, \mu^\star)$, but those prices are defined by two budget equalities that average over the *entire* horizon of (PE). A causal decoder at step t only has access to the prefix $(p_1, \hat{V}_1), \ldots, (p_t, \hat{V}_t)$, which is already an approximation of the true potential, not V_t itself. Therefore, $\theta^\star$ cannot be computed on the fly even setting approximation error aside: both sums in (PE) run over the full horizon $1, \ldots, n$, including positions the decoder has not yet reached. This subsection addresses this causal-deployment gap directly, assuming that some estimate $\hat{V}_t$ of the verifier potential is available. We treat $\hat{V}_t$ as fixed throughout this subsection, separating the price-tracking problem from the question of potential estimation. Appendix H derives the specific step-only estimator used in Algorithm 1. Write $\theta = (\lambda, \mu)^\top \in \Theta \subset \mathbb{R}^2$, $\ell_t(w) = \log p_t(w)$, $\phi_t(w) = (\ell_t(w), V_t(w))^\top$, and for a common parameter θ, define the tilted family

$$q_t^\theta(w) \;=\; \frac{\exp\big(\lambda\, \ell_t(w) + \mu\, V_t(w)\big)}{\sum_{w' \in \mathcal{W}} \exp\big(\lambda\, \ell_t(w') + \mu\, V_t(w')\big)} \;\propto\; p_t(w)^\lambda\, e^{\mu V_t(w)}, \tag{5}$$

Algorithm 1 J-EOS: Joint online tracking of (λ, μ) with online potential estimation.

Require: budgets (β, γ); base model p; verifier U; per-round rollout batch size m; rounds N; Newton tolerance ϵ; max inner iterations $J_{\max}$; initial $\theta_0 = (\lambda_0, 0)$ (plain temperature, no tilt)

$\hat{V}_t(w) \leftarrow 0$ for all t, w ▷ warm start: no verifier signal yet
for $r = 1, \ldots, N$ **do**
  $\theta \leftarrow \theta_{r-1}$ ▷ carry the previous round's converged prices
  for $t = 1, \ldots, n$ **do** ▷ inner loop: causal prefix tracking
    for $j = 1, \ldots, J_{\max}$ **do**
      $q_i^\theta(w \mid w^{i-1}) \leftarrow \dfrac{p_i(w)^\lambda \exp\big(\mu\, \hat{V}_i(w)\big)}{\sum_{w'} p_i(w')^\lambda \exp\big(\mu\, \hat{V}_i(w')\big)}$ for $i \le t$; evaluate $F_t(\theta)$ via (6)
      if $\|F_t(\theta)\|_2 \le \epsilon$ **then break**
      end if
      $\theta \leftarrow \theta - J_t(\theta)^{-1} F_t(\theta)$ ▷ Newton step using the exact Jacobian $DF_t(\theta)$
    end for
    $\theta_t \leftarrow \theta$; sample $w_t \sim q_t^{\theta_t}$
  end for
  freeze the round policy $b_{r,t} := q_t^{\theta_t}$ and draw m fresh i.i.d. sequences from b_r; collect verdicts U_j, $j \in [m]$
  set $\vartheta_r = (\lambda_r, \mu_r) := \theta_n$; if $\mu_r > 0$, update $\hat{V}_t(w) \leftarrow \frac{1}{\mu_r} \log \hat{Z}_{t+1}^{(r)}(w)$ on visited cells via (38) ▷ Appendix H
  $\theta_r \leftarrow \theta_n$ ▷ carry forward as next round's warm start
end for
return $\bar{\theta} := (\bar{\lambda}, \bar{\mu}) = \frac{1}{N} \sum_r \theta_r$; final trained table $\hat{V}$

which is exactly the optimal conditional (3) at $\theta = \theta^\star$. Define the two-dimensional cumulative constraint map

$$F_t(\theta) \;=\; \begin{pmatrix} \sum_{i=1}^t \big[\mathrm{CE}(q_i^\theta, p_i) - \gamma\big] \\ \sum_{i=1}^t \big[H(q_i^\theta) - \beta\big] \end{pmatrix}, \tag{6}$$

so the full-horizon KKT point satisfies $F_n(\theta^\star) = 0$. Note that (6) alone does not certify optimality, that is the tilted Gibbs form together with primal feasibility and complementary slackness established in Theorem 1. Here, it is specialized to the interior regime where both duals are strictly positive.

Prefix tracking. The causal analog of the length-n root equation is to solve the *prefix* KKT system

$$F_t(\theta_t) = 0, \qquad \text{and deploy } q_t = q_t^{\theta_t}. \tag{7}$$

Crucially, (7) is solved retrospectively with one trial parameter θ applied uniformly to $i \le t$ and it does not substitute the already-deployed $q_i^{\theta_i}$ back into the equation. Define the prefix residual of the *final* offline prices, $R_t := F_t(\theta^\star)$ and $R := \max_{1 \le t \le n} \|R_t\|_2$ (see Sima et al. (2026)), so $R_n = 0$ and R_t for $t < n$ measure how far $\theta^\star$ is from satisfying the length-t budgets.

Assumption 1 (Two-dimensional regularity)**.** Let $J_t := \int_0^1 DF_t\big(\theta_t + s(\theta^\star - \theta_t)\big)\, ds$. For every t at which (7) is solved, the smallest singular value $\sigma_{\min}(J_t) \ge c\, t$ for a constant $c > 0$ independent of t, where D represents the derivative.

Theorem 2 (Prefix-price tracking: Proof in Appendix G)**.** *Under Assumption 1,* $\|\theta_t - \theta^\star\|_2 \le \dfrac{\|R_t\|_2}{ct} \le \dfrac{R}{ct}$.

Assumption 2 (Uniform feature covariance bound)**.** There exists $M < \infty$ such that the largest eigenvalue $\lambda_{\max}\big(\mathrm{Cov}_{q_t^\theta}(\phi_t(W))\big) \le M$ for every relevant t and every θ on the segment joining θ_t and $\theta^\star$. This holds whenever $p_t(w) \ge \rho > 0$ and $|V_t(w)| \le V_{\max}$, with M depending only on $\log(1/\rho)$ and $V_{\max}$.

Theorem 3 (Sequence-law and pass@k stability. Proof in Appendix G)**.** *Let Q be the sequence law of the online policies $q_t^{\theta_t}$ and $Q^\star$ that of the offline optimum. Under Assumptions 1 and 2,*

$$\mathrm{KL}(Q \| Q^\star) \le \frac{M\pi^2}{12c^2} R^2, \qquad \mathrm{TV}(Q, Q^\star) \le \frac{\pi\sqrt{M}}{2\sqrt{6}\, c}\, R, \tag{8}$$

and, with $u_{\min} \le \min\{u_Q, u_{Q^\star}\}$,

$$\big|\mathrm{pass@}k(Q) - \mathrm{pass@}k(Q^\star)\big| \;\le\; \min\Big\{1,\; k(1-u_{\min})^{k-1}\,\tfrac{\pi\sqrt{M}}{2\sqrt{6}\,c}R\Big\}. \tag{9}$$

Remark 3 (Why KL, not summed TV). Summing per-step total variation directly gives $\mathrm{TV}(Q, Q^\star) \le \sum_t \mathrm{TV}(q_t^{\theta_t}, q_t^{\theta^\star})$, combining $\|\theta_t - \theta^\star\| = O(R/t)$ with $\sum_t 1/t = O(\log n)$ for an $O(R\log n)$ bound. Routing through the KL chain rule instead is sharper: KL is locally quadratic in the parameter error, so the relevant series is $\sum_t 1/t^2 = O(1)$, giving an $O(R)$ bound with no additional explicit logarithmic dependence on the horizon in (8). Furthermore, $R = \max_t \|F_t(\theta^\star)\|_2$ measures the maximal prefix imbalance of the final offline prices; Theorem 3 is a deterministic stability statement — small prefix residual implies small pass@k deviation rather than a regret bound in the online-learning sense, since V_t is held fixed throughout. Note also that if V_t is produced by a bounded head (e.g., a sigmoid output, $V_{\max} = 1$), Assumption 2's constant M is immediate and does not need to be separately verified.

An information-theoretic lower bound. The preceding guarantees isolate the deterministic price-tracking error for a fixed potential estimate. Independently of that construction, finite terminal verifier feedback imposes a statistical limitation on online adaptation within the tilted family. For an online algorithm $\mathcal{A}$, let q_r denote the feasible tilted policy selected at feedback round r, and define $\mathrm{Reg}_{\mathcal{A}}(N) := \sum_{r=1}^{N}\Big[\mathrm{pass@}k(q^\star) - \mathrm{pass@}k(q_r)\Big]$.

Theorem 4 (Verifier-feedback lower bound. Proof in Appendix I)**.** *Under the non-degeneracy and richness condition of Assumption 4 of Appendix I, there exists a finite family of feasible tilted instances such that every online algorithm selecting from that family using only terminal binary verifier feedback satisfies* $\sup_{\mathcal{I}} \mathbb{E}_{\mathcal{I}}[\mathrm{Reg}_{\mathcal{A}}(N)] \ge c\sqrt{N}$ *for all sufficiently large* N.

3.3 Calibrated Temperature Decoding (CTD)

Theorem 1 establishes the form of the optimum but requires the verifier potential. We now show that, under a checkable condition on the base model, the potential is not needed at all: the objective of (PE) becomes a function of the cross-entropy, the plausibility ceiling drops out, and the optimum is the tempering of p pinned by the entropy floor. The starting point is the *sequence-level* form.

Corollary 1 (Sequence-level optimum)**.** *Telescoping the per-token form of Theorem 1 over* t*, the optimum of* (PE) *is*

$$q^\star(w^n) \;\propto\; p(w^n)^{\lambda^\star}\exp\big(\mu^\star U(w^n)\big), \qquad p(w^n) = \prod_{t=1}^{n} p_t(w_t \mid w^{t-1}).$$

When $\lambda^\star > 0$*, this is equivalently* $q^\star(w^n) \propto \big(p(w^n)e^{\eta U(w^n)}\big)^{1/T}$ *with* $T = 1/\lambda^\star$ *and* $\eta = \mu^\star/\lambda^\star$.

Since the tilt acts on U through the whole sequence, the relevant calibration is between the verdict and the mean log-probability $\bar{\ell}(w^n) := \frac{1}{n}\sum_t \log p_t(w_t \mid w^{t-1}) = \frac{1}{n}\log p(w^n)$.

Definition 1 ((s,ε)-calibration)**.** The base model p is (s,ε)-calibrated w.r.t. U if there is a single model-global slope $s \ge 0$, an intercept $\bar{c}$, and a uniform residual bound $\varepsilon \ge 0$ such that

$$U(w^n) = s\,\bar{\ell}(w^n) + \bar{c} + \bar{r}(w^n), \qquad \sup|\bar{r}(w^n)| \le \varepsilon. \tag{10}$$

The slope s is a property of the model–verifier pair and not of any prefix. It quantifies how steeply confidence converts into correctness. (10) is a linearized, sequence-level instance of Platt scaling Platt (1999). The key observation is that the plausibility functional of (PE) is itself a mean log-probability: by the tower property, $-\frac{1}{n}\sum_t \mathbb{E}_{w^{t-1}\sim q}\big[\sum_w q_t \log p_t\big] = -\mathbb{E}_q[\bar{\ell}] =: \mathrm{CE}(q)$, and by the chain rule $\bar{H}(q) := \frac{1}{n}\sum_t \mathbb{E}[H(q_t)] = \frac{1}{n}H(q)$. Hence, under (10), $u_q = \bar{c} - s\,\mathrm{CE}(q) + \mathbb{E}_q[\bar{r}]$. Calibration turns pass@k maximization into cross-entropy minimization.

Theorem 5 (Perfect calibration: the entropy floor pins the temperature. Proof in Appendix J)**.** *Suppose* p *is* $(s,0)$*-calibrated with* $s > 0$*,* $p(w^n)$ *has a unique mode, and* $\beta \in (0, \log|\mathcal{W}|)$*. For* $a > 0$ *let* $q_a(w^n) \propto p(w^n)^a$ *and let* a_β *be the unique root of* $\bar{H}(q_a) = \beta$*. Then* (PE) *is feasible iff* $\gamma \ge \mathrm{CE}(q_{a_\beta})$*, and for every* k*, its solution has* sequence

law $q^\star(w^n) \propto p(w^n)^{a_\beta}$. *The ceiling γ plays no role beyond feasibility; in the notation of Theorem 1, $\lambda^\star = 0$ and $\mu^\star = na_\beta/s$. The autoregressive conditionals are*

$$q_t^\star(w \mid w^{t-1}) \propto p_t(w \mid w^{t-1})^{a_\beta}\, \Psi_t^{(a_\beta)}(w^{t-1}w), \quad \Psi_t^{(a)}(w^t) = \sum_{w_{t+1:n}\in\mathcal{W}^{n-t}} \prod_{s=t+1}^{n} p_s(w_s \mid w^{s-1})^a, \tag{11}$$

with $\Psi_n^{(a)} \equiv 1$. If the base conditionals are prefix-independent, $\Psi_t^{(a)}$ is token-free and the optimum is pure per-token temperature sampling

$$q_t^\star(w \mid w^{t-1}) = \frac{p_t(w \mid w^{t-1})^{a_\beta}}{\sum_{w'} p_t(w' \mid w^{t-1})^{a_\beta}}. \tag{12}$$

Two consequences arise: First, k enters only through β: with $\beta^\star(k)$ of Proposition 3, the effective temperature is $T^\star_{\text{eff}}(k) = 1/a_{\beta^\star(k)}$, and since $\beta \mapsto a_\beta$ is strictly decreasing, a non-decreasing $\beta^\star(k)$ induces a non-decreasing $T^\star_{\text{eff}}(k)$ in the calibrated regime. Second, only the *sign* of s enters the policy, not its magnitude.

Remark 4 (The collapse is sequence-level). $p(w^n)^a = \prod_t p_t(w_t \mid w^{t-1})^a$ factorizes, but its normalization does not: per-token temperature drops exactly the correction $\log \Psi_t^{(a)}(w^{t-1}w)$ of (11), which is token-constant under prefix independence and nearly so under low continuation variation.

Corollary 2 (Limited continuation variation)**.** *Let $q_{a,t}^{\text{temp}}(w \mid h) \propto p_t(w \mid h)^a$ and define $\delta_t(h) := \max_w \log \Psi_t^{(a)}(hw) - \min_w \log \Psi_t^{(a)}(hw)$, with $\delta_t := \sup_h \delta_t(h)$. Under perfect calibration,*

$$\mathrm{TV}\big(q_t^\star(\cdot \mid h), q_{a,t}^{\text{temp}}(\cdot \mid h)\big) \le \frac{1}{2}\Big(e^{\delta_t(h)} - 1\Big),$$

and hence $\mathrm{TV}(q^\star, q_a^{\text{temp}}) \le \frac{1}{2}\sum_{t=1}^n \big(e^{\delta_t} - 1\big)$. Therefore, the $\mathrm{pass@}k$ gap is at most

$$\Delta_{\Psi,k} := \big|\mathrm{pass@}k(q^\star) - \mathrm{pass@}k(q_a^{\text{temp}})\big| \le \min\left\{1, \frac{k}{2}\sum_{t=1}^n \big(e^{\delta_t} - 1\big)\right\} = O\left(k \sum_t \delta_t\right)$$

for small continuation variation. Prefix-independent conditionals correspond to $\delta_t = 0$.

Theorem 6 (Partial calibration: Proof in Appendix J)**.** *Under (s,ε)-calibration, let $\tilde q := q_{a_\beta}$ be the tempering of Theorem 5 and suppose $\gamma \ge \mathrm{CE}(\tilde q)$. Then (i) $0 \le u_{q^\star} - u_{\tilde q} \le \mathbb{E}_{q^\star}[\bar r] - \mathbb{E}_{\tilde q}[\bar r] \le 2\varepsilon$; (ii) $\big|\mathrm{pass@}k(q^\star) - \mathrm{pass@}k(\tilde q)\big| \le 2\varepsilon\, \sigma_k^+$, with $\sigma_k^+ := k(1 - u_{q^\star} + 2\varepsilon)^{k-1}$; (iii) if $2\varepsilon < u_{q^\star}$, σ_k^+ is strictly unimodal in k with peak near $1/(u_{q^\star} - 2\varepsilon)$, so explicit residual tilting pays off only in a bounded window of k.*

Algorithm 2 uses the regression only as a *certificate*: the Lower Confidence Bound (LCB) on s checks positive likelihood–correctness alignment, while $\hat\varepsilon$ is an empirical residual diagnostic used by the deployment gate; the exponent itself is set by the entropy floor. Note that $\hat\varepsilon$ is a practical residual diagnostic, not a certified estimator of the worst-case ε in Definition 1; Theorems 6–7 are stated in terms of the true ε, with no finite-sample guarantee on $\hat\varepsilon - \varepsilon$. Phase 3a is the zero-overhead CTD fast path. Phase 3b is used when the calibration certificate fails and hands decoding to a verifier-aware fallback that retains explicit continuation information; Section 4 instantiates this fallback with TMAD-GATED (Algorithm 3). Such a decoder can model residual signal that pure temperature omits. Concretely, drawing N independent prompt–completion pairs (x_i, w_i^n) with $x_i \sim \mathcal{D}$ and $w_i^n \sim p(\cdot \mid x_i)$, with labels $U_i \in \{0,1\}$ and scores $\bar\ell_i$ computed in one pass; OLS gives

$$\hat s = \frac{\sum_i (\bar\ell_i - \bar{\bar\ell})(U_i - \bar U)}{\sum_i (\bar\ell_i - \bar{\bar\ell})^2}, \qquad \hat{\bar c} = \bar U - \hat s\, \bar{\bar\ell}. \tag{13}$$

Phase 3a is ordinary temperature sampling, so CTD fast path adds no generation-time overhead.

Proposition 1 (Sequence-level identification of s)**.** *Under* (10) *the population OLS slope of U on $\bar\ell$ is $s + O(\varepsilon)$, and $\hat s$ concentrates around it at rate $O(1/\sqrt{N})$, with no step-level value estimation.*

Assumption 3 (Calibration-set regularity)**.** The N calibration pairs $(\bar\ell_i, U_i)$ are i.i.d. with $U \in \{0,1\}$ and (s,ε)-calibration (10). Score is bounded, $|\bar\ell| \le B := \log(1/\rho)$, and $s_{\text{spr}}^2 := \mathrm{Var}(\bar\ell) \ge s_0^2 > 0$.

Algorithm 2 Calibrated Temperature Decoding (CTD): certified fast path with verifier-aware fallback.

Require: base model p; budgets (β, γ); sample budget k; calibration set $\{(x_i, w_i^n, U_i)\}_{i=1}^N$; calibration-loss tolerance δ_{cal}; confidence level α; $\sigma_k^+ := k(1 - u_{q^\star} + 2\varepsilon)^{k-1}$

Phase 1 (offline, once per domain: certificate).

for each i compute $\bar{\ell}_i = \frac{1}{n}\sum_t \log p_t(\cdot)$ ▷ one forward pass

pooled OLS: $\hat{s} \leftarrow$(13); $\hat{\varepsilon} \leftarrow$ 95% quantile of $|U - \hat{s}\bar{\ell} - \hat{c}|$ on a held-out split

form a Lower Confidence Bound $\text{LCB}_\alpha(s)$ from the concentration bound of Theorem 7

$\hat{u} \leftarrow \frac{1}{N}\sum_i U_i$; $\hat{\sigma}_k^+ \leftarrow k(1 - \hat{u} + 2\hat{\varepsilon})^{k-1}$

Phase 2 (once per β).

solve $\hat{H}(a) = \beta$ by bisection, $\hat{H}(a)$ the occupancy-weighted per-token entropy of q_a^{temp} on calibration rollouts; set $\hat{a}$

check deployment feasibility $\text{CE}(q_{\hat{a}}^{\text{temp}}) \le \gamma$; if it fails, do not certify the CTD fast path

Phase 3 (generation, per prompt).

if $\text{LCB}_\alpha(s) > 0$ **and** $2\hat{\varepsilon}\hat{\sigma}_k^+ \le \delta_{\text{cal}}$ **then**

  3a (CTD): sample $w_t \sim q_t$, $q_t(w) \propto p_t(w \mid w^{t-1}, x)^{\hat{a}}$ ▷ plain temperature cost

else

  3b (fallback): use a verifier-aware decoder that retains explicit continuation information ▷ e.g., TMAD-GATED in Section 4

end if

Theorem 7 (Suboptimality of CTD. Proof in Appendix K)**.** *Under (s, ε)-calibration, let $\hat{q}_{\text{CTD}} := q_{\hat{a}}^{\text{temp}}$ be a Phase 3a policy satisfying $\bar{H}(\hat{q}_{\text{CTD}}) \ge \beta$ and $\text{CE}(\hat{q}_{\text{CTD}}) \le \gamma$, and define the structural cross-entropy excess $\Delta_{\text{CE}} := \text{CE}(\hat{q}_{\text{CTD}}) - \text{CE}(q_{a_\beta}) \ge 0$. Then, deterministically,*

$$0 \le \text{pass@}k(q^\star) - \text{pass@}k(\hat{q}_{\text{CTD}}) \le \sigma_k^{++}\big(\underbrace{s\,\Delta_{\text{CE}}}_{\textit{continuation}} + \underbrace{2\varepsilon}_{\textit{calibration floor}}\big), \tag{14}$$

where $\sigma_k^{++} := k\big(1 - u_{q^\star} + s\Delta_{\text{CE}} + 2\varepsilon\big)^{k-1}$ Under Assumption 3 with $s > C_0\varepsilon$, Phase 1 certifies with probability at least $1 - 8\exp\big(-N(s - C_0\varepsilon)^2/C_s^2\big)$, where C_0, C_s depend only on (B, s_0).

The proof mirrors Theorem 6(i): $\hat{q}_{\text{CTD}}$ meets the floor with equality, and $\text{CE}(q^\star) \ge \text{CE}(q_{a_\beta})$. Relative to a tilted-exponent analysis, three things disappear: no dual price, no horizon factor, and no $O(1/\sqrt{N})$ estimation term—the regression certifies but does not tune, so its error enters only through the exponentially small probability of a wrong sign. The structural term $s\Delta_{\text{CE}}$ is the cross-entropy price of dropping $\Psi^{(a)}$ at fixed entropy; it vanishes under prefix independence ($\delta_t \equiv 0$ in Corollary 2), where $q_a^{\text{temp}} = q_a$ for every a. The prediction is that in the calibrated regime, the right temperature is the entropy-matched one, and CTD should track tuned temperature sampling rather than beat it. Failure of the per-token feasibility check in Phase 2 means only that the CTD deployment approximation is not certified; exact infeasibility is the sequence-level condition of Theorem 5.

Remark 5 (CTD versus explicit online estimation). CTD certifies whether verifier outcomes are explained by the sequence log-probability; when they are, the potential is redundant, and temperature at the entropy-matched exponent is near-optimal at no generation-time cost. J-EOS assumes no such reduction: it estimates the continuation potential from terminal feedback while tracking both prices, at the cost of rollouts. Since our J-EOS analysis covers price tracking for a fixed potential estimate rather than the coupled procedure, we treat J-EOS as an analytical construction, and CTD plus a verifier-aware fallback. Section 4 instantiates this fallback with practical algorithm TMAD-GATED.

4 Experimental Results

We test the two practical decoders derived from theory on code generation. CTD tests the calibrated regime, where temperature alone should suffice (Theorem 5), while TMAD-GATED (Algorithm 3) tests whether an explicit continuation-success proxy for the potential V_t adds value beyond temperature (Theorem 1). We compare both against fixed-temperature sampling and AdapT, an adaptive temperature baseline reviewed in Appendix A.

TMAD-GATED (Algorithm 3, Appendix M) keeps the tilted-temperature family (T, η) of Theorem 1 stochastic at intervention points. The gate combines model confidence with structurally critical code positions; active steps use a

learned continuation-success proxy and short lookahead to select a local regime by an estimated-success/plausibility score.

Table 1: Pass@k (%) on HumanEval and MBPP, reported as mean $\pm$ standard deviation over independent runs. Within each run, methods use the same held-out subset. **Bold** marks the best mean in each column.

		HumanEval			MBPP		
Model	Method	pass@5	pass@10	pass@15	pass@5	pass@10	pass@15
Qwen2.5-1.5B	Fixed-T	46.2 ± 5.8	59.7 ± 7.8	64.8 ± 7.7	36.7 ± 1.1	52.0 ± 2.2	61.0 ± 2.2
	CTD	45.1 ± 5.5	59.7 ± 7.0	68.0 ± 7.7	38.2 ± 1.4	56.1 ± 7.2	62.7 ± 12.2
	AdapT	46.2 ± 2.3	59.7 ± 3.5	66.4 ± 6.6	41.5 ± 3.2	58.3 ± 1.1	67.6 ± 1.1
	TMAD-Gated	$\mathbf{47.6 \pm 4.5}$	$\mathbf{65.3 \pm 3.3}$	$\mathbf{72.2 \pm 3.4}$	$\mathbf{54.5 \pm 0.6}$	$\mathbf{69.7 \pm 2.3}$	$\mathbf{76.6 \pm 2.2}$
Qwen3-4B	Fixed-T	54.4 ± 4.5	63.5 ± 7.6	67.8 ± 6.8	61.6 ± 0.6	71.2 ± 0.1	$\mathbf{75.8 \pm 1.1}$
	CTD	50.6 ± 7.6	62.6 ± 10.5	68.4 ± 4.6	59.7 ± 2.5	68.3 ± 1.5	72.7 ± 1.1
	AdapT	52.6 ± 5.0	62.5 ± 4.0	69.1 ± 5.6	60.9 ± 0.2	70.5 ± 0.3	74.2 ± 1.1
	TMAD-Gated	$\mathbf{55.4 \pm 7.5}$	$\mathbf{66.5 \pm 10.0}$	$\mathbf{71.9 \pm 9.9}$	$\mathbf{66.3 \pm 1.9}$	$\mathbf{71.8 \pm 1.2}$	74.2 ± 1.1
CodeLlama-7B	Fixed-T	23.3 ± 4.2	34.2 ± 3.1	40.2 ± 5.8	26.0 ± 2.1	41.3 ± 1.2	49.0 ± 1.4
	CTD	22.5 ± 3.1	34.2 ± 4.1	41.3 ± 4.2	27.3 ± 0.5	40.2 ± 0.8	47.6 ± 1.0
	AdapT	24.1 ± 1.6	35.2 ± 2.2	$\mathbf{43.2 \pm 1.8}$	28.0 ± 1.3	42.3 ± 1.0	50.5 ± 1.2
	TMAD-Gated	$\mathbf{25.2 \pm 4.8}$	$\mathbf{36.0 \pm 3.9}$	42.3 ± 4.0	$\mathbf{30.3 \pm 1.7}$	$\mathbf{44.0 \pm 1.2}$	$\mathbf{52.1 \pm 1.4}$

TMAD-Gated. A lightweight continuation-success head is trained on fixed-temperature rollouts. At active positions, short lookahead refines its score for the highest-probability candidates, giving a practical proxy $\hat{S}_t$ for the Bellman potential V_t (see (63) in Appendix M). We sample from $q_{T,\eta}(w \mid w^{t-1}) \propto \big(p_t(w \mid w^{t-1}) e^{\eta \hat{S}_t(w|w^{t-1})}\big)^{1/T}$, while inactive positions use conservative low-temperature sampling.

Evaluation protocol and results. We evaluate HumanEval and MBPP with Qwen2.5-Coder-1.5B, Qwen3-4B-Base, and CodeLlama-7B, reporting pass@k for $k \in \{5, 10, 15\}$. Fixed-T's temperature is chosen via a grid search over $T \in [0.2, 0.8]$ maximizing pass@5, reused at pass@10 and pass@15. For each benchmark, we randomly partition the problems into a training pool (for calibration and head training) and a disjoint test set of 64 problems, on which all four methods are evaluated with the same pipeline (details in Appendix M; Figure 4 visualizes the Qwen results). TMAD-GATED is best in 16 of 18 settings, with its clearest gains on MBPP. As predicted by Theorem 7, CTD generally tracks tuned temperature sampling rather than consistently beating it, indicating that explicit continuation information provides the main additional gains.

Concluding Remarks and Discussion. We showed that pass@k maximization under an entropy floor and a cross-entropy ceiling is solved by the tilted temperature $q^\star \propto (p\, e^{\eta V})^{1/T}$, with ordinary temperature sampling as the calibrated special case. We present both online and offline solutions to this problem and validate them on three models and two code benchmarks. The remaining open problem relates to the estimation of the verifier potential V_t at scale, since better potential estimators are the most direct route to closing the gap to the optimum.

References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. Mirostat: A neural text decoding algorithm that directly controls perplexity, 2021. URL `https://arxiv.org/abs/2007.14966`.

Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

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

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

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. arXiv:2003.07892.

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

Kaifeng He, Mingwei Liu, Chong Wang, Zike Li, Yanlin Wang, Xin Peng, and Zibin Zheng. Adadec: A uncertainty-guided lookahead decoding framework for llm-based code generation, 2026. URL `https://arxiv.org/abs/2506.08980`.

John Hewitt, Christopher D. Manning, and Percy Liang. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP*, 2022. arXiv:2210.15191.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1904.09751.

Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. Args: Alignment as reward-guided search. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 17755–17773, 2024. URL `https://proceedings.iclr.cc/paper_files/paper/2024/file/4d1344317478ad99ff5f4e414aeab689-Paper-Conference.pdf`.

Tomasz Korbak, Ethan Perez, and Christopher L. Buckley. RL with KL penalties is better viewed as bayesian inference. In *Findings of the Association for Computational Linguistics: EMNLP*, 2022. arXiv:2205.11275.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2305.20050.

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

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.18290.

Jin Sima, Nikolaos Papagiannis, Ananth Grama, and Wojciech Szpankowski. Temperature sampling is optimal: An information-theoretic foundation of LLM decoding. 2026.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2408.03314.

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 *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2203.11171.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*, 2024.

Yuqi Zhu, Jia Li, Ge Li, YunFei Zhao, Zhi Jin, and Hong Mei. Hot or cold? adaptive temperature sampling for code generation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 437–445, 2024.

Yuqi Zhu, Ge Li, Xue Jiang, Jia Li, Hong Mei, Zhi Jin, and Yihong Dong. Uncertainty-guided chain-of-thought for code generation with llms, 2025. URL `https://arxiv.org/abs/2503.15341`.

Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2008.

Appendix

A Expanded Related Work

A recent line of work treats the number of samples drawn at decoding as a resource to be scaled, in analogy with parameters and pretraining tokens. Brown et al. (2024) show that coverage—the fraction of problems solved by *any* of k samples, i.e. $\mathrm{pass@}k$—grows log-linearly over several orders of magnitude in k, and Wu et al. (2024) chart the compute-optimal frontier between model size and sample count. Snell et al. (2025) make the allocation adaptive, spending more inference compute on harder prompts. These studies establish $\mathrm{pass@}k$ as the operational objective but treat the sampling distribution as fixed (typically the base model at a hand-tuned temperature); our contribution is to ask which decoding distribution *maximizes* $\mathrm{pass@}k$ under an explicit compute budget, and to show that the answer is the tilted-temperature as in (3).

Verifiers and selection. A complementary thread improves the *selection* step rather than the proposal distribution. Outcome-level verifiers trained on solution correctness (Cobbe et al., 2021), process reward models that score intermediate steps (Lightman et al., 2024), and best-of-n reranking against a learned reward (Nakano et al., 2021) all re-weight a fixed pool of samples. Our verifier signal plays a different role: rather than being used only to select among completed samples, it enters the proposal distribution through the continuation potential. In J-EOS, this potential is estimated from terminal verifier feedback, while under the calibration conditions of Section 3.3 its effect can be absorbed into the tempered sequence law. Self-consistency (Wang et al., 2023) can be read as a verifier-free aggregator and is recovered as a special case when the verifier is uninformative.

Temperature and truncation decoding. Heuristic decoding modifications—nucleus sampling (Holtzman et al., 2020) and truncation-as-desmoothing (Hewitt et al., 2022)—reshape the next-token distribution to trade diversity against quality. These transformations are justified empirically; we instead derive the optimal reshaping from a constrained $\mathrm{pass@}k$ objective. Under perfect calibration the optimum reduces to a pure power of the base sequence law, and ordinary per-token temperature sampling closely approximates this optimum when the remaining-horizon correction has limited variation across candidate tokens at each prefix.

Tilted distributions and KL-regularized control. The optimum $q^\star \propto \left(p\, e^{\eta V}\right)^{1/T}$ is an exponentially tilted (Gibbs) distribution, the same form that arises from the maximum-entropy principle (Jaynes, 1957; Cover and Thomas, 2006) and from KL-regularized control. Casting decoding as inference in a soft-optimal MDP (Ziebart et al., 2008; Levine, 2018) and viewing the KL penalty in RLHF as Bayesian conditioning (Korbak et al., 2022) both produce tilted policies;

Rafailov et al. (2023) exploit the closed form of the tilt to bypass explicit reward modeling. We share the tilted-distribution structure but work at inference time under a cross-entropy (compute) budget rather than a training-time KL penalty, which is what pins the temperature $T = 1/\lambda^\star$ to the budget.

Calibration. Our reduction from verifier estimation to temperature rests on calibration of the base model's log-probabilities. Temperature scaling and Platt scaling (Platt, 1999; Guo et al., 2017) are the standard post-hoc calibrators we adopt for the affine model of Section 3.3; calibration of pretrained transformers (Desai and Durrett, 2020) and the finding that language models are partly aware of their own correctness (Kadavath et al., 2022) motivate the (s, ε)-calibration assumption. Theorem 6 quantifies how the residual miscalibration ε degrades $\mathrm{pass@}k$, and Theorem 7 the finite-sample cost of estimating the calibration slope s.

Literature on practical algorithms. A separate line of work adapts the sampling rule online rather than fixing it in advance. Mirostat (Basu et al., 2021) targets a fixed perplexity level, while locally typical sampling (Meister et al., 2023) and η-sampling (Hewitt et al., 2022) truncate using token-level information content or entropy. More recent methods adapt using confidence or structure: AdapT (Zhu et al., 2024) raises sampling temperature at line-initial tokens, while AdaDec (He et al., 2026) combines an entropy threshold with lookahead reranking. Its final token selection is deterministic, so repeated decoding adds no sampling diversity and pass@k collapses to pass@1. ARGS (Khanov et al., 2024) similarly performs reward-guided greedy selection, and Uncert-CoT (Zhu et al., 2025) triggers multi-path reasoning above an uncertainty threshold.

B Proof of Lemma 1 (Bellman Consistency)

Recall from Section 2 that, for prices $\lambda^\star \geq 0$ and $\mu^\star > 0$, a step $t \in \{1, \ldots, n\}$, a prefix w^{t-1} and a token w (so $w_t = w$),

$$V_t(w \mid w^{t-1}) \;=\; \frac{1}{\mu^\star} \log \sum_{w_{t+1:n} \in \mathcal{W}^{\,n-t}} \prod_{s=t+1}^{n} p_s(w_s \mid w^{s-1})^{\lambda^\star}\, e^{\mu^\star U(w^n)}, \qquad w^n = (w^{t-1}, w, w_{t+1:n}), \tag{15}$$

with the convention that at $t = n$ the continuation set $\mathcal{W}^0$ is the single empty tuple and the product is empty. It is convenient to name the partition sum inside the logarithm: for a length-t prefix w^t $(1 \leq t \leq n)$ set

$$\Phi(w^t) \;:=\; \sum_{w_{t+1:n} \in \mathcal{W}^{\,n-t}} \prod_{s=t+1}^{n} p_s(w_s \mid w^{s-1})^{\lambda^\star}\, e^{\mu^\star U(w^n)}, \qquad \Phi(w^n) := e^{\mu^\star U(w^n)}, \tag{16}$$

so that (15) reads $\Phi(w^t) = \exp\big(\mu^\star V_t(w_t \mid w^{t-1})\big)$.

Lemma 1 (Bellman consistency, restated). *For any $\lambda^\star \geq 0$, $\mu^\star > 0$, step $t \in \{1, \ldots, n-1\}$, prefix w^{t-1} and token w,*

$$V_t(w \mid w^{t-1}) = \frac{1}{\mu^\star} \log \sum_{w' \in \mathcal{W}} p_{t+1}(w' \mid w^{t-1}w)^{\lambda^\star} \exp\big(\mu^\star V_{t+1}(w' \mid w^{t-1}w)\big),$$

with terminal condition $V_n(w \mid w^{n-1}) = U(w^{n-1}, w)$; equivalently $V_t = \frac{T}{\eta} \log Z_{t+1}(w^{t-1}w)$ with

$$Z_{t+1}(w^{t-1}w) = \sum_{w' \in \mathcal{W}} \Big(p_{t+1}(w' \mid w^{t-1}w)\, e^{\eta V_{t+1}(w' \mid w^{t-1}w)}\Big)^{1/T}.$$

The recursion involves only $(p, U, \lambda^\star, \mu^\star)$, so V exists, is unique, and is obtained by one backward sweep over prefix states.

Proof. **Well-posedness.** Fix t and w^t. Since $\rho \leq p_s(\cdot \mid w^{s-1}) \leq 1$ and $U \in [0, 1]$, every summand of (16) lies in $\big[\rho^{\lambda^\star(n-t)},\, e^{\mu^\star}\big]$, and the index set $\mathcal{W}^{\,n-t}$ is finite and nonempty. Hence

$$0 \;<\; \rho^{\lambda^\star(n-t)} \;\leq\; \Phi(w^t) \;\leq\; |\mathcal{W}|^{\,n-t} e^{\mu^\star} \;<\; \infty, \tag{17}$$

so $V_t = (\mu^\star)^{-1} \log \Phi(w^t)$ is finite for every (t, w^{t-1}, w), with

$$-\frac{\lambda^\star(n-t)}{\mu^\star} \log \frac{1}{\rho} \;\leq\; V_t(w \mid w^{t-1}) \;\leq\; 1 + \frac{n-t}{\mu^\star} \log |\mathcal{W}|.$$

(The two-sided bound (17) is what makes the log-space form legitimate.)

Recursion. Fix $t \in \{1, \dots, n-1\}$ and a length-t prefix $w^t = (w^{t-1}, w)$. Separate the first continuation token $w_{t+1} = w'$ and factor out $p_{t+1}(w' \mid w^t)^{\lambda^\star}$, which is constant in $w_{t+2:n}$:

$$\begin{aligned}\Phi(w^t) &= \sum_{w' \in \mathcal{W}} p_{t+1}(w' \mid w^t)^{\lambda^\star} \underbrace{\sum_{w_{t+2:n}} \prod_{s=t+2}^{n} p_s(w_s \mid w^{s-1})^{\lambda^\star} e^{\mu^\star U(w^n)}}_{=\,\Phi(w^t w')} \\ &= \sum_{w' \in \mathcal{W}} p_{t+1}(w' \mid w^t)^{\lambda^\star}\, \Phi(w^t w'),\end{aligned}$$

where the brace equals $\Phi(w^t w')$ because for $s \geq t+2$ the factors $p_s(w_s \mid w^{s-1})$ and the verifier $U(w^n)$ depend on w_{t+1} only through the now-fixed conditioning prefix $w^t w'$. Substituting $\Phi(w^t w') = e^{\mu^\star V_{t+1}(w' \mid w^t)}$ and taking $\mu^{\star -1} \log$ yields the stated recursion. At $t = n$ the empty-product convention gives $\Phi(w^n) = e^{\mu^\star U(w^n)}$, hence $V_n(w \mid w^{n-1}) = U(w^{n-1}, w)$.

Equivalent (T, η) form. With $\lambda^\star = 1/T$ and $\mu^\star = \eta/T$ (so $\eta > 0$ whenever $\mu^\star > 0$),

$$p_{t+1}(w' \mid w^{t-1}w)^{\lambda^\star} e^{\mu^\star V_{t+1}(w' \mid w^{t-1}w)} = \Big(p_{t+1}(w' \mid w^{t-1}w)\, e^{\eta V_{t+1}(w' \mid w^{t-1}w)}\Big)^{1/T},$$

so the sum in the recursion is exactly $Z_{t+1}(w^{t-1}w)$ — the per-node normalizer of the tilted-temperature step of Theorem 1 — and $\mu^{\star -1} = T/\eta$ gives $V_t = \frac{T}{\eta} \log Z_{t+1}(w^{t-1}w)$.

Existence, uniqueness, backward sweep. The closed form (15) exhibits V explicitly, so it exists. The recursion expresses V_t through V_{t+1} alone and the terminal condition fixes $V_n = U$; reading the layers $t = n, n-1, \dots, 1$ determines every V_t exactly once. Any V' obeying the recursion and the terminal condition equals V by downward induction on t (they agree at $t = n$; if they agree at $t+1$, the recursion forces agreement at t), giving uniqueness. As the right-hand side involves only $(p, U, \lambda^\star, \mu^\star)$ and no policy q, V is a deterministic function of the instance and the two prices, computed by a single backward pass of n layers with no self-consistency condition to solve. ■

V_t is a free energy, not an ordinary conditional average. The potential V_t is a log-partition quantity over tilted continuations rather than an ordinary conditional expectation of the terminal verifier. This distinction is important because the optimal decoder of Theorem 1 tilts directly on V_t. Four facts summarize its structure.

(i) At the terminal step V_t reduces to the verifier. When $t = n$ the continuation is empty and

$$V_n(w \mid w^{n-1}) = U(w^{n-1}, w).$$

(ii) V_t is a log-partition function. Fix a committed pair $w^t = (w^{t-1}, w)$, let $S_t(w^t) := \sum_{w_{t+1:n}} \prod_{s>t} p_s(w_s \mid w^{s-1})^{\lambda^\star}$ be the total tempered mass of its continuations, and let $\nu_t(\cdot \mid w^t)$ be the probability measure on $\mathcal{W}^{n-t}$ proportional to $\prod_{s>t} p_s(w_s \mid w^{s-1})^{\lambda^\star}$. Then (2) reads

$$\begin{aligned}V_t(w \mid w^{t-1}) &= \frac{1}{\mu^\star} \log S_t(w^t) + \frac{1}{\mu^\star} \log \mathbb{E}_{\nu_t}\big[e^{\mu^\star U(w^n)}\big] \\ &= \frac{1}{\mu^\star} \log S_t(w^t) + \mathbb{E}_{\nu_t}[U] + \frac{\mu^\star}{2} \mathrm{Var}_{\nu_t}(U) + O(\mu^{\star 2}).\end{aligned} \tag{18}$$

When $\lambda^\star = 1$ (equivalently $T = 1$), ν_t is the base-model continuation law and $S_t \equiv 1$, so as $\mu^\star \to 0$,

$$V_t(w \mid w^{t-1}) \longrightarrow \mathbb{E}_{w_{t+1:n} \sim p}\big[U(w^n) \mid w^t\big].$$

At any other temperature S_t need not equal one and the first term in (18) remains. This is harmless for the decoder: only the product $\mu^\star V_t$ enters the optimum, and by (18) it converges to $\log S_t$ as $\mu^\star \downarrow 0$.

(iii) V_t is not a probability. From (2), $p_s^{\lambda^\star} \leq 1$, $p_s \geq \rho$, and $U \in [0, 1]$ give the explicit range

$$-\frac{\lambda^\star (n-t)}{\mu^\star} \log \frac{1}{\rho} \;\leq\; V_t(w \mid w^{t-1}) \;\leq\; 1 + \frac{(n-t)}{\mu^\star} \log |\mathcal{W}|,$$

so V_t need not lie in $[0, 1]$.

Remark 6 (Degenerate prices, and what $\eta = 0$ actually gives). The lemma assumes $\mu^\star > 0$, equivalently $\eta > 0$. At $\mu^\star = 0$ the definition (15) is a $0/0$ and V_t is undefined; what survives is the product $\mu^\star V_t$, which is the only way V_t enters the optimum. Taking the limit $\mu^\star \downarrow 0$ in (16) gives $\mu^\star V_t \to \log S_t(w^t)$, with $S_t(w^t) = \sum_{w_{t+1:n}} \prod_{s>t} p_s(w_s \mid w^{s-1})^{\lambda^\star}$ the tempered continuation mass of Section 2, so as $\mu^\star \downarrow 0$ the tilted step of Theorem 1 becomes

$$q_t^\star(w \mid w^{t-1}) \;\propto\; p_t(w \mid w^{t-1})^{1/T}\; S_t(w^{t-1}w).$$

This is temperature sampling at the *sequence* level — telescoping (Corollary 1) gives $q^\star(w^n) \propto p(w^n)^{1/T}$ — but it is *not* per-token temperature sampling: the extra factor $S_t(w^{t-1}w)$ steers toward branches whose continuation carries cheap entropy, which is exactly the effect Corollary 3 isolates. The two coincide iff $S_t(w^{t-1}w)$ does not depend on w, i.e. at $n = 1$, at $T = 1$, or when the base conditionals are prefix-independent (the i.i.d.-token regime of Sima et al. (2026)). At $n \geq 2$ with prefix-dependent conditionals the factor is not constant.

C Full proof of Theorem 1

Here we give it in full, in particular the passage from the Lagrangian to the per-step cost-to-go the recursion (21) below, which is the principle of optimality specialised to the entropy-regularised program (PE), and the strong-duality step that certifies the resulting stationary point as the constrained optimum.

Throughout, a policy is $q = (q_1, \ldots, q_n)$ with $q(w^n) = \prod_{t=1}^n q_t(w_t \mid w^{t-1})$, and $q(w^{t-1}) := \prod_{s<t} q_s(w_s \mid w^{s-1})$ is the occupancy of the prefix w^{t-1} induced by q. We write $\langle q_t, c\rangle := \sum_w q_t(w)c(w)$.

C.1 Step 0: Lagrangian relaxation

Since $\mathrm{pass@}k(q) = 1 - (1 - u_q)^k$ is strictly increasing in $u_q := \mathbb{E}_{w^n \sim q}[U(w^n)]$, program (PE) is equivalent to maximising u_q under the same two constraints. Attach multipliers $\mu_E \geq 0$ to the entropy floor $\frac{1}{n}\sum_t \mathbb{E}_{w^{t-1}\sim q}[H(q_t)] \geq \beta$ and $\lambda_0 \geq 0$ to the plausibility ceiling $-\frac{1}{n}\sum_t \mathbb{E}_{w^{t-1}\sim q}[\sum_w q_t \log p_t] \leq \gamma$. The Lagrangian is

$$\mathcal{L}_0(q) = u_q + \mu_E\Big(\tfrac{1}{n}\textstyle\sum_t \mathbb{E}[H(q_t)] - \beta\Big) - \lambda_0\Big(-\tfrac{1}{n}\textstyle\sum_t \mathbb{E}[\sum_w q_t \log p_t] - \gamma\Big).$$

Dropping the additive constant $-\mu_E\beta + \lambda_0\gamma$ (which does not affect the maximiser in q) and renaming $\lambda := \lambda_0/n$,

$$\mathcal{L}(q) = E_{w^n\sim q}[U(w^n)] + \sum_{t=1}^n \mathbb{E}_{w^{t-1}\sim q}\big[\rho_t(w^{t-1}, q_t)\big], \tag{19}$$

$$\rho_t(w^{t-1}, q_t) := \tfrac{\mu_E}{n} H\big(q_t(\cdot \mid w^{t-1})\big) + \lambda \sum_w q_t(w \mid w^{t-1}) \log p_t(w \mid w^{t-1}), \tag{20}$$

which is the Lagrangian displayed in the main text. It is a finite-horizon control problem: terminal reward $U(w^n)$ at the leaves and per-step reward ρ_t, a concave functional of the action distribution q_t. The $1/n$ rescaling is "harmless" in the precise sense that it only relabels the dual variable; the optimal λ recovered below is $1/n$ times the multiplier of the ceiling as written in (PE). It is illustrated in Figure 1.

C.2 Step 1: The recursion

The one detail suppressed in the main text is why (19) collapses to a node-wise recursion. This is the principle of optimality; we prove it for the present objective, whose per-step "reward" is a functional of the whole conditional q_t rather than of a sampled action.

Lemma 3 (Principle of optimality)**.** *Let $J_{n+1}(w^n) := U(w^n)$ and, for $t = n, n-1, \ldots, 1$ and every prefix w^{t-1},*

$$J_t(w^{t-1}) := \max_{q_t \in \Delta(\mathcal{W})} \Big\{\rho_t(w^{t-1}, q_t) + \sum_{w\in\mathcal{W}} q_t(w \mid w^{t-1})\, J_{t+1}(w^{t-1}w)\Big\}. \tag{21}$$

Then $\max_q \mathcal{L}(q) = J_1(\varnothing)$, and any policy that attains the maximum in (21) *at every node is a global maximiser of* (19).

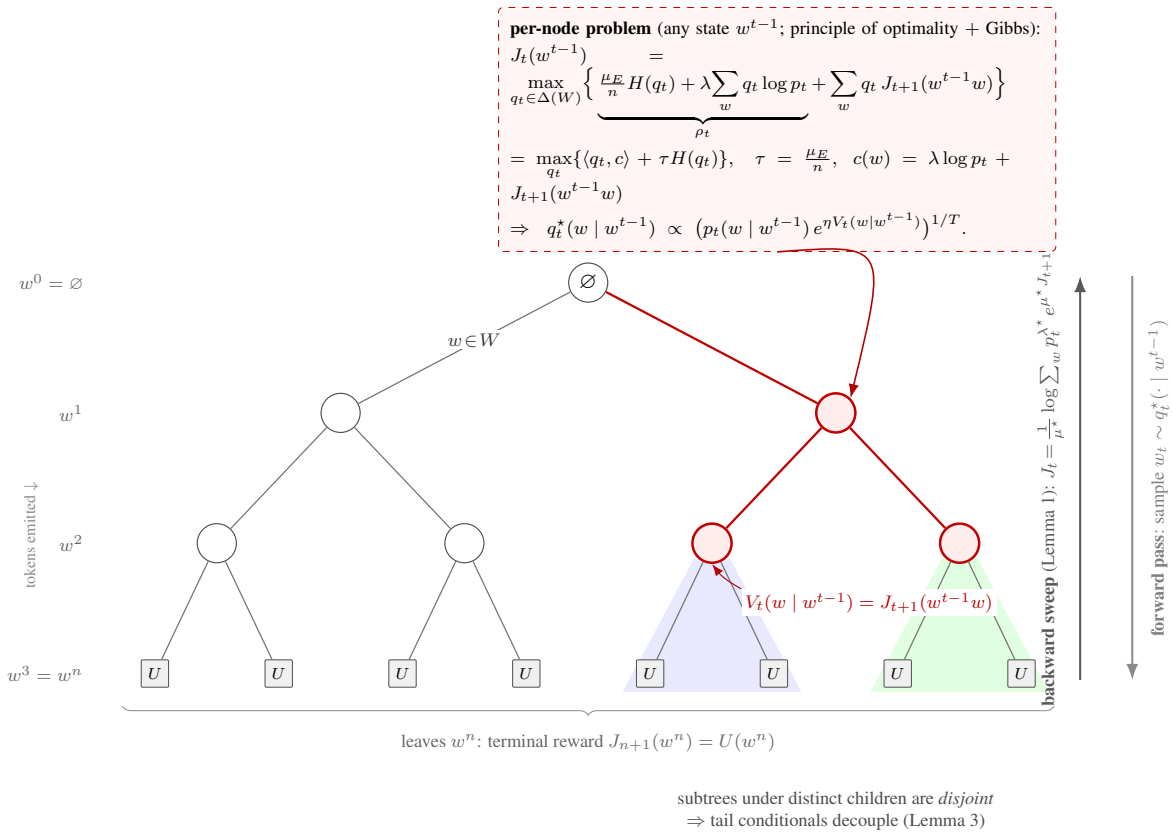


Figure 1: The decision tree behind the proof of Theorem 1. Program (PE) is a finite–horizon control problem on the token tree. A *node* is a committed prefix (a state) w^{t-1}; an *edge* is a token $w \in W$, so every node has $|W|$ children (drawn as 2 for legibility) and the tree has depth n. A *leaf* is a complete sequence w^n carrying the terminal reward $J_{n+1}(w^n) = U(w^n)$; every step contributes the per-step reward $\rho_t = \frac{\mu_E}{n} H(q_t) + \lambda \sum_w q_t \log p_t$. The proof is *two decoupled passes*. *Backward* (right, Lemma 1): the cost-to-go $J_t(w^{t-1})$—equivalently the potential $V_t(w \mid w^{t-1}) := J_{t+1}(w^{t-1}w)$ that lives on the *child*—is formed from the children by a single log-sum-exp, $J_t = \frac{1}{\mu^\star} \log \sum_w p_t^{\lambda^\star} e^{\mu^\star J_{t+1}}$; this uses only $(p, U, \lambda^\star, \mu^\star)$, no policy and no fixed point. *Forward* (left): sample $w_t \sim q_t^\star(\cdot \mid w^{t-1})$. The subtrees hanging off distinct children (shaded blue/green) are *disjoint*, so the tail conditionals decouple across them—this is exactly what lets the global $\max$ over q split into an independent per-node $\max$ (principle of optimality, Lemma 3). At each node that per-node problem is one Gibbs program (Lemma 2), solved by the tilted-temperature rule $q_t^\star \propto \big(p_t\, e^{\eta V_t}\big)^{1/T}$ (dashed box). Because V_t is a free-energy continuation value, the backward (J) and forward (u) sweeps stay decoupled and exact (Corollary 3).

Proof. For a fixed policy q and a prefix w^{t-1}, let

$$G_t^q(w^{t-1}) := \mathbb{E}_{w_{t:n} \sim q}\Big[U(w^n) + \sum_{s=t}^{n} \rho_s(w^{s-1}, q_s) \;\Big|\; w^{t-1}\Big], \qquad G_{n+1}^q(w^n) := U(w^n),$$

the conditional tail objective, the expectation being over $w_t, \ldots, w_n$ drawn from $q_t, \ldots, q_n$ started at w^{t-1}. By the tower property applied to (19), $\mathcal{L}(q) = G_1^q(\varnothing)$. We show $J_t(w^{t-1}) = \max_{q_t,\ldots,q_n} G_t^q(w^{t-1})$ by downward induction on t; the case $t = 1$ is the assertion.

Base. $G_{n+1}^q = U = J_{n+1}$, with no maximisation.

Step. Assume the claim at $t+1$ and fix w^{t-1}. The summand ρ_t is determined by q_t given the prefix, so conditioning on the next token $w_t = w$ and using the tower property,

$$G_t^q(w^{t-1}) = \rho_t(w^{t-1}, q_t) + \sum_{w \in \mathcal{W}} q_t(w \mid w^{t-1})\, G_{t+1}^q(w^{t-1}w). \tag{22}$$

For each w the term $G_{t+1}^q(w^{t-1}w)$ depends only on the conditionals $(q_{t+1}, \ldots, q_n)$ restricted to the sub-tree rooted at $w^{t-1}w$, and these sub-trees are pairwise disjoint across distinct w. Hence the tail conditionals may be optimised

independently within each sub-tree, and the coefficients $q_t(w \mid w^{t-1}) \geq 0$ in (22) do not couple them:

$$\begin{aligned}\max_{q_t,\dots,q_n} G_t^q(w^{t-1}) &= \max_{q_t} \Big\{ \rho_t(w^{t-1}, q_t) + \sum_w q_t(w \mid w^{t-1}) \max_{q_{t+1:n}} G_{t+1}^q(w^{t-1}w) \Big\} \\ &= \max_{q_t} \Big\{ \rho_t + \sum_w q_t \, J_{t+1}(w^{t-1}w) \Big\} = J_t(w^{t-1}),\end{aligned}$$

where the second equality is the induction hypothesis and the last is (21). The interchange of $\max_{q_t}$ with the inner maxima is licit because, once each $\max_{q_{t+1:n}} G_{t+1}^q(w^{t-1}w) = J_{t+1}(w^{t-1}w)$ is a constant in q_t, the outer bracket is a function of q_t alone. Assembling the node maximisers into a policy yields a global maximiser. ■

Substituting ρ_t from (19) into (21) writes the recursion in the form used in the main text,

$$J_t(w^{t-1}) = \max_{q_t \in \Delta(\mathcal{W})} \sum_w q_t(w) \big[\lambda \log p_t(w \mid w^{t-1}) + J_{t+1}(w^{t-1}w) \big] + \tfrac{\mu_E}{n} H(q_t), \qquad J_{n+1} = U.$$

C.3 Step 2: Per-node Gibbs solution

The bracket $c(w) := \lambda \log p_t(w \mid w^{t-1}) + J_{t+1}(w^{t-1}w)$ does not involve q_t, and $\rho_t = \langle q_t, \lambda \log p_t \rangle + \frac{\mu_E}{n} H(q_t)$, so the node problem is $\max_{q_t} \{ \langle q_t, c \rangle + \tau H(q_t) \}$ with $\tau = \mu_E / n > 0$. Lemma 2 (the Gibbs variational principle) applies verbatim and gives

$$q_t^\star(w \mid w^{t-1}) = \frac{e^{c(w)/\tau}}{\sum_{w'} e^{c(w')/\tau}}, \qquad J_t(w^{t-1}) = \tau \log \sum_w e^{c(w)/\tau}.$$

Since $c(w)/\tau = \frac{n}{\mu_E} \lambda \log p_t(w \mid w^{t-1}) + \frac{n}{\mu_E} J_{t+1}(w^{t-1}w) = \lambda^\star \log p_t + \mu^\star J_{t+1}$ with $\lambda^\star := n\lambda/\mu_E$ and $\mu^\star := n/\mu_E$ (so $\tau = 1/\mu^\star$),

$$q_t^\star(w \mid w^{t-1}) \propto p_t(w \mid w^{t-1})^{\lambda^\star} \exp\big(\mu^\star J_{t+1}(w^{t-1}w)\big), \tag{23}$$

$$J_t(w^{t-1}) = \tfrac{1}{\mu^\star} \log \sum_w p_t(w \mid w^{t-1})^{\lambda^\star} \exp\big(\mu^\star J_{t+1}(w^{t-1}w)\big). \tag{24}$$

C.4 Step 3: Identification with the potential of Lemma 1

Set $V_t(w \mid w^{t-1}) := J_{t+1}(w^{t-1}w)$. The left identity in (23) is (3)–(4) with temperature $T = 1/\lambda^\star$ and tilt $\eta = \mu^\star/\lambda^\star$. The terminal value is $V_n(w \mid w^{n-1}) = J_{n+1}(w^{n-1}w) = U(w^{n-1}w) = U(w^n)$, and the identity in (24), rewritten through V, is

$$\tfrac{1}{\mu^\star} \log Z_t(w^{t-1}), \qquad Z_t(w^{t-1}) = \sum_w p_t(w \mid w^{t-1})^{\lambda^\star} \exp\big(\mu^\star V_t(w \mid w^{t-1})\big),$$

i.e. the terminal condition and backward recursion of Lemma 1. That recursion involves only $(p, U, \lambda^\star, \mu^\star)$ and no policy, so V_t is its potential: the optimum is one backward sweep for V followed by the forward policy (4), with no self-consistent pair $(q^\star, V^\star)$ to solve.

C.5 Step 4: Convexity, Slater, strong duality, complementary slackness

It remains to certify that the Lagrangian stationary point is the constrained optimum and that the multipliers are the two active constraints. Pass to causal occupancy-measure coordinates: a policy is equivalent to a state–action visitation $d = (d_t)_t$, $d_t(w^{t-1}, w) := q(w^{t-1})\, q_t(w \mid w^{t-1})$, subject to the linear causal-consistency (Bellman-flow) constraints $\sum_w d_t(w^{t-1}, w) = \sum_{w'} d_{t-1}(w^{t-2}, w')\, \mathbf{1}[(w^{t-2}, w') = w^{t-1}]$ and $d_t \geq 0$; the feasible set $\mathcal{D}$ is a convex compact polytope. In these coordinates:

- $u_q = \sum_{w^n} d_n(w^{n-1}, w_n)\, U(w^n)$ and the plausibility functional $\frac{1}{n} \sum_t \sum_{w^{t-1}, w} d_t \log p_t$ are *linear* in d;
- the causal entropy $\frac{1}{n} \sum_t \sum_{w^{t-1}, w} d_t \log \frac{\sum_{w'} d_t(w^{t-1}, w')}{d_t(w^{t-1}, w)}$ is a sum of conditional entropies, each *concave* in d.

Hence (PE) maximises a linear objective over $\mathcal{D}$ under one concave ($\geq$) constraint and one linear ($\leq$) constraint: a convex program on a convex compact set. Slater's condition holds whenever $\beta < \log|\mathcal{W}|$ (the uniform policy is strictly entropy-feasible) and γ is not at its binding-tight extreme, so a strictly feasible d exists. Strong duality then applies: at the optimal multipliers (λ, μ_E) the unconstrained maximiser of $\mathcal{L}$ — which by Steps 1–3 is the tilted-temperature policy (3)–(4) — solves (PE). Complementary slackness gives

$$\mu_E\big(\bar{H}(q^\star) - \beta\big) = 0, \qquad \lambda\big(\gamma - \mathrm{CE}(q^\star, p)\big) = 0.$$

Thus a slack constraint has zero multiplier. In the interior case where both multipliers are positive, both constraints hold with equality and pin (λ, μ_E)—equivalently (T, η)—through the budgets (β, γ). ■

The optimum is a single backward sweep. By Lemma 1, the potential V_t in (4) depends only on $(p, U, \lambda^\star, \mu^\star)$. Hence, for fixed dual prices, it is obtained by a single backward sweep, after which the optimal forward policy is given directly by (4); no policy–value fixed point is required.

Corollary 3 (Free-energy decomposition)**.** *For $t < n$, with $q^\star$ from* (4) *and* $\mathrm{CE}(q, p) := -\sum_w q(w) \log p(w)$*,*

$$V_t(w \mid w^{t-1}) = \underbrace{\mathbb{E}_{w' \sim q^\star_{t+1}(\cdot \mid w^{t-1} w)}\big[V_{t+1}(w' \mid w^{t-1} w)\big]}_{\textit{expectation backup of } V} + \frac{1}{\mu^\star}\Big[H\big(q^\star_{t+1}\big) - \lambda^\star\, \mathrm{CE}\big(q^\star_{t+1}, p_{t+1}\big)\Big].$$

Thus V_t is an expectation backup augmented by an entropy–plausibility free-energy correction. Since $\mathrm{CE}(q, p) = H(q) + \mathrm{KL}(q\|p)$*, the correction can be written as*

$$\frac{1}{\mu^\star}\Big[(1 - \lambda^\star) H\big(q^\star_{t+1}\big) - \lambda^\star \mathrm{KL}\big(q^\star_{t+1} \| p_{t+1}\big)\Big],$$

so its sign is not fixed in general.

Proof. Take $\log$ of (4) at step $t+1$ and average under $q^\star_{t+1}$. Since $\sum_{w'} q^\star_{t+1} \log q^\star_{t+1} = -H(q^\star_{t+1})$, we obtain

$$-H(q^\star_{t+1}) = -\lambda^\star \mathrm{CE}(q^\star_{t+1}, p_{t+1}) + \mu^\star \mathbb{E}[V_{t+1}] - \log Z_{t+1}.$$

Solving for $\log Z_{t+1}$ and using $V_t = \mu^{\star -1} \log Z_{t+1}$ from Lemma 1 gives the claim. ■

D Prompt-Aware and Robust Formulations of (PE)

Program (PE) is written for a fixed prompt: $p_t(w \mid w^{t-1})$ already conditions on x, so its optimum is a per-prompt object. In deployment a single decoder is applied across a prompt distribution, and the natural objectives — average and worst-case — must be written down explicitly. This appendix does that for three formulations and shows that each is solved by the tilted-temperature family of Theorem 1 applied prompt-wise. The formulations differ only in how the two prices (T, η) are pinned: locally at each prompt, globally across the prompt distribution, or globally at a least-favorable mixture.

Let $\mathcal{X}$ be the prompt space with reference distribution $\mathcal{D}$; take $\mathcal{X}$ finite (or compact, with the usual continuity, so that the minima below are attained). For each $x \in \mathcal{X}$ the model provides conditionals $p_t(\cdot \mid w^{t-1}, x) \in \Delta(\mathcal{W})$ with $p_t \geq \rho > 0$, and a possibly x-dependent verifier $U(\cdot, x) : \mathcal{W}^n \to [0, 1]$. A decoder is a measurable map

$$\Phi : \ (x, w^{t-1}) \ \longmapsto \ q_t(\cdot \mid w^{t-1}, x) \in \Delta(\mathcal{W}), \qquad t = 1, \ldots, n,$$

i.e. it chooses a sampler at every prompt and prefix. Write the per-prompt quantities

$$u_q(x) := \mathbb{E}_{w^n \sim q(\cdot \mid x)}\big[U(w^n, x)\big], \qquad \bar{H}_q(x) := \frac{1}{n} \sum_t \mathbb{E}_{w^{t-1} \sim q}\, H\big(q_t(\cdot \mid w^{t-1}, x)\big),$$

$$\mathrm{CE}_q(x) := -\frac{1}{n} \sum_t \mathbb{E}_{w^{t-1} \sim q}\Big[\textstyle\sum_w q_t \log p_t\Big].$$

The observation that drives everything below: $u_q(x)$, $\bar{H}_q(x)$ and $\mathrm{CE}_q(x)$ depend on the decoder only through its restriction $q(\cdot \mid x)$ to prompt x. Distinct prompts are coupled only if a constraint or the objective ties them together.

For prices $\lambda > 0$, $\mu > 0$ let $V_t^x(w \mid w^{t-1}; \lambda, \mu)$ denote the verifier potential of Lemma 1 built from the pair $\big(p(\cdot \mid x), U(\cdot, x)\big)$ at those prices: it involves no policy, and is obtained by one backward sweep over prefix states per prompt. We write $\varpi \in \Delta(\mathcal{X})$ for a mixture over prompts, reserving w for tokens. We propose three prompt-level formulations:

(A) Average objective, per-prompt budgets.

$$\max_{\Phi} \ \mathbb{E}_{x\sim\mathcal{D}}\big[u_q(x)\big] \quad \text{s.t.} \quad \bar{H}_q(x) \geq \beta(x), \ \ \mathrm{CE}_q(x) \leq \gamma(x) \ \text{ for } \mathcal{D}\text{-a.e. } x. \tag{A}$$

(B) Average objective, pooled budgets.

$$\max_{\Phi} \ \mathbb{E}_{x\sim\mathcal{D}}\big[u_q(x)\big] \quad \text{s.t.} \quad \mathbb{E}_{x\sim\mathcal{D}}\big[\bar{H}_q(x)\big] \geq \beta, \ \ \mathbb{E}_{x\sim\mathcal{D}}\big[\mathrm{CE}_q(x)\big] \leq \gamma. \tag{B}$$

One diversity budget and one plausibility budget are shared across the whole prompt distribution — the deployment regime, where a single temperature is spent over the benchmark. This is the formulation the experiments of Section 4 instantiate: one cross-entropy budget anchored at a reference temperature, one deployed T, applied across all tasks.

(C) Worst-case prompt (robust / distributionally robust).

$$\max_{\Phi} \min_{x\in\mathcal{X}} u_q(x) \quad \text{s.t. per-prompt or pooled budgets} \qquad = \qquad \max_{\Phi} \min_{\varpi\in\Delta(\mathcal{X})} \mathbb{E}_{x\sim\varpi}\big[u_q(x)\big]. \tag{C}$$

The equality rewrites the worst prompt as the worst mixture (the inner objective is linear in ϖ, so its minimum is attained at a vertex); $\min_\varpi$ over all of $\Delta(\mathcal{X})$ is the distributionally robust objective with the largest ambiguity set, and any smaller set $\mathcal{K} \ni \mathcal{D}$ interpolates between (B) ($\mathcal{K} = \{\mathcal{D}\}$) and (C).

Remark 7 (The optimum is a *rule*, computed in two stages). It is tempting to read "averaging over x" in (B) as eliminating the prompt from the solution. It does not. Pooling eliminates x from the *constraints* and from the scalar objective, but the decision variable is the whole decoder $\Phi : x \mapsto q(\cdot \mid x)$, and the optimum is a rule, not a single distribution. Equivalently: (T, η) is a global *operator* and $\big(p(\cdot \mid x), U(\cdot, x)\big)$ is its prompt-local *argument*. At fixed global multipliers the Lagrangian integrand at prompt x depends only on $q(\cdot \mid x)$, so the inner maximization is pointwise and its argmax responds to the local pair even though the multipliers do not. Pooling flattens the operator, never the argument; the only way $q^\star(\cdot \mid x)$ could be x-free is the degenerate case in which $p(\cdot \mid x), U(\cdot, x)$ are constant in x. (Compare cross-validated ridge: the penalty is chosen from the data distribution with no test point in hand; the prediction still depends on the realized x.)

The two stages are therefore: *design* (offline, x random and only $\mathcal{D}$ known) solves the two budget equations for the scalars (T, η) from a calibration sample of $\mathcal{D}$ alone; *inference* (online, x realized) runs the backward sweep of Lemma 1 on $\big(p(\cdot \mid x), U(\cdot, x)\big)$ at those prices to obtain V^x, then samples (25). Since V depends on the prices as well as on the prompt, the sweep in (B) uses the same global $(\lambda^\star, \mu^\star)$ at every prompt, whereas in (A) each prompt's sweep uses its own.

Theorem 8 (Invariance)**.** *Assume $\lambda^\star > 0$ and $\mu^\star > 0$. In each of* (A)*,* (B)*,* (C) *the optimal decoder is, at every x, prefix w^{t-1} and token w,*

$$q_t^\star(w \mid w^{t-1}, x) = \frac{\big(p_t(w \mid w^{t-1}, x)\, \exp\big(\eta\, V_t^x(w \mid w^{t-1})\big)\big)^{1/T}}{Z_t(w^{t-1}, x; T, \eta)}, \tag{25}$$

the tilted-temperature family of Theorem 1 applied prompt-wise, with V^x the potential of Lemma 1 for the pair $\big(p(\cdot \mid x), U(\cdot, x)\big)$ at the prices in force. The formulations differ only in how (T, η) are pinned:

- *(A) $(T, \eta) = \big(T(x), \eta(x)\big)$, solving the two budget equations $\bar{H}_q(x) = \beta(x)$, $\mathrm{CE}_q(x) = \gamma(x)$ separately at each prompt; the potential V^x is built at that prompt's prices.*
- *(B) (T, η) is a single global pair, solving the two pooled budget equations $\mathbb{E}_\mathcal{D}[\bar{H}_q] = \beta$, $\mathbb{E}_\mathcal{D}[\mathrm{CE}_q] = \gamma$; every V^x is built at those global prices.*

(C) Per-prompt budgets: (C) *is solved by the optimum of* (A)*, and robustness is vacuous.* Pooled budgets: (T, η) *is the single global pair pinned by the two budget equations evaluated under a least-favorable mixture* $\varpi^\star \in \Delta(\mathcal{X})$*, and* $q^\star$ *equalizes the per-prompt success on its support,* $u_{q^\star}(x) = \min_{x'} u_{q^\star}(x')$ *for all* $x \in \operatorname{supp} \varpi^\star$*.*

Proof. Throughout, write the per-prompt Lagrangian integrand at prices (λ, μ_E),

$$\mathcal{L}_x\big(q(\cdot \mid x)\big) := u_q(x) + \sum_{t=1}^{n} \mathbb{E}_{w^{t-1}\sim q}\Big[\tfrac{\mu_E}{n} H(q_t) + \lambda \textstyle\sum_w q_t \log p_t\Big], \tag{26}$$

which is exactly the objective appearing in the proof of Theorem 1 for the fixed prompt x: a finite-horizon control problem with terminal reward $U(\cdot, x)$ and per-step reward $\lambda \mathbb{E}_{q_t}[\log p_t] + \frac{\mu_E}{n} H(q_t)$. By that proof — Lemma 2 applied node by node to the optimal cost-to-go — its unique maximizer is (25) with

$$\lambda^\star = \frac{n\lambda}{\mu_E}, \qquad \mu^\star = \frac{n}{\mu_E}, \qquad T = \frac{1}{\lambda^\star} = \frac{\mu_E}{n\lambda}, \qquad \eta = \frac{\mu^\star}{\lambda^\star} = \frac{1}{\lambda}, \tag{27}$$

and V^x the Lemma 1 potential of $\big(p(\cdot \mid x), U(\cdot, x)\big)$ at $(\lambda^\star, \mu^\star)$: it is the optimal cost-to-go itself, $V_t^x(w \mid w^{t-1}) = J_{t+1}(w^{t-1}w)$, produced by the dynamic program and depending on no policy. Also, in occupancy-measure coordinates $u_q(x)$ and $\mathrm{CE}_q(x)$ are linear and $\bar{H}_q(x)$ is concave, so each per-prompt program is convex on a convex compact set, and Slater's condition holds whenever $\beta < \log|\mathcal{W}|$ and γ is not binding-tight.

(A). Objective and constraints are both collections of per-prompt terms, and every term at x involves only $q(\cdot \mid x)$. Hence (A) separates into independent copies of (PE), one per x; each is solved by Theorem 1, whose duals $(\lambda^\star(x), \mu^\star(x))$ — and hence $\big(T(x), \eta(x)\big) = \big(1/\lambda^\star(x), \mu^\star(x)/\lambda^\star(x)\big)$ — are pinned by that prompt's two active constraints.

(B). Introduce scalar multipliers $\mu_E \geq 0$ (pooled entropy floor) and $\lambda \geq 0$ (pooled cross-entropy ceiling). The Lagrangian is

$$\mathcal{L}(\Phi; \lambda, \mu_E) = \mathbb{E}_{x\sim\mathcal{D}}\big[\mathcal{L}_x(q(\cdot \mid x))\big] - \mu_E\beta + \lambda\gamma.$$

At fixed (λ, μ_E) the expectation is of terms each depending on a single restriction $q(\cdot \mid x)$, so $\max_\Phi \mathcal{L}$ decouples across prompts, $\max_\Phi \mathbb{E}_x[\mathcal{L}_x] = \mathbb{E}_x[\max_{q(\cdot|x)} \mathcal{L}_x]$, the pointwise maximizers being measurable in x by (25) (a continuous function of the conditionals and of V^x, itself a continuous function of the conditionals through n log-sum-exps). Each inner problem is (26), so its solution is (25) with the prices (27) — *global*, because (λ, μ_E) are global, while the tilted object $\big(p(\cdot \mid x), V^x\big)$ is local. Convexity plus Slater gives strong duality, so the maximizer of $\mathcal{L}$ at the optimal multipliers solves (B), and (λ, μ_E) are fixed by complementary slackness at the two pooled active constraints $\mathbb{E}_\mathcal{D}[\bar{H}_q] = \beta$, $\mathbb{E}_\mathcal{D}[\mathrm{CE}_q] = \gamma$.

(C), per-prompt budgets. In (A) both the objective and the constraints separate, so its optimum maximizes $u_q(x)$ *pointwise*, subject to that prompt's budgets. A decoder that maximizes every $u_q(x)$ simultaneously maximizes $\min_x u_q(x)$ as well; hence the (A)-optimum solves (C) and nothing is lost by ignoring the robust objective. Prompts do not compete: each has its own budget.

(C), pooled budgets. Now prompts compete for one budget and the problem is a genuine saddle point. Pass to occupancy-measure (causal) coordinates: a decoder is equivalent to a causal visitation $\omega = (\omega_x)_{x\in\mathcal{X}}$, in which $u_q(x) = \langle \omega_x, U_x \rangle$ is linear in ω, the pooled cross-entropy is linear, and the pooled causal entropy is concave; the feasible set Ω is therefore convex and compact. The rewritten objective

$$G(\omega, \varpi) := \sum_x \varpi(x) \langle \omega_x, U_x \rangle$$

is bilinear: linear (hence concave) in ω on Ω, and linear — hence convex and continuous — in ϖ on the compact simplex $\Delta(\mathcal{X})$. Sion's minimax theorem applies, so $\max_\omega \min_\varpi G = \min_\varpi \max_\omega G$ and a saddle point $(\omega^\star, \varpi^\star)$ exists. For fixed ϖ, the inner $\max_\omega G(\cdot, \varpi)$ is exactly formulation (B) with $\mathcal{D}$ replaced by ϖ; hence its solution is (25) with the *single global* pair (T, η) pinned by the budgets under ϖ. Evaluating at $\varpi^\star$ gives the stated form. Finally, optimality of $\varpi^\star$ in the outer min is a linear program over $\Delta(\mathcal{X})$ with cost $u_{q^\star}(x)$; its optimal support lies in $\arg\min_x u_{q^\star}(x)$, and complementary slackness forces $u_{q^\star}(\cdot)$ to be constant $\big(= \min_{x'} u_{q^\star}(x')\big)$ on $\operatorname{supp} \varpi^\star$ — the equalization claim. ■

E Adaptive Optimization: Piecewise-in-Horizon Budgets

Program (PE) uses one diversity floor β and one plausibility ceiling γ over the full horizon. To model position-adaptive decoding, we let these budgets take a small number D of values along the horizon. The main structural fact is that the optimum remains in the tilted-temperature family: each block receives its own pair (T_d, η_d). The global result of Theorem 1 is recovered at $D = 1$.

This extension is useful for interpreting practical adaptive decoders. A position-dependent temperature schedule can exploit coarse variation across the horizon, while a value-aware tilted decoder can additionally react to the state of the current prefix. The block model isolates the first effect.

E.1 Block program and optimum

Fix cut points

$$0 = t_0 < t_1 < \cdots < t_D = n, \qquad B_d := \{t_{d-1}+1, \ldots, t_d\}, \qquad n_d := |B_d|.$$

Each block carries its own diversity and plausibility budgets (β_d, γ_d). The block-constrained problem is

$$\max_q \; u_q \quad \text{s.t.} \quad \frac{1}{n_d} \sum_{t \in B_d} \mathbb{E}_{w^{t-1} \sim q} H\big(q_t(\cdot \mid w^{t-1})\big) \geq \beta_d, \tag{28}$$

and

$$-\frac{1}{n_d} \sum_{t \in B_d} \mathbb{E}_{w^{t-1} \sim q} \Big[\sum_w q_t(w \mid w^{t-1}) \log p_t(w \mid w^{t-1}) \Big] \leq \gamma_d, \qquad d = 1, \ldots, D. \tag{PE-D}$$

As in (PE), maximizing $\mathrm{pass@}k$ is equivalent to maximizing the one-sample success u_q for fixed budgets.

Let $\mu_d > 0$ and $\lambda_{0,d} > 0$ denote the active block multipliers and define

$$\lambda_d := \frac{\lambda_{0,d}}{n_d}, \qquad \lambda_d^\star := \frac{n_d \lambda_d}{\mu_d}, \qquad \mu_d^\star := \frac{n_d}{\mu_d}, \qquad T_d := \frac{1}{\lambda_d^\star}, \qquad \eta_d := \frac{\mu_d^\star}{\lambda_d^\star}.$$

Write $d(t)$ for the block containing step t, set $J_{n+1}(w^n) := U(w^n)$, and define the backward recursion

$$J_t(w^{t-1}) = \frac{1}{\mu_{d(t)}^\star} \log \sum_{w \in W} p_t(w \mid w^{t-1})^{\lambda_{d(t)}^\star} \exp\Big(\mu_{d(t)}^\star J_{t+1}(w^{t-1}w)\Big). \tag{29}$$

Let

$$V_t(w \mid w^{t-1}) := J_{t+1}(w^{t-1}w).$$

Theorem 9 (D-block tilted temperature)**.** *Assume the block budgets are jointly attainable and all block entropy floors are active. Then the optimum of* (PE-D) *is*

$$q_t^\star(w \mid w^{t-1}) \propto p_t(w \mid w^{t-1})^{\lambda_{d(t)}^\star} \exp\Big(\mu_{d(t)}^\star V_t(w \mid w^{t-1})\Big), \tag{30}$$

or equivalently

$$q_t^\star(w \mid w^{t-1}) \propto \Big(p_t(w \mid w^{t-1}) e^{\eta_{d(t)} V_t(w \mid w^{t-1})}\Big)^{1/T_{d(t)}}.$$

Thus the optimum remains tilted-temperature, with a pair (T_d, η_d) for each horizon block.

Proof sketch. The proof is the proof of Theorem 1 with position-dependent dual prices. The Lagrangian still decomposes as a finite-horizon control problem; only the entropy and log-likelihood coefficients are now piecewise constant in t. Applying the Gibbs variational principle at each node gives (29) and (30). Strong duality and complementary slackness then identify the block prices with the active budgets. ■

E.2 What changes at block boundaries

For $D = 1$, the backward recursion can be flattened into the single sequence-level free energy used in the main text. With different block prices, that flattening generally fails across a block boundary: the continuation computed in one block is re-tempered by the prices of the preceding block.

For example, with two blocks and a boundary at t_1, for $t \in B_1$,

$$V_t = \frac{1}{\mu_1^\star} \log \sum_{w_{t+1:t_1}} \prod_{s=t+1}^{t_1} p_s^{\lambda_1^\star} \left(\sum_{w_{t_1+1:n}} \prod_{s=t_1+1}^{n} p_s^{\lambda_2^\star} e^{\mu_2^\star U(w^n)} \right)^{\mu_1^\star/\mu_2^\star}. \tag{31}$$

The inner and outer log-sum-exp operators merge only when the corresponding diversity prices agree. Thus the Bellman recursion remains simple, but the sequence law acquires seam corrections.

Proposition 2 (Sequence-level block correction)**.** *Under Theorem 9,*

$$q^\star(w^n) \propto \left(\prod_{d=1}^{D} \prod_{t \in B_d} p_t(w_t \mid w^{t-1})^{\lambda_d^\star} \right) \exp\left(\mu_D^\star U(w^n) + \sum_{d=1}^{D-1} (\mu_d^\star - \mu_{d+1}^\star) J_{t_d+1}(w^{t_d}) \right). \tag{32}$$

When all $\mu_d^\star$ coincide, the seam corrections vanish. If, in addition, all $\lambda_d^\star$ coincide, (32) *reduces to the sequence-level law of Corollary 1.*

Proof. Multiplying the conditionals in (30) yields

$$q^\star(w^n) = \left(\prod_t p_t^{\lambda_{d(t)}^\star} \right) \exp\left(\sum_{t=1}^{n} \mu_{d(t)}^\star (J_{t+1} - J_t) \right).$$

The sum telescopes within each block. Collecting the remaining boundary terms gives (32). ■

E.3 Connection to sample budget and practical adaptive decoding

For a fixed block schedule $\{(\beta_d, \gamma_d)\}_{d=1}^D$, the optimizer remains independent of k, because $\mathrm{pass@}k(q) = 1 - (1 - u_q)^k$ is monotone in u_q. As in the global case, k affects the decoder only through how the diversity budgets are selected. Hence the "larger k wants hotter" mechanism can be applied blockwise: larger sample budgets may justify more diversity, while the block structure determines where that diversity is spent.

The distinction between block adaptation and value-aware adaptation is also useful experimentally. A block temperature schedule can react to position but not to the realized prefix within a block. By contrast, the continuation potential $V_t(w \mid w^{t-1})$ is state dependent. This gives a direct interpretation of the practical comparisons in Section 4: AdapT represents position-dependent temperature adaptation, while TMAD-Gated uses similar intervention opportunities but additionally incorporates confidence and a continuation-success proxy. The block model therefore isolates the benefit available from coarse position-dependent temperature schedules; any remaining advantage from explicit continuation information corresponds to structure that a fixed block schedule cannot express.

Under perfect calibration, a block-temperature decoder remains especially simple when the diversity price is shared across blocks: the verifier contribution folds into the block exponents, while unequal diversity prices produce the seam terms in (32). More generally, those seam corrections require explicit continuation-value information rather than a temperature schedule alone.

F How to choose β

Where k enters the decoder. Since $\mathrm{pass@}k(q) = 1 - (1 - u_q)^k$ is strictly increasing in u_q, for any fixed budgets (β, γ) the maximizer of (PE) is the single tilted-temperature policy of Theorem 1 and is *independent of* k. The sample budget therefore enters the decoder only through the choice of entropy floor β: "larger k wants hotter" is not a property of the optimal decoder at a fixed operating point, but of how the diversity budget is spent as k grows. Making that

dependence explicit requires an exogenous rule for selecting β as a function of k; a natural one maximizes $\mathrm{pass@}k$ regularized by a strictly convex, strictly decreasing diversity-cost S,

$$\beta^\star(k) := \operatorname*{arg\,max}_{\beta\in[0,\log|\mathcal{W}|]} \big\{1-(1-u(\beta))^k - S(\beta)\big\}, \qquad \text{e.g. } S(\beta) = c\,(\log|\mathcal{W}|-\beta)^2,$$

where $u(\beta)$ is the optimal success at floor β and $\lambda^\star(\beta)$ the corresponding cross-entropy dual. Under this rule larger k favors a higher floor but, as one should expect from the non-monotonicity of the $\mathrm{pass@}k$ sensitivity σ_k (Theorem 6(iii)), only once the operating point is in the high-coverage regime; below it the comparative static reverses. The result below records this; note that S is a modeling choice, so the statement concerns the induced budget schedule, not the optimum of (PE) itself.

Proposition 3 (Budget-selection comparative static)**.** *Assume $u : [0,\log|\mathcal{W}|] \to (0,1)$ is C^1 and strictly decreasing, S is C^1, strictly decreasing and strictly convex, the dual $\lambda^\star(\beta) > 0$ is non-increasing in β on the budget range under consideration, and each $\beta^\star(k)$ is a non-degenerate interior maximizer. Then at every such k,*

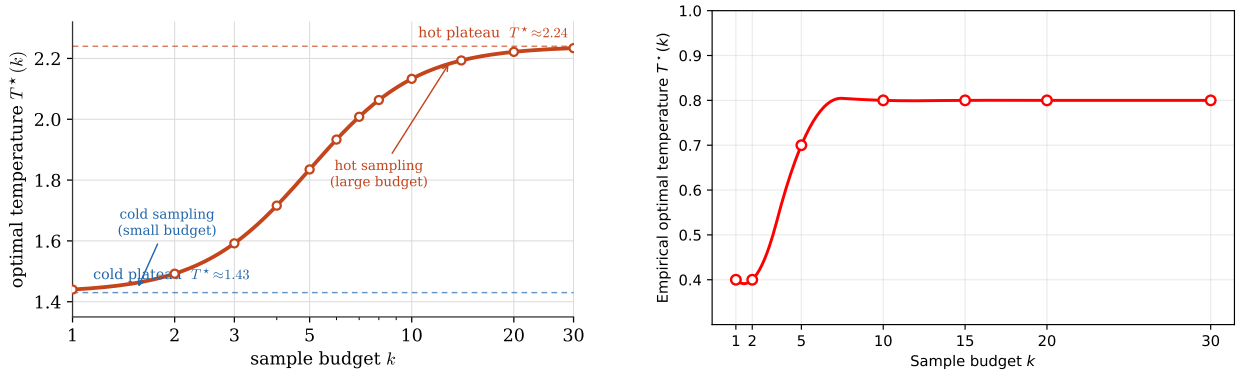
$$\operatorname{sign}\tfrac{d}{dk}\beta^\star(k) = \operatorname{sign}\big(\mathrm{pass@}k(\beta^\star(k)) - (1-e^{-1})\big).$$

Consequently, on any budget range where the operating point clears the coverage line $\mathrm{pass@}k(\beta^\star(k)) \geq 1-e^{-1}$ (equivalently $(1-u(\beta^\star(k)))^k \leq e^{-1}$), the floor $\beta^\star(k)$—and hence the induced temperature $T^\star(k) := 1/\lambda^\star(\beta^\star(k))$—is non-decreasing, strictly at interior points where the inequality is strict; below the crossover $k_0 = -1/\log(1-u^\star) \approx 1/u^\star$ it is non-increasing. Since $\beta^\star(k) \to \log|\mathcal{W}|$ and $\mathrm{pass@}k(\beta^\star(k)) \to 1$ as $k \to \infty$, all sufficiently large k lie in the increasing regime.

We prove Proposition 3 next. But first we recall the budget-selection objective

$$G_k(\beta) := \underbrace{1-(1-u(\beta))^k}_{=:F_k(\beta)} - S(\beta), \qquad \beta^\star(k) := \operatorname*{arg\,max}_{\beta\in[0,\log|\mathcal{W}|]} G_k(\beta),$$

with $u : [0,\log|\mathcal{W}|] \to (0,1)$ the optimal success at floor β (C^1, strictly decreasing), S the C^1, strictly decreasing, strictly convex diversity cost, and $T^\star(k) = 1/\lambda^\star(\beta^\star(k))$ with $\lambda^\star(\cdot) > 0$ non-increasing on the budget range under consideration. It is convenient to write $v(\beta) := 1-u(\beta) \in (0,1)$, so v is strictly increasing with $v'(\beta) = -u'(\beta) > 0$. We treat the budget $k \geq 1$ as a continuous parameter; the integer statement follows by restriction, and a discrete cross-check. Figure 2 illustrates the comparative static synthetically and on a real language model.



(a) Synthetic illustration.

(b) Empirical HumanEval sweep.

Figure 2: Temperature as a function of sample budget. (a) For a synthetic instance, we select the entropy-floor point maximizing $\mathrm{pass@}k$ under the diversity cost S and report the induced temperature $T^\star(k)$. **(b)** On 50 HumanEval problems with Qwen2.5-Coder-1.5B, we select the best fixed temperature from $T \in [0.4, 1.2]$ for each sample budget. Both show the non-decreasing temperature–budget relation predicted by Proposition 3.

The exact comparative static. Both derivatives below are elementary:

$$F_k'(\beta) = -k\, v(\beta)^{k-1} v'(\beta) < 0, \qquad \partial_k F_k'(\beta) = -v'(\beta)\, v(\beta)^{k-1}\big(1 + k \ln v(\beta)\big). \tag{33}$$

The first shows F_k is strictly decreasing in β (a higher entropy floor lowers coverage); the second is obtained by differentiating $k\, v^{k-1} = k\, e^{(k-1)\ln v}$ in k.

Lemma 4 (Sign of the comparative static)**.** *At a non-degenerate interior maximiser $\beta^\star(k)$ (i.e. $G_k'(\beta^\star) = 0$ and $G_k''(\beta^\star) < 0$), the selection $\beta^\star(\cdot)$ is C^1 near k and*

$$\operatorname{sign} \frac{d\beta^\star}{dk} = \operatorname{sign}\big(\partial_k F_k'(\beta^\star)\big) = \operatorname{sign}\big(k(-\ln v^\star) - 1\big) = \operatorname{sign}\big(\text{pass@}k(\beta^\star) - (1 - e^{-1})\big), \qquad v^\star := 1 - u(\beta^\star(k)).$$

Proof. Write the stationarity map $\Phi(\beta, k) := \partial_\beta G_k(\beta) = F_k'(\beta) - S'(\beta)$. By hypothesis $\Phi(\beta^\star, k) = 0$ and $\partial_\beta \Phi(\beta^\star, k) = G_k''(\beta^\star) < 0 \neq 0$, so the implicit function theorem gives a C^1 selection $\beta^\star(\cdot)$ with

$$\frac{d\beta^\star}{dk} = -\frac{\partial_k \Phi(\beta^\star, k)}{\partial_\beta \Phi(\beta^\star, k)} = -\frac{\partial_k F_k'(\beta^\star)}{G_k''(\beta^\star)},$$

where $\partial_k \Phi = \partial_k F_k'$ because S is k-free. As $G_k''(\beta^\star) < 0$, the sign equals that of $\partial_k F_k'(\beta^\star)$. By (33), and since $v' > 0$, $v^{k-1} > 0$,

$$\operatorname{sign}\big(\partial_k F_k'(\beta^\star)\big) = \operatorname{sign}\big(-(1 + k \ln v^\star)\big) = \operatorname{sign}\big(k(-\ln v^\star) - 1\big).$$

Finally $k(-\ln v^\star) - 1 > 0 \iff (v^\star)^k < e^{-1} \iff \text{pass@}k(\beta^\star) = 1 - (v^\star)^k > 1 - e^{-1}$. ■

Corollary 4 (Hotter sampling for larger k)**.** *Under the stated regularity, on any budget interval $\mathcal{K}$ on which the optimal operating point clears the coverage line,*

$$\text{pass@}k\big(\beta^\star(k)\big) \;\geq\; 1 - e^{-1} \qquad \iff \qquad \big(1 - u(\beta^\star(k))\big)^k \leq e^{-1} \qquad (k \in \mathcal{K}),$$

the floor $\beta^\star(k)$ is non-decreasing on $\mathcal{K}$, strictly at every interior k where the inequality is strict; consequently $T^\star(k) = 1/\lambda^\star(\beta^\star(k))$ is non-decreasing on $\mathcal{K}$, strictly at those k. Moreover this regime always obtains for all sufficiently large k: $\beta^\star(k) \to \log|\mathcal{W}|$ and $\text{pass@}k(\beta^\star(k)) \to 1$ as $k \to \infty$.

Proof. On $\mathcal{K}$ the coverage inequality is exactly $(v^\star)^k \leq e^{-1}$, i.e. $k(-\ln v^\star) - 1 \geq 0$, so Lemma 4 gives $\frac{d}{dk}\beta^\star(k) \geq 0$ there, with strict inequality where $(v^\star)^k < e^{-1}$. Integrating the C^1 selection over $\mathcal{K}$, $\beta^\star$ is non-decreasing (strict at interior strict points). Since $\lambda^\star(\cdot)$ is non-increasing, $\beta \mapsto 1/\lambda^\star(\beta)$ is non-decreasing, and composing with the non-decreasing $\beta^\star(\cdot)$ gives the claim for $T^\star$.

For the large-k statement, fix any $\beta_0 < \log|\mathcal{W}|$. Because $u > 0$ on $[0, \log|\mathcal{W}|]$, $\min_{[\beta_0, \log|\mathcal{W}|]} u =: u_0 > 0$ and hence $F_k(\beta) = 1 - (1 - u(\beta))^k \to 1$ uniformly on $[\beta_0, \log|\mathcal{W}|]$. Thus $G_k \to 1 - S$ uniformly there, and since S is strictly decreasing its maximiser over $[\beta_0, \log|\mathcal{W}|]$ is $\log|\mathcal{W}|$; as β_0 is arbitrary, $\beta^\star(k) \to \log|\mathcal{W}|$. Then $u(\beta^\star(k)) \to u(\log|\mathcal{W}|) > 0$, so $\text{pass@}k(\beta^\star(k)) = 1 - (1 - u(\beta^\star(k)))^k \to 1 \geq 1 - e^{-1}$, placing all large k in the regime. ■

Remark 8 (Where the naive global claim fails, and why it is expected). Below the coverage line the comparative static *reverses*: by Lemma 4, $\frac{d}{dk}\beta^\star(k) < 0$ whenever $\text{pass@}k(\beta^\star(k)) < 1 - e^{-1}$. The crossover budget is

$$k_0(u^\star) = \frac{1}{-\ln(1 - u^\star)} \;\approx\; \frac{1}{u^\star} \quad (u^\star \to 0),$$

which is exactly the maximiser of the pass@k sensitivity $\sigma_k = k(1-u)^{k-1}$ appearing in Theorem 6(iii). At a base success rate $u^\star \approx 0.2$ one has $k_0 \approx 4.5$, so at $k = 1, \ldots, 4$ the optimal floor can *decrease* in k before the monotone regime takes over at $k \gtrsim 5$; a grid search over the quadratic cost $S(\beta) = c(\log|\mathcal{W}| - \beta)^2$ reproduces this reversal precisely at the $1 - e^{-1}$ threshold. The non-monotonicity is therefore not a defect of the selection rule but the same low-coverage effect the paper already isolates for σ_k; the folklore "larger k wants hotter" is a theorem precisely on the high-coverage side of k_0.

Choosing the budgets β and γ. The two budgets decouple, which makes them easy to set in practice. The plausibility ceiling γ fixes the *temperature scale*: on the tempered family $q_a \propto p^a$ the per-token cross-entropy $\mathrm{CE}(q_a, p)$ is strictly monotone in the exponent $a = 1/T$ (proof of Theorem 5), so $\gamma \leftrightarrow T$ is a bijection. We therefore anchor γ at a reference temperature T_0 trusted for the model and domain, $\gamma := \mathrm{CE}(q_{1/T_0}, p)$, computed with a single forward pass over the calibration set and no sampling; γ then reads as "q may drift from p by at most the per-token cross-entropy T_0 induces," and is loosened to explore or tightened for precision. The entropy floor β fixes *diversity* and carries the budget dependence: following Proposition 3, we set $\beta^\star(k) = \arg\max_{\beta \in [0, \log|W|]} \{ 1 - (1 - u(\beta))^k - S(\beta) \}$ with a strictly convex, decreasing diversity cost $S(\beta) = c(\log|W| - \beta)^2$, whose maximizer is non-decreasing in k on the high-coverage regime—cold floors for small k, hot floors for large k. The single constant c trades coverage against diversity and is fixed once per domain on validation pass@k; absent an optimization, any schedule rising from $H(p)$ toward $\log|W|$ reproduces the same behavior. Feasibility requires only joint attainability: the base model realizes $(\beta, \gamma) = (H(p), H(p))$, so any $\beta \le H(p)$ with $\gamma \ge H(p)$ admits $q = p$, and driving β above $H(p)$ demands a γ loose enough to license the hotter optimum. Both quantities are occupancy-weighted per-token averages, hence estimated from a handful of calibration prompts at one forward pass each—exactly what CTD's Phase 1 computes.

G Proof of Theorems 2 and 3

Throughout this appendix $\theta = (\lambda, \mu)^\top \in \Theta \subset \mathbb{R}^2$, $\ell_t(w) := \log p_t(w \mid w^{t-1})$, $V_t(w) := V_t(w \mid w^{t-1})$, $\phi_t(w) := (\ell_t(w), V_t(w))^\top$, and

$$q_t^\theta(w) \;=\; \frac{\exp\big(\lambda\, \ell_t(w) + \mu\, V_t(w)\big)}{\sum_{w' \in \mathcal{W}} \exp\big(\lambda\, \ell_t(w') + \mu\, V_t(w')\big)} \;\propto\; p_t(w)^\lambda\, e^{\mu V_t(w)},$$

as in Section 3.2, with cumulative constraint map

$$F_t(\theta) \;=\; \begin{pmatrix} \sum_{i=1}^t \big(\mathrm{CE}(q_i^\theta, p_i) - \gamma\big) \\ \sum_{i=1}^t \big(H(q_i^\theta) - \beta\big) \end{pmatrix},$$

so that the offline active KKT point $\theta^\star = (\lambda^\star, \mu^\star)$ satisfies $F_n(\theta^\star) = 0$, the prefix tracker satisfies $F_t(\theta_t) = 0$ for every t, and $R_t := F_t(\theta^\star)$, $R := \max_{1 \le t \le n} \|R_t\|_2$ as in the main text. Throughout, V_t is held fixed with respect to θ, isolating the tracking question addressed here from the separate question of how V_t is estimated.

G.1 Proof of Theorem 2 (Prefix-price tracking)

Recall Assumption 1: for every t at which (7) is solved, $J_t := \int_0^1 DF_t\big(\theta_t + s(\theta^\star - \theta_t)\big)\, ds$ satisfies $\sigma_{\min}(J_t) \ge ct$ for a constant $c > 0$ independent of t.

Proof. By the fundamental theorem of calculus along the segment joining θ_t and $\theta^\star$,

$$F_t(\theta^\star) - F_t(\theta_t) \;=\; \int_0^1 DF_t\big(\theta_t + s(\theta^\star - \theta_t)\big)\, ds\; (\theta^\star - \theta_t) \;=\; J_t\,(\theta^\star - \theta_t).$$

Since $F_t(\theta_t) = 0$ and $F_t(\theta^\star) = R_t$ by definition, this reads $R_t = J_t(\theta^\star - \theta_t)$. Taking norms and using $\|J_t x\|_2 \ge \sigma_{\min}(J_t)\|x\|_2$,

$$\|R_t\|_2 \;\ge\; \sigma_{\min}(J_t)\, \|\theta^\star - \theta_t\|_2 \;\ge\; ct\, \|\theta_t - \theta^\star\|_2,$$

by Assumption 1. Rearranging, and using $\|R_t\|_2 \le R$ for every t, gives $\|\theta_t - \theta^\star\|_2 \le \|R_t\|_2/(ct) \le R/(ct)$. ■

Remark 9. No global uniqueness of the root of F_t is assumed. The bound only requires that the particular root θ_t selected by the tracker and the offline point $\theta^\star$ are joined by a segment on which the averaged Jacobian J_t is uniformly nonsingular in the sense of Assumption 1; this is exactly what the mean-value step above uses.

G.2 A conditional KL perturbation lemma

The bridge from the parameter-space guarantee of Theorem 2 to a distributional and, ultimately, a $\mathrm{pass@}k$ guarantee goes through the local geometry of the exponential family q_t^θ, not through summing total-variation distances directly (the KL-versus-TV discussion in Section 3.2). For $\theta \in \Theta$ let $A_t(\theta) := \log \sum_{w \in \mathcal{W}} \exp\big(\theta^\top \phi_t(w)\big)$ be the log-partition function of q_t^θ, so that $\nabla A_t(\theta) = \mathbb{E}_{q_t^\theta}[\phi_t(W)]$ and $\nabla^2 A_t(\theta) = \mathrm{Cov}_{q_t^\theta}(\phi_t(W))$.

Lemma 5 (Conditional KL perturbation)**.** *Under Assumption 2,*

$$D_{\mathrm{KL}}\big(q_t^{\theta_t} \,\|\, q_t^{\theta^\star}\big) \;\le\; \frac{M}{2}\,\|\theta_t - \theta^\star\|_2^2.$$

Proof. For an exponential family with natural parameter θ and log-partition A_t, the KL divergence between two members is exactly the Bregman divergence of A_t:

$$D_{\mathrm{KL}}\big(q_t^{\theta_t} \,\|\, q_t^{\theta^\star}\big) = A_t(\theta^\star) - A_t(\theta_t) - \nabla A_t(\theta_t)^\top(\theta^\star - \theta_t).$$

By Taylor's theorem with integral remainder,

$$A_t(\theta^\star) - A_t(\theta_t) - \nabla A_t(\theta_t)^\top(\theta^\star - \theta_t) = \int_0^1 (1-s)\,(\theta^\star - \theta_t)^\top \nabla^2 A_t\big(\theta_t + s(\theta^\star - \theta_t)\big)(\theta^\star - \theta_t)\,ds.$$

Since $\nabla^2 A_t(\theta) = \mathrm{Cov}_{q_t^\theta}(\phi_t(W))$, Assumption 2 bounds its largest eigenvalue by M at every point on the segment, so the integrand is at most $(1-s)\,M\,\|\theta^\star - \theta_t\|_2^2$; integrating over $s \in [0,1]$ contributes a factor $\int_0^1 (1-s)\,ds = \frac{1}{2}$, giving the claim. ■

G.3 Proof of Theorem 3 (Sequence-law and pass@k stability)

Let Q denote the autoregressive sequence law induced by the online conditionals $\{q_t^{\theta_t}\}_{t=1}^n$, and $Q^\star$ the sequence law induced by the offline optimum $\{q_t^{\theta^\star}\}_{t=1}^n$.

Proof. KL bound. By the chain rule for relative entropy on autoregressive sequence laws,

$$D_{\mathrm{KL}}(Q \,\|\, Q^\star) \;=\; \sum_{t=1}^n \mathbb{E}_Q\Big[D_{\mathrm{KL}}\big(q_t^{\theta_t}(\cdot \mid W^{t-1}) \,\big\|\, q_t^{\theta^\star}(\cdot \mid W^{t-1})\big)\Big].$$

Applying Lemma 5 termwise, then Theorem 2 ($\|\theta_t - \theta^\star\|_2 \le R/(ct)$),

$$D_{\mathrm{KL}}(Q \,\|\, Q^\star) \;\le\; \frac{M}{2}\sum_{t=1}^n \|\theta_t - \theta^\star\|_2^2 \;\le\; \frac{M}{2}\sum_{t=1}^n \frac{R^2}{c^2 t^2} \;=\; \frac{MR^2}{2c^2}\sum_{t=1}^n \frac{1}{t^2} \;\le\; \frac{M\pi^2}{12\,c^2}\,R^2,$$

using $\sum_{t=1}^\infty 1/t^2 = \pi^2/6$.

TV and success-probability bounds. Pinsker's inequality, $\mathrm{TV}(Q, Q^\star) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(Q \| Q^\star)}$, gives $\mathrm{TV}(Q, Q^\star) \le \frac{\pi\sqrt{M}}{2\sqrt{6}\,c}\,R$. Since $U : \mathcal{W}^n \to [0,1]$ is bounded,

$$|u_Q - u_{Q^\star}| \;=\; \big|\mathbb{E}_Q[U] - \mathbb{E}_{Q^\star}[U]\big| \;\le\; \mathrm{TV}(Q, Q^\star) \;\le\; \frac{\pi\sqrt{M}}{2\sqrt{6}\,c}\,R.$$

$\mathrm{pass@}k$ *bounds.* With $g_k(u) := 1 - (1-u)^k$, $g_k'(u) = k(1-u)^{k-1} \in [0,k]$ for $u \in [0,1]$, so g_k is k-Lipschitz and $|\mathrm{pass@}k(Q) - \mathrm{pass@}k(Q^\star)| = |g_k(u_Q) - g_k(u_{Q^\star})| \le k\,|u_Q - u_{Q^\star}|$; combined with $\mathrm{pass@}k \in [0,1]$ this gives the stated $\big|\mathrm{pass@}k(Q) - \mathrm{pass@}k(Q^\star)\big| \le \min\big\{1,\ k\,\frac{\pi\sqrt{M}}{2\sqrt{6}\,c}R\big\}$. For the sharper form, the mean value theorem gives ξ between u_Q and $u_{Q^\star}$ with $|g_k(u_Q) - g_k(u_{Q^\star})| = k(1-\xi)^{k-1}|u_Q - u_{Q^\star}|$; since $k(1-u)^{k-1}$ is decreasing in u and $\xi \ge u_{\min} := \min\{u_Q, u_{Q^\star}\}$, this is at most $k(1-u_{\min})^{k-1}|u_Q - u_{Q^\star}|$, giving $\big|\mathrm{pass@}k(Q) - \mathrm{pass@}k(Q^\star)\big| \le k(1-u_{\min})^{k-1}\,\frac{\pi\sqrt{M}}{2\sqrt{6}\,c}\,R$. ■

Remark 10 (Reading R). $R = \max_{t \le n} \|F_t(\theta^\star)\|_2$ measures the maximal prefix imbalance of the *final, offline* prices: how far $\theta^\star$ is from satisfying the length-t truncated budgets, at the worst truncation point. Theorem 3 is a deterministic stability statement – small prefix residual implies small $\mathrm{pass@}k$ deviation – rather than a regret bound in the online-learning sense, since V_t is held fixed throughout this appendix; Appendix H addresses the separate problem of estimating V_t from terminal verifier feedback. Because $\mathrm{pass@}k \in [0, 1]$, any bound exceeding one is vacuous, which is why the loose form of the bound is explicitly truncated at 1.

H Online Estimation of the Verifier Potential in J-EOS

J-EOS needs an estimate of the Bellman potential $V_t(w \mid w^{t-1})$ from terminal pass/fail feedback. After the causal price-tracking pass of round r, it has a sequence of prices $\theta_{r,t} = (\lambda_{r,t}, \mu_{r,t})$, $t = 1, \ldots, n$. We freeze the resulting J-EOS decoder and use it as an importance-sampling behavior policy. This appendix states the corresponding rollout identity, the step-only estimator used by Algorithm 1, and the approximation introduced by collapsing prefixes into a table indexed only by (t, w).

Target prices and frozen behavior. At the end of round r, let

$$\vartheta_r = (\lambda_r, \mu_r) := \theta_{r,n}, \qquad \mu_r > 0,$$

be the price pair whose Bellman potential we want to estimate. Freeze the within-round prices $\{\theta_{r,t}\}_{t=1}^n$ and define the behavior policy

$$b_{r,t}(w \mid w^{t-1}) := \frac{p_t(w \mid w^{t-1})^{\lambda_{r,t}} \exp\big(\mu_{r,t} \widehat{V}_t^{\mathrm{prev}}(w)\big)}{\sum_{a \in \mathcal{W}} p_t(a \mid w^{t-1})^{\lambda_{r,t}} \exp\big(\mu_{r,t} \widehat{V}_t^{\mathrm{prev}}(a)\big)}. \tag{34}$$

After freezing (34), draw a fresh batch of m independent rollouts from b_r. Using a fresh batch is important: the causal trajectory used to construct $\theta_{r,1:n}$ is adaptive and is not, in general, an i.i.d. draw from the subsequently frozen behavior policy.

For a fixed target pair $\vartheta = (\lambda, \mu)$, define the Bellman partition function after committing to (w^{t-1}, w) by

$$Z_{t+1}^{\vartheta}(w^{t-1}w) := \sum_{w_{t+1:n}} \left[\prod_{s=t+1}^{n} p_s(w_s \mid w^{s-1})^{\lambda} \right] e^{\mu U(w^n)}, \qquad V_t^{\vartheta}(w \mid w^{t-1}) = \frac{1}{\mu} \log Z_{t+1}^{\vartheta}(w^{t-1}w). \tag{35}$$

This is exactly the potential of Lemma 1, evaluated at the current target prices ϑ_r.

Proposition 4 (Off-policy rollout identity for J-EOS)**.** *Fix a round r and target prices $\vartheta_r = (\lambda_r, \mu_r)$ with $\mu_r > 0$. For every t, prefix w^{t-1}, and token w, if the behavior policy has full support, then*

$$Z_{t+1}^{\vartheta_r}(w^{t-1}w) = \mathbb{E}_{w_{t+1:n} \sim b_r} \left[e^{\mu_r U(w^n)} \prod_{s=t+1}^{n} \frac{p_s(w_s \mid w^{s-1})^{\lambda_r}}{b_{r,s}(w_s \mid w^{s-1})} \,\middle|\, w^t \right]. \tag{36}$$

Consequently, $V_t^{\vartheta_r}(w \mid w^{t-1}) = \mu_r^{-1} \log Z_{t+1}^{\vartheta_r}(w^{t-1}w)$.

Proof. Expand the expectation in (36). The continuation probability under b_r is $\prod_{s>t} b_{r,s}(w_s \mid w^{s-1})$; it cancels the denominator of the likelihood ratio, leaving $\prod_{s>t} p_s(w_s \mid w^{s-1})^{\lambda_r} e^{\mu_r U(w^n)}$. Summing over all continuations gives (35). ■

Step-only table used by J-EOS. Let $w_{(1)}^n, \ldots, w_{(m)}^n \overset{\text{iid}}{\sim} b_r$ be the fresh rollout batch and $U_{(j)} = U(w_{(j)}^n)$. For every visited cell (t, w) define

$$X_{(j)}^{(r,t)} := e^{\mu_r U_{(j)}} \prod_{s=t+1}^{n} \frac{p_s(w_{(j),s} \mid w_{(j)}^{s-1})^{\lambda_r}}{b_{r,s}(w_{(j),s} \mid w_{(j)}^{s-1})}, \tag{37}$$

$$\widehat{Z}_{t+1}^{(r)}(w) := \frac{1}{N_{r,t}(w)} \sum_{j:\, w_{(j),t} = w} X_{(j)}^{(r,t)}, \qquad \widehat{V}_t^{(r)}(w) := \frac{1}{\mu_r} \log \widehat{Z}_{t+1}^{(r)}(w), \tag{38}$$

where $N_{r,t}(w) = \#\{j : w_{(j),t} = w\}$. Cells with $N_{r,t}(w) = 0$ retain their previous value (or are smoothed with a stated pseudocount); they must not be assigned the logarithm of an empty sum.

Equation (38) is the update intended at the potential-update line of Algorithm 1. It requires storing the behavior probability $b_{r,s}(w_{(j),s} \mid w_{(j)}^{s-1})$ for each sampled token, but it does not require an oracle value of $(\lambda^\star, \mu^\star)$.

What the step-only table identifies. Because $\widehat{V}_t$ is indexed by (t, w) rather than the full prefix, the estimator cannot recover the exact prefix-conditioned potential in general. Conditioning only on the event $w_t = w$ gives

$$\bar{Z}_{t+1}^{(r)}(w) := \mathbb{E}_{w^{t-1}\sim b_r(\cdot|w_t=w)}\Big[Z_{t+1}^{\vartheta_r}(w^{t-1}w)\Big] = \mathbb{E}_{w^{t-1}\sim b_r(\cdot|w_t=w)}\Big[e^{\mu_r V_t^{\vartheta_r}(w|w^{t-1})}\Big], \tag{39}$$

and therefore

$$\bar{V}_t^{(r)}(w) := \frac{1}{\mu_r}\log \bar{Z}_{t+1}^{(r)}(w). \tag{40}$$

Thus the table converges to a behavior-weighted *soft average* of the prefix-conditioned potential. This is an approximation intrinsic to the $O(n|\mathcal{W}|)$ step-only representation, not a Monte-Carlo error. It disappears when $V_t^{\vartheta_r}(w \mid w^{t-1})$ is prefix-independent, and it can be reduced by conditioning the value estimator on a richer state or learned hidden representation.

Theorem 10 (Per-round consistency of the J-EOS estimator)**.** *Fix a round r and suppose that the m rollouts are i.i.d. from the frozen behavior policy b_r, that $\mathbb{P}_{b_r}(W_t = w) \geq \rho_\star > 0$ for all cells of interest, that $p_s(a \mid h) \geq \rho > 0$, and that $b_{r,s}(a \mid h) \geq \kappa > 0$ on the same support. Set*

$$\Delta_{r,t} := e^{\mu_r}\kappa^{-(n-t)} - \rho^{\lambda_r(n-t)}. \tag{41}$$

For $\delta \in (0,1)$, if $m \geq \frac{8}{\rho_\star}\log\frac{2n|\mathcal{W}|}{\delta}$, then with probability at least $1-\delta$, simultaneously for all visited cells,

$$\Big|\widehat{Z}_{t+1}^{(r)}(w) - \bar{Z}_{t+1}^{(r)}(w)\Big| \leq \Delta_{r,t}\sqrt{\frac{\log(4n|\mathcal{W}|/\delta)}{m\rho_\star}}. \tag{42}$$

Consequently $\widehat{V}_t^{(r)}(w)$ is consistent for $\bar{V}_t^{(r)}(w)$. Its logarithm produces the usual downward Jensen bias,

$$\mathbb{E}\Big[\widehat{V}_t^{(r)}(w)\Big] = \bar{V}_t^{(r)}(w) - \frac{\mathrm{Var}(\widehat{Z}_{t+1}^{(r)}(w))}{2\mu_r\,\bar{Z}_{t+1}^{(r)}(w)^2} + o\Big(\mathrm{Var}(\widehat{Z}_{t+1}^{(r)}(w))\Big). \tag{43}$$

Proof. Conditional on $W_t = w$, the fresh rollouts are i.i.d. and the importance-weighted quantity in (37) has mean $\bar{Z}_{t+1}^{(r)}(w)$ by Proposition 4 followed by the prefix marginalization in (39). Moreover, $U \in [0,1]$ implies $e^{\mu_r U} \in [1, e^{\mu_r}]$, while $p_s^{\lambda_r}/b_{r,s} \in [\rho^{\lambda_r}, \kappa^{-1}]$. Hence each weight lies in $[\rho^{\lambda_r(n-t)}, e^{\mu_r}\kappa^{-(n-t)}]$, an interval of width $\Delta_{r,t}$. A multiplicative Chernoff bound gives $N_{r,t}(w) \geq \frac{1}{2}m\rho_\star$ simultaneously with probability at least $1-\delta/2$ under the stated sample-size condition. Conditional on that event, Hoeffding's inequality applied to each cell, followed by a union bound, gives (42). Finally, $z \mapsto \mu_r^{-1}\log z$ is smooth on the positive range above; Jensen's inequality gives the sign of the bias and a second-order Taylor expansion gives (43). ■

Remark 11 (This does not by itself prove joint J-EOS convergence). Appendix G proves stability of the price tracker with the potential held fixed. The theorem above proves consistency of one potential update with the target prices and behavior policy frozen within a round. Combining these two facts is a two-timescale tracking argument; it requires an additional assumption controlling how quickly ϑ_r and the induced potential move between rounds. Therefore the present theory supports J-EOS as a per-round price-tracking plus potential-estimation procedure, but it should not be stated as a full no-regret or global-convergence theorem unless that extra argument is added.

Remark 12 (Boundary case $\mu_r = 0$). The representation $V = \mu^{-1}\log Z$ is undefined at $\mu = 0$. Since J-EOS is initialized at $\mu_0 = 0$, the first round should retain the zero-potential warm start until the tracked solution enters the interior ($\mu_r > 0$), or else the implementation should separately track the log-partition product μV near the boundary. The interior analysis in Section 3.2 and Appendix G already excludes the degenerate boundary case.

I Verifier-feedback lower bound: Proof of Theorem 4

The argument is Le Cam's two–point method: we exhibit two instances whose $\mathrm{pass@}k$ curves agree everywhere except on two disjoint tilt windows, where one is higher and the other lower by $\Theta(\Delta)$. Distinguishing them under terminal bandit verifier feedback costs $\Omega(1/\Delta^2)$ observations, each misallocated round incurs regret $\Omega(\Delta)$, and balancing $\Delta \asymp N^{-1/2}$ yields cumulative regret $\Omega(\sqrt{N})$. The result is independent of any particular online update rule.

Setup and reduction to tilt regret. At rollout $r = 1, \ldots, N$ the algorithm $\mathcal{A}$ picks a tilt $\hat{\eta}_r \in [0, \eta_{\max}]$ measurable w.r.t. the past, draws the k sequences of round r i.i.d. from the offline optimum $q^\star_{\hat{\eta}_r}$ of Theorem 1 (with λ adjusted to meet the entropy floor), and observes only the binary verdicts $U_j \in \{0, 1\}$, $j \in [k]$. Write $u_q := \mathbb{E}_{w^n \sim q}[U(w^n)]$ and

$$f(\eta) := \mathrm{pass@}k(q^\star_\eta) = 1 - \big(1 - u_{q^\star_\eta}\big)^k, \qquad \eta^\star \in \operatorname*{arg\,max}_{\eta \in [0, \eta_{\max}]} f(\eta). \tag{44}$$

Under the protocol the round-r sampler is $q_r = q^\star_{\hat{\eta}_r}$, so the regret of $\mathcal{A}$ is exactly the tilt regret

$$\mathrm{Reg}_{\mathcal{A}}(N) = N\,\mathrm{pass@}k(q^\star) - \sum_{r=1}^{N} \mathrm{pass@}k(q_r) = \sum_{r=1}^{N} \big(f(\eta^\star) - f(\hat{\eta}_r)\big), \tag{45}$$

since $\mathrm{pass@}k(q^\star) = f(\eta^\star)$ for the family we construct. The hard family is chosen so that the feasible tilted policies are parameterized by this single scalar tilt, hence lower-bounding (45) is algorithm-independent within the stated feedback/action model.

To host the perturbation we require the base instance to be non-degenerate and the verifier family rich, which is exactly the hypothesis referenced by Theorem 4.

Assumption 4 (Non-degenerate, rich verifier family)**.** There is a base instance $(p, U, \mathcal{W}, n, k)$ whose curve f in (44) is C^2 on $[0, \eta_{\max}]$ with $|f''| \le M$, $f'(\eta^\star) = 0$, and $f(\eta^\star) \in (c_0, 1 - c_0)$ for absolute constants $M, c_0 > 0$. Moreover the family $\big\{\eta \mapsto \mathbb{P}_{q^\star_\eta}(S) : S \subseteq \mathcal{W}^n \text{ measurable}\big\}$ is, after scaling, L^∞-dense in the smooth $\eta_{\max}$-supported perturbations of $[0, \eta_{\max}]$.

The smoothness and interior-optimum conditions follow from Theorem 1 (the tilted optimum, hence f, is a smooth function of η); $f(\eta^\star) \in (c_0, 1 - c_0)$ excludes the trivial regime where bandit feedback is uninformative; the richness clause is what lets us realize a prescribed curve perturbation at the verifier level (Lemma 7).

The hard pair. *Bumps.* Fix $\eta_- = \eta_{\max}/3$, $\eta_+ = 2\eta_{\max}/3$. Let $b : \mathbb{R} \to [0, 1]$ be smooth and compactly supported with $b(0) = 1$ and $\operatorname{supp} b \subseteq [-\eta_{\max}/6, \eta_{\max}/6]$, and set $\psi(\eta) := b(\eta - \eta_+) - b(\eta - \eta_-)$. The supports $B_- = [\eta_{\max}/6, \eta_{\max}/2]$ and $B_+ = [\eta_{\max}/2, 5\eta_{\max}/6]$ are disjoint (they meet only at the endpoint $\eta_{\max}/2$, where both bumps vanish), so $\psi(\eta_+) = +1$, $\psi(\eta_-) = -1$, and $\|\psi\|_\infty \le 1$.

Curves. Take the base verifier to have a *flat* curve $f_0 \equiv c$ with $c \in (c_0, 1 - c_0)$. For $\varepsilon \in \{+1, -1\}$ define the target

$$f_\varepsilon(\eta) = c + \varepsilon\,\Delta\,\psi(\eta), \qquad \Delta \in (0, c_0/2). \tag{46}$$

A flat base makes every gap inequality below hold directly, with no slope condition on f_0.

Realization in $\mathrm{pass@}k$ *space.* The map $u \mapsto 1 - (1 - u)^k$ is nonlinear, so the perturbation must be designed one layer down, on the per-sample rate. Let $u_0(\eta) = u_{q^\star_\eta}$ under the base verifier, so $1 - c = (1 - u_0(\eta))^k$; flatness makes u_0 constant in η. To move the curve to (46) the per-sample rate must move by g_ε solving $(1 - u_0 - g_\varepsilon)^k = (1 - u_0)^k - \varepsilon\Delta\psi$, i.e.

$$g_\varepsilon(\eta) = \frac{\varepsilon\,\Delta\,\psi(\eta)}{J} + O(\Delta^2), \qquad J := k\,(1 - u_0)^{k-1} = k\,(1 - c)^{(k-1)/k}. \tag{47}$$

Since $(k - 1)/k \in (0, 1)$ and $1 - c \in (c_0, 1 - c_0)$ we have $kc_0 \le J \le k$, hence $\|g_\varepsilon\|_\infty \le \Delta/(kc_0) + O(\Delta^2)$. The budget k enters (46) *only* through this $O(1/k)$ rescaling of the per-sample rate; the target curves are k-free except through c_0. The next lemma records the two-sided scale of u_0 that this induces; it is what keeps the divergence bound below independent of k.

Lemma 6 (Per-sample scale)**.** *For the flat base,* $u_0 = 1 - (1 - c)^{1/k}$ *is constant in* η *and*

$$\frac{c_0}{k} \;\le\; u_0 \;\le\; c, \qquad 1 - u_0 \;\ge\; c_0, \qquad \textit{so} \qquad J = k(1 - u_0)^{k-1} \in [kc_0,\ k].$$

Proof. Flatness gives $(1-u_0)^k = 1-c$, so $u_0 = 1-(1-c)^{1/k}$ is constant. Since $1-c \in (0,1)$ and $1/k \le 1$ we have $(1-c)^{1/k} \ge 1-c$, whence $u_0 \le c$ and $1-u_0 \ge 1-c \ge c_0$. For the lower bound, the map $x \mapsto \varphi(x) := (1-(1-c)^x)/x$ is nonincreasing on $(0,1]$: with $t := x \log \frac{1}{1-c} > 0$ one has $\varphi(x) = \log \frac{1}{1-c} \cdot (1-e^{-t})/t$, and $t \mapsto (1-e^{-t})/t$ is decreasing while t increases in x. Hence $ku_0 = \varphi(1/k) \ge \varphi(1) = 1-(1-c) = c$, i.e. $u_0 \ge c/k \ge c_0/k$. Finally $J = k(1-u_0)^{k-1} = k(1-c)^{(k-1)/k} \in [k(1-c), k] \subseteq [kc_0, k]$. ■

Lemma 7 (Realizability)**.** *Under Assumption 4, for every $\delta \in (0, \frac{1}{2}]$ there is $\Delta_0 > 0$ such that for all $\Delta \le \Delta_0$ there exist measurable $h_\varepsilon : \mathcal{W}^n \to [-c_0/2, c_0/2]$ with $U_\varepsilon := U_0 + h_\varepsilon \in [0,1]$ and*

$$\left\| \eta \mapsto \mathbb{E}_{q_\eta^\star}[h_\varepsilon] - g_\varepsilon \right\|_\infty \;\le\; \delta \, \|g_\varepsilon\|_\infty .$$

Consequently the realized curve is $f_\varepsilon(\eta) = c + \varepsilon\Delta\psi(\eta) + e_\varepsilon(\eta)$ with $\|e_\varepsilon\|_\infty \le \delta'\Delta$, where $\delta' = \delta/c_0 + O(\Delta) \to 0$ as $\delta \to 0$.

Proof. By (47), g_ε is smooth, $\eta_{\max}$-supported with $\|g_\varepsilon\|_\infty = O(\Delta/k)$. Normalize $\bar{g}_\varepsilon := g_\varepsilon/\|g_\varepsilon\|_\infty$ (smooth, $\eta_{\max}$-supported, unit sup-norm). By the richness clause, scaled differences of $\eta \mapsto \mathbb{P}_{q_\eta^\star}(S)$ are L^∞-dense in such functions; pick S_+, S_- and a scale so that $\beta\,(\mathbb{P}_{q_\eta^\star}(S_+) - \mathbb{P}_{q_\eta^\star}(S_-))$ approximates $\bar{g}_\varepsilon$ to sup-norm δ, and set $h_\varepsilon := \|g_\varepsilon\|_\infty\,\beta(\mathbf{1}_{S_+} - \mathbf{1}_{S_-})$, so $\mathbb{E}_{q_\eta^\star}[h_\varepsilon] = g_\varepsilon + (\text{error } \le \delta\|g_\varepsilon\|_\infty)$ and $\|h_\varepsilon\|_\infty = O(\Delta/k) \le c_0/2$ for Δ small. To keep $U_\varepsilon = U_0 + h_\varepsilon \in [0,1]$ with $U_0 \in \{0,1\}$, take $S_+ \subseteq \{U_0 = 0\}$ and $S_- \subseteq \{U_0 = 1\}$ (both carry $q_\eta^\star$-mass since $c \in (c_0, 1-c_0)$), so $h_\varepsilon \ge 0$ where $U_0 = 0$ and $h_\varepsilon \le 0$ where $U_0 = 1$; replacing U_ε by the randomized verifier that outputs 1 w.p. $U_\varepsilon(w^n)$ makes it binary without changing the per-sample rate. Finally the per-sample error propagates to the curve through $u \mapsto 1-(1-u)^k$: since $\mathrm{d}f = J\,\mathrm{d}u$ with $J \le k$ and $J \ge kc_0$, the pass@k error is $|e_\varepsilon| \le J \cdot \delta\|g_\varepsilon\|_\infty + O(\Delta^2) \le k \cdot \delta \cdot \frac{\Delta}{kc_0} + O(\Delta^2) = (\delta/c_0)\Delta + O(\Delta^2)$, i.e. $\delta' = \delta/c_0 + O(\Delta)$. ■

(The factor $J \asymp k$ amplifies the $O(\Delta/k)$ per-sample error into an $O(\Delta)$ curve error; this is why the realization is designed and controlled at the per-sample scale, and why δ' comes out k-free.)

Fix δ so that $\delta' \le 1/4$ and absorb e_ε into the effective gap $\Delta' := \frac{3}{4}\Delta$, written Δ below; (46) is then exact up to this constant factor.

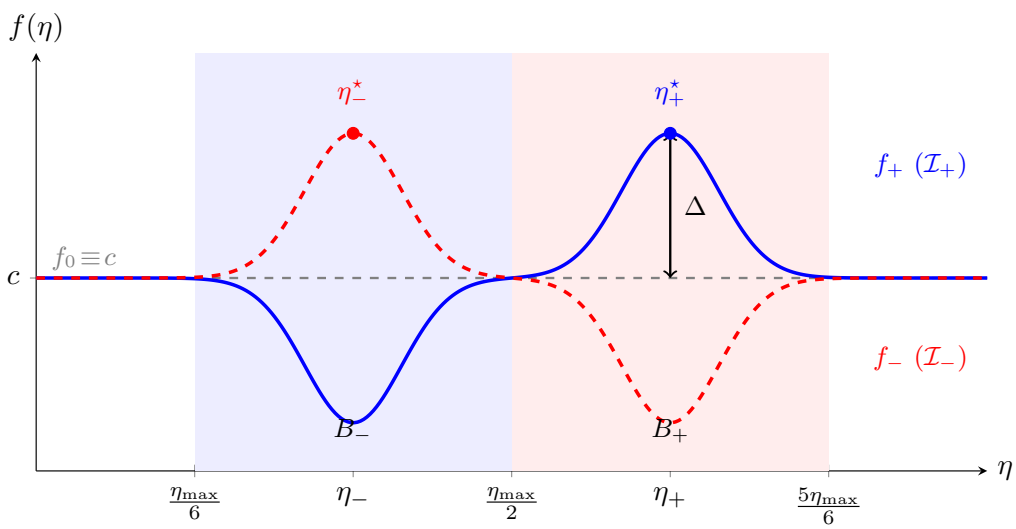


Figure 3: The hard pair. The two pass@k curves coincide outside $B_- \cup B_+$ and differ by Δ inside; f_+ peaks in B_+ and f_- in B_-. Under bandit feedback the learner reads the round's k verdicts (equivalently, up to constants, the pass bit $\Pi_r \sim \mathrm{Ber}(f_\varepsilon(\hat{\eta}_r))$), so the two worlds are statistically close (KL $\le C_0\Delta^2$ per round) yet have optima in different bumps.

Gap structure (Figure 3). With f_0 flat and $\psi \ge 0$ on B_+, $\psi \le 0$ on B_-, the maximizer of f_ε lies in B_ε with $f_\varepsilon(\eta_\varepsilon^\star) = c + \Delta$. For every $\eta \in B_-$, $f_+(\eta) = c - \Delta\, b(\eta - \eta_-) \le c$, so

$$f_+(\eta_+^\star) - f_+(\eta) = \Delta\big(1 + b(\eta - \eta_-)\big) \ge \Delta \qquad \forall\, \eta \in B_-, \tag{48}$$

and symmetrically for f_- with $B_\pm$ swapped. Off both bumps $f_\varepsilon \equiv c$, so any off-bump query has regret exactly Δ.

Closing the argument. *Occupancy regret.* Let $N_- = \sum_r \mathbf{1}\{\hat\eta_r \in B_-\}$, $N_+ = \sum_r \mathbf{1}\{\hat\eta_r \in B_+\}$, and $N_\circ = N - N_- - N_+$. By (48) and the exact off-bump regret Δ,

$$\mathrm{Reg}_{\mathcal{A}} \ge \Delta\big(N_- + \tfrac{1}{2}N_\circ\big) \text{ on } \mathcal{I}_+, \qquad \mathrm{Reg}_{\mathcal{A}} \ge \Delta\big(N_+ + \tfrac{1}{2}N_\circ\big) \text{ on } \mathcal{I}_-, \tag{49}$$

where $(N_- + \frac{1}{2}N_\circ) + (N_+ + \frac{1}{2}N_\circ) = N$ exactly.

KL divergence. Round r produces the k i.i.d. verdicts $U_1^{(r)}, \dots, U_k^{(r)} \sim \mathrm{Ber}\big(u_\varepsilon(\hat\eta_r)\big)$, where $u_\varepsilon(\eta) := u_{q_\eta^\star}$ under the realized verifier U_ε of Lemma 7; these verdicts are the learner's only observation, and $\hat\eta_r = \hat\eta_r(\text{past})$ is the same measurable map under both worlds. By the chain rule for the interaction and independence of the k draws within a round,

$$\mathrm{KL}\big(\mathbb{P}_+^N \,\big\|\, \mathbb{P}_-^N\big) = \sum_{r=1}^{N} \mathbb{E}_+\Big[\, k\,\mathrm{KL}\big(\mathrm{Ber}(u_+(\hat\eta_r)) \,\big\|\, \mathrm{Ber}(u_-(\hat\eta_r))\big)\Big].$$

It remains to bound the per-round divergence. Write $d(\eta) := u_+(\eta) - u_-(\eta)$. By (47) and Lemma 7, $d(\eta) = 2\Delta\psi(\eta)/J + \rho(\eta)$ where the remainder ρ collects the $O(\Delta^2)$ linearization term of (47) and the Lemma 7 error, both $O(\Delta/k)$; choosing $\Delta \le \Delta_\star$ makes $\|\rho\|_\infty \le \frac{1}{2}\cdot 2\Delta/J$, so with $\|\psi\|_\infty \le 1$ and $J \ge kc_0$ (Lemma 6),

$$|d(\eta)| \;\le\; \frac{3\Delta}{J} \;\le\; \frac{3\Delta}{kc_0}, \qquad \eta \in [0, \eta_{\max}].$$

By Lemma 6 and $\|g_\varepsilon\|_\infty,\ \|\text{error}\|_\infty = O(\Delta/k)$, for $\Delta \le \Delta_\star$ we also have $u_-(\eta) \in [c_0/(2k),\, 1 - c_0/2]$, hence $u_-(1-u_-) \ge c_0^2/(4k)$. Using $\mathrm{KL}(\mathrm{Ber}(p)\|\mathrm{Ber}(q)) \le (p-q)^2/(q(1-q))$,

$$k\,\mathrm{KL}\big(\mathrm{Ber}(u_+(\eta)) \,\big\|\, \mathrm{Ber}(u_-(\eta))\big) \;\le\; \frac{k\,d(\eta)^2}{u_-(1-u_-)} \;\le\; \frac{k\cdot 9\Delta^2/(k^2c_0^2)}{c_0^2/(4k)} \;=\; \frac{36}{c_0^4}\,\Delta^2 \;=:\; C_0\,\Delta^2,$$

uniformly in η (hence in $\hat\eta_r$), with $C_0 = 36/c_0^4$ depending only on c_0 — the budget k cancels, precisely because $u_0 = \Theta(1/k)$ (Lemma 6) shrinks the per-round information at the same rate the per-sample gap does. Therefore

$$\mathrm{KL}\big(\mathbb{P}_+^N \,\big\|\, \mathbb{P}_-^N\big) \;\le\; C_0\, N\, \Delta^2. \tag{50}$$

Remark (bandit sufficiency). The pass bit $\Pi_r = 1 - \prod_{j=1}^k (1 - U_j^{(r)}) \sim \mathrm{Ber}(f_\varepsilon(\hat\eta_r))$ is a deterministic function of the round's verdicts, so by data processing $\mathrm{KL}(\mathrm{Ber}(f_+)\|\mathrm{Ber}(f_-)) \le k\,\mathrm{KL}(\mathrm{Ber}(u_+)\|\mathrm{Ber}(u_-))$; conversely both sides are $\Theta(\Delta^2)$ because the flat base forces $u_0 = \Theta(1/k)$ (successes are rare, so Π_r is a sufficient statistic up to constants). Thus (50) is unchanged up to the constant C_0 whether the learner reads all k verdicts or only the round's pass bit; we bound the former, which is the richer of the two feedback models.

Bretagnolle–Huber. Let $A := \{N_- + \frac{1}{2}N_\circ \ge N/4\}$; since the two counts sum to N, on A^c we have $N_+ + \frac{1}{2}N_\circ > 3N/4$. By the Bretagnolle–Huber inequality (see Bretagnolle and Huber (1979)) and (50),

$$\mathbb{P}_+(A) + \mathbb{P}_-(A^c) \ge \tfrac{1}{2}\exp\big(-\mathrm{KL}(\mathbb{P}_+^N\|\mathbb{P}_-^N)\big) \ge \tfrac{1}{2}\exp\big(-C_0 N\Delta^2\big).$$

Choose $\Delta = (2C_0N)^{-1/2} \wedge (c_0/2)$. For $N \ge N_0$, where $N_0 = N_0(c_0, k, M)$ is large enough that both $\Delta = (2C_0N)^{-1/2}$ (so the minimum is the first term) and $\Delta \le \Delta_\star$ hold, we get $C_0N\Delta^2 = \frac{1}{2}$ and

$$\mathbb{P}_+(A) + \mathbb{P}_-(A^c) \;\ge\; \tfrac{1}{2}e^{-1/2} \;\ge\; 0.30. \tag{51}$$

By (49), on A the $\mathcal{I}_+$-regret is $\ge \Delta N/4$ and on A^c the $\mathcal{I}_-$-regret is $\ge \Delta N/4$. Passing to the maximum via the average and using (51),

$$\max\{\mathbb{E}_+[\mathrm{Reg}], \mathbb{E}_-[\mathrm{Reg}]\} \ge \tfrac{1}{2}\big(\mathbb{E}_+[\mathrm{Reg}] + \mathbb{E}_-[\mathrm{Reg}]\big) \ge \frac{\Delta N}{8}\big(\mathbb{P}_+(A) + \mathbb{P}_-(A^c)\big) \ge \frac{0.30\,\Delta N}{8}.$$

Substituting $\Delta = (2C_0N)^{-1/2}$, so that $\Delta N = \sqrt{N/(2C_0)}$,

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta\big[\mathrm{Reg}_{\mathcal{A}}(N)\big] \;\ge\; \max\{\mathbb{E}_+[\mathrm{Reg}], \mathbb{E}_-[\mathrm{Reg}]\} \;\ge\; \frac{0.30}{8\sqrt{2C_0}}\sqrt{N} \;=:\; c\sqrt{N}$$

for all $N \ge N_0$, with $c = 0.30/(8\sqrt{2C_0}) > 0$ depending only on c_0. This is Theorem 4. ■

Remark 13 (Interpretation). Dividing the cumulative lower bound by N gives a worst-case average pass@k discrepancy of order $\Omega(N^{-1/2})$. This statistical barrier is separate from the deterministic prefix-tracking discrepancy controlled by Theorem 3: even if price tracking were exact, finite terminal verifier feedback still prevents uniformly faster adaptation over the hard family.

J Proof of Theorems 5 and 6

Throughout we use the *sequence-level* calibration of Definition 1,

$$U(w^n) = s\,\bar{\ell}(w^n) + \bar{c} + \bar{r}(w^n), \qquad \bar{\ell}(w^n) = \tfrac{1}{n}\log p(w^n), \qquad \sup_{w^n} |\bar{r}(w^n)| \le \varepsilon, \tag{52}$$

and write $\mathrm{CE}(q) := -\frac{1}{n}\mathbb{E}_q[\log p(w^n)]$, $\bar{H}(q) := \frac{1}{n}\sum_t \mathbb{E}_{w^{t-1}\sim q}\, H\big(q_t(\cdot \mid w^{t-1})\big)$. Every distribution on $\mathcal{W}^n$ factorizes autoregressively, so (PE) is an optimization over $\Delta(\mathcal{W}^n)$; we use this freely. Since $p_t \ge \rho > 0$, $p(w^n) \ge \rho^n > 0$ on all of $\mathcal{W}^n$.

Lemma 8 (Sequence-level form of the budgets)**.** *For every policy q:* (a) $-\frac{1}{n}\sum_t \mathbb{E}_{w^{t-1}\sim q}\big[\sum_w q_t \log p_t\big] = \mathrm{CE}(q)$*;* (b) $\bar{H}(q) = \frac{1}{n}H(q)$*;* (c) *under* (52)*,* $u_q = \bar{c} - s\,\mathrm{CE}(q) + \mathbb{E}_q[\bar{r}]$.

Proof. (a) $\mathbb{E}_q[\log p(w^n)] = \sum_t \mathbb{E}_q[\log p_t(w_t \mid w^{t-1})] = \sum_t \mathbb{E}_{w^{t-1}\sim q}\big[\sum_w q_t(w \mid w^{t-1}) \log p_t(w \mid w^{t-1})\big]$ by the tower property. (b) Chain rule of entropy. (c) Take $\mathbb{E}_q$ in (52) and use $\mathbb{E}_q[s\bar{\ell}] = \frac{s}{n}\mathbb{E}_q \log p = -s\,\mathrm{CE}(q)$. ■

Lemma 9 (Tempered family)**.** *For $a \ge 0$ let $q_a(w^n) \propto p(w^n)^a$ and $A(a) := \log \sum_{w^n} p(w^n)^a$. Then:*

(i) $H(q_a) = A(a) - aA'(a)$ *and* $\frac{d}{da}H(q_a) = -a\,\mathrm{Var}_{q_a}(\log p)$.

(ii) *If $p(w^n)$ has a unique mode, then $H(q_0) = n\log|\mathcal{W}|$, $H(q_a) \to 0$ as $a \to \infty$, and $a \mapsto H(q_a)$ is continuous and strictly decreasing on $(0,\infty)$. Hence for every $\beta \in (0, \log|\mathcal{W}|)$ there is a unique $a_\beta > 0$ with $\bar{H}(q_{a_\beta}) = \beta$, and $\beta \mapsto a_\beta$ is continuous and strictly decreasing.*

(iii) (Entropy-constrained CE minimality.) *For every $q \in \Delta(\mathcal{W}^n)$ with $\bar{H}(q) \ge \beta$, $\mathrm{CE}(q) \ge \mathrm{CE}(q_{a_\beta})$, with equality iff $q = q_{a_\beta}$.*

Proof. (i) $\log q_a = a\log p - A(a)$ and $A'(a) = \mathbb{E}_{q_a}[\log p]$, so $H(q_a) = -\mathbb{E}_{q_a}[a\log p - A(a)] = A - aA'$; differentiating, $\frac{d}{da}H(q_a) = -aA''(a) = -a\mathrm{Var}_{q_a}(\log p)$. (ii) q_0 is uniform on $\mathcal{W}^n$ and q_a concentrates on the unique mode as $a \to \infty$. A unique mode with $|\mathcal{W}^n| \ge 2$ makes $\log p$ non-constant, and q_a has full support, so $\mathrm{Var}_{q_a}(\log p) > 0$ and (i) gives strict decrease. The claims on a_β follow from continuity and strict monotonicity, using (b) of Lemma 8. (iii) For any q, $0 \le \mathrm{KL}(q \,\|\, q_{a_\beta}) = -H(q) - a_\beta\mathbb{E}_q[\log p] + A(a_\beta)$. Hence, using $H(q) \ge n\beta = H(q_{a_\beta})$ and (i),

$$a_\beta\, n\,\mathrm{CE}(q) = -a_\beta\mathbb{E}_q[\log p] \;\ge\; H(q) - A(a_\beta) \;\ge\; H(q_{a_\beta}) - A(a_\beta) = -a_\beta\mathbb{E}_{q_{a_\beta}}[\log p] = a_\beta\, n\,\mathrm{CE}(q_{a_\beta}).$$

Divide by $a_\beta n > 0$. Equality forces $\mathrm{KL}(q\|q_{a_\beta}) = 0$. ■

Part (iii) uses no calibration: it is the maximum-entropy characterization of tempering, read as a cross-entropy minimization.

Proof of Theorem 5. With $\varepsilon = 0$, Lemma 8(c) gives $u_q = \bar{c} - s\,\mathrm{CE}(q)$. Since $\mathrm{pass@}k(q) = 1 - (1-u_q)^k$ is strictly increasing in u_q and $s > 0$, for every k program (PE) is equivalent to minimizing $\mathrm{CE}(q)$ over $\mathcal{F} := \{q : \bar{H}(q) \ge \beta,\ \mathrm{CE}(q) \le \gamma\}$. By Lemma 9(iii), $\inf\{\mathrm{CE}(q) : \bar{H}(q) \ge \beta\} = \mathrm{CE}(q_{a_\beta})$, attained only at q_{a_β}. If $\gamma < \mathrm{CE}(q_{a_\beta})$, then $\mathcal{F} = \emptyset$. If $\gamma \ge \mathrm{CE}(q_{a_\beta})$, then $q_{a_\beta} \in \mathcal{F} \subseteq \{\bar{H} \ge \beta\}$, so it is the unique minimizer over $\mathcal{F}$. This proves feasibility iff $\gamma \ge \mathrm{CE}(q_{a_\beta})$; γ enters only through feasibility.

Relation to Theorem 1. With $\lambda^\star = 0$ and $\mu^\star = na_\beta/s$, the sequence law of Corollary 1 in $(\lambda^\star, \mu^\star)$ form is $p^{\lambda^\star} e^{\mu^\star U} \propto e^{\mu^\star \frac{s}{n}\log p} = p^{a_\beta}$. This is the boundary case of Theorem 1 in which the ceiling is slack (its multiplier vanishes) and the entropy multiplier is $\mu_E = n/\mu^\star = s/a_\beta > 0$. The (T,η) parametrization degenerates there, and the exponent is carried by $\mu^\star$ alone.

Conditionals. Write $p(w^n) = p(w^t)\prod_{s>t} p_s(w_s \mid w^{s-1})$ and marginalize the future:

$$q_t^\star(w \mid w^{t-1}) = \frac{\sum_{w_{t+1:n}} p(w^n)^{a_\beta}}{\sum_{w,w_{t+1:n}} p(w^n)^{a_\beta}} \;\propto\; p_t(w \mid w^{t-1})^{a_\beta}\, \Psi_t^{(a_\beta)}(w^{t-1}w), \tag{53}$$

the prefix factor $p(w^{t-1})^{a_\beta}$ cancelling, with $\Psi_t^{(a)}$ as in (11) and $\Psi_n^{(a)} \equiv 1$. Under prefix-independent conditionals, $\Psi_t^{(a)} = \prod_{s>t} \sum_{w_s} p_s(w_s)^a$ does not depend on the prefix or on w, leaves the normalizer, and (53) reduces to (12).

At $s = 0$, $u_q \equiv \bar{c}$ and every feasible q is optimal. Finally, $\beta \mapsto a_\beta$ is strictly decreasing by Lemma 9(ii), which is the monotonicity used after Theorem 5. ■

Remark 14 (Why sequence-level scoring). Lemma 8(c) is exact because the calibration score $\bar{\ell}$ is the same functional as the plausibility ceiling, by Lemma 8(a). A token-level hypothesis $V_t(w) = s \log p_t(w) + c_t$ does not yield it. Under (52) with $\varepsilon = 0$, the potential of Lemma 1 at the prices above is $V_t(w \mid w^{t-1}) = \bar{c} + \frac{s}{n} \log p(w^{t-1}) + \frac{s}{n} \log p_t(w \mid w^{t-1}) + \frac{1}{\mu^\star} \log \Psi_t^{(a_\beta)}(w^{t-1}w)$: affine in $\log p_t$ with slope s/n, plus the remaining-horizon term. Sequence-level scoring is also exactly what (13) estimates.

Lemma 10 (Sharp mean-difference inequality)**.** *For probability measures P, Q on a common space and any bounded g,* $\left|\mathbb{E}_P[g] - \mathbb{E}_Q[g]\right| \le \left(\sup g - \inf g\right) \mathrm{TV}(P, Q)$.

Proof. For any constant c, $\mathbb{E}_P[g] - \mathbb{E}_Q[g] = \sum_x (P - Q)(x)(g(x) - c)$. Take $c = \frac{1}{2}(\sup g + \inf g)$, so $|g - c| \le \frac{1}{2}(\sup g - \inf g)$, and use $\sum_x |P(x) - Q(x)| = 2\mathrm{TV}(P, Q)$. ■

Proof of Theorem 6. Let $\tilde{q} := q_{a_\beta}$, which is defined for $s \ge 0$ since it depends only on (p, β).

(i) By Lemma 9(ii), $\bar{H}(\tilde{q}) = \beta$, and $\mathrm{CE}(\tilde{q}) \le \gamma$ by hypothesis, so $\tilde{q}$ is feasible and $u_{q^\star} \ge u_{\tilde{q}}$. The optimum $q^\star$ is feasible, so $\bar{H}(q^\star) \ge \beta$, and Lemma 9(iii) gives $\mathrm{CE}(q^\star) \ge \mathrm{CE}(\tilde{q})$. By Lemma 8(c) and $s \ge 0$,

$$u_{q^\star} - u_{\tilde{q}} = s\big[\mathrm{CE}(\tilde{q}) - \mathrm{CE}(q^\star)\big] + \mathbb{E}_{q^\star}[\bar{r}] - \mathbb{E}_{\tilde{q}}[\bar{r}] \;\le\; \mathbb{E}_{q^\star}[\bar{r}] - \mathbb{E}_{\tilde{q}}[\bar{r}] \;\le\; 2\varepsilon.$$

Lemma 10 sharpens the last step to $2\varepsilon\, \mathrm{TV}(q^\star, \tilde{q})$.

(ii) The map $u \mapsto 1 - (1-u)^k$ has derivative $k(1-u)^{k-1}$, positive and non-increasing in u. By the mean value theorem there is $\xi \in [u_{\tilde{q}}, u_{q^\star}]$ with $\mathrm{pass@}k(q^\star) - \mathrm{pass@}k(\tilde{q}) = k(1-\xi)^{k-1}(u_{q^\star} - u_{\tilde{q}})$. Since $\xi \ge u_{\tilde{q}} \ge u_{q^\star} - 2\varepsilon$ by (i), $k(1-\xi)^{k-1} \le \sigma_k^+$. Combined with (i), the gap lies in $[0, 2\varepsilon\sigma_k^+]$.

(iii) For $v \in (0,1)$ let $\sigma_v(k) := k(1-v)^{k-1}$, $k > 0$. Then $\frac{d}{dk} \log \sigma_v(k) = \frac{1}{k} + \log(1-v)$, which is positive for $k < k_v := -1/\log(1-v)$ and negative after. So σ_v is strictly unimodal with maximizer k_v, and $\sigma_v(k) \to 0$ as $k \to 0^+$ and as $k \to \infty$. With $v = u_{q^\star} - 2\varepsilon \in (0,1)$ we get $\sigma_k^+ = \sigma_v(k)$, peaking at $k_v = 1/v + O(1)$. Hence for every $\delta > 0$ the set $\{k : 2\varepsilon\sigma_k^+ > \delta\}$ is a bounded (possibly empty) interval, and outside it $\tilde{q}$ is δ-optimal by (ii). ■

K Proof of Theorem 7: suboptimality and certification of CTD

We keep the notation of Appendix J. The deployed policy is $\hat{q}_{\mathrm{CTD}} = q_{\hat{a}}^{\mathrm{temp}}$, per-token temperature at the Phase-2 exponent. The regression of Phase 1 does *not* enter the policy: $\hat{a}$ depends only on (p, β). The proof therefore has a deterministic part, the $\mathrm{pass@}k$ bound, and a stochastic part, the probability that Phase 1 certifies.

Part A: the deterministic bound. The proof uses only the following property of Phase 2:

$$\bar{H}(\hat{q}_{\mathrm{CTD}}) \;\ge\; \beta, \qquad \mathrm{CE}(\hat{q}_{\mathrm{CTD}}) \le \gamma, \tag{54}$$

which holds with equality in the entropy when $\hat{a}$ is the exact root. Then:

(a) $\hat{q}_{\mathrm{CTD}}$ is feasible, so $u_{q^\star} \ge u_{\hat{q}_{\mathrm{CTD}}}$.

(b) By Lemma 9(iii) applied to $\hat{q}_{\mathrm{CTD}}$, $\Delta_{\mathrm{CE}} = \mathrm{CE}(\hat{q}_{\mathrm{CTD}}) - \mathrm{CE}(q_{a_\beta}) \ge 0$, with equality iff $\hat{q}_{\mathrm{CTD}} = q_{a_\beta}$.

(c) Applied to $q^\star$, which is feasible, the same lemma gives $\mathrm{CE}(q^\star) \ge \mathrm{CE}(q_{a_\beta})$.

By Lemma 8(c) and $s \ge 0$,

$$u_{q^\star} - u_{\hat{q}_{\mathrm{CTD}}} = s\big[\mathrm{CE}(\hat{q}_{\mathrm{CTD}}) - \mathrm{CE}(q^\star)\big] + \mathbb{E}_{q^\star}[\bar{r}] - \mathbb{E}_{\hat{q}_{\mathrm{CTD}}}[\bar{r}] \;\le\; s\big[\mathrm{CE}(\hat{q}_{\mathrm{CTD}}) - \mathrm{CE}(q_{a_\beta})\big] + 2\varepsilon = s\Delta_{\mathrm{CE}} + 2\varepsilon.$$

As in Theorem 6(ii), the mean value theorem with $\xi \ge u_{q^\star} - (s\Delta_{\mathrm{CE}} + 2\varepsilon)$ gives $0 \le \mathrm{pass@}k(q^\star) - \mathrm{pass@}k(\hat{q}_{\mathrm{CTD}}) \le \sigma_k^{++}(s\Delta_{\mathrm{CE}} + 2\varepsilon)$, which is (14).

When $\Delta_{\mathrm{CE}} = 0$. Under prefix-independent conditionals, $\Psi_t^{(a)}$ is token-free for every a, so by (53) $q_a^{\mathrm{temp}} = q_a$ for all a. Then $\bar{H}(\hat{q}_{\mathrm{CTD}}) = \beta$ forces $\hat{a} = a_\beta$ by Lemma 9(ii), and $\Delta_{\mathrm{CE}} = 0$. In general, Δ_{CE} is the cross-entropy price, at fixed entropy, of dropping the remaining-horizon factor $\Psi_t^{(a)}$.

Remark 15 (Monte-Carlo root in Phase 2). Part A needs only (54), not an exact root. If Phase 2 estimates $\bar{H}(q_a^{\mathrm{temp}})$ to accuracy τ and bisects on the conservative target $\beta + \tau$, then (54) holds on the accuracy event, and the extra slack enters the bound only through a larger Δ_{CE}.

Part B: certification probability. Write $s_{\mathrm{spr}}^2 = \mathrm{Var}(\bar{\ell}) \geq s_0^2$ and $s_{\mathrm{pop}} := \mathrm{Cov}(\bar{\ell}, U)/\mathrm{Var}(\bar{\ell})$.

Population slope (Proposition 1). By (52), $\mathrm{Cov}(\bar{\ell}, U) = s\mathrm{Var}(\bar{\ell}) + \mathrm{Cov}(\bar{\ell}, \bar{r})$. Since $|\bar{r}| \leq \varepsilon$, Cauchy–Schwarz gives $|\mathrm{Cov}(\bar{\ell}, \bar{r})| \leq s_{\mathrm{spr}}\varepsilon$. Hence

$$|s_{\mathrm{pop}} - s| \leq \varepsilon/s_0 =: C_0\varepsilon, \qquad |s_{\mathrm{pop}}| \leq \tfrac{\sqrt{\mathrm{Var}U}}{s_{\mathrm{spr}}} \leq \tfrac{1}{2s_0}. \tag{55}$$

Concentration. The empirical slope (13) is $\hat{s} = \widehat{\mathrm{Cov}}/\widehat{\mathrm{Var}}$, a function of four empirical means: those of $\bar{\ell} \in [-B, 0]$, $U \in \{0, 1\}$, $\bar{\ell}U \in [-B, 0]$, and $\bar{\ell}^2 \in [0, B^2]$. By Hoeffding's inequality and a union bound, with probability at least $1 - \delta/2$,

$$|\widehat{\mathrm{Cov}} - \mathrm{Cov}| \vee |\widehat{\mathrm{Var}} - \mathrm{Var}| \leq \eta_M := c_1\bar{B}^2\sqrt{\tfrac{\log(8/\delta)}{N}}, \qquad \bar{B} := \max\{1, B\},$$

with c_1 absolute. On this event, if $\eta_M \leq s_0^2/2$, then $\widehat{\mathrm{Var}} \geq s_0^2/2$, and

$$|\hat{s} - s_{\mathrm{pop}}| = \Big|\tfrac{\widehat{\mathrm{Cov}} - \mathrm{Cov}}{\widehat{\mathrm{Var}}} + s_{\mathrm{pop}}\tfrac{\mathrm{Var} - \widehat{\mathrm{Var}}}{\widehat{\mathrm{Var}}}\Big| \leq \frac{2\eta_M(1 + |s_{\mathrm{pop}}|)}{s_0^2} \leq \frac{C_s}{2}\sqrt{\tfrac{\log(8/\delta)}{N}}, \qquad C_s := \frac{4c_1\bar{B}^2}{s_0^2}\Big(1 + \frac{1}{2s_0}\Big).$$

Sign. Suppose $s > C_0\varepsilon$ and choose $\delta := 8\exp\big(-N(s - C_0\varepsilon)^2/C_s^2\big)$, so that $\frac{C_s}{2}\sqrt{\log(8/\delta)/N} = \frac{1}{2}(s - C_0\varepsilon)$. The side condition holds automatically: by (55),

$$\eta_M = c_1\bar{B}^2(s - C_0\varepsilon)/C_s \leq c_1\bar{B}^2|s_{\mathrm{pop}}|/C_s < s_0^2/8.$$

On the event, $\hat{s} \geq s_{\mathrm{pop}} - \frac{1}{2}(s - C_0\varepsilon) \geq \frac{1}{2}(s - C_0\varepsilon) > 0$, using $s_{\mathrm{pop}} \geq s - C_0\varepsilon$. Hence

$$\mathbb{P}(\hat{s} \leq 0) \;\leq\; \delta/2 \;\leq\; 8\exp\big(-N(s - C_0\varepsilon)^2/C_s^2\big),$$

with $C_0 = 1/s_0$ and C_s depending only on (B, s_0). This is the second claim of Theorem 7. ■

Remark 16 (What the estimate does and does not control). Unlike an analysis in which $\hat{s}$ sets the exponent, the policy here does not depend on $\hat{s}$. As a result there is no $O(1/\sqrt{N})$ term in (14) and no horizon factor to cancel. The estimate $(\hat{s}, \hat{\varepsilon})$ serves two purposes: it decides whether the calibrated regime applies, and $\hat{\varepsilon}$ provides an empirical residual diagnostic for the deployment gate.

L The optimal value $\mathrm{pass@}k(q^\star)$

Theorem 1 fixes the *form* of the optimal decoder; here we compute the *value* it attains, $\mathrm{pass@}k(q^\star) = 1-(1-u_{q^\star})^k$ with $u_{q^\star} = \mathbb{E}_{w^n \sim q^\star}[U(w^n)]$. This provides an exact yardstick against which both estimators are scored, without requiring Monte Carlo estimation of the optimum. Two facts organize the computation. The potential V of Lemma 1 depends only on $(p, U, \lambda^\star, \mu^\star)$ and on no policy, so nothing here requires a fixed point; and since the tilt carries V while the value averages U — objects that differ by a free energy (Corollary 3) — $u_{q^\star}$ is *not* a single softmax recursion but the second of two decoupled backward sweeps. At the root the sequence-level tilt (Corollary 1), $q^\star(w^n) \propto p(w^n)^{\lambda^\star}e^{\mu^\star U(w^n)}$, makes $(U, \log p)$ an exponential family, which pins $u_{q^\star}$ in closed form. We then specialize twice: to a binary verifier, where the value is a logistic in the tilt, and to the count-threshold instances, where it is an $O(nM)$ dynamic program.

Two backward sweeps. Index the potential by the completed prefix, $V_t(w^t) := V_t(w_t \mid w^{t-1})$, so $V_n(w^n) = U(w^n)$ and $V_0(\varnothing) := \frac{1}{\mu^\star}\log Z_1$. Lemma 1 reads, for $0 \leq t \leq n-1$,

$$V_t(w^t) = \tfrac{1}{\mu^\star}\log Z_{t+1}(w^t), \qquad Z_{t+1}(w^t) = \sum_{w' \in \mathcal{W}} p_{t+1}(w' \mid w^t)^{\lambda^\star}e^{\mu^\star V_{t+1}(w^t w')} = e^{\mu^\star V_t(w^t)}, \tag{56}$$

and the policy is

$$q_{t+1}^\star(w' \mid w^t) = p_{t+1}(w' \mid w^t)^{\lambda^\star}\, e^{\mu^\star(V_{t+1}(w^t w') - V_t(w^t))}. \tag{57}$$

Equation (57) is worth reading slowly: the optimal decoder is the tempered base model reweighted by the *increment* of the potential, the term $\mu^\star V_t(w^t)$ acting as the log-normalizer at the node. Both objects in (56) are computable from the instance alone.

Proposition 5 (Exact value: two decoupled sweeps)**.** *Define the pass cost-to-go* $u_{n+1}(w^n) = U(w^n)$, $u_t(w^{t-1}) = \sum_w q_t^\star(w \mid w^{t-1})\, u_{t+1}(w^{t-1}w)$. *Then* $\mathrm{pass@}k(q^\star) = 1 - (1 - u_1(\varnothing))^k$. *Sweep* (56) *does not involve* u *and the value backup does not feed back into* V*; each is a finite acyclic pass, so the pair is exact in two sweeps with no iteration.*

Proof. The backup is the tower property for the fixed sampler $q^\star$; by Lemma 1, V (hence $q^\star$) is a deterministic function of $(p, U, \lambda^\star, \mu^\star)$, so (56) closes without u, and the k draws are i.i.d. under $q^\star$. ■

The two sweeps must be kept apart. Folding them into the single recursion $u_t \propto \sum_w p_t^{\lambda^\star} e^{\mu^\star u_{t+1}} u_{t+1}$ is invalid once $V^\star \neq J$: by Corollary 3, $V_{t-1} = \mathbb{E}_{q_t^\star}[V_t] + \frac{1}{\mu^\star}[H(q_t^\star) - \lambda^\star \mathrm{CE}(q_t^\star, p_t)]$ exceeds its own backup by a free energy, so V is not the pass probability. The two objects are distinct in general; at the terminal step the free-energy continuation reduces directly to U, while degenerate-price limits should be interpreted through the product $\mu^\star V_t$ as in Appendix B. The first sweep is a log-sum-exp over children, the second a plain average; only the first sits in an exponent.

Free-energy form. Let $\mathcal{A}(\lambda, \mu) := \log \sum_{w^n} p(w^n)^\lambda e^{\mu U(w^n)}$ be the cumulant generating function of the pair $(\log p, U)$ over $\mathcal{W}^n$, so $\mathcal{A}(\lambda^\star, \mu^\star) = \log Z_1 = \mu^\star V_0(\varnothing)$ by (56) at $t = 0$: the root potential *is* the sequence free energy, obtained by one sweep with no enumeration of $\mathcal{W}^n$.

Lemma 11 (Dual gradients)**.** *For* $q^{\lambda,\mu} \propto p^\lambda e^{\mu U}$, $\partial_\mu \mathcal{A} = u_{q^{\lambda,\mu}}$, $\partial_\lambda \mathcal{A} = \mathbb{E}[\log p] = -\overline{\mathrm{CE}}$, $\partial_\mu^2 \mathcal{A} = \mathrm{Var}(U) \geq 0$, $\partial_\lambda^2 \mathcal{A} = \mathrm{Var}(\log p) \geq 0$*; in particular* $u_{q^\star} = \partial_\mu \mathcal{A}|_{(\lambda^\star, \mu^\star)}$.

Proof. $q^{\lambda,\mu} = \exp(\lambda \log p + \mu U - \mathcal{A})$; differentiating $\sum q^{\lambda,\mu} = 1$ once gives the means, twice the variances, and $\log p(w^n) = \sum_t \log p_t$ gives $-\overline{\mathrm{CE}}$. ■

So the value is a dual gradient of the same free energy the decoder normalizes by; the free energy is convex and the value is non-decreasing in the tilt. The next identity converts this into a quantity the root-solve already produces.

Proposition 6 (Value = root potential minus budget)**.** *Here* $\bar{H} := \sum_{t=1}^n \mathbb{E}_{q^\star}[H(q_t^\star)]$ *and* $\overline{\mathrm{CE}} := -\sum_{t=1}^n \mathbb{E}_{q^\star} \sum_w q_t^\star(w \mid w^{t-1}) \log p_t(w \mid w^{t-1})$ *denote the unnormalized horizon totals. With* $\frac{1}{n}\bar{H} = \beta$ *and* $\frac{1}{n}\overline{\mathrm{CE}} = \gamma$ *the active budgets of (PE),*

$$u_{q^\star} = V_0(\varnothing) - \frac{n}{\mu^\star}\big(\beta - \lambda^\star \gamma\big), \qquad \mathrm{pass@}k(q^\star) = 1 - (1 - u_{q^\star})^k. \tag{58}$$

Proof. Averaging $\log$ (57) under $q_t^\star(\cdot \mid w^{t-1})$ gives Corollary 3; taking the $q^\star$-occupancy expectation and summing over t telescopes to $V_0(\varnothing) = u_{q^\star} + \frac{1}{\mu^\star}[\bar{H} - \lambda^\star \overline{\mathrm{CE}}]$, using $V_n = U$. Substitute $\bar{H} = n\beta$, $\overline{\mathrm{CE}} = n\gamma$. ■

Every term on the right comes from the same root-solve that returns (T, η), so comparing (58) with the forward value sweep of Proposition 5 is a free implementation check: the two must agree to machine precision.

Binary verifier: a logistic in the tilt. When $U \in \{0, 1\}$ the sequence space splits into a passing and a failing part, and only the tempered mass of each survives. For $\lambda > 0$ put

$$P_\lambda := \sum_{w^n :\, U(w^n)=1} p(w^n)^\lambda, \qquad Q_\lambda := \sum_{w^n :\, U(w^n)=0} p(w^n)^\lambda, \qquad \ell_\lambda := \log \frac{P_\lambda}{Q_\lambda}. \tag{59}$$

The scalar ℓ_λ is the *tempered log-odds of passing*: writing $\pi_\lambda(w^n) \propto p(w^n)^\lambda$ for pure sequence-level tempering at $T = 1/\lambda$ (no verifier tilt), one has $\mathbb{P}_{\pi_\lambda}[U = 1] = P_\lambda/(P_\lambda + Q_\lambda)$ and hence $\ell_\lambda = \mathrm{logit}\mathbb{P}_{\pi_\lambda}[U = 1]$. At $\lambda = 1$ it is the base model's own log-odds of solving the task; ℓ_λ is finite exactly when both classes carry mass, which is the only non-degeneracy needed below. Two cautions: ℓ_λ is a single number attached to the whole instance and is sequence-level (it grows like n through $\log p$), so it is not a per-token quantity; and despite the shared letter it is unrelated to the per-sequence score $\bar{\ell}(w^n) = \frac{1}{n} \log p(w^n)$ of Definition 1, which reappears in Corollary 5.

Theorem 11 (Exact value, binary verifier)**.** *Let $U \in \{0,1\}$ and $\sigma(x) = 1/(1+e^{-x})$. Then $\mathcal{A}(\lambda,\mu) = \log(Q_\lambda + e^\mu P_\lambda)$, $u_{q^{\lambda,\mu}} = \sigma(\mu + \ell_\lambda)$, and*

$$u_{q^\star} = \sigma(\mu^\star + \ell_{\lambda^\star}), \qquad \text{pass@}k(q^\star) = 1 - \big(1 - \sigma(\mu^\star + \ell_{\lambda^\star})\big)^k, \tag{60}$$

with exact sensitivities $\partial_\mu u = u(1-u) > 0$ and $\partial_\lambda u = u(1-u)\big(\mathbb{E}_{\pi_\lambda^+}[\log p] - \mathbb{E}_{\pi_\lambda^-}[\log p]\big)$, where $\pi_\lambda^\pm$ are p^λ restricted to the passing / failing sequences and normalized.

Proof. Split $\mathcal{A}$ by the value of U; then $u = \partial_\mu \mathcal{A} = e^\mu P_\lambda/(Q_\lambda + e^\mu P_\lambda) = \sigma(\mu + \ell_\lambda)$, $\partial_\mu u = \sigma' = u(1-u)$, and $\partial_\lambda u$ follows by chaining through ℓ_λ with $\partial_\lambda \log P_\lambda = \mathbb{E}_{\pi_\lambda^+}[\log p]$. ■

Read (60) as an additive decomposition on the logit scale: temperature sets the prior log-odds $\ell_{\lambda^\star}$ and the verifier tilt shifts it by exactly $\mu^\star$. The two prices therefore act on the same axis but through different channels — $\mu^\star$ enters with unit slope, while $\lambda^\star$ enters only through how tempering redistributes mass between the passing and failing classes, which is the difference of means in $\partial_\lambda u$ and can vanish.

Corollary 5 (The λ-sensitivity is CTD's regression slope)**.** *With $\bar\ell = \frac{1}{n}\log p$, $\partial_\lambda u = n\,u(1-u)\big(\mathbb{E}_{\pi_\lambda^+}[\bar\ell] - \mathbb{E}_{\pi_\lambda^-}[\bar\ell]\big)$; at $\lambda = 1$ this difference equals $s\mathrm{Var}(\bar\ell)/(\bar u(1-\bar u)) + O(\varepsilon)$ under (s,ε)-calibration (point-biserial), the population form of CTD's OLS slope $\hat s$. A near-zero slope therefore indicates weak local sensitivity of the value to temperature, consistent with flat CTD curves on HumanEval and MBPP.*

(Without binariness the logistic is lost, but $u_{q^\star} = \partial_\mu \mathcal{A}$ and $\partial_\mu u = \mathrm{Var}_{q^\star}(U) \geq 0$ still makes the value increasing in the tilt.)

Count-threshold benchmarks. For these count-threshold instances, the exact optimum is available and can be used to score both estimators. The instance carries, at each step t, a *good set* $G_t \subseteq \mathcal{W}$: the tokens that count as progress at that step (in this construction, the tokens whose merit clears the threshold; G_t is part of the specification, not learned). A sequence passes iff at least M of its n tokens are good,

$$U(w^n) = \mathbf{1}\big[\,|\{t : w_t \in G_t\}| \geq M\,\big], \tag{61}$$

so $M = 1$ is "at least one correct step" and $M = n$ is pure AND. Assume in addition that both the base conditionals and the good sets are prefix-free, $p_t(\cdot \mid w^{t-1}) = p_t(\cdot)$ and G_t depending on t alone. Then no feature of a prefix matters except the running count $c_t := |\{s \leq t : w_s \in G_s\}|$, which is a sufficient statistic for both sweeps: V and u collapse from functions on $\mathcal{W}^t$ to functions of $c \in \{0,\dots,t\}$. Writing the tempered good and bad masses at step t as $P_t^{(\lambda)} = \sum_{w\in G_t} p_t(w)^\lambda$, $Q_t^{(\lambda)} = \sum_{w\notin G_t} p_t(w)^\lambda$ and $C_t^{(\lambda)} = P_t^{(\lambda)} + Q_t^{(\lambda)}$, the two sweeps of Proposition 5 become

$$S_t(c) = \tfrac{1}{\mu^\star}\log\big(P_{t+1}^{(\lambda^\star)} e^{\mu^\star S_{t+1}(c+1)} + Q_{t+1}^{(\lambda^\star)} e^{\mu^\star S_{t+1}(c)}\big),$$

$$g_t(c) = \frac{P_{t+1}^{(\lambda^\star)} e^{\mu^\star S_{t+1}(c+1)}}{P_{t+1}^{(\lambda^\star)} e^{\mu^\star S_{t+1}(c+1)} + Q_{t+1}^{(\lambda^\star)} e^{\mu^\star S_{t+1}(c)}},$$

$$v_t(c) = g_t(c)\,v_{t+1}(c+1) + (1 - g_t(c))\,v_{t+1}(c), \qquad S_n(c) = v_n(c) = \mathbb{1}[c \geq M], \qquad u_{q^\star} = v_0(0).$$

Here $S_t(c)$ is the potential of a state with count c after t tokens, $g_t(c)$ is the probability that the optimal decoder emits a *good* token from that state, and $v_t(c)$ is the pass value. The division of labour of Proposition 5 is visible in the displays: S log-sum-exps, v averages, and only S appears in the exponent of g. Cost is $O(nM)$ time and $O(M)$ memory, all k at once, with no Monte-Carlo.

For pure AND ($M = n$) the count collapses to alive/dead. A dead branch still carries the continuation free energy $e^{\mu^\star S_{t+1}^{\mathrm{d}}} = \prod_{s>t+1} C_s^{(\lambda^\star)} \neq 1$ (that is the *potential*; the pass *value* on a dead branch is 0), and

$$u_{q^\star} = \prod_{t=0}^{n-1} g_t^\star, \qquad g_t^\star = \frac{P_{t+1}^{(\lambda^\star)} e^{\mu^\star S_{t+1}^{\mathrm{a}}}}{P_{t+1}^{(\lambda^\star)} e^{\mu^\star S_{t+1}^{\mathrm{a}}} + Q_{t+1}^{(\lambda^\star)} e^{\mu^\star S_{t+1}^{\mathrm{d}}}}, \tag{62}$$

$S_n^{\mathrm{a}} = 1,\ S_n^{\mathrm{d}} = 0$. Replacing $e^{\mu^\star S_{t+1}^{\mathrm{d}}}$ by 1 mis-scales the bad-token weight in (57), so the recovered decoder is not $q^\star$ and its deployed pass@k falls short of the optimum — by an order of magnitude when the discarded product of normalizers is far from 1; reference curves built on it must be recomputed. On a dead branch (57) reduces to pure temperature $p^{\lambda^\star}$, so the optimum is state-dependent even at $M = n$ and the prefix-free Poisson–binomial score $u_q = \mathbb{P}[\mathrm{PoiBin}(g_1,\dots,g_n) \geq M]$ does not substitute for it.

M Additional Experimental Details

In this appendix we provide additional experimental results and details for our practical Algorithm 3.

Algorithm 3 TMAD-Gated: continuation-value fitting and gated decoding.

Require: base model p, verifier U, reference temperature $T^\star$, calibration prompts, gate g, conservative temperature T_{conf}, regime set $\mathcal{R} = \{(T_j, \eta_j)\}_{j=1}^{J}$

Offline: collect fixed-$T^\star$ rollouts and fit $\hat{S}_\theta(h, w)$ to predict eventual verifier success.
for $t = 1, \ldots, n$ **do**
  compute $p_t(\cdot \mid w^{t-1})$ and mark t active if confidence is low or the position is structurally critical
  if t is inactive **then**
    sample conservatively, $w_t \sim q_{\text{conf}}$ with $q_{\text{conf}}(w) \propto p_t(w \mid w^{t-1})^{1/T_{\text{conf}}}$
  else
    score candidate nucleus with $\hat{S}_t$ and refine the top candidates by short lookahead (63)
    choose $(T, \eta) \in \mathcal{R}$ using an estimated-success/plausibility score
    sample $w_t \sim q_{T,\eta}$, where $q_{T,\eta}(w) \propto \big(p_t(w) e^{\eta \hat{S}_t(w)}\big)^{1/T}$
  end if
end for
return w^n

M.1 Continuation-success head training and diagnostics

We provide here only the additional implementation details for the continuation-success head used by TMAD-GATED. Training data is collected via *fixed-temperature* rollouts: for each training problem we sample k completions at the reference temperature $T^\star$ (i.e. $\lambda_{\text{ref}} = 1/T^\star$, with no value-based tilting), and label every token position in a rollout with the rollout's realized binary outcome $U \in \{0, 1\}$ (unit-test pass/fail). To concentrate training signal where it is most informative, we restrict attention to problems for which the pooled draws contain both at least one pass and one fail, rank them by the empirical variance $\hat{p}(1 - \hat{p})$ of their pass rate $\hat{p}$, and use the top-K problems to build the training buffer.

The head takes as input $h_t = [h_t^{(L)}; \bar{h}_t] \in \mathbb{R}^{2d}$, the concatenation of the final-layer hidden state $h_t^{(L)}$ with the mean of the last four transformer-layer hidden states $\bar{h}_t$ at position t. The head is a low-rank bilinear function over the vocabulary,

$$\hat{S}_\theta(h_t, w) = \sigma\big(a_w^\top (B h_t) + b_w\big), \qquad A \in \mathbb{R}^{|\mathcal{V}| \times r},\ B \in \mathbb{R}^{r \times 2d}, \tag{63}$$

where a_w is the row of A corresponding to token w, $r \ll d$ is the bottleneck rank, and σ is the sigmoid. Parameters $\theta = (A, B, b)$ are trained with binary cross-entropy between $\hat{S}_\theta(h_t, w_t)$ and the rollout label U, using a class-balancing positive weight, Adam, and gradient clipping. We hold out a disjoint subset of *problems*, rather than token positions, for early stopping and select the checkpoint with the lowest held-out binary cross-entropy.

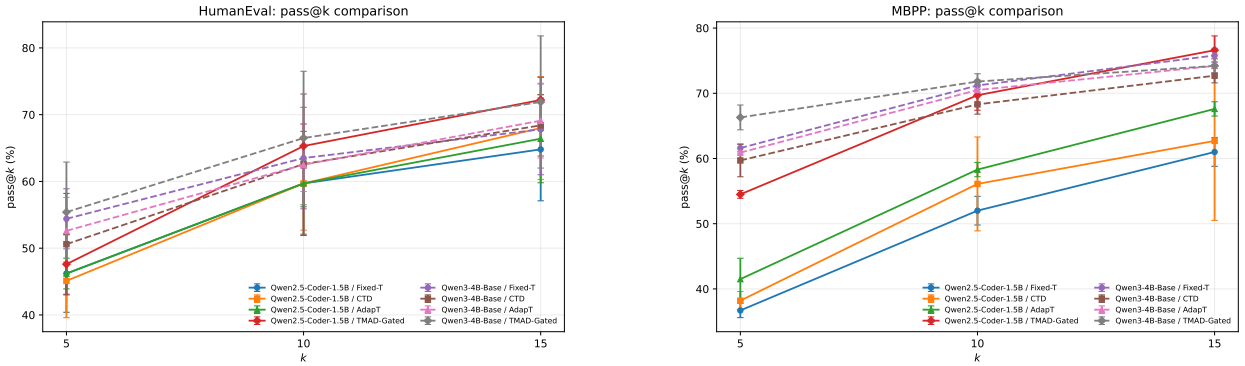
TMAD-Gated decoding details. We use nucleus sampling with $p = 0.95$. A position is active when the maximum base-model token probability is below 0.90 or when it is a structurally critical code position under the same newline/indentation state used by our ADAPT reproduction. Inactive positions use conservative sampling at $T_{\text{conf}} = 0.40$. At an active position, the value head scores the full nucleus. We select from the finite regime grid

$$T \in \{0.75T^\star,\ T^\star,\ 1.5T^\star\}, \qquad \eta \in \{0,\ 1,\ 3\}.$$

For each regime, the verifier tilt is scaled by the normalized entropy of the current base distribution, and the regime is scored by the Lagrangian-style criterion

$$\mathbb{E}_q[\hat{S}_t^{\text{LA}}] \ + \ \kappa\, \mathbb{E}_q[\log p_t], \qquad \kappa = 0.6,$$

where $\hat{S}_t^{\text{LA}}$ denotes the lookahead-refined continuation score. The highest-scoring regime is selected and a token is sampled from its stochastic tilted-temperature distribution. Thus the current implementation uses a soft plausibility penalty rather than a hard per-step budget-feasibility filter.



(a) HumanEval. **(b)** MBPP.

Figure 4: Pass@k curves for the final controlled comparison. Mean pass@k for $k \in \{5, 10, 15\}$ across repeated runs, corresponding to Table 1. Error bars indicate the standard deviation across runs.

Because $\hat{S}_t$ is a practical proxy for continuation success rather than an exact estimate of the free-energy potential V_t, we evaluate it directly on held-out problems using ROC-AUC. This measures whether the learned score ranks token–prefix states that eventually lead to successful completions above those leading to failures. The held-out split is disjoint at the problem level from the data used to train the head.

Table 2: Held-out ROC-AUC of the continuation-success head $\hat{S}_\theta$, reported as mean $\pm$ standard deviation across runs. Higher is better; AUC measures the practical proxy, not direct error in V_t.

Model	HumanEval AUC	MBPP AUC
Qwen2.5-Coder-1.5B	0.632 ± 0.022	0.641 ± 0.007
Qwen3-4B-Base	0.648 ± 0.009	0.628 ± 0.092
CodeLlama-7B	0.646 ± 0.072	0.639 ± 0.033

M.2 Held-out evaluation split

For each model and benchmark, a disjoint subset of problems is reserved for calibration and continuation-head training, and all controlled decoding methods are evaluated on the same remaining held-out problems. Consequently, the primary matched comparison is among fixed-temperature sampling, our ADAPT reproduction, CTD, and TMAD-GATED. In all reported pass@k comparisons, methods within a model–benchmark pair are evaluated on the same held-out subset.