Information-Theoretic Foundations of Online Temperature Sampling

Jin Sima
Purdue University
simaj@purdue.edu

Nikolaos Papagiannis
Purdue University
npapagia@purdue.edu

Ananth Grama
Purdue University
ayg@purdue.edu

Wojciech Szpankowski
Purdue University
szpan@purdue.edu

Abstract

Repetitive, degenerate generation is a persistent failure of LLM decoding, and the standard remedies (temperature, nucleus, and top-k sampling) are static heuristics. By Kac's lemma and the AEP, a stationary source takes time exponential in its entropy to repeat a block. We therefore maximize the entropy of the sampling distributions under one cross-entropy budget that keeps the output plausible under the model. Offline, the solution is temperature sampling, with the inverse temperature as the budget's Lagrange multiplier. *Our contribution is the online problem*: the conditionals $p_1, p_2, \ldots$ are revealed one at a time, and the sampling distribution q_t must be chosen causally from $p_1, \ldots, p_t$ while the budget couples the entire horizon. The optimal temperature is thus a global shadow price that must be estimated from a prefix. We propose Entropic Online Sampling (EOS), which re-solves for the temperature on each prefix. We prove it loses $O(R \log n)$ entropy against the offline optimum while overshooting the budget by $O(R \log n)$, with R an instance-dependent drift parameter. A budget-tightened variant is exactly feasible at the same rate. Conversely, every exactly feasible online algorithm must lose $\Omega(R)$, so EOS is optimal up to a $\log n$ factor. When the conditionals are drawn from a finite mixture, $R = \tilde{O}(\sqrt{n})$. Experiments confirm that EOS tracks the offline temperature and beats fixed-temperature sampling in entropy.

1 Introduction

Despite the empirical success of large language models (LLMs) (Brown et al., 2020; Radford et al., 2019), their decoding procedures remain poorly understood theoretically. A central failure mode is *degenerate generation*: repetitive, low-entropy, or vacuous output. Greedy and beam search exacerbate it (Holtzman et al., 2020; Meister and Cotterell, 2023), and the standard remedies (e.g., temperature scaling, nucleus sampling, top-k truncation, repetition penalties) are heuristics adopted for their empirical utility rather than derived from any objective. Behind these heuristics lies a natural optimization problem. Write $p_t(\cdot) = P(\cdot \mid w^{t-1})$ for the model's conditional distribution at step t and q_t for the distribution actually sampled from. Diversity is measured by $\sum_{t \le n} H(q_t)$ (i.e., equivalently, for stationary ergodic sources, by the recurrence-time exponent) and plausibility by the cross-entropy of the generated stream against the model, $-\sum_{t \le n} \sum_w q_t(w) \log p_t(w)$, which we require to stay below a budget γn. Maximizing the former subject to the latter is a convex program, and its solution is the tempered (Gibbs) family, that is, temperature sampling.

The *offline* problem just described is a specialization of the classical maximum-entropy principle: a single linear constraint yields a single Lagrange multiplier, hence a single temperature. We record it in Section 3.1 only because we need it as a comparator and we do not claim it as a contribution beyond bookkeeping. What is new in the present paper is estimating that shadow price *online*, from a prefix, without access to the horizon.

The contribution of this paper is the *online* problem, in which $p_1, p_2, \ldots$ are revealed sequentially and q_t must be a function of $p_1, \ldots, p_t$ alone. What distinguishes it from standard online convex optimization (Cesa-Bianchi and Lugosi, 2006; Orabona, 2019) is not the sequential structure but the constraint: there is a single cross-entropy budget for the entire horizon, coupling all n rounds, rather than one constraint per round. A causal algorithm must therefore decide how much of the budget to spend early, without knowing what the remaining rounds will look like. This is a genuine information-theoretic obstruction, and quantifying it is the substance of the paper.

One point requires care and we make it explicit throughout: *which* algorithms are being compared. An algorithm may satisfy the global budget exactly, or may satisfy only a running version of it at every prefix and end the horizon with residual slack. These are different classes, and a lower bound proved for one says nothing about the other. We therefore state all online results with the feasibility class named, and we give both an exactly feasible algorithm and a tradeoff lower bound that interpolates between the classes.

Our contributions. We list our contributions next: **1.** *An explicit online protocol* (Section 2), specifying when the policy commits, what it observes, the comparator, and the minimax quantifier order used by the lower bound. We define the exactly feasible class and the so called V-feasible class, and identify the instance-dependent drift parameter R that governs everything. We emphasize that R is defined through the offline optimum and is therefore *not* observable to a causal algorithm.

2. *The EOS algorithm and a bicriteria guarantee* (Theorem 5). EOS chooses λ_t by solving, at each prefix, the one-dimensional monotone equation that the offline optimum would solve on the data seen so far. It loses at most $O(R \log n)$ entropy relative to the offline optimum and overshoots the global budget by at most $O(R \log n)$. We report the slack as a first-class quantity rather than suppressing it.

3. *An exactly feasible variant* (Theorem 6). Running EOS against the tightened budget $\gamma - \varepsilon$ with $\varepsilon = \Theta(R \log n/n)$ absorbs the terminal slack and yields *exact* global feasibility, at an additional entropy cost of $\lambda^* \varepsilon n = O(R \log n)$ by a sensitivity (envelope) argument, on every instance whose slack the reserve covers.

4. *Lower bounds and the entropy–feasibility frontier* (Theorems 3 and 4). No exactly feasible online algorithm achieves entropy loss $o(R)$. More generally, for a violation budget $V \geq 0$ the attainable loss obeys $\Omega(R) - O(V)$, recovering the previous statement at $V = 0$.

5. *When is R small* (Theorem 7). If the p_t are drawn independently from a fixed set of k distributions, then simultaneously over all prefixes $R = O\big(\log(1/\rho)\sqrt{n(k + \log(n/\delta))}\big)$ with probability $1 - \delta$, so the entropy loss of EOS is $\tilde{O}(\sqrt{n}) = o(n)$.

Related work. That maximum-likelihood decoding degenerates is well documented. Holtzman et al. (2020) introduced nucleus sampling as an empirical fix; Meister and Cotterell (2023) read beam search post hoc in terms of uniform information density. Temperature scaling, top-k sampling (Fan et al., 2018), and repetition penalties (Li et al., 2016) are motivated by intuition rather than derived; unlikelihood training (Welleck et al., 2020) targets repetition at training time under a static objective. Adaptive decoders such as Mirostat (Basu et al., 2020) and entropy-based dynamic temperature (Zhang et al., 2024) adjust sampling at each step toward a per-token target surprise or entropy; neither poses a single budget over the horizon or comes with an optimality guarantee. Information-theoretic analyses of LLMs have been largely descriptive: calibration and uncertainty estimation (Guo et al., 2017; Ovadia et al., 2019), the train/generate mismatch (Meister and Cotterell, 2023), entropy collapse (Holtzman et al., 2020).

On the algorithmic side, online convex optimization and mirror descent (Cesa-Bianchi and

Lugosi, 2006; Orabona, 2019) give tight regret bounds for per-round constraints. Constrained and long-term-constraint OCO is closer, but there the constraints are revealed per round (or the feasible set is fixed in advance) and regret is measured against a running average of constraint violations. Our single budget instead couples all n rounds through one shadow price λ^* identified only in hindsight, so per-round dual-ascent updates do not by themselves yield a feasible temperature. Online universal learning (Wu et al., 2026) extends online techniques to information-theoretic objectives, but without a budget coupled across the horizon.

The closest setting is online resource allocation and online linear programming, where a single budget is shared across the horizon and its shadow price is learned online, by dual (mirror) descent (Balseiro et al., 2020) or by re-solving on the observed prefix (Agrawal et al., 2014; Li and Ye, 2022). EOS is in the spirit of these re-solving schemes, specialized to a single resource with an entropy reward: the response to a price is available in closed form (tempering), so the online decision is one-dimensional and the price is a temperature. Whereas those works assume stochastic (typically i.i.d.) inputs, we parameterize performance by the instance-dependent drift R, which gives guarantees for non-stationary sequences and a matching $\Omega(R)$ lower bound in the exactly feasible class.

The paper is organized as follows: Section 2 fixes notation and states the protocol. Section 2.1 records what entropy buys us (recurrence times) and proves the equivalence between the probability floor and the cross-entropy constraint. Section 3.1 is the offline warm-up. Section 3.2 contains the online results. Section 3.4 bounds R. Section 4 reports experiments.

2 Setup of Online Protocol

Convention on logarithms. Throughout, log denotes the *natural* logarithm and entropies are in nats; the plausibility set is $\mathcal{W}_\gamma^n = \{w^n : P(w^n) \geq e^{-\gamma n}\}$. Every statement holds verbatim in base 2 after dividing γ by $\log 2$. (This convention removes the $1/\ln 2$ factors from the Karush–Kuhn–Tucker conditions and makes the identity $g_t'(\lambda) = \mathrm{Var}[\log 1/p_t]$ exact rather than approximate; it matters for the constants below.)

Let $\mathcal{W}$ be the token alphabet, $|\mathcal{W}| < \infty$, and let $\mathcal{W}^n$ be the set of length-n token sequences. Let

$$p_t(w) \;=\; P(w \mid w^{t-1}) \in \Delta(\mathcal{W})$$

be the conditional distribution produced by the model given the prefix $w^{t-1} = (w_1, \ldots, w_{t-1})$, so that $P(w^n) = \prod_{t=1}^n P(w_t \mid w^{t-1}) = \prod_{t=1}^n p_t(w_t)$. This is the exact chain rule, not an independence assumption. We write $\mathcal{W}_\gamma^n = \{w^n : P(w^n) \geq e^{-\gamma n}\}$ for the set of *plausible* sequences. Tokens are drawn as $w_t \sim q_t$ where $q_t = \phi_t(p_1, \ldots, p_t) \in \Delta(\mathcal{W})$, so the sampled sequence has law $\prod_t q_t(w_t)$. Top-k, top-p, and temperature sampling are all instances of this.

Conditionals. In our analysis $p_1, \ldots, p_n$ are chosen by the environment of the protocol and are unaffected by the sampled tokens. Under autoregressive feedback, the sequence-level optimum $Q^*(w^n) \propto P(w^n)^\lambda$ has conditionals $Q^*(w_t \mid w^{t-1}) \propto p_t(w_t)^\lambda Z_{t+1}(w^t)$ with $Z_{t+1}(w^t) = \sum_{w_{t+1}^n} \prod_{s>t} p_s(w_s)^\lambda$, which reduces to temperature sampling exactly when Z_{t+1} does not depend on w_t. Our guarantees are stated for this regime; Algorithm 1 runs unchanged under feedback and is evaluated there in Section 4.

Assumption 1 (Bounded p_t)**.** There is $\rho > 0$ with $p_t(w) \geq \rho$ for all $t \in [n]$, $w \in \mathcal{W}$.

This is the standard consequence of top-k / nucleus pre-filtering followed by renormalization (EOS then acts on this support), and it is what makes the cross-entropy per round bounded:

$0 \le -\sum_w q(w)\log p_t(w) \le \log(1/\rho)$ for every $q \in \Delta(\mathcal{W})$. Furthermore, for $\lambda \ge 0$ define the tempered distribution and the two scalar functions we use throughout:

$$q_t^\lambda(w) = \frac{p_t^\lambda(w)}{\sum_{w'\in\mathcal{W}} p_t^\lambda(w')}, \tag{1}$$

$$f_t(\lambda) = H\Big(q_t^\lambda\Big), \qquad g_t(\lambda) = \gamma + \sum_{w\in\mathcal{W}} q_t^\lambda(w)\log p_t(w). \tag{2}$$

Thus $-\sum_w q_t^\lambda(w)\log p_t(w) = \gamma - g_t(\lambda)$ is the per-round cross-entropy, and $g_t(\lambda) \ge 0$ means round t is within budget.

The offline optimization problems. The *offline* problem, in which all of $p_1, \dots, p_n$ are known in advance, is

$$\max_{q_t\in\Delta(\mathcal{W}),\, t\in[n]} \sum_{t=1}^{n} H(q_t) \tag{3}$$

$$\text{s.t.} \qquad -\sum_{t=1}^{n}\sum_{w\in\mathcal{W}} q_t(w)\log p_t(w) \;\le\; \gamma n. \tag{4}$$

Its value is denoted $\mathrm{OPT}(\gamma n)$ and its solution Q_t^{off}.

The online protocol. We state the protocol explicitly, since every subsequent claim depends on the quantifier order. The horizon n, the budget γ, the alphabet $\mathcal{W}$ and the floor ρ are public. The algorithm $\mathcal{A} = (\phi_1, \dots, \phi_n)$ is *committed ex ante*, before any p_t is revealed.

> For $t = 1, \dots, n$: the environment reveals $p_t \in \Delta(\mathcal{W})$; the algorithm outputs $q_t^{\mathcal{A}} = \phi_t(p_1, \dots, p_t)$; the token $w_t \sim q_t^{\mathcal{A}}$ is emitted. The environment is adversarial in the sense that the sequence $p_1, \dots, p_n$ is chosen by an adversary that knows $\mathcal{A}$ but must produce p_t before seeing $q_t^{\mathcal{A}}$; equivalently, lower bounds are stated as $\exists\{p_t\}$ after $\forall\mathcal{A}$.

The online optimization problem then can be formulated as follows:

$$\begin{aligned} &\max_{q_t=\phi_t(\{p_i\}_{i=1}^t)\in\Delta(\mathcal{W}):\, t\in[n]} \sum_{t=1}^{n} H(q_t) \\ &\text{s.t.} \quad -\sum_{t=1}^{n}\sum_{w\in\mathcal{W}} q_t(w)\log p_t(w) \le \gamma n. \end{aligned} \tag{5}$$

Performance is measured by the *entropy loss*

$$\sum_t H(Q_t^{\mathrm{off}}) - \sum_t H(q_t^{\mathcal{A}})$$

against the offline optimum of (4) on the same instance, together with the *global slack* $\big[-\sum_t\sum_w q_t^{\mathcal{A}}(w)\log p_t(w) - \gamma n\big]_+$. The offline optimum is the honest comparator: it upper-bounds the value of every causal policy, so a guarantee stated against it is stronger than the same guarantee stated against a causal reference.

Definition 1 (Feasibility classes)**.** Let $V \geq 0$. An online algorithm $\mathcal{A}$ is *V-feasible* if for every instance

$$-\sum_{t=1}^{n} \sum_{w \in \mathcal{W}} q_t^{\mathcal{A}}(w) \log p_t(w) \;\leq\; \gamma n + V.$$

0-feasible algorithms are called *exactly feasible*. We write $\mathcal{A}_V$ for the class of V-feasible algorithms; $\mathcal{A}_0 \subseteq \mathcal{A}_V$ for all V.

Definition 2 (Drift parameter)**.** With Q^{off} the offline optimum of (4), set

$$R_t \;=\; \Big| \sum_{i=1}^{t} \sum_{w \in \mathcal{W}} Q_i^{\text{off}}(w) \log p_i(w) + \gamma t \Big| \;=\; \Big| \sum_{i=1}^{t} g_i(\lambda^*) \Big| \tag{6}$$

and $R \;=\; R(n) \;=\; \max_{t \in [n]} R_t$, where λ^* is defined in (9).

We note that R_t measures how far the offline-optimal policy is, at prefix t, from having spent exactly a t/n fraction of its budget. If the p_t are exchangeable in the strong sense that every prefix looks like the whole, R is small; if the instance front-loads or back-loads its hard rounds, R is large.

Remark 1 (R is not observable)**.** R is defined through Q^{off}, hence through λ, hence through the entire horizon. It is an *instance-complexity* parameter, not a quantity available to a causal algorithm. All our upper bounds are stated as "$O(R \log n)$" with R appearing only in the analysis; the one place where the algorithm would need to know R is the tightening parameter of Theorem 6, and we address that explicitly in Remark 4.

In passing we observe that both the offline program and the online algorithm are defined through the root of a monotone scalar equation, and that root need not exist for every γ. We record the condition.

Assumption 2 (Strict feasibility on every prefix)**.** For every $t \in [n]$, the budget γ satisfies

$$\begin{aligned} \frac{1}{t} \sum_{i=1}^{t} \log \frac{1}{\max_w p_i(w)} \;&<\; \gamma \\ \leq\; \frac{1}{t} \sum_{i=1}^{t} \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \log \frac{1}{p_i(w)}. \end{aligned} \tag{7}$$

The left inequality says the budget is achievable at all (even the greedy policy $\lambda \to \infty$ has cross-entropy $\frac{1}{t}\sum_i \log(1/\max_w p_i)$, so a smaller γ is infeasible); the right says it is binding (at $\lambda = 0$, i.e. uniform sampling, the cross-entropy is already at or above budget, so $\lambda^* \geq 0$). If the right inequality fails the constraint is inactive and uniform sampling is optimal — a degenerate case we exclude. Under Assumptions 1 and 2 the map $\lambda \mapsto \sum_{i \leq t} g_i(\lambda)$ is continuous and nondecreasing, negative at $\lambda = 0$ and positive as $\lambda \to \infty$, so a root exists; Assumption 3 below makes it unique.

Assumption 3 (Curvature)**.** There is $c > 0$ such that for all $t \in [n]$,

$$\sum_{i=1}^{t} \mathrm{Var}_{Q_i^{\text{off}}} \Big[\log \tfrac{1}{p_i(w)} \Big] \;\geq\; ct. \tag{8}$$

Since $g_t'(\lambda) = \mathrm{Var}_{q_t^\lambda}[\log(1/p_t)]$ (Proposition 3), this says the prefix constraint functions have positive slope at λ^*: the temperature is identified at rate $1/t$ from the prefix. It fails only if the conditionals are asymptotically deterministic or asymptotically uniform. Figure 3 gives empirical support.

2.1 Entropy controls recurrence time

We first make precise the sense in which maximizing entropy provably delays repetition and prompts diversity. This is a classical theorem, and we state it as such rather than as an analogy but we also state its scope, because it applies literally only in the stationary specialization.

Proposition 1 (Recurrence-time theorem; Wyner and Ziv (1989), Ornstein and Weiss (1993))**.** *Let $\{W_i\}$ be a stationary ergodic source with entropy rate H, and let $\mathcal{R}_n$ denote the recurrence time of the length-n prefix, i.e. the first $m > 0$ with $W_{m+1}^{m+n} = W_1^n$. Then*

$$\frac{1}{n}\log \mathcal{R}_n \;\longrightarrow\; H \qquad \textit{almost surely.}$$

Consequently the time before the generated stream repeats a length-n block grows as $e^{nH(1+o(1))}$ almost surely: higher entropy provably yields exponentially longer recurrence times.

The intuition is Kac's lemma (Kac, 1947) — for a stationary ergodic source the mean return time to a state is the reciprocal of its probability — combined with the AEP (Cover and Thomas, 2006), by which a typical n-block has probability e^{-nH}. However, see (Szpankowski, 1993) for a rephrasing of this problem from the suffix-tree point of view.

Remark 2 (Scope)**.** Proposition 1 applies verbatim to fixed-temperature sampling in the regime of Theorem 7, where the generated stream is i.i.d. and hence stationary ergodic; there, maximizing $\sum_t H(q_t)$ maximizes the recurrence exponent. In the general non-stationary setting the AEP half of the argument survives and is exactly what Theorem 1 establishes non-asymptotically, while the Kac "return time = 1/probability" step does not. We therefore adopt $\sum_t H(q_t)$ as the objective on the strength of the stationary case and analyze it on its own terms; we do not claim a recurrence-time theorem in the general case.

The constraint we actually impose is the cross-entropy budget. The next result says it is equivalent, up to $O(\sqrt{n\log(1/\delta)})$, to the probability floor $w^n \in \mathcal{W}_\gamma^n$ which is the constraint one would write down from first principles. In the language of information geometry, the budget defines a cross-entropy (KL) ball around the model and the tempered family (1) is the corresponding I-projection geodesic; the "floor" is not an ad hoc device.

Theorem 1. *Let $p_1, \ldots, p_n$ be given, w^n sampled sequentially from $q_1, \ldots, q_n$, and let Assumption 1 hold. Write $L = \log(1/\rho)$.*

(i) If

$$\sum_{t\in[n]}\sum_{w} q_t(w)\log p_t(w) + \gamma n \;\geq\; -L\sqrt{\tfrac{n}{2}\log\tfrac{2}{\delta}},$$

then

$$\Pr\Big(P(w^n) \;\geq\; e^{-\gamma n - 2L\sqrt{(n/2)\log(2/\delta)}}\Big) \;\geq\; 1-\delta.$$

(ii) Conversely, if $\Pr(P(w^n) \geq e^{-\gamma n - \Delta_n}) \geq 1-\delta$ then

$$\sum_{t\in[n]}\sum_{w} q_t(w)\log p_t(w) + \gamma n$$
$$\geq\; -\Delta_n - L\sqrt{\tfrac{n}{2}\log\tfrac{2}{\delta}} - \delta n L.$$

Choosing $\delta = O(n^{-1/2})$, the constraint $w^n \in \mathcal{W}_\gamma^n$ and the cross-entropy constraint (4) agree to within $O(\sqrt{n\log n})$, i.e. they are equivalent at the level of the exponent.

3 Main Results

In this section we formulate our main results starting with a warm-up exercise regarding the offline optimization which we use as a comparator. Then we solve the online optimization problem and propose Algorithm 1 that we prove achieves asymptotically the performance of the online optimization.

3.1 Offline Warm-Up: the Comparator

To start, we discuss briefly the offline optimization problem since we use it as a comparator for the online optimization.

Proposition 2 (Offline optimum is fixed-temperature sampling)**.** *Under Assumptions 1 and 2, the solution of* (4) *is* $Q_t^{\text{off}} = q_t^{\lambda^*}$ *as in* (1)*, where* $\lambda^* \geq 0$ *solves*

$$\begin{aligned} -\sum_{t\in[n]}\sum_{w\in\mathcal{W}} q_t^{\lambda^*}(w)\log p_t(w) \;&=\; \gamma n, \\ \textit{equivalently}\quad \sum_{t\in[n]} g_t(\lambda^*) &= 0. \end{aligned} \tag{9}$$

Under Assumption 3 the root is unique. The optimal inverse temperature is common to all t*:* $\lambda^* = 1/T^*$*.*

The proof is the Karush–Kuhn–Tucker system for a concave objective under one linear constraint, and is given in Section C for completeness. We underline the reading: *one* global constraint yields *one* multiplier, hence *one* temperature. This is why temperature sampling — as opposed to any other reweighting of p_t — is the correct functional form, and it is also why the online problem is nontrivial: a causal algorithm must estimate a global shadow price from a prefix.

Lemma 1 (Sensitivity / envelope)**.** *Let* $\mathrm{OPT}(b)$ *denote the optimal value of* (4) *with the budget* γn *replaced by* b*. Then* OPT *is concave and nondecreasing on its domain, and* $\mathrm{OPT}'(b) = \lambda^*(b)$*, the optimal multiplier at budget* b*. In particular, for* $b' \leq b$*,*

$$0 \;\leq\; \mathrm{OPT}(b) - \mathrm{OPT}(b') \;\leq\; \lambda^*(b')\,(b - b').$$

Lemma 1 is the workhorse of Section 3.2: it converts every statement about *budget* into a statement about *entropy* at the exchange rate λ^*.

3.2 The Online Problem

Next, we provide a solution to the *online optimization problem.*

Theorem 2 (Clairvoyant benchmark)**.** *For a fixed instance* $p_1, \ldots, p_n$*, the online program* (5) *attains the same optimal value and optimizer as the offline program* (4)*: the optimum is*

$$q_t(w) = Q_t^{\text{on}} = \frac{p_t^{\lambda^{\text{on}}}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda^{\text{on}}}(w)}, \quad t \in [n], \tag{10}$$

with a single *exponent* $\lambda^{\text{on}} = \lambda^*$ *common to all* t*, fixed by the global budget*

$$-\sum_{t\in[n]}\sum_{w\in\mathcal{W}} \frac{p_t^{\lambda^{\text{on}}}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda^{\text{on}}}(w)} \log p_t(w) = \gamma n. \tag{11}$$

This is a clairvoyant benchmark, not an admissible policy under the protocol (which commits $\mathcal{A}$ before the instance); the difficulty of the online problem is purely informational*, since λ^* is a function of the entire horizon and cannot be computed from a prefix.*

Proof. For a fixed instance $(p_1, \ldots, p_n)$ each q_t is trivially a (constant) function of $p_1, \ldots, p_t$, so the causal constraint in (5) does not shrink the feasible set relative to (4); the two programs coincide. Proposition 2 then yields the optimizer $q_t^{\lambda^*}$ with a *single* multiplier λ^* (one budget, one Lagrange multiplier, one temperature), and complementary slackness under Assumption 2 forces the budget to bind, giving (11). A per-round exponent λ_t^{on} would leave the allocation of cross-entropy across rounds undetermined; since $f_t'(\lambda) = -\lambda\, g_t'(\lambda)$ (Proposition 3), equalizing the marginal entropy-per-nat-of-budget across t is exactly the common-λ^* solution, so no per-round choice can improve $\sum_t H(q_t)$. □

3.3 Entropic Online Sampling

Unfortunately, computing Q_t^{on} requires the horizon-dependent optimal temperature λ^* and is therefore not available to a causal algorithm. Therefore, we propose a simpler and efficient algorithm, called Entropic Online Sampling (EOS) Algorithm 1. It is the natural causal surrogate for Proposition 2: at each prefix, solve the offline problem restricted to the data seen so far. The rest of this section is devoted to proving EOS is asymptotically optimal.

Equation (12) of Algorithm 1 is one-dimensional with a nonincreasing left-hand side, so bisection or Newton's method solves it to precision ϵ in $O(\log(1/\epsilon))$ evaluations. Each evaluation costs $O(t|\mathcal{W}|)$, since every past round must be re-tempered at the new λ. The clipping step is what makes the algorithm well defined when Assumption 2 fails on a short prefix — a case that does occur early in generation, when a few near-deterministic rounds can put the prefix budget out of reach. Note the contrast with the offline solution: EOS uses a single exponent λ_t across all rounds $i \le t$ of the prefix, updated as t grows, whereas λ^* is a single exponent fixed with hindsight.

Algorithm 1 Entropic Online Sampling (EOS)

Require: budget γ; cap Λ

1: **for** $t = 1, \ldots, n$ **do**

2: observe p_t

3: solve for $\lambda_t \in [0, \Lambda]$ the monotone scalar equation

$$-\sum_{i\in[t]} \sum_{w\in\mathcal{W}} q_i^{\lambda_t}(w) \log p_i(w) = \gamma t, \tag{12}$$

setting $\lambda_t = 0$ if $\sum_{i\in[t]} g_i(\lambda) > 0$ for all $\lambda \in [0, \Lambda]$, and $\lambda_t = \Lambda$ if it is < 0 throughout

4: emit $w_t \sim q_t^{\lambda_t}$

5: **end for**

We need the following estimate of λ_t to prove asymptotic optimality of Algorithm 1.

Lemma 2 (Bounded multiplier)**.** *Under Assumptions 1 and 2, $\lambda_t, \lambda^* \ge 0$; if (8) holds with q_i^λ for all $\lambda \in [0, \Lambda]$, $\Lambda := (\log(1/\rho) - \gamma)/c$, then $\lambda_t, \lambda^* \le \Lambda$. Otherwise Λ is any upper bound on λ^* (Algorithm 1 clips λ_t to $[0, \Lambda]$).*

3.4 Asymptotic Optimality of EOS

We start with a general lower bound proved in Appendix F.

Theorem 3 (Lower bound, $\mathcal{A}_0$)**.** *Let γ and $|\mathcal{W}|$ satisfy*

$$-\frac{(|\mathcal{W}|-1)\log\left(\frac{1-e^{-\gamma/3}}{|\mathcal{W}|-1}\right)}{|\mathcal{W}|}+\frac{\gamma}{3|\mathcal{W}|}>\frac{4\gamma}{3}$$

and $\log|\mathcal{W}| \;\geq\; \frac{4\gamma}{3}$, and let $R \leq \gamma n/6$. Then for every exactly feasible online algorithm $\mathcal{A} \in \mathcal{A}_0$ there exists an instance $\{p_t\}_{t\in[n]}$ with drift parameter $R(n) = R$ such that

$$\sum_{t=1}^{n} H(Q_t^{\text{off}}) - \sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \;\geq\; \kappa\, R, \qquad \kappa \;=\; \frac{\gamma^3}{216\, G^2}, \tag{13}$$

and $G := \left(\frac{4\gamma}{3} + \log\frac{1}{1-e^{-\gamma/3}}\right)^2$.

The construction is a pair of instances agreeing on a common prefix of length $3R/\gamma$ and diverging afterwards; because $\mathcal{A}$ commits before seeing the divergence, it must misallocate budget on one of the two, and Lemma 1 converts the misallocation into entropy.

The entropy–feasibility frontier. Theorem 3 constrains $\mathcal{A}_0$. It says nothing about an algorithm that enforces only a running constraint and finishes with residual slack, since such an algorithm is not in $\mathcal{A}_0$. The right object is a frontier.

Theorem 4 (Pareto lower bound)**.** *Fix $V \geq 0$. Under the hypotheses of Theorem 3, for every V-feasible online algorithm $\mathcal{A} \in \mathcal{A}_V$ there exists an instance with drift parameter R such that*

$$\sum_{t=1}^{n} H(Q_t^{\text{off}}) - \sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \;\geq\; \kappa R \;-\; \Lambda V, \tag{14}$$

with κ as in (13) *and $\Lambda \geq$ the tail multipliers λ_x^* (Section G).*

At $V = 0$ this is Theorem 3. For $V > 0$ it quantifies the exchange rate: enlarging the feasible set by V nats of budget buys at most ΛV nats of entropy, and the bound is nonvacuous as long as $V < \kappa R/\Lambda$. The proof (Section G) is a one-line modification of Section F: slack V loosens the tail subproblem from $-R - B$ to $-R - B - V$, and Lemma 1 bounds the resulting gain by ΛV.

Theorem 5 (Bicriteria guarantee for EOS)**.** *Let q_t^{on} be the output of Algorithm 1 and let Assumptions 1 to 3 hold. Then there is a constant $C = C(\rho, c, \Lambda, |\mathcal{W}|)$ such that*

$$\sum_{t=1}^{n} H(Q_t^{\text{off}}) - \sum_{t=1}^{n} H(q_t^{\text{on}}) \;\leq\; C\, R \log n, \tag{15}$$

$$-\sum_{t=1}^{n}\sum_{w\in\mathcal{W}} q_t^{\text{on}}(w)\log p_t(w) - \gamma n \;\leq\; C\, R \log n. \tag{16}$$

We report (16) as a first-class quantity: EOS is $O(R\log n)$-feasible, not exactly feasible, and the pair (15)–(16) is a point on the frontier of Theorem 4 rather than a comparison against Theorem 3. The next subsection removes the slack.

We now establish a feasible upper bound showing that our Algorithm 1 achieves asymptotic optimality.

Theorem 6 (EOS with budget tightening)**.** *Fix $\varepsilon \in (0, \gamma)$ and run Algorithm 1 with γ replaced by $\gamma_\varepsilon := \gamma - \varepsilon$ in* (12)*. Let $R_\varepsilon := R(n; \gamma_\varepsilon)$ be the drift parameter at the tightened budget and let Assumptions 1 to 3 hold at γ_ε. With C as in Theorem 5 and cap $\Lambda_\varepsilon \geq \lambda^*(\gamma_\varepsilon)$ ($= \Lambda + \varepsilon/c$ under Lemma 2):*

(i) (Exact feasibility.) *If $\varepsilon n \geq C R_\varepsilon \log n$, the output satisfies the global constraint of* (4) *exactly: $-\sum_t \sum_w q_t(w) \log p_t(w) \leq \gamma n$, i.e. it is exactly feasible on every such instance.*

(ii) (Entropy.) *Against the offline optimum at the* original *budget,*

$$\sum_{t=1}^{n} H(Q_t^{\text{off}}) - \sum_{t=1}^{n} H(q_t) \;\leq\; C R_\varepsilon \log n \;+\; \Lambda_\varepsilon\, \varepsilon n.$$

In particular, ε with $\varepsilon n = C R_\varepsilon \log n$ (a fixed point, since R_ε depends on ε), when it exists, gives an exactly feasible algorithm with entropy loss at most $(1 + \Lambda_\varepsilon)\, C R_\varepsilon \log n = O(R_\varepsilon \log n)$.

The proof (Section E) is short: (i) is (16) applied at budget γ_ε, whose slack of at most $C R_\varepsilon \log n$ is absorbed by the εn nats held in reserve; (ii) combines (15) at γ_ε with Lemma 1, which prices the reserve at $\lambda^* \leq \Lambda_\varepsilon$ nats of entropy per nat of budget.

Remark 3 (Why the terminal multiplier cannot do this)**.** A tempting alternative is to leave EOS unchanged and fix the final multiplier λ_n against the realized prefix so that the global constraint holds with equality. This does not work. The accumulated slack after $n - 1$ rounds is $\Theta(R \log n)$, which for $R = \Theta(\sqrt{n})$ is $\Theta(\sqrt{n} \log n)$, whereas a single round can move the cross-entropy sum by at most $\log(1/\rho) = O(1)$. The correction must be spread over $\Theta(R \log n)$ rounds — or, as here, built into the budget from the start.

Remark 4 (Tuning ε without knowing R)**.** By Remark 1, R_ε is not observable online. Three options. (a) In the regime of Section 3.4, $R_\varepsilon = O(\sqrt{n(k + \log(n/\delta))})$ with probability $1 - \delta$, so $\varepsilon = \Theta(\sqrt{(k + \log(n/\delta))/n}\, \log n)$ is a valid a priori choice. (b) Use the observable proxy $\hat{R}_t = \max_{i \leq t} |\sum_{j \leq i} g_j(\lambda_t)|$, which lower-bounds R up to the estimation error $|\lambda_t - \lambda^*| = O(R/(ct))$, and tighten adaptively; this costs a further constant factor. (c) A doubling schedule on ε, restarting when the running slack exceeds the reserve.

Corollary 1 (Optimality up to a logarithm)**.** *Given a bound on R_ε, within the exactly feasible class $\mathcal{A}_0$, the attainable entropy loss is at least κR (Theorem 3) and at most $O(R_\varepsilon \log n)$ (Theorem 6). The two agree up to a single factor of $\log n$ (and up to the difference between R and R_ε, see Remark 5).*

Remark 5 (R versus R_ε)**.** Tightening the budget perturbs the comparator, so Theorem 6 is stated in terms of R_ε. To first order the perturbation cancels: writing $A_t(\lambda) = \sum_{i \leq t} g_i(\lambda)$, the tightened root satisfies $\lambda^*(\gamma_\varepsilon) - \lambda^*(\gamma) \approx \varepsilon/c$, and

$$A_t^{\varepsilon}\big(\lambda^*(\gamma_\varepsilon)\big) \;\approx\; A_t\big(\lambda^*(\gamma)\big) + \frac{\varepsilon}{c} \sum_{i \leq t} g_i' - \varepsilon t \;\approx\; A_t\big(\lambda^*(\gamma)\big),$$

since $\sum_{i \leq t} g_i' \approx ct$. The residual is controlled by the inhomogeneity of the curvature across the horizon, and vanishes when the g_i' are exchangeable. In the regime of Section 3.4 the two are of the same order, since the bound of Theorem 7 holds uniformly in γ over compact subsets of the feasible range; so nothing is lost in the headline corollary $\tilde{O}(\sqrt{n})$. We do not claim $R_\varepsilon = \Theta(R)$ in general.

When Is the Drift Parameter R Small? In the worst case $R = \Theta(n)$ and the guarantees are vacuous, as they must be: Theorem 3 exhibits such instances. The question is whether $R = o(n)$ under conditions resembling real decoding. We analyze the simplest stationarity proxy: the p_t are drawn independently from a finite set of k distributions. We present this as a first tractable case within our regime of Section 2; it does not capture the dependence of p_t on previously sampled tokens.

Lemma 3 (ℓ_1 concentration; Bretagnolle–Huber-Carol inequality, see van der Vaart and Wellner (1996))**.** *Let p_t be drawn independently with $\Pr(p_t = p^{(i)}) = P_i$, $i \in [k]$, and let $\hat{P}_i = |\{t \le n : p_t = p^{(i)}\}|/n$. Then with probability at least $1-\delta$,*

$$\sum_{i\in[k]} \left|\hat{P}_i - P_i\right| \;\le\; \sqrt{\frac{2\big(k\log 2 + \log\frac{1}{\delta}\big)}{n}}.$$

Proof. Write $\sum_i |\hat{P}_i - P_i| = 2\sup_{A\subseteq[k]}\big(\sum_{i\in A}\hat{P}_i - \sum_{i\in A} P_i\big)$. For fixed A, one-sided Hoeffding gives $\Pr(\sum_{i\in A}(\hat{P}_i - P_i) \ge \epsilon) \le e^{-2n\epsilon^2}$; a union bound over the 2^k subsets and solving $2^k e^{-2n\epsilon^2} = \delta$ bounds the supremum by $\sqrt{(k\log 2 + \log\frac{1}{\delta})/(2n)}$, and doubling yields the claim. □

Theorem 7 (R is $\tilde{O}(\sqrt{n})$, uniformly over prefixes)**.** *In the setting of Lemma 3, with probability at least $1-\delta$,* simultaneously for all $t \in [n]$,

$$\begin{aligned} R \;&=\; \max_{t\in[n]} R_t \;\le\; 2\log\tfrac{1}{\rho}\sqrt{2n\Big(k\log 2 + \log\tfrac{n+1}{\delta}\Big)} \\ &=\; O\Big(\log\tfrac{1}{\rho}\sqrt{n\big(k+\log\tfrac{n}{\delta}\big)}\Big). \end{aligned}$$

Consequently, by Theorems 5 and 6, the entropy loss of EOS and of its exactly feasible variant is $\tilde{O}(\sqrt{n}) = o(n)$.

4 Experiments

We validate three predictions: (i) EOS stays close to the offline optimum and above budget-feasible temperatures fixed in advance; (ii) the curvature condition Assumption 3; (iii) that EOS tracks λ^*. The comparison in (i) is motivated by Proposition 2: a fixed temperature is optimal only if it equals $1/\lambda^*$, which depends on the whole horizon, so a temperature chosen in advance either exceeds the budget or leaves budget unspent, which by Lemma 1 costs entropy at rate λ^*.

We use ELI5 (Fan et al., 2019) and Alpaca (Taori et al., 2023) with Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024), $\gamma = 0.9$ and a probability cutoff $\rho = 10^{-4}$ (realizing Assumption 1). The fixed baseline is $T = 0.7$, a common default that spends well below the budget (e.g. 0.35 nats per token in the autoregressive run below), so it is feasible but leaves budget unspent. On both datasets EOS tracks the offline optimum with a small gap and exceeds $T = 0.7$ in entropy (Figure 1), λ_t converges to λ^*, and the cumulative variance grows linearly (Figure 3, Appendix I).

Autoregressive regime. We also run Algorithm 1 unchanged on genuine autoregressive generation, which the analysis does not cover: 200 ELI5 questions, $n = 1024$ tokens, against a fixed-temperature grid and standard truncation samplers, compared at matched realized cross-entropy (Section I). EOS stays close to the fixed-temperature entropy frontier, losing 0.006, 0.037

and 0.088 nats per token at $\gamma = 0.5, 0.9, 1.3$, which reflects the small loss to the offline comparator that Theorem 5 allows. Yet it is markedly less repetitive at the same cross-entropy (Figure 2). At essentially equal cross-entropy (0.50), EOS with $\gamma = 0.5$ repeats 2.0% of 16-word blocks (with at most 10% edits) against 3.4% for nucleus sampling with $p = 0.95$. Against the fixed-temperature frontier at $\gamma = 0.9$, seq-rep-4 falls from 0.068 to 0.043, repeated 16-word blocks from 1.6% to 0.9%, near-duplicate sentences from 3.4% to 2.1%, and cross-sample 4-gram overlap from 0.019 to 0.011, with 95% intervals excluding zero at all budgets and at every edit tolerance (Figure 4, Appendix I). Under feedback, plain EOS overshoots its budget: its realized cross-entropy is 1.01 at $\gamma = 0.9$ and 1.53 at $\gamma = 1.3$. The tightened variant corrects the mean, with a fifth of the per-prompt spread of a calibrated fixed temperature (Table 1, Appendix I).

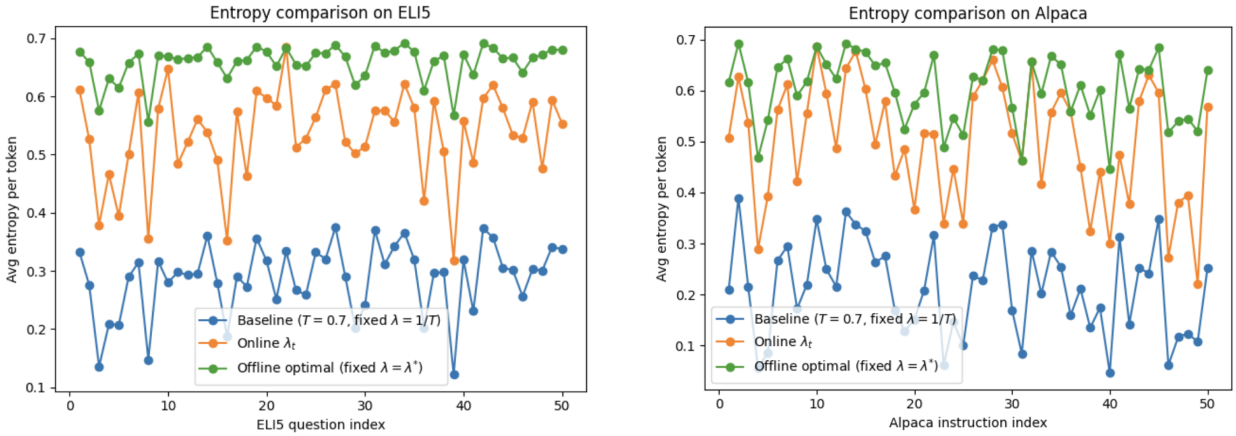

Figure 1: Exogenous regime: per-token entropy on ELI5 (left) and Alpaca (right). The offline optimum (green) is highest; EOS (orange) tracks it with a small gap, as Proposition 2 and Theorem 5 predict; the fixed temperature $T = 0.7$ (blue) is budget-feasible but leaves budget unspent.

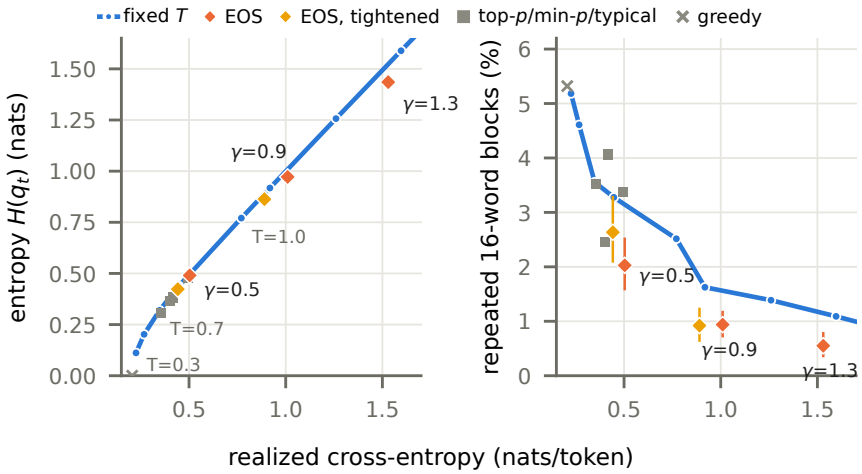

Figure 2: Autoregressive regime (Llama-3-8B-Instruct, ELI5). Entropy (left) and repeated 16-word blocks with $\leq$ 10% edits (right) against realized cross-entropy; the curve is fixed temperature ($T \in [0.3, 1.15]$). Bars are 95% bootstrap intervals over prompts.

5 Concluding Remarks

We formulated decoding as online entropy maximization under a single cross-entropy budget coupled across the horizon. EOS attains $O(R \log n)$ entropy loss with $O(R \log n)$ slack, tightening its budget makes it exactly feasible (given a drift bound) at the same order, and no exactly feasible online algorithm can lose $o(R)$; in the i.i.d.-mixture regime $R = \tilde{O}(\sqrt{n})$. Empirically, EOS also reduces repetition under autoregressive feedback, which our analysis does not cover. Open directions are closing the $\log n$ gap, extending the analysis to autoregressive feedback and its future-partition-function tilt (Section 2), and selecting γ from a downstream *pass@k* (Chen et al., 2021) target.

References

Agrawal, S., Wang, Z., and Ye, Y. (2014). A dynamic near-optimal algorithm for online linear programming. *Operations Research*, 62(4):876–890.

Balseiro, S., Lu, H., and Mirrokni, V. (2020). Dual mirror descent for online allocation problems. In *International Conference on Machine Learning*, pages 613–628. PMLR.

Basu, S., Ramachandran, G. S., Keskar, N. S., and Varshney, L. R. (2020). Mirostat: A neural text decoding algorithm that directly controls perplexity. *arXiv preprint arXiv:2007.14966*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience.

Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., and Auli, M. (2019). Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*.

Fan, A., Lewis, M., and Dauphin, Y. (2018). Hierarchical neural story generation. *Association for Computational Linguistics*.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. *International Conference on Machine Learning*.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. *International Conference on Learning Representations*.

Kac, M. (1947). On the notion of recurrence in discrete stochastic processes. *Bulletin of the American Mathematical Society*, 53(10):1002–1010.

Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., and Jurafsky, D. (2016). A diversity-promoting objective function for neural conversation models. *North American Chapter of the Association for Computational Linguistics*.

Li, X. and Ye, Y. (2022). Online linear programming: Dual convergence, new algorithms, and regret bounds. *Operations Research*, 70(5):2948–2966.

Meister, C. and Cotterell, R. (2023). If beam search is the answer, what was the question? *Transactions of the Association for Computational Linguistics*, 11:152–171.

Orabona, F. (2019). Modern online learning: A short survey. *arXiv preprint arXiv:1912.13213*.

Ornstein, D. S. and Weiss, B. (1993). Entropy and data compression schemes. *IEEE Transactions on Information Theory*, 39(1):78–83.

Ovadia, Y., Factor, E., Molchanov, Y., Ritter, D., Bixby, R., Nado, Z., Sculley, D., Nowozin, S., Duchi, J., Ghahramani, Z., and Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Szpankowski, W. (1993). Asymptotic properties of data compression and suffix trees. *IEEE Transactions on Information Theory*, 39(5):1647–1659.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. Bretagnolle–Huber–Carol inequality, Proposition A.6.6.

Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., and Weston, J. (2020). Neural text generation with unlikelihood training. *International Conference on Learning Representations*.

Wu, C., Grama, A., and Szpankowski, W. (2026). Online universal learning from information-theoretic perspective. *Foundation and Trends in Communication and Information Theory*, 23:225–443.

Wyner, A. D. and Ziv, J. (1989). Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Transactions on Information Theory*, 35(6):1250–1258.

Zhang, S., Bao, Y., and Huang, S. (2024). Edt: Improving large language models' generation by entropy-based dynamic temperature sampling. *arXiv preprint arXiv:2403.14541*.

A Basic Facts

Proposition 3. *Let Assumption 1 hold and let f_t, g_t be as in (2). Then, with $L = \log(1/\rho)$,*

$$g_t'(\lambda) = \mathrm{Var}_{q_t^\lambda}\left[\log \tfrac{1}{p_t(w)}\right] \in [0, L^2], \qquad f_t'(\lambda) = -\lambda\, g_t'(\lambda) \in [-\Lambda L^2,\, 0], \qquad |g_t''(\lambda)| \;\le\; L^3.$$

Proof. Since $q_t^\lambda(w) \propto e^{\lambda \log p_t(w)}$ is a one-parameter exponential family with natural parameter λ and sufficient statistic $\log p_t(w)$, the cumulant generating function $\psi_t(\lambda) = \log \sum_w p_t^\lambda(w)$ satisfies $\psi_t' = \mathbb{E}_{q_t^\lambda}[\log p_t]$ and $\psi_t'' = \mathrm{Var}_{q_t^\lambda}[\log p_t]$. As $g_t(\lambda) = \gamma + \psi_t'(\lambda)$, we get $g_t' = \psi_t'' = \mathrm{Var}[\log(1/p_t)] \ge 0$; the range bound follows since $\log(1/p_t) \in [0, L]$ under Assumption 1, so the variance is at most $L^2/4 \le L^2$. Also $g_t'' = \psi_t''' = \mathbb{E}[(\log p_t - \mathbb{E}\log p_t)^3]$, bounded in absolute value by L^3. For f_t, $H(q_t^\lambda) = \psi_t(\lambda) - \lambda\psi_t'(\lambda)$, whence $f_t'(\lambda) = -\lambda\psi_t''(\lambda) = -\lambda g_t'(\lambda)$, which lies in $[-\Lambda L^2, 0]$ for $\lambda \in [0, \Lambda]$. □

Proof of Lemma 2. Fix t and put $A_t(\lambda) = \sum_{i \le t} g_i(\lambda)$. Nonnegativity of λ_t and λ^* is the right-hand inequality of Assumption 2. Under the additional hypothesis, Proposition 3 gives $A_t' \ge ct$ on $[0, \Lambda]$. At $\lambda = 0$, $g_i(0) = \gamma - \frac{1}{|\mathcal{W}|}\sum_w \log \frac{1}{p_i(w)} \ge \gamma - L$, so $A_t(\Lambda) \ge A_t(0) + ct\Lambda \ge -t(L-\gamma) + t(L-\gamma) = 0$; since A_t is nondecreasing, its root lies in $[0, \Lambda]$. The same argument with $t = n$ bounds λ^*. □

Proof of Lemma 1. $\mathrm{OPT}(b)$ is the value of a concave maximization over a convex set whose constraint is relaxed linearly in b; concavity and monotonicity are immediate, and the derivative equals the optimal multiplier by standard Lagrangian sensitivity (the multiplier is unique under Assumption 3, so OPT is differentiable). The displayed inequality is concavity plus $\mathrm{OPT}' \ge 0$. □

B Proof of Theorem 1

Proof. Let $X_t = \log p_t(w_t) + \gamma$ and $D_t = X_t - \mathbb{E}_{q_t}[X_t \mid w^{t-1}]$, so that $S_t = \sum_{i \le t} D_i$ is a martingale. Under Assumption 1, $\log p_t(w_t)$ takes values in $[-L, 0]$, so D_t lies in an interval of length at most L. By Azuma–Hoeffding for martingales with bounded ranges,

$$\Pr\big(|S_n| \ge \epsilon\big) \;\le\; 2\exp\!\left(-\frac{2\epsilon^2}{nL^2}\right),$$

so with probability at least $1-\delta$, $|S_n| \le L\sqrt{(n/2)\log(2/\delta)}$. Since

$$\log P(w^n) + \gamma n = \sum_{t \in [n]} X_t = \sum_{t \in [n]} \sum_w q_t(w) \log p_t(w) + \gamma n + S_n,$$

part (i) follows. For (ii), take expectations: on the event of probability $1-\delta$ we have $\log P(w^n) \ge -\gamma n - \Delta_n$, and on its complement $\log P(w^n) \ge -nL$ deterministically, so $\mathbb{E}[\log P(w^n)] \ge -\gamma n - \Delta_n - \delta n L$. Since $\mathbb{E}[S_n] = 0$, $\mathbb{E}[\log P(w^n)] = \mathbb{E}\big[\sum_t \sum_w q_t(w) \log p_t(w)\big]$. Since the p_t are exogenous, the q_t do not depend on w^{t-1}, so the right side is the left side of (ii), which gives the claim (the $L\sqrt{\cdot}$ term is not even needed). □

C Proof of Theorem 2

Proof. The objective $\sum_t H(q_t)$ is concave and the constraint is linear, so (4) is convex; Assumption 2 gives a Slater point, so strong duality and the KKT conditions hold. With multipliers $\lambda \ge 0$ for the

budget and μ_t for normalization, the Lagrangian is

$$\mathcal{L} = \sum_t \mu_t \Big(1 - \sum_w q_t(w)\Big) - \sum_t H(q_t) - \lambda \Big(\sum_t \sum_w q_t(w) \log p_t(w) + \gamma n \Big),$$

and $\partial \mathcal{L} / \partial q_t(w) = \log q_t(w) + 1 - \lambda \log p_t(w) - \mu_t$. Setting this to zero gives $q_t(w) \propto p_t^\lambda(w)$, i.e. $q_t = q_t^\lambda$. Note that λ does not depend on t: there is a single constraint, hence a single multiplier, hence a single temperature. Complementary slackness and Assumption 2 force the constraint to bind, which is (9); uniqueness follows from strict monotonicity of $\lambda \mapsto \sum_t g_t(\lambda)$ under Assumption 3 (Proposition 3). □

D Proof of Theorem 5

Proof. Write $A_t(\lambda) = \sum_{i \le t} g_i(\lambda)$, let $\tilde{\lambda}_t$ be a root of A_t, so that $\lambda_t = \Pi_{[0,\Lambda]}(\tilde{\lambda}_t)$ by (12) and the clipping in Algorithm 1, and note $|A_t(\lambda^*)| \le R$ by Definition 2.

Step 1: λ_t tracks λ^*. By Taylor's theorem with Lagrange remainder,

$$A_t(\lambda) = A_t(\lambda^*) + \Big(\sum_{i \le t} g_i'(\lambda^*) \Big)(\lambda - \lambda^*) + L_t(\lambda), \qquad |L_t(\lambda)| \le \tfrac{1}{2}\, t\, L^3\, (\lambda - \lambda^*)^2, \tag{17}$$

using $|g_i''| \le L^3$ from Proposition 3. Take $\lambda = \lambda^* + \frac{2R}{ct}$. By Assumption 3, $\sum_{i \le t} g_i'(\lambda^*) \ge ct$, hence

$$A_t\Big(\lambda^* + \tfrac{2R}{ct}\Big) \;\ge\; -R + 2R - \frac{2R^2 L^3}{c^2 t} \;>\; 0 \qquad \text{whenever } t \;\ge\; T_0 := \Big\lceil \tfrac{8RL^3}{c^2} \Big\rceil.$$

Symmetrically $A_t(\lambda^* - \frac{2R}{ct}) < 0$ for $t \ge T_0$. Since A_t is nondecreasing, a root $\tilde{\lambda}_t$ lies in $[\lambda^* - \frac{2R}{ct}, \lambda^* + \frac{2R}{ct}]$, and since $\lambda^* \in [0, \Lambda]$ the projection onto $[0, \Lambda]$ does not increase the distance to λ^*, so

$$\big|\lambda_t - \lambda^*\big| \;\le\; \frac{2R}{ct} \qquad \text{for all } t \ge T_0. \tag{18}$$

Step 2: entropy loss. Split the horizon at T_0. For $t < T_0$ bound each term crudely by the maximum possible entropy difference, $\log |\mathcal{W}|$; there are $T_0 = O(R)$ such terms, contributing $O(R \log |\mathcal{W}|)$. For $t \ge T_0$, by Proposition 3 and a second-order expansion of f_t,

$$f_t(\lambda^*) - f_t(\lambda_t) \;\le\; |f_t'(\lambda^*)|\, |\lambda_t - \lambda^*| + \tfrac{1}{2} \max_\lambda |f_t''(\lambda)|\, |\lambda_t - \lambda^*|^2,$$

with $|f_t'| \le \Lambda L^2$ and $|f_t''| = |g_t' + \lambda g_t''| \le L^2 + \Lambda L^3$. Using (18),

$$\sum_{t \ge T_0} \Lambda L^2 \frac{2R}{ct} \;=\; O(R \log n), \qquad \sum_{t \ge T_0} \tfrac{1}{2}(L^2 + \Lambda L^3) \frac{4R^2}{c^2 t^2} \;\le\; O\Big(\frac{R^2}{T_0}\Big) \;=\; O(R),$$

where the second sum uses $\sum_{t \ge T_0} t^{-2} \le 1/(T_0 - 1)$ and $T_0 = \Theta(R)$. Adding the three contributions gives (15).

Step 3: global slack. The slack is $-\sum_t g_t(\lambda_t)$. For $t < T_0$, $|g_t| \le L$, contributing $O(R)$. For $t \ge T_0$, expand g_t around λ^* and use $\sum_t g_t(\lambda^*)$ bounded by R in absolute value together with $|g_t'| \le L^2$, $|g_t''| \le L^3$ and (18):

$$\Big| \sum_{t \ge T_0} g_t(\lambda_t) \Big| \;\le\; R + \sum_{t \ge T_0} L^2 \frac{2R}{ct} + \sum_{t \ge T_0} \tfrac{1}{2} L^3 \frac{4R^2}{c^2 t^2} \;=\; O(R \log n),$$

which is (16). □

E Proof of Theorem 6

Proof. (i) Theorem 5 applied at budget γ_ε gives

$$-\sum_t \sum_w q_t(w) \log p_t(w) \;\le\; \gamma_\varepsilon n + CR_\varepsilon \log n \;=\; \gamma n - \varepsilon n + CR_\varepsilon \log n \;\le\; \gamma n$$

whenever $\varepsilon n \ge CR_\varepsilon \log n$.

(ii) By Theorem 5 at budget γ_ε, the entropy of the output is at least $\mathrm{OPT}(\gamma_\varepsilon n) - CR_\varepsilon \log n$. By Lemma 1 with $b = \gamma n$ and $b' = \gamma_\varepsilon n$,

$$\mathrm{OPT}(\gamma n) - \mathrm{OPT}(\gamma_\varepsilon n) \;\le\; \lambda^*(\gamma_\varepsilon n)\, \varepsilon n \;\le\; \Lambda_\varepsilon \varepsilon n$$

using Lemma 2. Adding the two displays gives the claim. Setting $\varepsilon = CR_\varepsilon \log n/n$ satisfies the hypothesis of (i) with equality and makes the second term $\Lambda_\varepsilon CR_\varepsilon \log n$. □

F Proof of the Lower Bound, Theorem 3

W.l.o.g. let $\mathcal{W} = [|\mathcal{W}|]$ and let M be the smallest integer with $-\frac{(M-1)\log \frac{1-e^{-\gamma/3}}{M-1}}{M} + \frac{\gamma}{3M} > \frac{4\gamma}{3}$ and $\log M \ge \frac{4\gamma}{3}$. For $\frac{1}{M} \le x \le 1$ set

$$d_x(w) = \begin{cases} x, & w = M \\ \frac{1-x}{M-1}, & w \in [M-1], \end{cases} \qquad C(\lambda, x) = -\sum_w \frac{d_x^\lambda(w)}{\sum_{w'} d_x^\lambda(w')} \log d_x(w).$$

Then $C(0, e^{-\gamma/3}) > \frac{4\gamma}{3}$ and $C(\infty, e^{-\gamma/3}) \le \frac{\gamma}{3} < \frac{2\gamma}{3}$, so by continuity there are $\lambda_1, \lambda_2 \ge 0$ with

$$C(\lambda_1, e^{-\gamma/3}) = \tfrac{4\gamma}{3}, \qquad C(\lambda_2, e^{-\gamma/3}) = \tfrac{2\gamma}{3}. \tag{19}$$

Since $C(\lambda, \frac{1}{M}) = \log M \ge \frac{4\gamma}{3}$ and $C(\lambda, 1) = 0$, there are $x_1, x_2 \in [\frac{1}{M}, 1]$ with

$$C(\lambda_1, x_1) = \tfrac{2\gamma}{3}, \qquad C(\lambda_2, x_2) = \tfrac{4\gamma}{3}, \tag{20}$$

and x_3, x_4 with $C(\lambda_1, x_3) = C(\lambda_2, x_4) = \gamma$. Define

$$p_{t,1} = \begin{cases} d_{e^{-\gamma/3}} & t \le \frac{3R}{\gamma} \\ d_{x_1} & \frac{3R}{\gamma} < t \le \frac{6R}{\gamma} \\ d_{x_3} & \text{else} \end{cases} \qquad p_{t,2} = \begin{cases} d_{e^{-\gamma/3}} & t \le \frac{3R}{\gamma} \\ d_{x_2} & \frac{3R}{\gamma} < t \le \frac{6R}{\gamma} \\ d_{x_4} & \text{else.} \end{cases} \tag{21}$$

A direct computation from (19)–(21) gives $\lambda_\ell^* = \lambda_\ell$ for $\ell \in \{1, 2\}$ and

$$\max_t \left| \sum_{i \le t} \sum_w Q_i^{\mathrm{off}}(w) \log p_i(w) + \gamma t \right| = R$$

on both instances, so both have drift parameter exactly R. Taking $\mathcal{W} = [M]$ (the remaining symbols removed by the pre-filtering behind Assumption 1), both instances satisfy Assumptions 1 to 3 with ρ, c depending only on (γ, M), not on R or n: every $p_{t,\ell}$ is one of five fixed full-support distributions d_x, and each round's variance under $q_t^{\lambda_\ell}$ is positive.

The two instances agree on $t \le 3R/\gamma =: \alpha$, so any online $\mathcal{A}$ produces the same $q_t^{\mathcal{A}}$ there. Writing $B := \sum_{t \le \alpha} \sum_w q_t^{\mathcal{A}}(w) \log p_{t,1}(w) + 2R$, so $R + B$ is the prefix value of $\mathcal{A}$ in the constraint of S

below, while the offline prefix value is $-R$ on instance 1 and $+R$ on instance 2. We treat $B \geq R/2$ on instance 2; the cases $B \leq -R/2$ and $|B| < R/2$ are handled at the end.

Let $S(\{p_i\}_{i\leq t}, x)$ be the value of $\max \sum_{i\leq t} H(q_i)$ subject to $\sum_{i\leq t}\sum_w q_i(w)\log p_i(w) + \gamma t \geq x$, attained by $q_i^{\lambda_x^*}$ where λ_x^* makes the constraint bind. Since $\mathcal{A} \in \mathcal{A}_0$ is globally feasible, its tail value is at least $-R-B$, and S is nonincreasing in x, so

$$\sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \;\leq\; S(\{p_i\}_{i\leq\alpha}, R+B) + S(\{p_i\}_{i>\alpha}, -R-B).$$

With $\lambda_R^* = \lambda_{-R}^* = \lambda_2$ (the offline optimum of instance 2 has a single multiplier, and its prefix and tail values are R and $-R$) and $f_t' = -\lambda g_t'$, the same chain of identities as in the main text yields

$$\sum_{t=1}^{n} H(Q_t^{\text{off}}) - \sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \;\geq\; \sum_{i\leq\alpha}\int_{\lambda_R^*}^{\lambda_{R+B}^*} (\lambda - \lambda_R^*)\, g_i'(\lambda)\, d\lambda. \tag{22}$$

A variance lower bound. On the common prefix, $p_i = d_{e^{-\gamma/3}}$, so $\log(1/p_i(w))$ takes only the two values $a = \gamma/3$ (at $w = M$) and $b = \log\frac{M-1}{1-e^{-\gamma/3}} \geq \log(M-1) \geq \frac{4\gamma}{3}$. Let $P = q_i^{\lambda}(M)$ and $\mu = \mathbb{E}_{q_i^{\lambda}}[\log(1/p_i)] = Pa + (1-P)b$. For $\lambda_R^* \leq \lambda \leq \lambda_{3R/2}^*$ one has $\mu \in [\frac{\gamma}{2}, \frac{2\gamma}{3}]$ (this is the definition of $\lambda_{3R/2}^*$ together with monotonicity of g). Then

$$1 - P = \frac{\mu - a}{b-a} \;\geq\; \frac{\gamma/6}{b-a}, \qquad P = \frac{b-\mu}{b-a} \;\geq\; \frac{2\gamma/3}{b-a},$$

so that

$$g_i'(\lambda) = \mathrm{Var}_{q_i^{\lambda}}\left[\log\tfrac{1}{p_i}\right] = P(1-P)(b-a)^2 \;\geq\; \frac{2\gamma}{3}\cdot\frac{\gamma}{6} \;=\; \frac{\gamma^2}{9}. \tag{23}$$

The gap in λ. Let $\Delta := \lambda_{3R/2}^* - \lambda_R^*$. By definition of these roots,

$$\int_{\lambda_R^*}^{\lambda_{3R/2}^*} \sum_{i\leq\alpha} g_i'(\lambda)\, d\lambda \;=\; \frac{R}{2},$$

and $\sum_{i\leq\alpha} g_i'(\lambda) \leq \alpha G$ with $G = (\frac{4\gamma}{3} + \log\frac{1}{1-e^{-\gamma/3}})^2$ an upper bound on each g_i'. Since $\alpha = 3R/\gamma$,

$$\Delta \;\geq\; \frac{R/2}{(3R/\gamma)G} \;=\; \frac{\gamma}{6G}. \tag{24}$$

Note that $\Delta = \Theta(1)$, not $\Theta(R)$: the prefix has $\Theta(R)$ rounds, each contributing curvature, so a constraint shift of $R/2$ moves the multiplier by only a constant.

Conclusion. Combining (22), (23) and (24), and using $\lambda_{R+B}^* \geq \lambda_{3R/2}^*$ (since $B \geq R/2$),

$$\sum_{t=1}^{n} H(Q_t^{\text{off}}) - \sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \;\geq\; \alpha\cdot\frac{\gamma^2}{9}\int_0^{\Delta} u\, du \;=\; \frac{3R}{\gamma}\cdot\frac{\gamma^2}{9}\cdot\frac{\Delta^2}{2} \;\geq\; \frac{R\gamma}{6}\cdot\frac{\gamma^2}{36G^2} \;=\; \frac{\gamma^3}{216\,G^2}\,R,$$

which is (13).

Remaining cases. Write $y \mapsto \lambda(y)$ for the prefix root at value y (nondecreasing, $\lambda(-R) = \lambda_1$, $\lambda(R) = \lambda_2$); the substitution $dy = \sum_{i \le \alpha} g_i'(\lambda)\, d\lambda$ turns (22) into $\int_{y^{\text{off}}}^{y^{\mathcal{A}}} \big(\lambda(y) - \lambda(y^{\text{off}})\big)\, dy$. (a) $B \le -R/2$, instance 2: on $[\lambda(R/2), \lambda(R)]$ one has $\mu \in [\frac{2\gamma}{3}, \frac{5\gamma}{6}]$, so the two-point computation gives $g_i' \ge \frac{\gamma}{3} \cdot \frac{\gamma}{2} \ge \frac{\gamma^2}{9}$, and $\lambda(R) - \lambda(R/2) \ge \gamma/(6G)$ as in (24); the conclusion is identical. (b) $|B| < R/2$, instance 1: here $y^{\mathcal{A}} = R + B > R/2$ and $y^{\text{off}} = -R$, so the loss is at least $\int_0^{R/2} \big(\lambda(y) - \lambda(-R)\big) dy \ge \frac{R}{2}\big(\lambda(0) - \lambda(-R)\big) \ge \frac{R}{2} \cdot \frac{R}{\alpha G} = \frac{\gamma R}{6G} \ge \kappa R$, using $\sum_{i \le \alpha} g_i' \le \alpha G$ and $36G \ge \gamma^2$. □

G Proof of Theorem 4

Proof. Run the construction of Section F verbatim. The only place exact feasibility was used is the decomposition $\sum_t H(q_t^{\mathcal{A}}) \le S(\{p_i\}_{i \le \alpha}, R + B) + S(\{p_i\}_{i > \alpha}, -R - B)$, which encodes that a prefix overspend of B must be repaid in the tail. If $\mathcal{A}$ is merely V-feasible, the tail budget is loosened by V and the correct decomposition is

$$\sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \;\le\; S(\{p_i\}_{i \le \alpha}, R + B) + S(\{p_i\}_{i > \alpha}, -R - B - V).$$

The value function $S(\cdot, x)$ is nonincreasing in x with $\partial S / \partial x = -\lambda_x^*$ (Lemma 1 applied to the tail subproblem), and $\lambda_x^* \le \Lambda$ by the hypothesis of Theorem 4, so

$$S(\{p_i\}_{i > \alpha}, -R - B - V) - S(\{p_i\}_{i > \alpha}, -R - B) \;\le\; \Lambda V.$$

Subtracting this from the bound of Theorem 3 gives (14). □

H Proof of Theorem 7

Proof. Let $g^{(i)}(\lambda) = \gamma + \sum_w q^{\lambda,(i)}(w) \log p^{(i)}(w)$ be the constraint function of component i, let $t_i = |\{j \le t : p_j = p^{(i)}\}|$, and let $\hat{P}_i = n_i/n$. By Assumption 1, $|g^{(i)}(\lambda)| \le \log(1/\rho) =: L$ for every $\lambda \ge 0$.

The key identity. By definition of λ^* (Proposition 2), $\sum_{t \le n} g_t(\lambda^*) = 0$; grouping the rounds by component,

$$\sum_{i \in [k]} n \hat{P}_i\, g^{(i)}(\lambda^*) \;=\; 0. \tag{25}$$

Consequence. For every t,

$$R_t = \Big| \sum_{i \in [k]} t_i\, g^{(i)}(\lambda^*) \Big| = \Big| \sum_{i \in [k]} \big(t_i - t\hat{P}_i\big) g^{(i)}(\lambda^*) \Big| \;\le\; L \sum_{i \in [k]} \big| t_i - t\hat{P}_i \big|,$$

where the second equality uses (25) scaled by t/n. By the triangle inequality,

$$\sum_i \big| t_i - t\hat{P}_i \big| \;\le\; \sum_i \big| t_i - tP_i \big| + t \sum_i \big| \hat{P}_i - P_i \big|.$$

Uniform concentration. Apply Lemma 3 at each sample size $t \in [n]$ with confidence $\delta/(n+1)$, and once more at $t = n$ for $\sum_i |\hat{P}_i - P_i|$. On the intersection event, which has probability at least $1 - \delta$ by a union bound over the $n + 1$ applications, we have simultaneously for all t

$$\sum_i |t_i - tP_i| \;\le\; \sqrt{2t\Big(k \log 2 + \log \tfrac{n+1}{\delta}\Big)} \quad \text{and} \quad t \sum_i |\hat{P}_i - P_i| \;\le\; t\sqrt{\tfrac{2(k \log 2 + \log \frac{n+1}{\delta})}{n}}.$$

Both are at most $\sqrt{2n(k\log 2 + \log\frac{n+1}{\delta})}$ for $t \le n$, so

$$R = \max_{t\in[n]} R_t \;\le\; 2L\sqrt{2n\Big(k\log 2 + \log\tfrac{n+1}{\delta}\Big)},$$

as claimed. □

Remark 6. The union bound over prefixes is what converts a statement about the empirical frequencies at a single sample size into the bound on $R = \max_t R_t$ that the theorem actually needs; it costs $\log(1/\delta) \to \log(n/\delta)$ and preserves the $\tilde{O}(\sqrt{n})$ rate.

I Experimental Details

Curvature. Figure 3 shows the cumulative variance under the offline temperature in the exogenous ELI5 experiment.

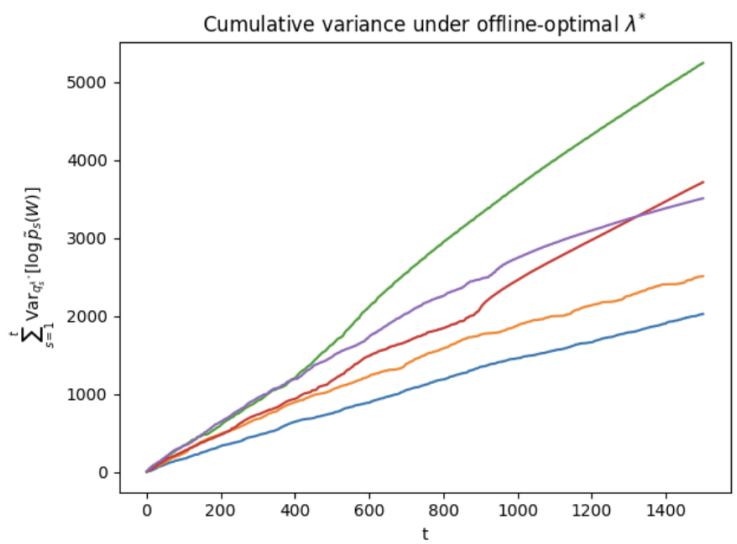

Figure 3: Cumulative variance $\sum_{i\le t} \mathrm{Var}_{q_i^{\lambda^*}}[\log(1/p_i)]$ against t. Approximate linearity is direct evidence for Assumption 3; c may be read off as the slope.

Autoregressive setup. Meta-Llama-3-8B-Instruct (bf16) generates $n = 1024$ tokens for 200 ELI5 questions, 4 samples per prompt and method, on one AMD MI210 GPU per run, with end-of-sequence tokens suppressed so every method is compared on the same horizon. The support at each step is $\mathcal{W}_t = \{w : p_t(w) \ge \rho\}$, $\rho = 10^{-4}$, capped at the 256 most likely tokens, and $\Lambda = 50$. λ_t solves (12) to tolerance $10^{-6}t$ by Newton's method warm-started at λ_{t-1}, using that $\lambda \mapsto \sum_{i\le t} g_i(\lambda)$ has exact derivative $\sum_{i\le t} \mathrm{Var}_{q_i^\lambda}[\log(1/p_i)]$ (so the left-hand side of (12) has derivative $-\sum_{i\le t} \mathrm{Var}_{q_i^\lambda}[\log(1/p_i)]$), with a bisection safeguard (about two evaluations of the prefix sum per token). Methods: EOS with $\gamma \in \{0.5, 0.9, 1.3\}$; its tightened variant (Theorem 6) with $\varepsilon = 0.1\gamma$; a re-solving heuristic that spreads the remaining budget $\gamma n - \sum_{i<t} \mathrm{CE}_i$ evenly over the remaining rounds (not covered by our theory); fixed temperature $T \in \{0.3, 0.5, 0.7, 0.85, 1.0, 1.03, 1.06, 1.09, 1.12, 1.15, 1.3\}$; nucleus $p \in \{0.9, 0.95\}$ (Holtzman et al., 2020); min-$p \in \{0.05, 0.1\}$; locally typical sampling (0.9); and

greedy decoding. For every method we log $H(q_t)$ and $\mathrm{CE}_t = -\sum_w q_t(w) \log p_t(w)$ against the full model distribution, from the raw logits, so the realized slack is on the same scale for all methods.

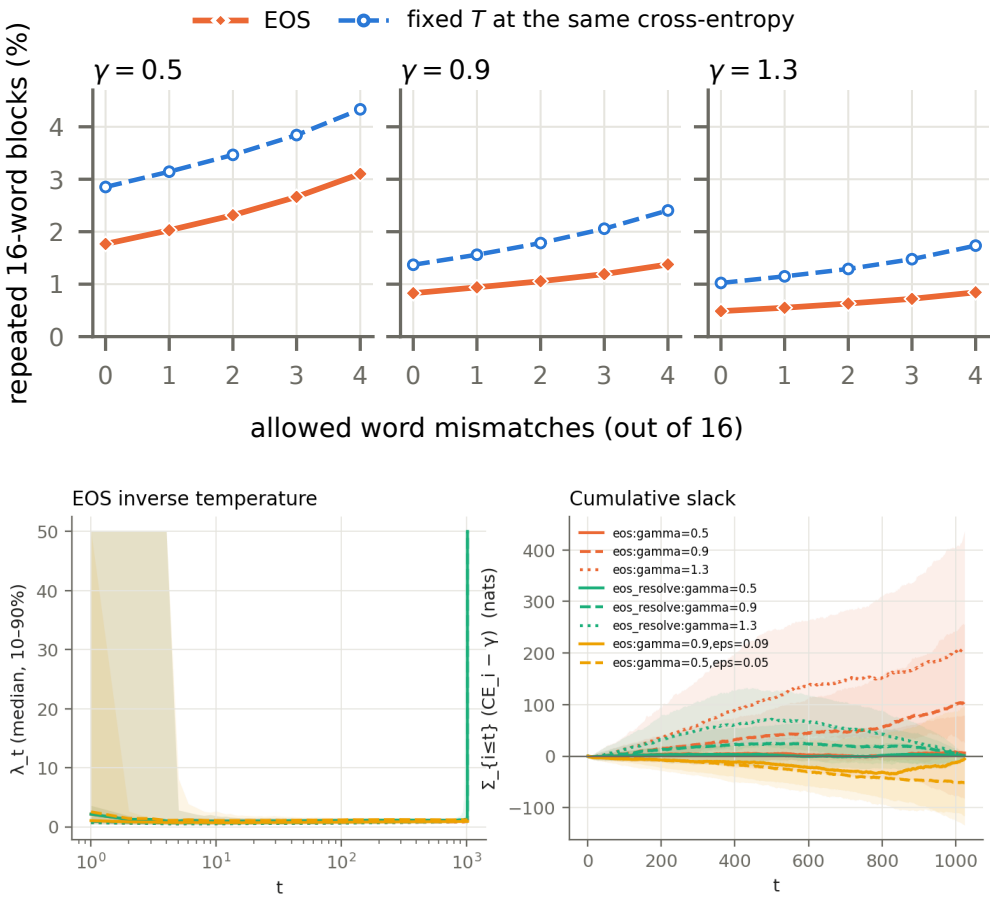

Figure 4: Top: repeated 16-word blocks as the number of allowed word mismatches grows, for EOS and for fixed temperature at the same realized cross-entropy; other samplers are compared at a single tolerance in Figure 2. Bottom: EOS inverse temperature λ_t and cumulative overshoot $\sum_{i \le t}(\mathrm{CE}_i - \gamma)$ (median, 10–90% band).

Repetition metrics. Approximate repetition has no single definition, so we report a family of metrics indexed by a tolerance, each computed on model tokens and on normalized words (lower-cased, digits collapsed, punctuation removed). (i) *Exact recurrence*: Λ_t, the length of the longest string starting at t that already occurs earlier (Wyner and Ziv, 1989; Ornstein and Weiss, 1993); Proposition 1 predicts $\Lambda_t \approx \log t/H$, and we report the entropy-rate estimate $\hat{H} = [\mathrm{mean}_t\,(\Lambda_t + 1)/\log t]^{-1}$, averaged over positions $2 \le t \le 0.95n$ whose match is not truncated by the end of the output. (ii) *Approximate recurrence*: for $m \in \{8, 16, 32\}$ and $\tau \in [0, 0.3]$, the fraction of positions whose length-m block is within Hamming distance τm of an earlier block; $\tau = 0$ is exact repetition. (iii) *Loops and n-grams*: fraction of the output in tandem repeats (period ≤ 64, at least three periods), seq-rep-n and rep-ℓ (Welleck et al., 2020), and LZ78 and zstd compression ratios. (iv) *Sentence level*: fraction of sentences whose normalized edit similarity to an earlier sentence exceeds $\theta = 0.8$. Diversity across the samples of a prompt (pairwise 4-gram Jaccard overlap) is reported separately. The headline approximate-repetition metric uses $m = 16$ and $\tau = 0.1$, i.e.

repeated 16-word blocks with at most one mismatched word.

Comparison protocol. Entropy and repetition both fall as the budget tightens, so all comparisons are at matched realized cross-entropy: (a) the *frontier*, every method in the plane of realized cross-entropy against a metric, with the fixed-T curve; (b) the *offline comparator*, the fixed-T curve interpolated at each prompt's own realized cross-entropy, whose gap to EOS estimates the entropy loss of Theorem 5; (c) *budget adherence*, a single temperature calibrated to meet γ on average over a held-out 25% of prompts, evaluated on the rest. Intervals are 95% bootstrap intervals over prompts (2000 resamples).

Additional results. At $\gamma = 0.9$ the match-length entropy-rate estimate rises from 2.63 (fixed temperature at the same cross-entropy) to 2.83 nats per word, and LZ78 entropy from 3.71 to 3.78, so the reduction in repetition is also visible in compression-based measures. The reduction holds at every edit tolerance, from exact repeats to blocks differing in a quarter of their words (Figure 4, top). Nucleus and locally typical sampling are no less repetitive than fixed temperature at matched cross-entropy, and min-p (0.05) only slightly. The default $T = 0.7$ spends 0.35 nats per token (entropy 0.32), leaving most of the $\gamma = 0.9$ budget unspent. The overshoot of plain EOS grows with the budget (realized cross-entropy 0.50, 1.01, 1.53 at $\gamma = 0.5, 0.9, 1.3$); the tightened variant and the re-solving heuristic correct it (Table 1). A single temperature calibrated to γ has a larger mean excess than plain EOS, because its per-prompt cross-entropy varies almost five times as much.

Table 1: Per-prompt budget adherence on 150 held-out prompts, $\gamma = 0.9$. CE is realized cross-entropy (nats/token); sd is across prompts; excess is $\frac{1}{|\mathcal{P}|}\sum_p [\mathrm{CE}_p - \gamma]_+$ (prompts within budget count as zero).

	mean CE	sd	% over	excess	H
fixed T=1.03 (calibrated)	0.982	0.332	49	0.166	0.981
EOS	1.012	0.071	95	0.112	0.973
EOS, tightened	0.890	0.068	39	0.021	0.863
re-solving (heuristic)	0.900	0.010	53	0.001	0.859