All Repetition and No Creativity Makes GenAI a Dull Boy

Jin Sima [1] **Nikolaos Papagiannis** [1] **Ananth Grama** [1] **Wojciech Szpankowski** [1]

Abstract

Degeneracy in Large Language Models (LLMs) refers to the problem of repetitive low-quality, high-probability tokens, which often lead to loss of meaning and coherence. This remains a persistent failure mode in current generative models. While next-token heuristics such as temperature scaling, nucleus sampling, and top-k sampling are widely used in practice, their effectiveness lacks a unifying theoretical foundation. In this paper, we present theoretical underpinnings of temperature-based sampling and related heuristics, identify when and why they mitigate degeneracy, and use this analysis to propose a new adaptive temperature control algorithm. We establish a principled connection between repetitiveness and the entropy of the sampling distribution. Leveraging Kac's Lemma, we show that the expected time to repetition is asymptotically proportional to the entropy of the sampled-token distribution, and prove that maximizing this entropy yields an optimal *offline* sampling strategy. We extend this insight to more realistic settings with only partial knowledge of evolving model distributions, to formulate an *online* optimization objective that maximizes the *average online entropy*, while constraining next-token selection to those with baseline model probability. Using this formulation, we derive an adaptive temperature-control algorithm and establish theoretical guarantees on its stability and performance. Empirical evaluations support our theoretical foundations and demonstrate that the proposed method not only recovers standard sampling heuristics as special cases, but achieves superior performance in non-repetitive sequence generation compared to strong prior baselines.

[1]Purdue University, W. Lafayette, U.S.A. Correspondence to: Wojciech Szpankowski <szpan@purdue.edu>, Jin Sima <simaj@purdue.edu>.

Preprint. January 29, 2026.

1. Introduction

Large language models (LLMs) have achieved remarkable success in natural language understanding, code generation, image synthesis, and related applications (Radford et al., 2019; Brown et al., 2020). Despite these advances, a key limitation remains: models trained offline on static corpora often fail to maintain reliable and diverse outputs in dynamic, real-world environments. Degenerate behaviors such as repetitiveness, model collapse, and weakly-grounded predictions frequently emerge when the model encounters distributions that differ from its training data (Holtzman et al., 2020; Meister & Cotterell, 2023). Such degeneration has been widely experimentally observed in neural text generation systems and linked to low-entropy decoding and highly concentrated token distributions (Holtzman et al., 2020; Welleck et al., 2020).

Existing offline-trained LLMs cannot adapt in real time to such shifts, limiting their applicability in timely, context-aware, and varied generation. Current mitigation strategies such as temperature scaling, top-k/nucleus sampling, repetition penalties, and n-gram blocking promote diversity in controlled settings, but are largely static and do not respond well to evolving input distributions (Fan et al., 2018; Li et al., 2016). Consequently, these heuristics provide limited solutions to long-horizon degeneracy and distribution drift.

This motivates the need for robust *online* models that dynamically adapt their output distributions while balancing fidelity, diversity, and informativeness. From an online learning perspective, generation can be viewed as a sequential decision process under uncertainty, naturally connecting to classical frameworks in online prediction and adaptive learning (Cesa-Bianchi & Lugosi, 2006; Orabona, 2019). By optimizing generation strategies to maintain high entropy and representational diversity, we show that online adaptation directly addresses degenerate behaviors This highlights the fact that diversity is important for expressive, non-repetitive, and robust sequence generation.

Efficient online adaptation of LLMs is a major challenge, as naïve approaches risk catastrophic forgetting, instability, or quality degradation from noisy updates unless carefully controlled. When aggregated over time, these factors potentially render LLMs largely unusable. Recent approaches, such as online fine-tuning and adaptive optimization, highlight

both the promise and complexity of dynamic adaptation. Integrating formalisms from online learning with generative modeling presents significant opportunities, improving efficiency over batch retraining, ensuring robustness in non-stationary environments, and leveraging mathematical tools for rigorous performance guaranties. This interplay between the foundations of online learning and concepts of generative modeling forms the core motivation for our work.

Our Contributions. We conceptualize a generative model as an autoregressive sequence generator. At each discrete time instant t, a conditional probability distribution for token w, $p_t := P(w|w^{t-1})$, is computed by the generative model based on previously sampled tokens $w^{t-1} := (w_1, \ldots, w_{t-1})$. This conditional distribution is used to construct a *sampling distribution* $q_t(w)$, from which the next token w_t is selected. Our goal is to construct the sampling distribution that yields optimal generation performance, which in turn requires a clear optimization criterion.

In addition to high likelihood, generative models must also optimize for expressiveness, or equivalently, the avoidance of repetition. Informally, a desirable generative model is one in which a typical token sequence w^n does not repeat too quickly (Holtzman et al., 2020; Meister & Cotterell, 2023). To formalize this notion, we argue that the recurrence time (or return time) $R(w^n)$, which characterizes the average number of tokens (or time) to return to the original state, be as large as possible. A rigorous foundation for this concept is provided by Kac's Lemma (Kac, 1947), which states that for stationary ergodic processes W, the recurrence time satisfies $R(w^n) \sim 1/P(w^n)$, where $P(w^n)$ is the probability of observing the sequence of tokens w^n of length n. Moreover, for a typical token sequence w^n, the asymptotic equipartition property (AEP) in information theory (Cover & Thomas, 2006) tells us that $P(w^n) \sim 2^{-nH(W)}$, where $H(W)$ denotes the entropy rate of the underlying process generating w^n. Consequently, $R(w^n) \sim 2^{nH(W)}$, indicating that maximizing recurrence time is asymptotically equivalent to maximizing entropy of the generative process.

In our setting, distribution W corresponds to the sampling distribution q_t. Our objective is therefore to maximize the entropy $H(q_t)$. At the same time, the sampling distribution should produce "reasonable tokens" whose probability is above a prescribed threshold. In practice, the model probability of a sampled token sequence decays exponentially with sequence length n. It is desirable that the exponent be upper bounded by some γ to avoid tokens with very small probability, as implemented in current AI systems. Hence, the model probability of a "reasonable" token sequence is lower bounded by the threshold $2^{-n\gamma}$. In Theorem 3.1, we prove that this condition is equivalent to

$$-\sum_{t\in[n]}\sum_{w} q_t(w)\,\log P\big(w \mid w^{t-1}\big) \le \gamma n.$$

In Theorem 3.2, we show that the problem of maximizing the entropy of the generative process under the probability threshold constraint yields an optimal sampling distribution of the form of temperature sampling:

$$q_t(w) = \frac{P^{\,\lambda^*}\big(w \mid w^{t-1}\big)}{\sum_{w'} P^{\,\lambda^*}(w' \mid w^{t-1})},$$

where $\lambda^* = 1/T^*$ where T^* is the optimal temperature.

In practice, however, temperature is typically chosen heuristically (e.g., $T = 0.7$), without any guarantee of matching the optimal value T^*. Beyond this mismatch, more fundamental limitations arise: existing offline-trained large language models are unable to adapt their sampling distributions in real time in response to distributional shifts or evolving contexts. This lack of adaptability restricts their effectiveness in applications that require timely, context-aware, and diverse generation. To address these challenges, we next introduce an *online* formulation of the constrained entropy maximization problem. Informally, in the *online entropy optimization problem*, we consider the objective:

$$\max_{\{q_t\}_{t=1}^n} \sum_{t=1}^{n} H(q_t)\,, \tag{1}$$

with sampling distribution q_t depending *only* on the probability distributions $\{p_i\}_{i=1}^t$ observed until time t, subject to probability threshold constraint $w^n \in \mathcal{W}_\gamma^n := \{w^n : P(w^n) \ge 2^{-n\gamma}\}$ for $\gamma > 0$, and $\mathcal{W}^n$ being the set of all sequences of tokens of length n. Please see (2)-(3) in Section 2 for a formal statement of this optimization problem.

Our goal is to solve this problem in full generality and to develop an implementable and computationally efficient Entropic Online Sampling (EOS) algorithm (Algorithm 1). We begin by analyzing the online optimization problem formalized in Section 2 and its solution in Lemma 3.3. Although this lemma provides a complete characterization of the optimal solution, the resulting strategy is not directly implementable since it requires knowledge of all token probability distributions a-priori. To address this limitation, we design an approximate online algorithm (Algorithm 1) that is computationally tractable. We prove in Theorems 3.4 and 3.5 that this algorithm is asymptotically optimal, i.e., the sampling distribution it produces converges to the optimal online sampling distribution. Finally, in Section 4, we validate our approach empirically using real-world data, demonstrating that our EOS algorithm consistently outperforms existing heuristic methods.

Relevant Literature. Research on large language models and text generation has demonstrated impressive capabilities across a wide range of tasks. A growing body of work has shown that standard maximum-likelihood training combined with deterministic or low-entropy decoding often leads to

degenerate outputs, characterized by repetition, blandness, and model collapse. Early analyses identified beam search and greedy decoding as major contributors to these failures, motivating strategies designed to preserve diversity (Holtzman et al., 2020; Meister & Cotterell, 2023).

A significant body of literature proposes static decoding heuristics to mitigate degeneracy, including temperature scaling, top-k sampling, nucleus (top-p) sampling, and repetition penalties. These techniques have proven effective in controlled settings but operate independently of the data-generation process and cannot adapt to nonstationary or adversarial input streams (Fan et al., 2018; Li et al., 2016). Alternative training-time approaches, such as unlikelihood training and diversity-promoting objectives, penalize repetitive patterns directly, but still rely on offline optimization and fixed distributions (Welleck et al., 2020; Li et al., 2016).

From a theoretical perspective, degeneracy can be interpreted as a collapse of the model's predictive entropy and effective support. Information-theoretic analyses of language models have linked expressive generation to entropy, uncertainty calibration, and distributional smoothness. Recent work has examined the mismatch between likelihood-based training objectives and generation-time decoding, emphasizing that standard objectives do not guarantee desirable sequence-level properties (Meister & Cotterell, 2023).

Separately, the online learning and adaptive prediction literature provides a principled framework for decision-making under non-stationarity. Classical results in online convex optimization, mirror descent, and adversarial prediction establish regret guarantees and adaptive behavior in dynamic environments (Cesa-Bianchi & Lugosi, 2006; Orabona, 2019; Wu et al., 2025). While these ideas have been successfully applied to bandits, control, and reinforcement learning, their integration into autoregressive sequence generation remains limited. Some early work explores neural networks as online learners, but without explicit guarantees on diversity or non-repetition (Graves et al., 2014).

More recently, robustness and distribution shift have emerged as central challenges for foundation models. Studies on dataset shift, uncertainty estimation, and calibration highlight the brittleness of large models when deployed in evolving environments (Ovadia et al., 2019; Guo et al., 2017). However, these results primarily focus on predictive accuracy rather than generative diversity.

In contrast to prior approaches, our work bridges degenerate outputs, information-theoretic diversity, and optimization in an online setting. We formulate online non-repetitive generation as an adaptive sequential process and propose entropy-aware strategies that dynamically adjust generation behavior in response to observed data, providing both theoretical guarantees and practical relevance for robust LLMs.

2. Problem Setup

Let $\mathcal{W}$ denote the alphabet for the tokens and let $\mathcal{W}^n$ be set of the sequences of tokens of length n. Let $p_t(w) = P(w|w^{t-1}) \in \Delta(\mathcal{W})$, be the probability distribution generated by the model (which can be viewed as the last layer of the transformer) when the input token sequence is w^{t-1}, where $\Delta(\mathcal{W})$ is the set of all probability distributions over $\mathcal{W}$. The probability distribution $P(w^n) = \prod_{t=1}^n P(w_t|w^{t-1}) = \prod_{t=1}^n p_t(w_t)$ represents the model's estimate of the probability distribution over token sequences $w^n \in \mathcal{W}^n$, and can be obtained by the conditional distributions $P(w_t|w^{t-1})$, $t \in [n] = \{1, \ldots, n\}$ given by the model. Let $\mathcal{W}^n_\gamma = \{w^n : P(w^n) \geq 2^{-n\gamma}\}$ be the set of length n sequences with model probability at least $2^{-n\gamma}$. The set $\mathcal{W}^n_\gamma$ represents sequences that are considered *reasonable* by the model.

The goal is to sample sequences w^n in a sequential manner such that w^n is, with high probability, in $\mathcal{W}^n_\gamma$. Specifically, w_t is sampled sequentially for $t \in [n]$ according to a sampling distribution $q_t = \phi_t(\{p_i\}_{i=1}^t) \in \Delta(\mathcal{W})$, which is a function of prior distributions, p_i, known up to time t. Therefore, the distribution of a sampled sequence w^n is $\prod_{t=1}^n q_t(w_t)$. When $q_t = p_t$, w^n has distribution $P(w^n)$. Note that widely used sampling algorithms, such as top-k, top-p, or temperature sampling, are specific examples of q_t. In general, the conditional distribution $p_t(w) = P(w|w^{t-1})$ depends on the sampled tokens w^{t-1} in the past. However, since the behavior of such dependence is hard to characterize and the effect of the choice of token w_t on future conditional distributions $P(w_{t'}|w^{t'-1})$, $t' \geq t$, is unpredictable, we treat $p_t(w) = P(w|w^{t-1})$ as arbitrarily given.

One strategy to sample sequences with high model probability is to sample every token with the maximum conditional probability, i.e., $q_t(w) = 1$ if $w = \arg\max_{w \in \mathcal{W}} p_t(w)$ and 0 otherwise. However, this strategy generates deterministic sequences, which reflects lack of expressiveness or creativity, and results in degenerate (repeated) sequences. We are interested in the tradeoff between the model probability guarantee γ (which is the measure of goodness of generated sequences considered in this paper) and the creativity/ expressiveness of the sampled sequence, which is measured by the entropy of the sampling probability distributions q_t, $t \in [n]$, and is related to the repetitiveness of the sampled sequence. Specifically, we define the following *online optimization problem*

$$\max_{q_t = \phi_t(\{p_i\}_{i=1}^t) \in \Delta(\mathcal{W}): t \in [n]} \sum_{t=1}^n H(q_t) \tag{2}$$

$$s.t. \quad -\sum_{t=1}^n \sum_{w \in \mathcal{W}} q_t(w) \log p_t(w) \leq \gamma n, \tag{3}$$

where γ is a given parameter. We justify below that a

solution to the above optimization problem achieves the creativity-probability tradeoff considered in the paper. We will also provide a theoretical justification for the commonly used temperature sampling as the solution to the above online optimization problem.

Note that Problem (2) corresponds to *online* optimization of q_t, $t \in [n]$ in the sense that no knowledge of p_i, $i > t$ is assumed. Consider, in contrast, the following offline (static) problem, where $\{q_t\}_{t=1}^n$ are optimized with $\{p_t\}_{t=1}^n$ known:

$$\max_{q_t \in \Delta(\mathcal{W}): t \in [n]} \sum_{t=1}^{n} H(q_t)$$
$$s.t. \quad -\sum_{t=1}^{n} \sum_{w \in \mathcal{W}} q_t(w) \log p_t(w) \leq \gamma n \tag{4}$$

In Theorem 3.2, we show that the optimal solution to (4) is temperature sampling with fixed temperature. However, this may not hold for the online setting, as shown in Lemma 3.3. In Algorithm 1 we present an online learning algorithm for (2) and prove in Theorems 3.4 and 3.5 that the achieved entropy is less than the optimal entropy for the offline problem (4) by at most $o(n)$ with high probability.

3. Main Results

In this section, we present our main theoretical results. We first show in Theorem 3.1 that the probability threshold constraint for "reasonable words" $\mathcal{W}_\gamma^n$ is asymptotically equivalent to our condition (3). We present an offline optimal solution in Theorem 3.2 that is hard to implement in real-world settings. We then present an efficient online algorithm (Algorithm 1) that is asymptotically optimal under mild assumptions (Theorem 3.5).

3.1. Preliminary Results

We start by establishing three preliminary results that highlight key aspects of our online optimization formulation.

Theorem 3.1. *Let w^n be the sequence sampled by q_t, $t \in [n]$ sequentially, i.e., the probability of sampling w^n is $\prod_{t\in[n]} q_t(w_t)$. Then, if:*

$$\sum_{t\in[n]} \sum_{w\in\mathcal{W}} q_t(w) \log p_t(w) + \gamma n \geq -O(\sqrt{n \log \frac{1}{\delta}}), \tag{5}$$

we have:

$$\Pr(P(w^n) \geq 2^{-\gamma n - O(\sqrt{n \log \frac{1}{\delta}})}) \geq 1 - \delta. \tag{6}$$

Conversely, if (6) holds and $p_t(w) \geq \rho$ for $w \in \mathcal{W}$ and for some constant $\rho > 0$, we have:

$$\sum_{t\in[n]} \sum_{w\in\mathcal{W}} q_t(w) \log p_t(w) + (1-\delta)\gamma n$$
$$\geq -O(\sqrt{n \log \frac{1}{\delta}}) - \delta n (\log \frac{1}{\rho} + \gamma). \tag{7}$$

By choosing $\delta = O(\frac{1}{\sqrt{n}})$, the constraint $w^n \in \mathcal{W}_\gamma^n$ is asymptotically equivalent to (3) with high probability.

Proof. Define the random variables $X_t = \log P(w_t|w^{t-1}) + \gamma$, $t \in [n]$. Then, $\mathbb{E}_{q_t}[X_t] = \sum_{w\in\mathcal{W}} q_t(w) \log p_t(w) + \gamma$. Let $S_t = \sum_{i=1}^{t} (X_t - \mathbb{E}_{q_t}[X_t])$, for $t \in [n]$. Observe that S_t is a martingale. Moreover, $|S_{t+1} - S_t| \leq 2 \log \frac{1}{\rho}$. By the Azuma's inequality (Szpankowski, 2001), we have that:

$$\Pr(|S_t| \geq \epsilon) \leq \exp\left(\frac{-\epsilon^2}{4n \log \frac{1}{\rho}}\right).$$

Thus, with probability at least $1-\delta$, we have $|S_t| \leq \sqrt{4n \log \frac{1}{\rho} \log \frac{1}{\delta}}$. Therefore, with probability at least $1-\delta$,

$$\log P(w^n) + \gamma n = \sum_{t\in[n]} X_t$$
$$= \sum_{t\in[n]} \sum_{w\in\mathcal{W}} q_t(w) \log p_t(w) + n\gamma + S_t$$
$$\geq \sum_{t\in[n]} \sum_{w\in\mathcal{W}} q_t(w) \log p_t(w) + n\gamma - \sqrt{4n \log \frac{1}{\rho} \log \frac{1}{\delta}}.$$

In summary, if (5) holds, we have (6). Similarly, if (6) holds, we have:

$$\sum_{t\in[n]} \sum_{w\in\mathcal{W}} q_t(w) \log p_t(w) + n\gamma = \sum_{t\in[n]} \mathbb{E}_{q_t}[X_t]$$
$$\geq (1-\delta)(-\gamma n - O(\sqrt{n \log \frac{1}{\delta}})) - \delta n (\log \frac{1}{\rho} + \gamma),$$

which completes the proof. □

The next two results present solutions to the offline and online optimization problems, respectively.

Theorem 3.2. *The optimal solution for the offline optimization problem (4) is given by*

$$Q_t^{off}(w) = \frac{p_t^{\lambda^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda^*}(w)}, \tag{8}$$

where λ^* is the unique solution of

$$-\sum_{t\in[n]} \sum_{w\in\mathcal{W}} \frac{p_t^{\lambda^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda^*}(w)} \log p_t(w) = \gamma n. \tag{9}$$

Proof. The proof follows from the Karush–Kuhn–Tucker conditions for (4). Note that the objective function $\sum_{t\in[n]} H(q_t)$ is concave in q_t, $t \in [n]$, and the constraint is linear in q_t, $t \in [n]$. Hence, (4) is a convex optimization problem. Consider the Lagrangian:

$$\mathcal{L}(\lambda, \mu_1, \ldots, \mu_n) = \sum_{t\in[n]} \mu_t(1 - \sum_{w\in\mathcal{W}} q_t(w)) - \sum_{t\in[n]} H(q_t) - \lambda(\sum_{t\in[n]}\sum_{w\in\mathcal{W}} q_t(w)\log p_t(w) + \gamma n).$$

The partial derivative with respect to $q_t(w)$ is given by

$$\frac{\partial \mathcal{L}}{\partial q_t(w)} = \log q_t(w) + \frac{1}{\ln 2} - \lambda \log p_t(w) - \mu_t.$$

Solving $\frac{\partial \mathcal{L}}{\partial q_t(w)} = 0$, we have $q_t(w) = \exp\left(\mu_t - \frac{1}{\ln 2}\right) p_t^{\lambda}(w)$. Combining with $\sum_{w\in\mathcal{W}} q_t(w) = 1$, we have $q_t(w) = \frac{p_t^{\lambda}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda}(w)}$. Also, $q_t(w)$ satisfies the constraint in (4), hence (8) and (9) follow. □

Next we provide a solution to the optimization problem with the proof in Appendix A.

Lemma 3.3. *The optimal solution to the online problem (2) and (3) is given by:*

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

for some λ_t^{on}, $t \in [n]$, such that:

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

3.2. An Efficient Online Algorithm

Our solution to the online optimization problem is not algorithmically efficient and hard to realize in practice (i.e., in an online setting). To handle this, in the next two main results, we first establish a lower bound on the entropy loss for any online algorithm, when compared to the offline algorithm. We then design an approximate, efficient online Algorithm 1 that we prove is asymptotically online optimal.

We start with a general lower bound proved in Appendix B. Throughout the rest of the paper we will assume that $p_t(w) \geq \rho$ for some $\rho > 0$.

Theorem 3.4 (Lower Bound)**.** *For any online algorithm $\mathcal{A}$ that selects q_t as a function of $\{p_i\}_{i=1}^t$ and satisfies (3), let $q_t^{\mathcal{A}}$, $t \in [n]$, be the output of Algorithm $\mathcal{A}$. Then for any γ and $R \leq \frac{\gamma n}{6}$ satisfying:*

$$-\frac{(|\mathcal{W}|-1)\log(\frac{1-\exp(-\frac{\gamma}{3})}{|\mathcal{W}|-1})}{|\mathcal{W}|} + \frac{\gamma}{3|\mathcal{W}|} > \frac{4\gamma}{3}$$

Algorithm 1 Solving λ_t

Input: γ, n, p_t at time t.
Output: q_t at time t.
for $t = 1$ **to** n **do**
Let $q_t^{on}(w) = \frac{p_t^{\lambda_t}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda_t}(w)}$, where λ_t satisfies

$$\sum_{i\in[t]} -\sum_{w\in\mathcal{W}} \frac{p_i^{\lambda_t}(w)}{\sum_{w\in\mathcal{W}} p_i^{\lambda_t}(w)} \log p_i(w) = \gamma t. \tag{14}$$

end for

and $\log|\mathcal{W}| \geq \frac{4\gamma}{3}$, there exists a set of distributions p_t, $t \in [n]$, such that:

$$\sum_{t=1}^{n} H(Q_t^{off}) - \sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \geq \Omega(R), \tag{12}$$

where $R : \max_{t\in[n]} R_t$ and

$$R_t = \left|\sum_{i=1}^{t}\sum_{w\in\mathcal{W}} Q_i^{off}(w)\log p_i(w) + \gamma t\right| \tag{13}$$

for $t \in [n]$.

Sketch of the proof. We briefly describe the main idea of the proof leaving details to Appendix B. In addition to $R_t =: R_t^{off}$ defined above, we introduce $R_t^{\mathcal{A}} = \left|\sum_{i=1}^{t}\sum_{w\in\mathcal{W}} q_i^{\mathcal{A}}(w)\log p_i(w) + \gamma t\right|$. Next we construct two distributions $p_{t,1}$ and $p_{t,2}$ such that they are equal up to time $0 \leq t \leq \alpha = 3R/\gamma$ but different otherwise (see (24)). We will prove that $R_\alpha^{off} = -R$ for $p_t = p_{t,1}$ and $R_\alpha^{off} = R$ for $p_t = p_{t,2}$, which implies $|R_\alpha^{off} - R_\alpha^{\mathcal{A}}| = \Omega(R)$ either when $p_t = p_{t,1}$ or $p_t = p_{t,2}$. Finally, after some tedious algebra we show that:

$$\sum_{t=1}^{n} H(Q_t^{off}) - \sum_{t=1}^{n} H(q_t^{\mathcal{A}}) = \Omega(|R_\alpha^{off} - R_\alpha^{\mathcal{A}}|) = \Omega(R)$$

as needed. □

The lower bound tells us that the best achievable performance for any algorithm is bounded by R away from the offline optimal, which can go up to $O(n)$. However, as we prove in Theorem 3.8, there exist classes of p_t for which $R = o(n)$. To complete the picture, we need an upper bound on the entropy loss that is $o(n)$. That is, we need an online algorithm that approximates the optimal entropy for the offline setting well. We prove in Theorem 3.5 that our Algorithm 1 is asymptotically online optimal. However, before we derive an upper bound, we make some comments about Algorithm 1. Note that the λ_t, $t \in [n]$ computed in Algorithm 1 are obtained by solving the *offline optimization*

problem (4) given the probabilities p_i up to time $i = t$. They are not necessarily the same as the optimal λ_t^{on} (see Lemma 3.3). Compared to constraint (11), where the exponents λ_t^{on}, $t \in [n]$ are different for each time step t, the constraint (14) has the same exponent λ_t for all time steps $i \in [t]$.

Next, we present the second main result, showing that Algorithm 1 is asymptotically optimal, with proof in Appendix C.

Theorem 3.5. *Let q_t^{on}, $t \in [n]$, be the output of Algorithm 1. Let λ^* satisfy (9) such that $Q_t^{off} = \frac{p_t^{\lambda^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda^*}(w)}$ is the solution to the offline optimization problem (4). If there exists a constant c such that:*

$$\sum_{i=1}^{t} \mathrm{Var}_{Q_i^{off}}\left[\log \frac{1}{p_i(w)}\right] \geq ct, \tag{15}$$

then we have:

$$\sum_{t=1}^{n} H(Q_t^{off}) - \sum_{t=1}^{n} H(q_t^{on}) \leq O(R \log n) \tag{16}$$

$$\sum_{t=1}^{n} \sum_{w\in\mathcal{W}} Q_t^{off} \log p_t(w) + \gamma n \geq O(-R \log n). \tag{17}$$

where $R = \max_{t\in[n]} R_t$ with R_t defined in (13).

Remark 3.6. While in theory, bounded violation (17) of the constraint (4) can occur, we observe no violation of (4) in our experiments (see Section 4 and Figure 3 that verifies our assumption (15)). The reason for this is that the rate of p_t's that are close to singleton distributions becomes higher as t increases. Hence, λ_t's in Algorithm 1 are higher than λ^* and thus, constraint (4) is satisfied.

Finally, we study the behavior of R. In general R can be $O(n)$, however, for some p_t we show below that R reduces to $R = o(n)$. Consider p_t, $t \in [n]$, that are independently and randomly selected from a finite set of k distributions $\{p^{(1)}, \ldots, p^{(k)}\}$ such that $\Pr(p_t = p^{(i)}) = P_i$, $i \in [k]$ and $\sum_{i=1}^{k} P_i = 1$. While p_t, $t \in [n]$ in practice are not exactly i.i.d., they present some stationarity, which through our analysis for the i.i.d. case, enables Algorithm 1 to converge to a good solution. We start with the following lemma which is proved in Appendix D.

Lemma 3.7. *Let p_t be independently selected according to $\Pr(p_t = p^{(i)}) = P_i$, $i \in [k]$ and $\hat{P}_i = \frac{|\{t:p_t=p^{(i)}\}|}{n}$ $i \in [k]$ be the empirical distribution of p_t for $t \in [n]$. Then with probability at least $1-\delta$, the total variation between $\hat{P}_i$, $i \in [k]$ and P_i, $i \in [k]$, i.e., $\frac{\sum_{i\in[k]}(|\hat{P}_i - P_i)|}{2}$, is at most $\sqrt{\frac{(k+1)\log 2 + \log\frac{1}{\delta}}{2n}}$ for any k and n.*

We now present our last main result, which shows that for the aforementioned p_t, $t \in [n]$, we have $R = o(n)$.

Theorem 3.8. *Let p_t be independently selected according to $\Pr(p_t = p^{(i)}) = P_i$. Then $R \leq O\left(\sqrt{n \log \frac{1}{\delta}}\right)$ with probability at least $1-\delta$.*

Proof. Let $t_i = |\{j : j \leq t, p_j = p^{(i)}\}|$, $i \in [k]$ be the number of occurrences of $p^{(i)}$ up to time t. Then, from Lemma 3.7, we have $\sum_{i\in[k]} |t_i - tP_i| \leq \sqrt{2t((k+1)\log 2 + \log\frac{1}{\delta})}$, leading to:

$$\begin{aligned}
&\left|\sum_{i\in[k]} t_i \sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w) + t\gamma\right| \\
\leq &\sum_{i\in[k]} tP_i \left|\sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w) + t\gamma\right| \\
&+ \sum_{i\in[k]} |t_i - tP_i| \left|\sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w)\right| \\
\leq &\sqrt{\frac{2t^2((k+1)\log 2 + \log\frac{1}{\delta})}{n}}(L_{max}\log^2\frac{1}{\rho} - \gamma) \\
&+ \sqrt{2n((k+1)\log 2 + \log\frac{1}{\delta})}(L_{max}\log^2\frac{1}{\rho} - \gamma) \\
\leq &O(\sqrt{n\log\frac{1}{\delta}}),
\end{aligned}$$

where the last derivations follow from Proposition B.1 since $p_t(w) > \rho$. The full proof is presented in Appendix D. □

4. Experimental Results

We now present experimental results verifying our theoretical results on real data and show that Algorithm 1 *consistently outperforms* state-of-the-art solutions.

4.1. Experimental Setting

We use *Meta-Llama-3-8B-Instruct* for our experiments, conducted on AMD Instinct MI210 GPUs with 64 GB. Recall that $p_t(w) = P(w \mid w^{t-1})$ denotes the probability distribution produced by the final layer of the transformer model at time step t. For practical reasons, since candidate tokens with extremely small probabilities are unlikely to be selected during sequence generation, we do not store or utilize the full probability distribution. Instead, we apply a probability cutoff of 10^{-4}, retaining only tokens whose probabilities exceed this threshold. We denote this truncated probability as $\tilde{p}_t(w)$. To implement the online update of λ_t in Algorithm 1, we restrict λ_t to a finite set of candidate values. Specifically, we consider a uniformly spaced grid $\Lambda = \{\lambda^{(1)}, \ldots, \lambda^{(m)}\} \subset [0, \lambda_{\max}]$, with $\lambda_{\max} = 6$ in all experiments. The grid size is set equal to sequence length, i.e., $m = n$. At time step t, for each candidate λ in the grid,

the online algorithm evaluates the cumulative average:

$$\frac{-1}{t}\sum_{i=1}^{t}\sum_{w\in\mathcal{W}}\frac{\tilde{p}_i(w)^\lambda}{\sum_{w'\in\mathcal{W}}\tilde{p}_i(w')^\lambda}\log\tilde{p}_i(w),$$

and selects λ_t to be the closest point to the parameter γ.

To draw meaningful conclusions regarding our theoretical results and their applicability, the generation process outputs long-context responses. Therefore, a sequence length of $n = 1500$ tokens was selected for all experiments. In addition, datasets were chosen to facilitate the investigation of test cases that benefit from diverse, high-entropy responses. In particular, we present results on the following datasets:

1. ELI5 (Explain Like I'm Five) (Fan et al., 2019) is an open-ended explanatory question-answer dataset in which multiple long-form answers can be equally valid. This is well suited for studying diversity in generated responses.

2. Alpaca (Taori et al., 2023) is an instruction-following dataset consisting of diverse prompts paired with model-generated reference outputs. Many instructions admit multiple reasonable completions that differ in phrasing, structure, and emphasis, making the dataset suitable for evaluating diversity in instruction-conditioned text generation. Results are presented in Appendix E.

3. Natural Questions (Kwiatkowski et al., 2019) is a question-answer benchmark based on real user queries, where answers are grounded in factual content but may vary in length, level of detail, and supporting context. This variability allows for multiple valid responses and provides a complementary setting for analyzing diversity. Results are presented in Appendix F.

The primary objective of our experiments is to demonstrate the advantage of the proposed Algorithm 1 in terms of output diversity, as measured by entropy, compared to the commonly used heuristic temperature setting in text generation (typically $T = 0.7$, or equivalently $\lambda = 1/0.7$). For each experiment, we record the entropy of the generated sequence under both generation schemes and compare them to the optimal offline entropy characterized by Theorem 3.2, which can be computed retrospectively after full observation of the underlying token distributions from which the sequence is generated. The budget parameter γ is set to 0.7 in all experiments. In this context, we also illustrate the convergence of the sequence $\{\lambda_t\}_{n\geq t\geq 1}$ to the optimal value λ^* over the course of text generation. In addition, we also report results on the variance of log-probabilities under the induced sampling distribution verifying our assumption (15). These results provide evidence supporting the approximately linear behavior of the variance required for Theorem 3.5.

The experiments are divided into two categories. In the first category, a sequence is generated using the heuristic temperature setting $T = 0.7$ and the corresponding model-induced distributions p_t are recorded at each time step. Subsequently, the proposed online algorithm operates on these fixed distributions to produce the distributions q_t. By fixing p_t through a single baseline generation, this category of experiments enables controlled comparisons of the entropy achieved by the heuristic offline baseline and the online algorithm, without confounding effects arising from variability in the underlying model distributions. The second category of experiments allows independent generations for the heuristic baseline and the offline algorithm, illustrating the applicability of the proposed method in realistic text generation scenarios where model distributions are not controlled.

4.2. Experimental Results on Llama

Here we only report our experimental results for ELI5 dataset. Results for the Alpaca dataset are reported in Appendix E, while our findings for Natural Questions database can be found in Appendix F.

We present the experimental results on the ELI5 dataset, obtained from 50 randomly sampled questions under the fixed-distribution setting, shown in Figures 1, 2, and 3. Figure 1 compares the per-token entropy achieved by the heuristic temperature-based generation ($T = 0.7$), the proposed online algorithm, and the offline optimal policy characterized by Theorem 3.2. These results clearly demonstrate that the online algorithm consistently achieves higher entropy than the heuristic baseline, indicating substantially greater output diversity. Figure 2 illustrates the evolution of the online parameter λ_t over time for a representative sequence. As expected, the sequence $\{\lambda_t\}_{t=1}^n$ converges to the offline optimal value λ^* as more tokens are generated, providing empirical evidence for the stability and effectiveness of the proposed online update rule. Finally, Figure 3 reports the cumulative variance of the log-probabilities under the optimal sampling distribution, $\sum_{i=1}^t \mathrm{Var}_{q_i^{\lambda^*}}[\log\tilde{p}_i(W)]$, as a function of the sequence length t, averaged over five sampled generations. The approximately linear growth observed in all cases is consistent with the theoretical predictions, further validating our assumption (15).

In the second group of experiments, the heuristic baseline and the proposed online algorithm generate sequences *independently*, resulting in different underlying model-induced distributions $\{p_t\}$ across methods. This setting reflects realistic autoregressive text generation, and the objective of our experiments is to evaluate whether the entropy gains of the online algorithm persist under this more challenging regime, while maintaining the plausibility of generated outputs.

Figure 4 reports the per-token entropy achieved by the two methods under independent generations. Although the resulting curves naturally exhibit increased variability, the online algorithm achieves higher entropy on average, indicating improved output diversity even when the underlying

model distributions are not controlled.

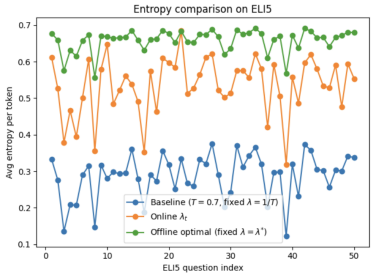

Figure 1. Comparison of per-token entropy for heuristic generation with $T = 0.7$, the proposed online algorithm, and the offline optimal policy.

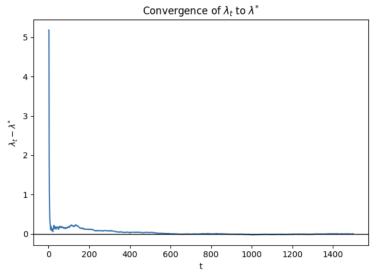

Figure 2. Convergence of the online parameter λ_t to the offline optimal value λ^* over the course of sequence generation.

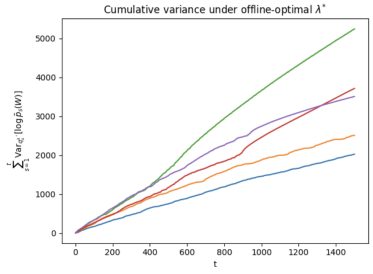

Figure 3. Scaling of the variance $\sum_{i=1}^{t} \mathrm{Var}_{q_i^{\lambda^*}}[\log \tilde{p}_i(W)]$ with respect to t for five sampled generations.

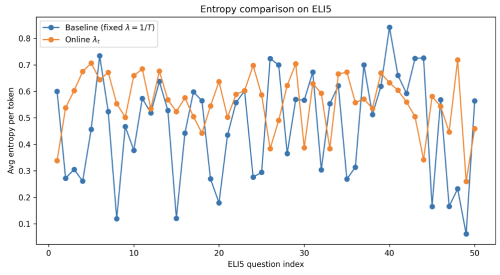

Figure 4. Per-token entropy under independent generations for the ELI5 dataset.

To assess plausibility, Figure 5 reports the score $\log_2 P(w^n) + \gamma n$. Despite increased variance across runs, the online algorithm consistently produces sequences with nonnegative plausibility scores ($\log_2 P(w^n) + \gamma n > 0$), suggesting that the observed entropy gains do not arise from implausible or low-probability outputs.

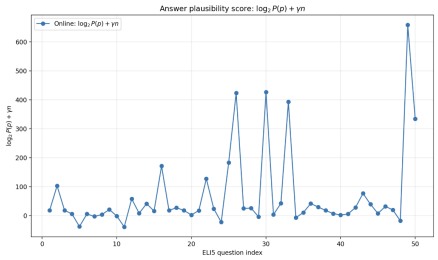

Figure 5. Plausibility score $\log_2 P(w^n) + \gamma n$ under independent generations for the ELI5 dataset.

To summarize our experimental findings across all datasets, we report aggregate entropy statistics for each generation scheme in Tables 1 and 2. Across all settings, the proposed online algorithm consistently achieves higher entropy than the heuristic baseline, while remaining close to the offline optimal benchmark.

Table 1. Summary of entropy statistics across datasets under the fixed-distribution setting. The gain column reports the difference between the online algorithm and the heuristic baseline ($T = 0.7$).

Dataset	$T = 0.7$		Online		Offline Opt	Gain
	Avg	Std	Avg	Std	Avg	Avg
ELI5	0.288	0.061	0.535	0.079	0.658	**0.247**
Alpaca	0.213	0.091	0.499	0.117	0.604	**0.286**
Natural Questions	0.167	0.099	0.447	0.141	0.587	**0.280**

Table 2. Summary of corresponding entropy statistics across datasets for independent generations.

Dataset	$T = 0.7$		Online		Offline Opt	Gain
	Avg	Std	Avg	Std	Avg	Avg
ELI5	0.470	0.196	0.562	0.107	0.667	**0.092**
Alpaca	0.493	0.201	0.606	0.183	0.656	**0.113**
Natural Questions	0.167	0.100	0.452	0.151	0.587	**0.285**

Concluding Remarks

We presented a new formulation for the problem of generating creative/ non-repetitive high-quality outputs from generative models by explicitly maximizing entropy of token probability distributions while guaranteeing threshold generation probability. Our formulation provides a theoretical justification of temperature-based sampling and related heuristics. We derived the optimal solution to the constrained optimization problem and an approximate algorithm for an online solution that asymptotically approaches the optimal solution. We present comprehensive experimental results supporting all theoretical claims.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning and Generative AI. There are many potential societal consequences of our work, including better, safer, and more efficient AI.

References

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

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

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

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

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

Graves, A., Wayne, G., and Danihelka, I. Neural networks as online learners. *arXiv preprint arXiv:1412.7753*, 2014.

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

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

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

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

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

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

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

Ovadia, Y. et al. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 2019.

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

Szpankowski, W. *Average Case Analysis of Algorithms on Sequences*. Wiley, New York, 2001.

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

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

Wu, C., Grama, A., and Szpankowski, W. Online universal learning from information-theoretic perspective, 2025. URL https://www.cs.purdue.edu/homes/spa/papers/fnt25.pdf.

A. Proof of Lemma 3.3

Proof. Following similar arguments in the proof of Theorem 3.2, we find that $q_t(w)$ of the form $\frac{p_t^{\lambda_t^{on}}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda_t^{on}}(w)}, t \in [n]$ maximizes $H(q_t)$ for fixed value of $\sum_{w\in\mathcal{W}} q_t(w)\log p_t(w) + \gamma$. Hence (10) follows. To show that (11) holds, suppose on the contrary the equality does not hold, we have that

$$\sum_{t\in[n]} - \sum_{w\in\mathcal{W}} \frac{p_t^{\lambda_t^{on}}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda_t^{on}}(w)} \log p_t(w) < \gamma n.$$

Note that the left hand side decreases with λ_t, $t \in [n]$. One can decrease λ_n^{on} so that

$$\sum_{t\in[n]} - \sum_{w\in\mathcal{W}} \frac{p_t^{\lambda_t^{on}}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda_t^{on}}(w)} \log p_t(w) \le \gamma n$$

still holds. Note that $H(Q_n^{on})$ decreases with λ_n^{on}. Hence, decreasing λ_n^{on} increases $H(Q_n^{on})$, contradicting to the optimality of Q_t^{on}, $t \in [n]$. Hence, we have (11). □

B. Proof of the Lower Bound Theorem 3.4

We first provide some facts about the constraint and the objective function in (2) and (3) that will be repeatedly used in the proofs.

Proposition B.1. *For $p_t(w) > \rho$, let*

$$f_t(\lambda) = - \sum_{w\in\mathcal{W}} \frac{p_t^{\lambda}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda}(w)} \log \frac{p_t^{\lambda}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda}(w)} \tag{18}$$

$$g_t(\lambda) = \sum_{w\in\mathcal{W}} \frac{p_t^{\lambda}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda}(w)} \log p_t(w) + \gamma. \tag{19}$$

Then we have $0 \le g_t'(\lambda) \le \log^2(\frac{1}{\rho})$ and $-L_{max}\log^2(\frac{1}{\rho}) \le f_t'(\lambda) \le 0$.

Proof. It can be verified that

$$g_t'(\lambda) = \mathrm{V}ar_{\frac{p_t^{\lambda}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda}(w)}} \left[\log \frac{1}{p_t(w)}\right] \tag{20}$$

$$f_t'(\lambda) = - \lambda g_t'(\lambda). \tag{21}$$

Hence, the proposition follows. □

Lemma B.2. *If $p_t(w) \ge \rho$ for some constant $\rho > 0$, then*

$$\lambda^* \le L_{max} \triangleq \max\{1, \frac{\sum_{t=1}^n H(p_t)}{n\gamma + \sum_{t=1}^n \log \max_{w\in\mathcal{W}} p_t^{\lambda^*}(w)}\},$$

where λ^* is given in Theorem 3.2. In practice, it is observed that $\sum_{t=1}^n H(p_t) = O(n)$ after filtering out small p_t entries using, e.g., top k or top-p sampling, and $n\gamma + \sum_{t=1}^n \log \max_{w\in\mathcal{W}} p_t^{\lambda^*}(w) = \Omega(n) > 0$ with γ properly selected. Hence, $L_{max} \le O(1)$.

Proof. Let

$$f_t(\lambda) = - \sum_{w\in\mathcal{W}} \frac{p_t^{\lambda}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda}(w)} \log \frac{p_t^{\lambda}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda}(w)}, \quad t \in [n]$$

which is the entropy of the distribution $q_t(w) = p_t^\lambda/(\sum_{w\in\mathcal{W}} p_t^\lambda)$. Then

$$\begin{aligned} f_t'(\lambda) =& -\lambda\Bigg(\sum_{w\in\mathcal{W}} \frac{p_t^{\lambda^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda^*}(w)} \log^2 p_t(w) - \Big(\sum_{w\in\mathcal{W}} \frac{p_t^{\lambda^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda^*}(w)} \log p_t(w)\Big)^2\Bigg) \\ =& -\lambda \mathrm{Var}_{\frac{p_t^{\lambda^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda^*}(w)}}\big[\log\frac{1}{p_t(w)}\big] \le 0. \end{aligned} \tag{22}$$

Thus we can write (2) as

$$\begin{aligned} \lambda^*\gamma n =& -\sum_{w\in\mathcal{W}} \frac{p_t^{\lambda^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda^*}(w)} \log p_t^{\lambda^*}(w) = \sum_{t=1}^n \Bigg(f_t(\lambda^*) - \log\Big(\sum_{w\in\mathcal{W}} p_t^{\lambda^*}(w)\Big)\Bigg) \\ \le& \sum_{t=1}^n \Big(f_t(\lambda^*) - \log\max_{w\in\mathcal{W}} p_t^{\lambda^*}(w)\Big) \\ \le& \sum_{t=1}^n H(p_t) - \lambda^* \sum_{t=1}^n \log\max_{w\in\mathcal{W}} p_t^{\lambda^*}(w) \end{aligned}$$

for $\lambda \ge 1$, where the last inequality follows from the facts that $f_t'(\lambda) \le 0$ and that $f_t(1) = H(p_t)$. Hence, we have

$$\lambda^* \le \max\{1, \frac{\sum_{t=1}^n H(p_t)}{n\gamma + \sum_{t=1}^n \log\max_{w\in\mathcal{W}} p_t^{\lambda^*}(w)}\}$$

which completes the proof of the lemma. □

Now we are ready to prove Theorem 3.4. W.L.O.G., let $\mathcal{W} = [|\mathcal{W}|]$. Let M be the smallest integer such that $-\frac{(M-1)\log(\frac{1-\exp(-\frac{\gamma}{3})}{M-1})}{M} + \frac{\gamma}{3M} > \frac{4\gamma}{3}$ and $\log M \ge \frac{4\gamma}{3}$. Let us define

$$d_x(w) = \begin{cases} x & w = M \\ \frac{1-x}{M-1} & w \in [M-1] \end{cases}$$

for $\frac{1}{M} \le x \le 1$ and let

$$C(\lambda, x) = -\sum_{w\in\mathcal{W}} \frac{d_x^\lambda(w)}{\sum_{w\in\mathcal{W}} d_x^\lambda(w)} \log d_x(w)$$

for $\lambda \ge 0$. Then

$$\begin{aligned} C(0, \exp(-\frac{\gamma}{3})) =& -\frac{(M-1)\log(\frac{1-\exp(-\frac{\gamma}{3})}{M-1})}{M} + \frac{\gamma}{3M} > \frac{4\gamma}{3}, \\ C(\infty, \exp(-\frac{\gamma}{3})) =& -\log\max_{w\in\mathcal{W}} d(w) \le \frac{\gamma}{3} < \frac{2\gamma}{3}. \end{aligned}$$

Hence, there exist non-negative λ_1 such that $C(\lambda_1, \exp(-\frac{\gamma}{3})) = \frac{4\gamma}{3}$ and non-negative λ_2 such that $C(\lambda_2, \exp(-\frac{\gamma}{3})) = \frac{2\gamma}{3}$, i.e.,

$$\begin{aligned} -\sum_{w\in\mathcal{W}} \frac{d_{\exp(-\frac{\gamma}{3})}^{\lambda_1}(w)}{\sum_{w\in\mathcal{W}} d_{\exp(-\frac{\gamma}{3})}^{\lambda_1}(w)} \log d_{\exp(-\frac{\gamma}{3})}(w) =& \frac{4\gamma}{3} \\ -\sum_{w\in\mathcal{W}} \frac{d_{\exp(-\frac{\gamma}{3})}^{\lambda_2}(w)}{\sum_{w\in\mathcal{W}} d_{\exp(-\frac{\gamma}{3})}^{\lambda_2}(w)} \log d_{\exp(-\frac{\gamma}{3})}(w) =& \frac{2\gamma}{3}. \end{aligned} \tag{23}$$

Moreover, note that $C(\lambda, \frac{1}{M}) = \log M \geq \frac{4\gamma}{3}$ and $C(\lambda, 1) = 0 \leq \frac{2\gamma}{3}$ for any $\lambda \geq 0$. Hence, there exist $\frac{1}{M} \leq x_1 \leq 1$ such that $C(\lambda_1, x_1) = \frac{2\gamma}{3}$ and $\frac{1}{M} \leq x_2 \leq 1$ such that $C(\lambda_2, x_2) = \frac{4\gamma}{3}$, i.e.,

$$
\begin{aligned}
-\sum_{w\in\mathcal{W}} \frac{d_{x_1}^{\lambda_1}(w)}{\sum_{w\in\mathcal{W}} d_{x_1}^{\lambda_1}(w)} \log d_{x_1}(w) &= \frac{2\gamma}{3} \\
-\sum_{w\in\mathcal{W}} \frac{d_{x_2}^{\lambda_2}(w)}{\sum_{w\in\mathcal{W}} d_{x_2}^{\lambda_2}(w)} \log d_{x_2}(w) &= \frac{4\gamma}{3}.
\end{aligned} \tag{24}
$$

Moreover, there exist $\frac{1}{M} \leq x_3, x_4 \leq 1$ such that $C(\lambda_1, x_3) = \gamma$ and $C(\lambda_2, x_4) = \gamma$. Consider the following two sequences of distributions

$$
\begin{aligned}
p_{t,1} &= \begin{cases} d_{\exp(-\frac{\gamma}{3})} & t \in [\frac{3R}{\gamma}] \\ d_{x_1} & t \in \{\frac{3R}{\gamma}+1, \ldots, \frac{6R}{\gamma}\} \\ d_{x_3} & t \in \{\frac{6R}{\gamma}+1, \ldots, n\} \end{cases}. \\
p_{t,2} &= \begin{cases} d_{\exp(-\frac{\gamma}{3})} & t \in [\frac{3R}{\gamma}] \\ d_{x_2} & t \in \{\frac{3R}{\gamma}+1, \ldots, \frac{6R}{\gamma}\} \\ d_{x_4} & t \in \{\frac{6R}{\gamma}+1, \ldots, n\}. \end{cases}
\end{aligned} \tag{25}
$$

According to Theorem 3.2, when $p_t = p_{t,\ell}, \ell \in \{1,2\}$, we have $Q_t^{\mathit{off}} = \frac{p_t^{\lambda_\ell^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda_\ell^*}(w)}$, where $\lambda_\ell^*, \ell \in \{1,2\}$, are unique solutions satisfying

$$
-\sum_{t\in[n]} \sum_{w\in\mathcal{W}} \frac{p_{t,\ell}^{\lambda_\ell^*}(w)}{\sum_{w\in\mathcal{W}} p_{t,\ell}^{\lambda_\ell^*}}(w) \log p_{t,\ell}(w) = \gamma n.
$$

Note that from (23), (24), and (25), we have

$$
\begin{aligned}
-\sum_{t\in[n]} \sum_{w\in\mathcal{W}} \frac{p_{t,1}^{\lambda_1}(w)}{\sum_{w\in\mathcal{W}} p_{t,1}^{\lambda_1}(w)} \log p_{t,1}(w) &= \frac{3R}{\gamma}\cdot\frac{4\gamma}{3} + \frac{3R}{\gamma}\cdot\frac{2\gamma}{3} + (n - \frac{6R}{\gamma})\gamma = n\gamma \\
-\sum_{t\in[n]} \sum_{w\in\mathcal{W}} \frac{p_{t,2}^{\lambda_2}(w)}{\sum_{w\in\mathcal{W}} p_{t,2}^{\lambda_2}(w)} \log p_{t,2}(w) &= \frac{3R}{\gamma}\cdot\frac{2\gamma}{3} + \frac{3R}{\gamma}\cdot\frac{4\gamma}{3} + (n - \frac{6R}{\gamma})\gamma = n\gamma
\end{aligned}
$$

and thus $\lambda_1^* = \lambda_1$ and $\lambda_2^* = \lambda_2$. In addition, we have

$$
|\sum_{i=1}^{t} \sum_{w\in\mathcal{W}} \frac{p_{i,\ell}^{\lambda_\ell^*}(w)}{\sum_{w\in\mathcal{W}} p_{i,\ell}^{\lambda_1^*}(w)} \log p_{i,\ell}(w)| = \begin{cases} \frac{\gamma(\frac{3R}{\gamma} - |t - \frac{3R}{\gamma}|)}{3} & t \in [\frac{6R}{\gamma}] \\ 0 & t \in \{\frac{6R}{\gamma}+1, \ldots, n\}. \end{cases}
$$

Therefore, we have

$$
\max_{t\in[n]} |\sum_{i=1}^{t} \sum_{w\in\mathcal{W}} Q_i^{\mathit{off}}(w) \log p_i(w)| = R
$$

both when $p_t = p_{t,1}, t \in [n]$, and $p_t = p_{t,2}, t \in [n]$.

In the following, we show that

$$
\sum_{t=1}^{n} H(Q_t^{\mathit{off}}) - \sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \geq \Omega(R)
$$

either when $p_t = p_{t,1}, t \in [n]$ or $p_t = p_{t,2}, t \in [n]$. Note that $p_t = d_{\exp(-\frac{\gamma}{3})}, t \in [\frac{3R}{\gamma}]$ both when $p_t = p_{t,1}, t \in [n]$ and when $p_t = p_{t,2}$. Hence, the sampling distributions $q_t^{\mathcal{A}}, t \in [\frac{3R}{\gamma}]$ when $p_t = p_{t,1}, t \in [n]$ are the same as $q_t^{\mathcal{A}}, t \in [\frac{3R}{\gamma}]$ when $p_t = p_{t,2}, t \in [n]$. Denote $Q_{t,\ell}^{\mathit{off}}$ as the optimal solution to (4) when $p_t = p_{t,\ell}, t \in [n], \ell \in \{1,2\}$. Then from (23) we have

$$
|\sum_{t=1}^{\frac{3R}{\gamma}} \sum_{w\in\mathcal{W}} Q_{t,1}^{off}(w) \log p_{t,1}(w) - \sum_{t=1}^{\frac{3R}{\gamma}} \sum_{w\in\mathcal{W}} Q_{t,2}^{off}(w) \log p_{t,2}(w)| = 2R,
$$

which implies

$$\max\Bigg\{|(\sum_{t=1}^{\frac{3R}{\gamma}}\sum_{w\in\mathcal{W}}Q_{t,1}^{off}(w)\log p_{t,1}(w)+3R)-(\sum_{t=1}^{\frac{3R}{\gamma}}\sum_{w\in\mathcal{W}}Q_t^{\mathcal{A}}(w)\log p_{t,1}(w)+3R)|,$$
$$|(\sum_{t=1}^{\frac{3R}{\gamma}}\sum_{w\in\mathcal{W}}Q_{t,2}^{off}(w)\log p_{t,2}(w)+3R)-(\sum_{t=1}^{\frac{3R}{\gamma}}\sum_{w\in\mathcal{W}}Q_t^{\mathcal{A}}(w)\log p_{t,1}(w)+3R)|\Bigg\}$$
$$\geq\frac{|(\sum_{t=1}^{\frac{3R}{\gamma}}\sum_{w\in\mathcal{W}}Q_{t,1}^{off}(w)\log p_{t,1}(w)+3R)-(\sum_{t=1}^{\frac{3R}{\gamma}}\sum_{w\in\mathcal{W}}Q_{t,2}^{off}(w)\log p_{t,2}(w)+3R)|}{2}=R.$$

W.L.O.G. assume that

$$|(\sum_{t=1}^{\frac{3R}{\gamma}}\sum_{w\in\mathcal{W}}Q_{t,1}^{off}(w)\log p_{t,1}(w)+3R)-(\sum_{t=1}^{\frac{3R}{\gamma}}\sum_{w\in\mathcal{W}}Q_t^{\mathcal{A}}(w)\log p_{t,1}(w)+3R)|$$
$$=|\sum_{t=1}^{\frac{3R}{\gamma}}\sum_{w\in\mathcal{W}}Q_{t,1}^{off}(w)\log p_{t,1}(w)-\sum_{t=1}^{\frac{3R}{\gamma}}\sum_{w\in\mathcal{W}}Q_t^{\mathcal{A}}(w)\log p_{t,1}(w)|\geq R \tag{26}$$

From (23), we have

$$\sum_{t=1}^{\frac{3R}{\gamma}}\sum_{w\in\mathcal{W}}Q_{t,1}^{off}(w)\log p_{t,1}(w)+3R=R \tag{27}$$

Define

$$B:=\sum_{t=1}^{\frac{3R}{\gamma}}\sum_{w\in\mathcal{W}}Q_t^{\mathcal{A}}(w)\log p_{t,1}(w)+2R$$

Then according to (26) and (27), we have either $B\geq R$ or $B\leq -R$. We prove that $\sum_{t=1}^n H(Q_t^{off})-\sum_{t=1}^n H(q_t^{\mathcal{A}})\geq\Omega(R)$ for the case when $p_t=p_{t,1}$, $t\in[n]$ and $B\geq R$. The proof of $\sum_{t=1}^n H(Q_t^{off})-\sum_{t=1}^n H(q_t^{\mathcal{A}})\geq\Omega(R)$ for case when $p_t=p_{t,1}$, $t\in[n]$ and $B\leq -R$ is similar.

Let $S(\{p_i\}_{i=1}^t,x)$ be the optimal value of the following optimization problem

$$\max_{q_i\in\Delta(\mathcal{W}):i\in[t]}\sum_{i=1}^t H(q_i)\qquad\text{s.t.}\qquad\sum_{i=1}^t\sum_{w\in\mathcal{W}}q_i(w)\log p_i(w)+\gamma t\geq x. \tag{28}$$

Similar to Theorem 3.2, we have

$$S(\{p_i\}_{i=1}^t,x)=\sum_{i=1}^t H(q_i^*)$$

where $q_i^*(w)=\frac{p_i^{\lambda_x^*}(w)}{\sum_{w\in\mathcal{W}}p_i^{\lambda_x^*}(w)}$, $i\in[t]$, and λ_x^* satisfy

$$\sum_{i=1}^t\sum_{w\in\mathcal{W}}\frac{p_i^{\lambda_x^*}(w)}{\sum_{w\in\mathcal{W}}p_t^{\lambda_x^*}(w)}\log p_i(w)+\gamma t=x.$$

Since $\sum_{t=1}^{\frac{3R}{\gamma}}\sum_{w\in\mathcal{W}}Q_t^{\mathcal{A}}(w)\log p_{t,1}(w)+3R=R+B$, we have $\sum_{t=\frac{3R}{\gamma}+1}^n\sum_{w\in\mathcal{W}}Q_t^{\mathcal{A}}(w)\log p_{t,1}(w)+n\gamma-3R=-R-B$ and thus

$$\sum_{t=1}^n H(q_t^{\mathcal{A}})\leq S(\{p_i\}_{i=1}^{\frac{3R}{\gamma}},R+B)+S(\{p_i\}_{i=\frac{3R}{\gamma}+1}^n,-R-B).$$

Let λ_R^*, λ_{R+B}^*, λ_{-R}^*, and λ_{-R-B}^* satisfy

$$\begin{aligned}
&\sum_{i=1}^{\frac{3R}{\gamma}} \sum_{w\in\mathcal{W}} \frac{p_i^{\lambda_R^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda_x^*}(w)} \log p_i(w) + 3R = R \\
&\sum_{i=1}^{\frac{3R}{\gamma}} \sum_{w\in\mathcal{W}} \frac{p_i^{\lambda_{R+B}^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda_{R+B}^*}(w)} \log p_i(w) + 3R = R + B \\
&\sum_{i=\frac{3R}{\gamma}+1}^{n} \sum_{w\in\mathcal{W}} \frac{p_i^{\lambda_{-R}^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda_{-R}^*}(w)} \log p_i(w) + n - 3R = -R \\
&\sum_{i=\frac{3R}{\gamma}+1}^{n} \sum_{w\in\mathcal{W}} \frac{p_i^{\lambda_{-R-B}^*}(w)}{\sum_{w\in\mathcal{W}} p_t^{\lambda_x^*}(w)} \log p_i(w) + n - 3R = -R - B.
\end{aligned} \tag{29}$$

Note that from (27) and the fact that $\sum_{t=1}^n \sum_{w\in\mathcal{W}} Q_t^{\mathit{off}}(w) \log p_t(w) + \gamma n = 0$, we have $\lambda_R^* = \lambda_{-R}^*$. Then

$$\begin{aligned}
&\sum_{t=1}^n H(Q_t^{\mathit{off}}) - \sum_{t=1}^n H(q_t^{\mathcal{A}}) \\
\geq& S(\{p_i\}_{i=1}^{\frac{3R}{\gamma}}, R) + S(\{p_i\}_{i=\frac{3R}{\gamma}+1}^n, -R) - S(\{p_i\}_{i=1}^{\frac{3R}{\gamma}}, R+B) - S(\{p_i\}_{i=\frac{3R}{\gamma}+1}^n, -R-B) \\
=& \sum_{i=1}^{\frac{3R}{\gamma}} f_t(\lambda_R^*) + \sum_{i=\frac{3R}{\gamma}+1}^{n} f_t(\lambda_{-R}^*) - \sum_{i=1}^{\frac{3R}{\gamma}} f_t(\lambda_{R+B}^*) - \sum_{i=\frac{3R}{\gamma}+1}^{n} f_t(\lambda_{-R-B}^*) \\
=& \sum_{i=1}^{\frac{3R}{\gamma}} \int_{\lambda_{R+B}^*}^{\lambda_R^*} f_t'(\lambda)d\lambda + \sum_{i=\frac{3R}{\gamma}+1}^{n} \int_{\lambda_{-R-B}^*}^{\lambda_{-R}^*} f_t'(\lambda)d\lambda \\
\overset{(a)}{=}& \sum_{i=1}^{\frac{3R}{\gamma}} \int_{\lambda_R^*}^{\lambda_{R+B}^*} \lambda g_t'(\lambda)d\lambda + \sum_{i=\frac{3R}{\gamma}+1}^{n} \int_{\lambda_{-R}^*}^{\lambda_{-R-B}^*} \lambda g_t'(\lambda)d\lambda \\
\overset{(b)}{\geq}& \sum_{i=1}^{\frac{3R}{\gamma}} \int_{\lambda_R^*}^{\lambda_{R+B}^*} \lambda g_t'(\lambda)d\lambda - \sum_{i=\frac{3R}{\gamma}+1}^{n} \int_{\lambda_{-R-B}^*}^{\lambda_{-R}^*} \lambda_{-R}^* g_t'(\lambda)d\lambda \\
\overset{(c)}{=}& \sum_{i=1}^{\frac{3R}{\gamma}} \int_{\lambda_R^*}^{\lambda_{R+B}^*} (\lambda - \lambda_R^*) g_t'(\lambda)d\lambda
\end{aligned} \tag{30}$$

where $f_t(\lambda)$ and $g_t(\lambda)$, $t \in [n]$, are defined in (18). Equality (a) follows from the fact that $f_t'(\lambda) = -\lambda g'(\lambda)$. Inequality (b) follows from the fact that $\lambda_{-R}^* \geq \lambda_{-R-B}^*$ since $B \geq R \geq 0$ and $g'(\lambda) \geq 0$. Equality (c) follows from the fact that $\lambda_R^* = \lambda_{-R}^*$ and the fact

$$\int_{\lambda_R^*}^{\lambda_{R+B}^*} g'(\lambda)d\lambda = \int_{\lambda_{-R-B}^*}^{\lambda_{-R}^*} g'(\lambda)d\lambda = B$$

according to (29).

In the following, we show that $\sum_{i=1}^{\frac{3R}{\gamma}} \int_{\lambda_R^*}^{\lambda_{R+B}^*} (\lambda - \lambda_R^*) g_t'(\lambda)d\lambda \geq \Omega(B) = \Omega(R)$. To this end, we first show that $g'(\lambda)$ is lower bounded by $\Omega(1)$ for $\lambda_R^* \leq \lambda \leq \lambda_{\frac{3R}{2}}^* \leq \lambda_{R+B}^*$, where $\lambda_{\frac{3R}{2}}^*$ satisfies

$$\sum_{i=1}^{\frac{3R}{\gamma}} \sum_{w\in\mathcal{W}} \frac{p_i^{\lambda_{\frac{3R}{2}}^*}(w)}{\sum_{w\in\mathcal{W}} p_i^{\lambda_{\frac{3R}{2}}^*}(w)} \log p_i(w) + 3R = \frac{3R}{2}.$$

Note that $g_t(\lambda)$, $t \in [n]$, is increasing in λ. Hence, we have

$$\frac{3R}{2} \leq -\sum_{i=1}^{\frac{3R}{\gamma}} \sum_{w \in \mathcal{W}} \frac{p_i^{\lambda}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda}(w)} \log p_i(w) \leq 2R$$

and thus

$$\frac{\gamma}{2} \leq -\sum_{w \in \mathcal{W}} \frac{p_i^{\lambda}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda}(w)} \log \frac{1}{p_i(w)} \leq \frac{2\gamma}{3}$$

for $\lambda_R^* \leq \lambda \leq \lambda_{\frac{3R}{2}}^* \leq \lambda_{R+B}^*$. Therefore, we have

$$\frac{\gamma}{2} \leq \mathbb{E}_{\frac{p_i^{\lambda}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda}(w)}} [\log \frac{1}{p_i(w)}] \leq \frac{2\gamma}{3}$$

for $t \in [\frac{3R}{\gamma}]$. Note that $g_i'(\lambda) = Var_{\frac{p_i^{\lambda}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda}(w)}} [\log \frac{1}{p_i(w)}]$ and that $\log \frac{1}{p_i(w)} = \frac{\gamma}{3}$ for $w = M$ and $\log \frac{1}{p_i(w)} \geq \log(M-1) \geq \frac{4\gamma}{3}$. Therefore, we have that

$$g_t'(\lambda) \geq \min\{(\frac{\gamma}{3} - \frac{2\gamma}{3})^2, (\frac{4\gamma}{4} - \frac{2\gamma}{3})^2\} \geq \frac{\gamma^2}{9}$$

Finally, note that $g_t'(\lambda) \leq \log^2 \frac{1}{p_t(1)} = \log^2 \frac{M}{1-\exp(-\frac{\gamma}{3})} \leq (\frac{4\gamma}{3} + \log \frac{1}{1-\exp(-\frac{\gamma}{3})})^2$. Hence,

$$\lambda_{\frac{3R}{2}}^* - \lambda_R^* \geq \frac{R}{2(\frac{4\gamma}{3} + \log \frac{1}{1-\exp(-\frac{\gamma}{3})})^2}. \tag{31}$$

From (30) and (31), we have

$$\sum_{t=1}^{n} H(Q_t^{off}) - \sum_{t=1}^{n} H(q_t^{\mathcal{A}}) \geq \sum_{i=1}^{\frac{3R}{\gamma}} \int_{\lambda_R^*}^{\lambda_{R+B}^*} (\lambda - \lambda_R^*) g_t'(\lambda) d\lambda \geq \frac{(\lambda_{\frac{3R}{2}}^* - \lambda_R^*)\frac{\gamma^2}{9}}{2} \geq \Omega(R).$$

The proof is complete.

C. Proof of Upper Bound Theorem 3.5

According to Algorithm 1 and the definition of R, we have

$$\sum_{t=1}^{n} \sum_{w \in \mathcal{W}} \frac{p_t^{\lambda_t}(w)}{\sum_{w \in \mathcal{W}} p_t^{\lambda_t}(w)} \log p_t(w) + \gamma n = 0$$

$$|\sum_{i=1}^{t} \sum_{w \in \mathcal{W}} \frac{p_i^{\lambda^*}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda^*}(w)} \log p_i(w) + \gamma t| \leq R.$$

Consider the Taylor expansion

$$\sum_{i=1}^{t} \sum_{w \in \mathcal{W}} \frac{p_i^{\lambda}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda}(w)} \log p_i(w) = \sum_{i=1}^{t} \sum_{w \in \mathcal{W}} \frac{p_i^{\lambda^*}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda^*}(w)} \log p_i(w) + \sum_{i=1}^{t} g_i'(\lambda^*)(\lambda - \lambda^*) + L(\lambda), \tag{32}$$

where $|L(\lambda)| \leq \frac{\max_{\lambda'} \sum_{i=1}^{t} g_i''(\lambda')(\lambda - \lambda^*)^2}{2} \leq \frac{\log^3 \frac{1}{\rho}}{2}(\lambda - \lambda^*)^2$ by the Langrange remainder bound. Note that $\sum_{i=1}^{t} \sum_{w \in \mathcal{W}} \frac{p_i^{\lambda}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda}(w)} \log p_i(w)$ is increasing in λ. Hence, for $\lambda \geq \lambda^* + \frac{2R}{ct}$ and $t \geq \frac{8R \log^3 \frac{1}{\rho}}{c^2}$, we have

$$\sum_{i=1}^{t} \sum_{w \in \mathcal{W}} \frac{p_i^{\lambda}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda}(w)} \log p_i(w) \geq \sum_{i=1}^{t} \sum_{w \in \mathcal{W}} \frac{p_i^{\lambda^* + \frac{2R}{ct}}(w)}{\sum_{w \in \mathcal{W}} p_i^{\lambda^* + \frac{2R}{ct}}(w)} \log p_i(w)$$

$$\geq -R + 2R - \frac{4R^2 \log^3 \frac{1}{\rho}}{c^2 t} > 0.$$

Similarly, for $\lambda \le \lambda^* - \frac{2R}{t}$ and $t \ge \frac{8R\log^3\frac{1}{\rho}}{c^2}$, we have

$$\begin{aligned}\sum_{i=1}^{t}\sum_{w\in\mathcal{W}}\frac{p_i^{\lambda}(w)}{\sum_{w\in\mathcal{W}}p_i^{\lambda}(w)}\log p_i(w) \le& \sum_{i=1}^{t}\sum_{w\in\mathcal{W}}\frac{p_i^{\lambda^*-\frac{2R}{ct}}(w)}{\sum_{w\in\mathcal{W}}p_i^{\lambda^*-\frac{2R}{ct}}(w)}\log p_i(w)\\ \le& R - 2R + \frac{4R^2\log^3\frac{1}{\rho}}{c^2 t} < 0.\end{aligned}$$

Therefore, we have

$$\lambda^* - \frac{2R}{ct} \le \lambda_t \le \lambda^* + \frac{2R}{ct}$$

for $t \ge \frac{8R\log^3\frac{1}{\rho}}{c^2}$. This implies

$$\begin{aligned}&\sum_{t=1}^{n} H(Q_t^{off}) - \sum_{t=1}^{n} H(q_t^{on})\\ =& \sum_{t=1}^{\frac{8R\log^3\frac{1}{\rho}}{c^2}} (f(\lambda^*) - f(\lambda_t)) + \sum_{t=\frac{8R\log^3\frac{1}{\rho}}{c^2}+1}^{n} (f(\lambda^*) - f(\lambda_t))\\ \overset{(a)}{\le}& L_{max}^2\log^2\frac{1}{\rho}\frac{8R\log^3\frac{1}{\rho}}{c^2} + \sum_{\frac{8R\log^3\frac{1}{\rho}}{c^2}+1}^{n} (f'(\lambda^*)|\lambda-\lambda_t| + \max_{\lambda} f''(\lambda)|\lambda-\lambda_t|^2)\\ \overset{(b)}{\le}& L_{max}^2\log^2\frac{1}{\rho}\frac{8R\log^3\frac{1}{\rho}}{c^2} + \sum_{\frac{8R\log^3\frac{1}{\rho}}{c^2}+1}^{n} (L_{max}\log^2\frac{1}{\rho}\frac{2R}{ct} + (L_{max}\log^3\frac{1}{\rho} + \log^2\frac{1}{\rho})\frac{2R}{ct})\\ \le& O(R\log n)\end{aligned}$$

where (a) follows from $f_t'(\lambda) = \lambda g_t'(\lambda) \le L_{max}\log^2\frac{1}{\rho}$ and the Taylor expansion. (b) follows from $f_t''(\lambda) = \lambda g_t''(\lambda) + g_t'(\lambda) \le L_{max}\log^3\frac{1}{\rho} + \log^2\frac{1}{\rho}$, where L_{max} is given in Lemma B.2. Thus, we have (16). Similarly,

$$\begin{aligned}&\sum_{t=1}^{n}\sum_{w\in\mathcal{W}} q_t^{on}\log p_t(w) + \gamma n = \sum_{t=1}^{n} g_t(\lambda_t)\\ \ge& \sum_{t=1}^{\frac{8R\log^3\frac{1}{\rho}}{c^2}} g_t(\lambda_t) + \sum_{t=\frac{8R\log^3\frac{1}{\rho}}{c^2}+1}^{n} g_t(\lambda^*) - \sum_{t=\frac{8R\log^3\frac{1}{\rho}}{c^2}+1}^{n} g_t'(\lambda^*)|\lambda_t-\lambda^*| - \sum_{t=\frac{8R\log^3\frac{1}{\rho}}{c^2}+1}^{n}\max_{\lambda} g_t''(\lambda)(\lambda_t-\lambda^*)^2\\ \ge& -\frac{8R\log^3\frac{1}{\rho}\log^2\frac{1}{\rho}}{c^2} - R - \sum_{t=1}^{n}\frac{2R\log^2\frac{1}{\rho}}{ct} - \sum_{t=1}^{n}\log^3\frac{1}{\rho}\frac{2R\log^2\frac{1}{\rho}}{ct} \ge O(-R\log n).\end{aligned}$$

Thus, we have (17) and this completes the proof.

D. Proof of Lemma 3.7 and Theorem 3.8

D.1. Proof of Lemma 3.7

Note that

$$\frac{\sum_{i\in[k]}|\hat{P}_i - P_i|}{2} = \sup_{A\subseteq[k]}\sum_{i\in A}\hat{P}_i - \sum_{i\in A}P_i.$$

Since p_t, $t \in [n]$ are i.i.d. generated, by Hoeffding's inequality we have

$$Pr(|\sum_{i\in A}\hat{P}_i - \sum_{i\in A}P_i| \geq \epsilon) \leq 2\exp(-2n\epsilon^2)$$

for any $A \subseteq [k]$ and $\epsilon > 0$. By the union bound, we have

$$Pr(\sup_{A\subseteq[k]}|\sum_{i\in A}\hat{P}_i - \sum_{i\in A}P_i| \geq \epsilon) \leq 2^{k+1}\exp(-2n\epsilon^2)$$

Let $2^{k+1}\exp(-2n\epsilon^2) = \delta$. We have Lemma 3.7.

D.2. Proof of Theorem 3.8

Proof. W.L.O.G., assume that nP_i, $i \in [k]$ is an integer. Let $\bar{\lambda}$ satisfy

$$\sum_{i\in[k]} nP_i \sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\bar{\lambda}}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\bar{\lambda}}} \log p^{(i)}(w) + n\gamma = 0.$$

Note that

$$\sum_{i\in[k]} n\hat{P}_i \sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w) + n\gamma = 0.$$

By Lemma 3.7, we have $\sum_{i\in[k]}|\hat{P}_i - P_i| \leq 2\sqrt{\frac{(k+1)\log 2 + \log\frac{1}{\delta}}{2n}}$ with probability at least $1-\delta$. Hence,

$$\begin{aligned}
&\left|\sum_{i\in[k]} nP_i \sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w) + n\gamma\right| \\
\leq &\sum_{i\in[k]} n\hat{P}_i \left|\sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w) + n\gamma\right| + \sum_{i\in[k]} n|P_i - \hat{P}_i| \left|\sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w)\right| \\
\leq &\sqrt{2n((k+1)\log 2 + \log\frac{1}{\delta})}(L_{max}\log^2\frac{1}{\rho} - \gamma),
\end{aligned}$$

where the last inequality follows from Proposition B.1 since $g_t(\lambda) \leq L_{max}\max_\lambda g_t'(\lambda) \leq L_{max}\log^2\frac{1}{\rho}$.

On the other hand, from Lemma 3.7, we have $\sum_{i\in[k]}|t_i - tP_i| \leq \sqrt{2t((k+1)\log 2 + \log\frac{1}{\delta})}$, leading to

$$\begin{aligned}
&\left|\sum_{i\in[k]} t_i \sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w) + t\gamma\right| \\
\leq &\sum_{i\in[k]} tP_i \left|\sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w) + t\gamma\right| \\
&+ \sum_{i\in[k]} |t_i - tP_i| \left|\sum_{w\in\mathcal{W}} \frac{(p^{(i)}(w))^{\lambda^*}}{\sum_{w\in\mathcal{W}}(p^{(i)}(w))^{\lambda^*}} \log p^{(i)}(w)\right| \\
\leq &\sqrt{\frac{2t^2((k+1)\log 2 + \log\frac{1}{\delta})}{n}}(L_{max}\log^2\frac{1}{\rho} - \gamma) \\
&+ \sqrt{2n((k+1)\log 2 + \log\frac{1}{\delta})}(L_{max}\log^2\frac{1}{\rho} - \gamma) \\
\leq &O(\sqrt{n\log\frac{1}{\delta}})
\end{aligned}$$

where the last lines follow from Proposition B.1 since $p_t(w) > \rho$. □

E. Experimental Results for Alpaca

Applying the same experimental setup to the Alpaca dataset yields qualitatively similar results, as shown in Figures 6– 10. In particular, in the fixed-distribution setting the online algorithm maintains higher entropy than the heuristic baseline, λ_t converges to λ^*, and the cumulative variance exhibits the predicted linear scaling, confirming the generality of our findings. Under independent generations, we observe analogous behavior: the online algorithm continues to attain higher entropy on average, while maintaining nonnegative plausibility scores as measured by $\log_2 P(W^n) + \gamma n$.

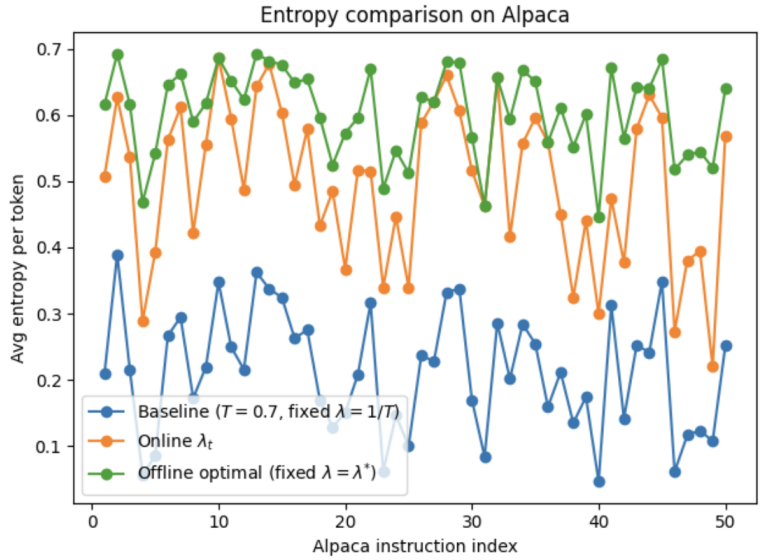

Figure 6. Comparison of per-token entropy for heuristic generation with $T = 0.7$, the proposed online algorithm, and the offline optimal policy.

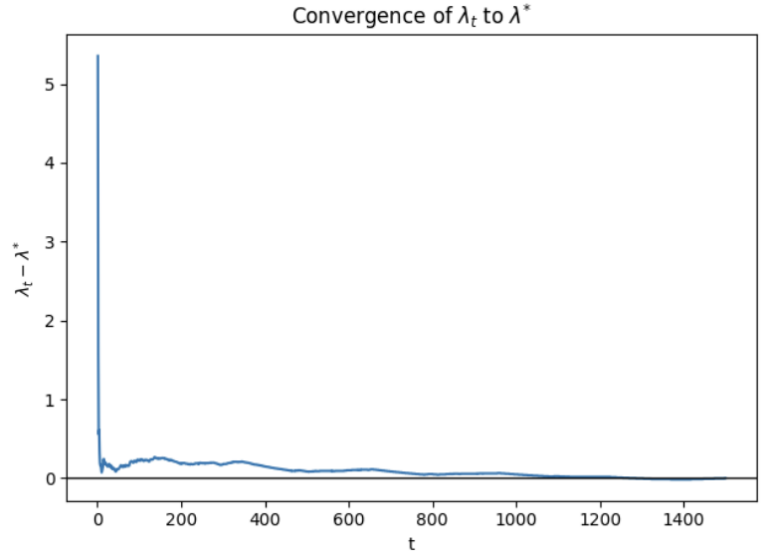

Figure 7. Convergence of the online parameter λ_t to the offline optimal value λ^* over the course of sequence generation.

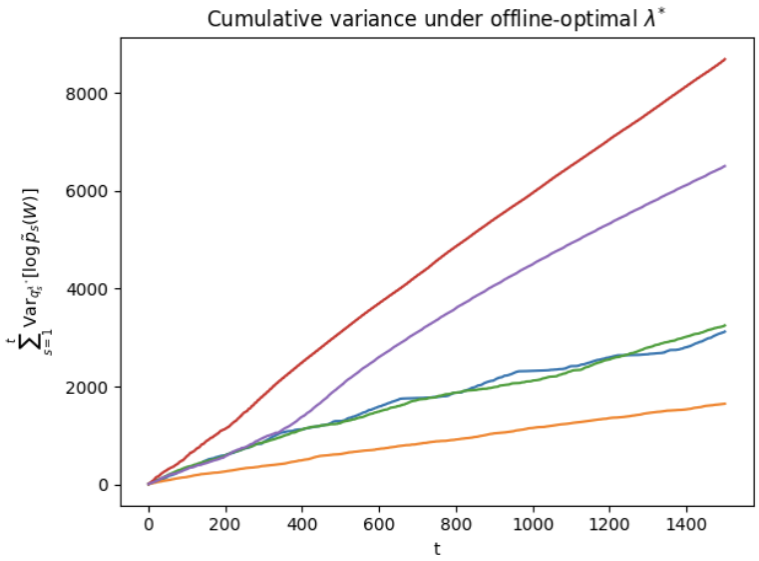

Figure 8. Scaling of the variance $\sum_{i=1}^{t} \mathrm{Var}_{q_i^{\lambda^*}}[\log \tilde{p}_{M,i}(W)]$ with respect to t for 5 sampled generations

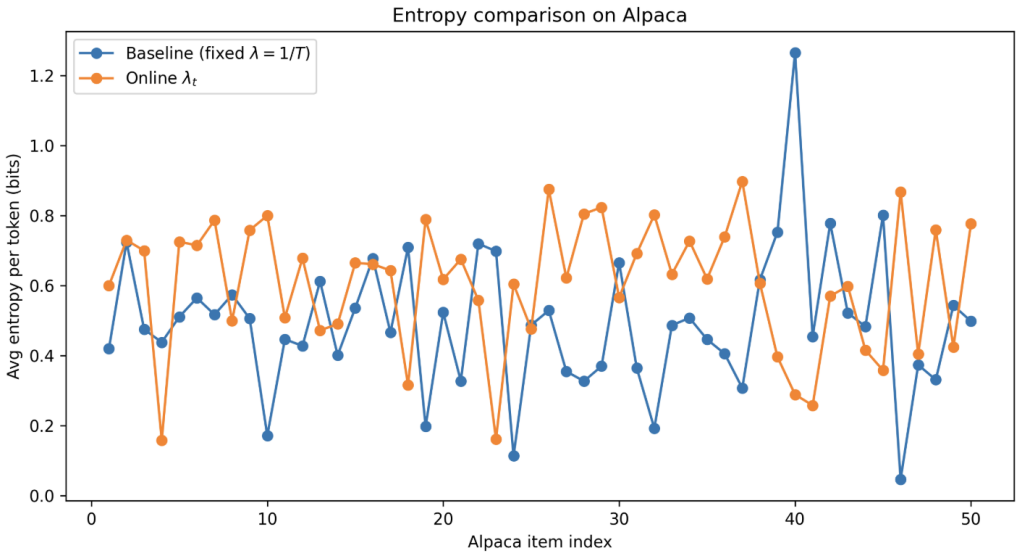

Figure 9. Per-token entropy under independent generations for the Alpaca dataset

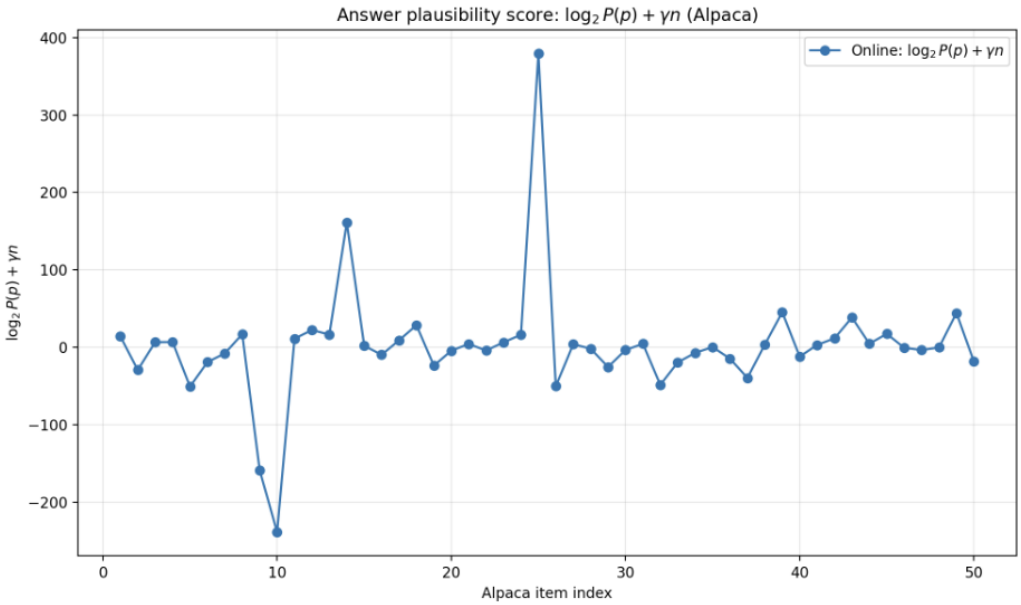

Figure 10. Plausibility score $\log_2 P(W^n) + \gamma n$ under independent generations for the Alpaca dataset

F. Natural Questions

As a final case study, we apply the same experimental setup to the Natural Questions dataset, providing an additional open-domain question-answering setting (see Figures 11- 15).

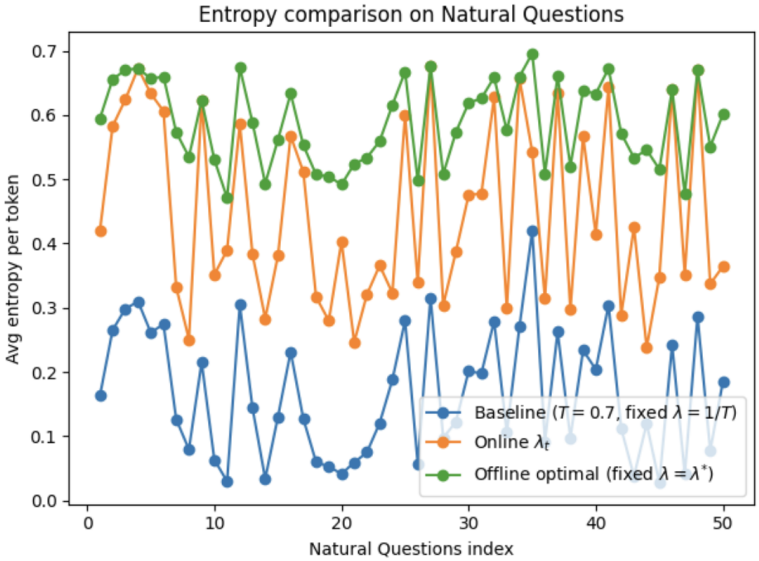

Figure 11. Comparison of per-token entropy for heuristic generation with $T = 0.7$, the proposed online algorithm, and the offline optimal policy.

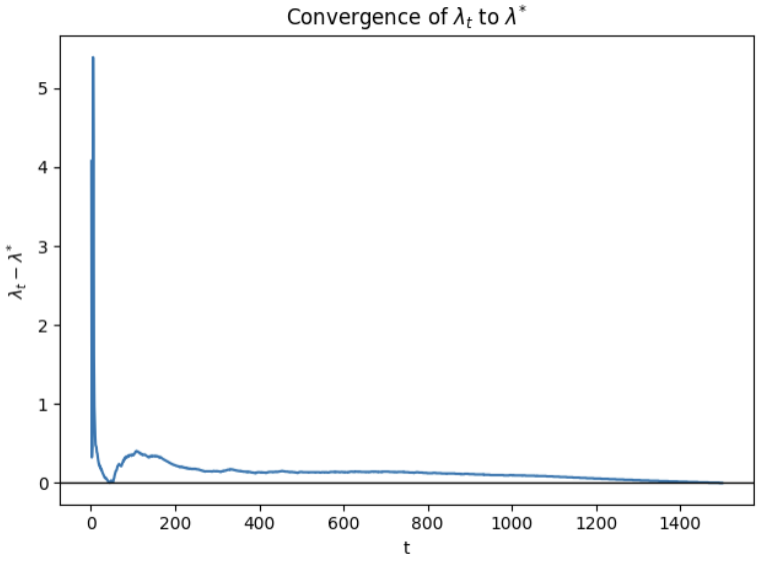

Figure 12. Convergence of the online parameter λ_t to the offline optimal value λ^* over the course of sequence generation.

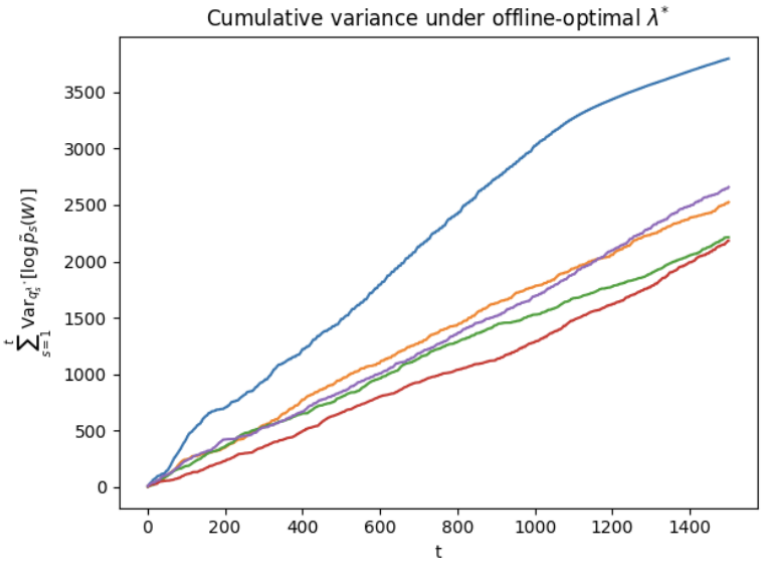

Figure 13. Scaling of the variance $\sum_{i=1}^{t} \mathrm{Var}_{q_i^{\lambda^*}}[\log \tilde{p}_{M,i}(W)]$ with respect to t for 5 sampled generations

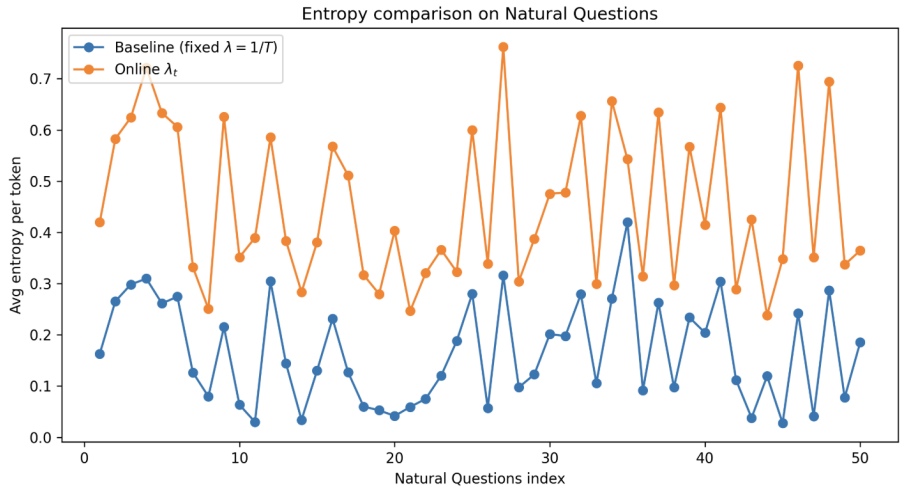

Figure 14. Per-token entropy under independent generations for the Natural Questions dataset

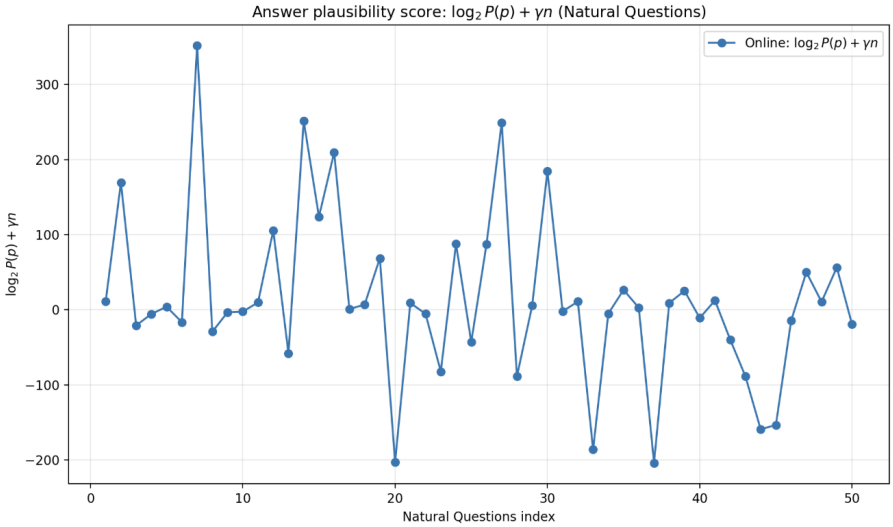

Figure 15. Plausibility score $\log_2 P(W^n) + \gamma n$ under independent generations for the Natural Questions dataset