# RECONCILING PRIVACY WITH PROVENANCE FOR LANGUAGE MODEL WATERMARKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Watermarking has great potential for regulating the use of language models (LMs), but its deployment has been marred by controversy, as many users argue that fine-grained tracking of LM usage violates their privacy. This tension reduces the effectiveness of watermarks, as users engage in evasion strategies such as paraphrasing that can remove much of the watermark signal. We show that provenance and privacy can be reconciled if we shift the goal of watermark design from tracking individuals to estimating aggregate usage. We develop locally differentially private watermarks which provably limit the evidence that any individual text can contribute to detection, while still enabling accurate estimation of population frequencies. We introduce the Red-Green Sequence (RGS) watermark, a quality-preserving and computationally efficient watermark that achieves accurate frequency estimation at high privacy levels across various models, sample sizes and generation entropies in settings that reflect realistic deployment scenarios. On C4 language modeling tasks, we attain $0.22\%$ MSE with $0.8\%$ bias at a low $\varepsilon$ of 1, while in more challenging settings like an essay writing task with very few text samples, we still find errors on the order of $1.1\%$ MSE with $0.39\%$ bias at $\varepsilon = 2$.

## 1 INTRODUCTION

Watermarking has emerged as a promising approach to enable content provenance through injecting recoverable signals into each piece of model generated text Kirchenbauer et al. (2023); Kuditipudi et al. (2024). Among provenance techniques, watermarking is unique in providing a false positive guarantee. If the detection key is kept secret, the false positive rate of falsely identifying a piece of text as being LLM generated can be precisely controlled via the generation mechanism (Zhao et al., 2023).

While provable guarantees for tracking individual texts are powerful for the targeted monitoring and regulation of AI-generated content, they have also faced significant pushback from users in recent times (Brickner-Wood, 2026). A major deployment by Anthropic (Anthropic, 2026) has led to accusations of surveillance and concentration of power, with increasing concerns that steganographic techniques may be used to spy upon users (Brand, 2026).

The tension between the desire for accurate tracking by regulators and users' desire for privacy has also led to adversarial dynamics, where users actively modify their text to evade watermark detection. Even simple strategies, like paraphrasing by another language model, can erase most of the signal from the watermarks (Krishna et al., 2023; Zhang et al., 2025). This leads to a paradoxical situation where the precision and accuracy of a watermark lead to user behavior that sabotages watermark efficacy.

Therefore, we ask if we can achieve many of the provenance goals of watermarking without violating individual privacy. We answer the question in the affirmative: we observe that many of the regulatory and monitoring goals of watermarking only require population level tracking, and in this case, it is possible to design privacy-preserving watermarks that enable accurate estimation of the prevalence while protecting individuals.

We develop the Red-Green Sequence (RGS) watermark, which is a locally differentially private watermark that provably limits the evidence that any individual text can contribute to detection, while enabling efficient estimation of the prevalence of watermarked texts in a population of texts.



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

Our watermark builds upon standard red-green style language model watermarks (Kirchenbauer et al., 2023), but carefully modifies it to be effective in the local differential privacy setting. We find that sequence-level watermarks, which are unappealing in the nonprivate setting, are highly efficient in injecting watermarks in private, low-$\varepsilon$ settings.

To detect these watermarks, we identify several interesting statistical problems in the bias-variance trade-offs of moment estimators for privacy-preserving watermarking, where the natural variability of watermark entropy across texts leads to different sampling and estimation choices compared to standard, randomized-response style estimators.

We implement this watermark on a range of scenarios, finding our frequency estimates to be highly accurate across a range of settings. In a standard language modeling task, we can attain low, $0.22\%$ MSE with $0.8\%$ bias at a low $\varepsilon$ of 1, while in more challenging settings like an essay writing task with very few text samples, we still find errors on the order of $1.1\%$ MSE with $0.39\%$ bias at $\varepsilon = 2$. These results suggest that privacy-preserving watermarking is possible and effective.

## 2 Related Work

Language model watermarks inject recoverable statistical signal into the generated text to enable detection of the presence of the watermark from the text. Prominent token-level watermarks work by reweighting token-level probabilities (Kirchenbauer et al., 2023) or coupling sampling with key randomness Aaronson (2022); Kuditipudi et al. (2024). Cryptographic techniques are used to improve robustness to token-level editing Christ & Gunn (2024). Semantic watermarks (Hou et al., 2024) use locality-sensitive hashing to improve robustness to paraphrasing. Among these, the closest to our work is Kirchenbauer et al. (2023), who use the same red-green split idea inject watermark signals, but our goal differs significantly from theirs in that we aim to precisely control the signal rather than maximizing it.

Our watermark guarantees privacy formally through local differential privacy (DP) (Dwork et al., 2006; Dwork & Roth, 2014), and the watermark method is inspired by randomized response (Warner, 1965), which enables efficient estimation of the prevalence of sensitive attributes while protecting individual respondent, achieving local DP guarantees. The efficiency of such locally private prevalence estimators in estimating of unknown population parameters have been studied by Duchi et al. (2013); Kairouz et al. (2014), who characterize optimal estimation rates and mechanisms. Frequency estimation for a fixed population under local DP is studied by Wang et al. (2017). Our setting departs from randomized response and generic local DP estimation by incorporating two important constraints: our private mechanism must enable signal to be recoverable from the output text alone, and the mechanism can only control the generation for the watermarked group but not the unwatermarked group.

## 3 Problem Setup

### 3.1 Watermarks for language models

The process of watermarking involves modifying the standard language model decoding process using a secret key $k$, so that correlations between the choice of the key and text output can be used to guarantee that the only way for a user to generate a piece of text is through the LLM decoder which uses $k$.

More formally, a watermark consists of a decoder $\Gamma$, which is an algorithm that maps a secret key $k \in \mathcal{K}$ and a distribution over completions $p_x$ to a sequence $y \in \mathcal{Y}$ stochastically, i.e. it specifies a new distribution over completions, usually different from the base model distribution $p_x$, and samples a sequence $y$ from that distribution. The decoder is designed to allow detectors to detect the presence of the watermark in a text sequence $y$ given the secret key $k$.

For example, suppose $\mathcal{Y}$ consists of only single-token sequences. Then, the Soft Red List watermark of Kirchenbauer et al. (2023) has the following decoder

$$\Pr[\Gamma(k; p_x) = y \mid k] = \frac{p_x(y)\exp(\delta \mathbf{1}\{y \in \mathcal{G}\})}{\sum_{y' \in \mathcal{Y}} p_x(y')\exp(\delta \mathbf{1}\{y' \in \mathcal{G}\})},$$

where $\mathcal{G}$ is the set of green tokens obtained by partitioning the token space into green and red tokens via a hash function with key $k$, and $\delta$ is a parameter that controls the strength of the watermark. The decoder biases the watermark towards outputting more green sequences than usual. Generalizing to longer sequences, the watermark applies the same bias towards green tokens at each token position, so a text with many green tokens is strong evidence for the watermark. Consequently, an LM user who is unwilling to disclose their LM usage might paraphrase the watermarked text before publishing, in the hope to reduce the number of green tokens and evade detection.

Importantly, watermarks are often designed to allow for high-powered detection from a single piece of text. In the case of a red-green list watermark, any moderately long piece of text will carry an exponentially increasing watermark signal, and many watermarking papers show extremely low p-values for high-entropy sequences Kirchenbauer et al. (2023); Kuditipudi et al. (2024).

The core claim of our work is that we can instead consider watermarks that more precisely control the watermarking signal, where individual texts still contribute some signal to an aggregate, population watermark frequency estimator, but no individual can be identified. To illustrate how we construct this estimator, we first illustrate a similar tension in survey design and how it can be addressed.

### 3.2 From individual detection to aggregate estimation: randomized response

Consider the scenario of reporting a sensitive attribute to a survey question, such as "have you ever used illegal drugs?". It has long been known that survey respondents can often alter their answers to avoid the possibility that their true, sensitive status may later be revealed. To address this, survey designers have developed techniques that incentivize respondents to answer truthfully while maintaining the integrity of the survey. One such technique is randomized response (RR) Warner (1965), which allows respondents to reveal sensitive information to interviewers with plausible deniability.

Consider a fixed pool of $n$ respondents whose sensitive attributes are $W_1, \ldots, W_n \in \{0, 1\}$, with $W_i = 1$ indicating that respondent $i$ has the sensitive attribute and $W_i = 0$ otherwise. Suppose that the true prevalence of the sensitive attribute in the pool is $\lambda = \frac{1}{n}\sum_{i=1}^{n} W_i$.

RR works as follows: with $W_i \in \{0, 1\}$ being the sensitive bit, the respondents engage independently with a randomization device with $\pi \in (\frac{1}{2}, 1)$ and output a noisy answer $H_i \in \{0, 1\}$

$$\Pr[H_i = 1 \mid W_i = 1] = \pi, \qquad \Pr[H_i = 1 \mid W_i = 0] = 1 - \pi.$$

RR provides plausible deniability to the respondents: if they had reported a sensitive answer $H = 1$, they can still claim that it was due to randomization since there is $1 - \pi$ probability that they are forced to report $H = 1$ even if their true status is $W = 0$.

Such plausible deniability can be formalized by the notion of local differential privacy (DP) (Dwork et al., 2006; Dwork & Roth, 2014). Consider a randomized mechanism which maps a private bit $W$ to a discrete output $Y \in \mathcal{Y}$. The mechanism is $\varepsilon$-locally differentially private if

$$\Pr[Y = y \mid w] \leq e^{\varepsilon} \Pr[Y = y \mid w'] \quad \text{for all } y \in \mathcal{Y} \text{ and } w, w' \in \{0, 1\}. \tag{1}$$

This condition bounds the likelihood ratio of observing any output under the two possible values of the sensitive bit, thus limiting the evidence that the output can reveal about the sensitive bit. The randomized response mechanism is locally $\ln \frac{\pi}{1-\pi}$-DP (Dwork & Roth, 2014).

While privacy is preserved as each person's output is only a noisy version of their true bit, aggregating the noisy bits still allows us to recover information about the mean of the true bits in the pool of $n$ respondents, $\lambda = \frac{1}{n}\sum_{i=1}^{n} W_i$. This is because randomization is precisely controlled by the probability $\pi$. Writing $\bar{H} = \frac{1}{n}\sum_{i=1}^{n} H_i$, we observe that

$$\mathbb{E}[\bar{H} \mid W] = \lambda\pi + (1 - \lambda)(1 - \pi)$$

Correspondingly, the moment estimator $\hat{\lambda}_{RR} = \frac{\bar{H} - (1-\pi)}{2\pi - 1}$ is unbiased and consistent for $\lambda$ (Warner, 1965).

### 3.3 Locally differentially private watermarks

We have illustrated how randomized response can preserve individual privacy and produce efficient estimation of aggregate prevalence in the survey setting. Inspired by this, we design a watermark that

is locally differentially private, and an efficient estimator for the prevalence of watermarked texts in a mixture of texts. First, we formalize the notion of local differential privacy for watermarks.

**Definition 1.** *For some $\varepsilon > 0$, a watermark decoder $\Gamma$ is $\varepsilon$-locally differentially private with respect to the base model distribution $p$, if for any key $k$ and any $y \in \mathcal{Y}$:*

$$e^{-\varepsilon} \leq \frac{\Pr[\Gamma(k; p) = y \mid k]}{p(y)} \leq e^{\varepsilon}, \tag{2}$$

We discuss the implication of the above definition in Appendix A.1.

# 4 The Red-Green-Sequence (RGS) Watermark

We now present a simple construction for a watermark that satisfies $\varepsilon$-local DP, called RGS. RGS closely follows the classic token-level red-green list construction of Kirchenbauer et al. (2023), but involves a few subtle changes that arise from our desire to fulfill local differential privacy.

## 4.1 Algorithm

The biggest difference between RGS and Kirchenbauer et al. (2023) is that RGS operates at the sequence level, rather than token level. For Kirchenbauer et al. (2023), the sequence-level watermarking is computationally expensive and unnatural. Identifying the set of potential green and red sequences requires a large Monte Carlo sample and provides little benefit, as injecting small distortions at each token generation step already provides exponentially strong detection signal.

In contrast, for differentially private watermarking, we have the opposite problem. Token-level watermarks lead to hard-to-control privacy leakage, as the information that we leak is a sequence of red-green marks, which is a high-dimensional bit string that is difficult to keep private. This challenge is analogous to the observation in locally differentially private frequency estimation that the optimal estimator is equivalent to a one-bit randomized response-style estimator (Duchi et al., 2013; Kairouz et al., 2014), as that is the minimal information that needs to be leaked to enable frequency estimation. Because of this, sequence-level watermarking is far more natural in our setting, and we encode exactly one red-green bit for each sequence. We begin by presenting a naive implementation of this approach which samples $M$ candidate samples, sorts them into green and red bins, and reweights their samples appropriately (Algorithm 1).

**Algorithm 1** RGS generation

**Require:** Prompt $x$, candidate count $M$, privacy budget $\varepsilon$, keyed hash $h_k$ with parameter $\pi_0$

1: Draw $\tilde{Y}_1, \ldots, \tilde{Y}_M \overset{\text{iid}}{\sim} LM(\cdot \mid x)$.
2: Deduplicate to $\{(Y_i, c_i)\}_{i=1}^{m}$; set $p_i \leftarrow c_i / M$.
3: $\mathcal{G} \leftarrow \{i : h_k(Y_i) = G\}$; $P_G \leftarrow \sum_{i \in \mathcal{G}} p_i$.
4: $Q_G \leftarrow \min\{e^{\varepsilon} P_G,\ 1 - e^{-\varepsilon}(1 - P_G)\}$.
5: For each $i \in [m]$, set

$$q_i \leftarrow \begin{cases} p_i Q_G / P_G, & i \in \mathcal{G}, \\ p_i (1 - Q_G)/(1 - P_G), & i \notin \mathcal{G}. \end{cases}$$

6: Draw $I \sim \text{Categorical}(q_1, \ldots, q_m)$.
7: **return** $Y_I$.

Comparing Algorithm 1 to the earlier token-level red-green watermark in Eq 3.1, we see that one major difference is in the definition of the reweighting term $Q_G/P_G$, which is not a fixed offset determined by $\delta$, but a term that is set to precisely attain our target privacy budget on either the green or red list probability ratio. With this choice of reweighting, we attain our local DP guarantee.

**Proposition 1** (RGS privacy)**.** *RGS is $\varepsilon$-DP with respect to the base model.*

*Proof.* The proof follows directly from our choice of reweighting weights, and we defer the proof to section A.5.1. □

One major drawback of sequence-level watermarks is that the requirement of sampling $M$ additional samples is highly wasteful (compared to token-level ones, which are essentially free). However, another difference between standard watermarking and our setting is that the watermarking strength per sequence is *very* low, and thus the difference in distribution before and after watermarking is very similar. This enables highly efficient rejection sampling algorithms for this setting, which down-weights all red completions by a multiplicative factor of $e^{-\varepsilon}$. In Appendix A.2, we give pseudocode for this rejection sampler, and show that it is $\varepsilon$-DP and requires few samples till acceptance.

We might view RGS as a type of randomized response (RR) for watermarking. While broadly this is correct, there are two major conceptual differences between RR and our setting that lead to complexities. First, while RR can randomize for both the group with the sensitive attribute and the group without, RGS can only randomize for the watermarked group, since the unwatermarked group is not generated by the watermark decoder.

Additionally, RGS uses a keyed PRF, where a fixed key always assigns the same color to the same sequence, so an analyst can recover the report from the observed text. The construction of the keyed hash function with parameter $\pi_0$ and associated assumptions are discussed in Appendix A.3. With the key fixed, the noise in the report now comes from the randomness in sampling the underlying LM. This comes with an important consequence: the extent to which the randomization can be controlled depends on the inherent properties of the underlying base model distribution. The higher the entropy, the better control we have over the randomization, and the more accurate the prevalence estimate can be. We discuss the implications on estimation and design choices in detail in Section 5.

# 5 Estimation

Having established privacy, we turn to the statistical objective of estimating watermark prevalence. Let $Y_1, \ldots, Y_n$ be a mixture of watermarked and unwatermarked texts, with $W_i \in \{0, 1\}$ indicating whether $Y_i$ is watermarked. We define $\lambda = \frac{1}{n}\sum_{i=1}^{n} W_i$ as the true prevalence of watermarked texts in the mixture. Finally, we assume conditional independence of the text given the watermark bit, $Y_i \mid W_i$ for all $i \in [n]$.

Denote the index sets of watermarked and unwatermarked texts as $I_1 = \{i : W_i = 1\}$, $I_0 = \{i : W_i = 0\}$. For each $i \in [n]$, define $p_i$ as the distribution over text sequences without watermarking: for $i \in I_0$, $p_i$ is the distribution of unwatermarked text, which need not be the model distribution. For $i \in I_1$, $p_i$ is the *underlying base model distribution* without the watermark decoder $\Gamma$. For example, if the text is generated by prompting the watermarked LM with a prompt $x$, then $p_i$ is the model distribution $p(\cdot|x)$.

## 5.1 Moment Estimator

The similarity of RGS to randomized response naturally motivates the use of a moment estimator for prevalence estimation. To construct this estimator, we first derive the expected fraction of green texts with and without the watermark.

For each $i \in [n]$ and any fixed key $k$, denote the probability of sampling a green sequence under the *unwatermarked* distribution $p_i$ as $g_i(k) = \sum_{y \in \mathcal{Y}} p_i(y)\mathbf{1}\{h_k(y) = G\}$. For watermarked texts $i \in I_1$, the probability of sampling a green sequence is upweighted, which we define as $q_i(k) = \Pr[h_k(\Gamma(k; p_i)) = G \mid k]$.

Let $H_i = \mathbf{1}\{h_k(Y_i) = G\}$ and $\bar{H} = \frac{1}{n}\sum_{i=1}^{n} H_i$. Under the conditional independence assumption, the indicators $H_i$ conditional on $W$ and $k$ are independent Bernoulli random variables with success probabilities $g_i(k)$ for $i \in I_0$ and $q_i(k)$ for $i \in I_1$ respectively. Therefore,

$$\mathbb{E}[\bar{H} \mid k, W] = \lambda \frac{1}{I_1} \sum_{i \in I_1} q_i(k) + (1 - \lambda)\frac{1}{I_0} \sum_{i \in I_0} g_i(k)$$

Note that these probabilities are generally not knowable without assumptions about the underlying distributions. RGS works around this by taking a direct a plug-in approach, replacing the unknown $g_i(k)$'s with the known expectation $\mathbb{E}_k[g_i(k)]$, which is equal to $\pi_0$ for all $i$, by the PRF construction.

Under Assumption 1, the following lemma establishes the distribution of $g_i$ under an ideal PRF. The proof is given in A.5.3.

**Lemma 1.** *Let* $\|p_i\|^2 = \sum_{y\in\mathcal{Y}} p_i(y)^2$ *and* $\langle p_i, p_j\rangle = \sum_{y\in\mathcal{Y}} p_i(y)p_j(y)$. *Under Assumption 1*

$$\mathbb{E}_k[g_i(k)] = \pi_0, \qquad \mathrm{Var}_k[g_i(k)] = \pi_0(1-\pi_0)\sum_{y\in\mathcal{Y}}\|p_i(y)\|^2$$
$$\mathrm{Cov}_k[g_i(k), g_j(k)] = \pi_0(1-\pi_0)\langle p_i, p_j\rangle$$

Lemma 1 links the second moments of $g_i$ with the diversity of the distributions $p_i$. We can see that $\|p_i\|^2$ is the probability that two independent draws from $p_i$ yield the same sequence, which is a measure of entropy of $p_i$, while $\langle p_i, p_j\rangle$ is the probability that independent draws from $p_i$ and $p_j$ produce the same sequence, which measures the "diversity" of the underlying distributions.

The variance identity in the lemma suggests that if the underlying distributions $p_i$ have high entropy, say $\|p_i\|_2^2 \to 0$, then the unwatermarked green probability for every piece of text and key converges to the null frequency $\pi_0$. Therefore, with $M \to \infty$ in both the parallel and the rejection samplers, the corresponding watermarked green probability is $\pi_1 = \min\{e^\varepsilon\pi_0, 1 - e^{-\varepsilon}(1-\pi_0)$ for the parallel sampler, and $\frac{e^\varepsilon\pi_0}{e^\varepsilon\pi_0+1-\pi_0}$ for the rejection sampler. In this case, we find that $\tilde{\lambda} = \frac{\bar{H}-\pi_0}{\pi_1-\pi_0}$ is an unbiased estimator of $\lambda$ conditional on $k$ and $W$, exactly mirroring the moment estimator for randomized response. We use the clipped version of this estimator: $\hat{\lambda}_{\mathrm{RGS}} = \mathrm{proj}_{[0,1]}(\tilde{\lambda})$.

On the other hand, when the underlying distributions do not have perfect entropy, our $g_i(k)$ have nonzero variance over the choice of $k$, and there is no guarantee that the average unwatermarked green frequency will converge to $\pi_0$, resulting in a biased estimator. We now characterize this estimator error that arises from finite watermark entropy and decompose it into bias and variance terms.

**Proposition 2** (Conditional MSE)**.** *Fix any key* $k$. *Fix* $W = (W_1, \ldots, W_n)$, *and suppose that* $Y_i \mid W_i$ *are conditionally independent. Let* $I_1 = \{i : W_i = 1\}$, $I_0 = \{i : W_i = 0\}$. *Define*

$$V(k) := \frac{1}{n}\sum_{i=1}^{n}(H_i - q_i), \qquad B(k) := \frac{1}{|I_1|}\sum_{i\in I_1}(q_i - \pi_1), \qquad U(k) := \frac{1}{|I_0|}\sum_{i\in I_0}(g_i - \pi_0)$$

*Then,*

$$\begin{aligned}
&\mathbb{E}[(\hat{\lambda}-\lambda)^2 \mid k, W] \\
=&\frac{1}{(\pi_1-\pi_0)^2}\left[\mathbb{E}[V(k)^2 \mid k] + (\lambda B(k) + (1-\lambda)U(k))^2\right] \\
\le&\frac{1}{(\pi_1-\pi_0)^2}\left[\frac{1}{4n} + (\lambda B(k) + (1-\lambda)U(k))^2\right].
\end{aligned} \tag{3}$$

We defer the proof to A.5.4.

While the variance term $V(k)$ vanishes as $O(1/n)$, the bias terms $B(k)$ and $U(k)$ are not guaranteed to vanish, since they depend on the fixed key $k$ and the underlying distributions of the watermarked and unwatermarked texts.

### 5.2 Null bias and distributional diversity

The first term, $U(k)$, measures the green sequence fraction *marginalized* across all the examples, which is a loose condition that can occur even if every single text $i$ was generated from a low-entropy language model.

To see this, observe that Lemma 1 shows,

$$\begin{aligned}
\mathbb{E}_k[U(k)] &= \frac{1}{|I_0|}\sum_{i\in I_0}(\mathbb{E}_k[g_i(k)] - \pi_0) = 0 \\
\mathrm{Var}_k[U(k)] &= \frac{\pi_0(1-\pi_0)}{|I_0|^2}\left(\sum_{i\in I_0}\|p_i\|^2 + \sum_{i,j\in I_0, i\neq j}\langle p_i, p_j\rangle\right)
\end{aligned} \tag{4}$$

In this case, we can see that the per-example entropies can be low (and thus $\|p_i\|^2$ can be high), but the variance is dominated by the cross terms $\langle p_i, p_j\rangle$ for $i \neq j$, which can be small as long as each low-entropy distribution $p_i$ is mutually distinct from another.

### 5.3 BIAS TERMS FOR WATERMARKED TEXTS

Unlike $U(k)$ which can go away under a set of $p_i$'s that are each low-entropy but collectively diverse, the bias term for the watermarked texts $B(k)$ could be persistent even with diverse population distributions. Consider the previous example where the $p_i$'s are all concentrated on a single sequence that differ across $i$'s. For a fixed key, for each $i$ we have $g_i(k) = 1$ or $g_i(k) = 0$ under both samplers, and the corresponding $q_i(k)$ is either 1 or 0, whose average need not equal $\pi_1$.

More generally, both samplers biases the model towards green sequences by taking the original base model green probability $g_i(k)$ and upweighting it via a concave function. Jensen's inequality then implies a non-positive bias in expectation over $k$. The formal statement and proof for this bias and proof can be found in Fact 2 in the appendix. More precisely, we can decompose $B(k)$ into the linear error induced by prompt homogeneity of the watermarked pool, and the concavity error induced by low-entropy, which we show can be persistent even with enough diversity between prompts. We use the rejection sampler for the following analysis.

**Lemma 2.** *Denote $f(x) = \frac{e^\varepsilon x}{e^\varepsilon x + 1 - x}$ as the upweighting function used by the rejection sampler. The negative biases persist in expectation:*

$$\mathbb{E}_k[B(k)] \le -(e^\varepsilon - 1) f'(\pi_0) e^{-\varepsilon} \pi_0 (1 - \pi_0) \frac{1}{m} \sum_{i=1}^{m} \|p_i\|_2^2.$$

The proof is deferred to A.5.5.

Despite persistent bias, there exist design choices that can help us mitigate this bias through a bias-variance tradeoff. We show that adjusting the hash function $h_k$ to have a larger null green fraction $\pi_0$ can reduce the bias at the expense of variance, enabling asymptotic consistency of our estimates.

Let $d = \pi_1 - \pi_0$. Taking expectation over the key $k$ in Equation 3, we can decompose the error to

$$\begin{aligned} &\mathbb{E}_k[(\tilde{\lambda} - \lambda)^2 \mid W] \\ =&\frac{1}{d^2}\mathbb{E}_k[V(k)^2] + \frac{\lambda^2}{d^2}\mathbb{E}_k[B(k)^2] + \frac{(1-\lambda)^2}{d^2}\mathbb{E}_k[U(k)^2] + \frac{2\lambda(1-\lambda)}{d^2}\mathbb{E}_k[B(k)U(k)] \end{aligned} \tag{5}$$

The following simplifies the last 3 terms into a more manageable form as long as we have sufficiently many pieces of text (i.e. $|I_0| \to \infty$ and $|I_1| \to \infty$) with sufficient entropy across individuals $i$. The proof is deferred to A.5.8.

**Proposition 3.** *Assume*

$$\|\bar{p}_0\|_2^2 := \frac{1}{|I_0|^2} \sum_{i,j \in I_0} \langle p_i, p_j \rangle \to 0, \qquad \|\bar{p}_1\|_2^2 := \frac{1}{|I_1|^2} \sum_{i,j \in I_1} \langle p_i, p_j \rangle \to 0.$$

*Then*

$$\mathbb{E}_k[(\tilde{\lambda} - \lambda)^2 \mid W] \le \frac{1}{d^2}\mathbb{E}_k[V(k)^2] + \frac{\lambda^2}{d^2}(\mathbb{E}_k[B(k)])^2 + R,$$

*where the remainder $R \to 0$ as $\|\bar{p}_0\|_2^2$ and $\|\bar{p}_1\|_2^2 \to 0$.*

The first $V(k)$ term vanishes at rate $O(1/n)$ and thus we can take $d \to 0$ at any rate $o(1/n)$. We can further show that the second term decreases as we decrease $d$, allowing us to trade off bias for variance.

**Lemma 3.** *Let $d = \pi_1 - \pi_0$. Then, $\frac{|E_k[B(k)]|}{d}$ is non-increasing in $\pi_0$.*

The proof is deferred to A.5.6.

In practice, we find that for large sample settings, the choice of optimal hash $h_k$ is to target a null fraction $\pi_0$ that is often significantly higher than what would be optimal for the corresponding randomized response setting.

# 6 EXPERIMENTS

We now evaluate the empirical effectiveness of the RGS watermark at estimating the prevalence of watermarked texts in a mixture of watermarked and unwatermarked texts. To assess our estimator in both high-entropy and low-entropy settings, we consider a general language modeling task as well as a much harder coding and essay writing task inspired by real-world instruction-following tasks.

Throughout the experiment, we consider modern models and instruction-tuned models, consisting of Qwen3-1.7B, Qwen3-4B, Qwen3-8B, and Llama-3.1-8B-Instruct. For computational reasons, we use the RGS rejection sampler in Algorithm 2 to generate watermarked completions, with a maximum of $M = 16$ candidates, and the keyed hash is implemented as a 128-bit BLAKE2b hash (Aumasson, 2015) of the generated token sequence. We use $\pi_0 = 0.5$ throughout this section.

## 6.1 INTERNET TEXT (C4)

We use prompts derived from the Colossal Clean Crawled Corpus (C4), a collection of cleaned web text Raffel et al. (2019). We use the first 500 prompts from the processed C4 dataset in Pan et al. (2024) as our prompt set. For each prompt, we sample base model completions with temperature=1 and top-p=1, and apply RGS rejection sampling with $M = 16$ to construct watermarked responses from a disjoint set of candidates. We use the first 50 tokens per completion for watermarking generation and detection. We combine completions across the 500 prompts into disjoint watermarked

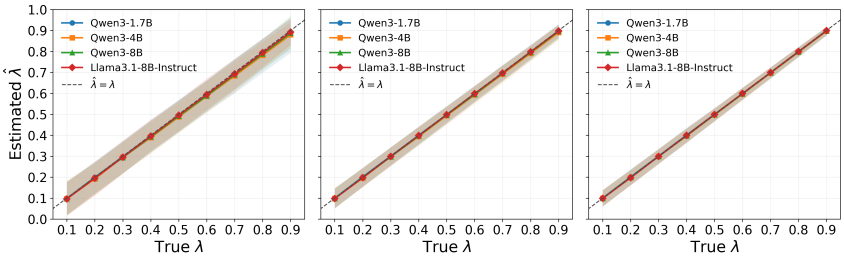


Figure 1: RGS estimates $\hat{\lambda}$ against true $\lambda$ for various models on C4, with $\varepsilon = 1, 2, 3$ from left to right. Markers show the mean estimate and the shaded bands show the 5 to 95th percentiles.

and unwatermarked pools. For each $\lambda \in \{0.1, 0.2, \ldots, 0.9\}$, we construct 50 mixtures of total size $n = 2000$ with $n\lambda$ watermarked and $n(1 - \lambda)$ unwatermarked texts sampled uniformly without replacement from each pool. For each mixture, we test RGS generation and estimation with 3 different privacy budgets ($\varepsilon \in \{1, 2, 3\}$), and repeat the same test using a set of 20 randomly sampled keys from the space of 128-bit keys.

**Accurate frequency estimates with private watermarks.** We find that differentially private watermarks provide highly accurate estimates of frequency across a range of $\lambda$, $\varepsilon$, and $n$ values. Figure 1 shows our estimates of the frequency $\lambda$ as we vary the true frequency $\lambda$ across a wide range. Our estimates show low bias across the entire range, with average absolute biases of $0.0085$, $0.0034$, and $0.0015$ for $\varepsilon = 1, 2, 3$ respectively. The 5-95% percentile bands are generally tight, with average half-widths of $0.0752$, $0.0405$, and $0.0304$ for $\varepsilon = 1, 2, 3$.

These findings hold across a range of $\varepsilon$ values, and our estimator remains accurate even at a fairly stringent $\varepsilon = 1$ value. The main challenges of the estimator are maintaining low relative error at small true $\lambda$, but such an estimator may be inherently difficult to construct given the sampling noise inherent to randomized response style frequency estimates.

## 6.2 INSTRUCTION FOLLOWING

We then test RGS on a much more challenging setting, such as a small number of students sharing the same homework prompt, where the underlying distributions $p_i$'s are low-entropy, and the prompts are highly homogeneous. Throughout this section, we use $\varepsilon = 2$. Our setting consists of two datasets: essay and coding.

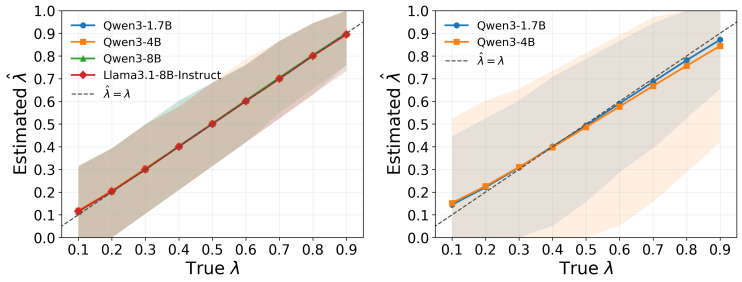


Figure 2: RGS estimates $\hat{\lambda}$ against true $\lambda$ for various models at $\varepsilon = 2$, with the essay dataset on the left and the coding dataset on the right. Markers show mean estimates and shaded bands show the 5 to 95th percentiles of the estimates.

For the essay setting, we randomly select 20 prompts from the OUTFOX dataset consisting of argumentative essay prompts Koike et al. (2024). Each mixture we test consists of $100\lambda$ texts sampled from RGS decoder, and $100(1-\lambda)$ from the same base model, all independently sampled under the same prompt. For the coding setting, we randomly select 20 problems from the LiveCodeBench dataset Jain et al. (2024) and construct the same one-prompt mixture as the essay setting. We enable the thinking mode for Qwen3-1.7B and Qwen3-4B and allow them to emit at most 256 thinking tokens before the final answer, which reduces entropy and more accurately reflects realistic LLM use.

**Accurate frequency estimates for instruction following settings.** We find that even in this extreme stress test where there is only a *single* prompt, and only 100 total samples, we are able to obtain accurate frequency estimates at $\varepsilon = 2$ (Figure 2). The essay setting shows low bias and low overall error across the range of $\lambda$, likely due to the higher per-prompt entropy. In contrast, the coding setting shows higher bias and variance, with visually distinguishable deviation from the unbiased $x = y$ line at low $\lambda$ values. Despite this, the estimates have relatively low bias, and it is promising that even extreme cases where entropy is near zero, we are able to obtain such estimates.

### 6.3 Empirical effectiveness of increasing $\pi_0$

We found that empirically, the bias-variance tradeoff strategy proposed in Section 5.3 by increasing $\pi_0$ is highly effective. The results and detailed discussion is deferred to A.4.

## 7 Discussion

In this work, we focused on the core challenges behind locally differentially private watermarking, and show that careful construction of a sequence level, randomized-response style watermark can balance privacy and efficacy. However, there remain important future work for this line of work, such as imbuing mild levels of robustness, since currently the sequence-level hashing used by RGS is sensitive to minor edits. Possible mitigations include using locality sensitive hashing Hou et al. (2024) or incorporating error correcting codes Christ & Gunn (2024). Another direction for future work is preventing privacy leakage across time via repeated inference, as currently our watermark does not guarantee privacy over unlimited queries by a single user.

### AI use statement

We used AI tools for generating synthetic datasets. We have not used AI to help develop theoretical models or conceptual frameworks. We have used AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES


Scott Aaronson. My ai safety lecture for ut effective altruism, Dec 2022. URL https://scottaaronson.blog/?p=6823.

Anthropic. How claude's text watermark works. https://www.anthropic.com/news/claude-text-watermark, 2026. Accessed: 2026-09-26.

J-P. Aumasson. Rfc 7693: The blake2 cryptographic hash and message authentication code (mac), 2015.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. *Concentration Inequalities*, pp. 208–240. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-28650-9. doi: 10.1007/978-3-540-28650-9_9. URL https://doi.org/10.1007/978-3-540-28650-9_9.

Brand. Spymarks, not watermarks. https://brand.io/article/spymarks/, 2026. Accessed: 2026-09-26.

Brady Brickner-Wood. The mark of the machine, Aug 2026. URL https://www.newyorker.com/culture/infinite-scroll/the-mark-of-the-machine.

Miranda Christ and Sam Gunn. Pseudorandom Error-Correcting Codes, June 2024. URL http://arxiv.org/abs/2402.09370. arXiv:2402.09370 [cs.CR].

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438, 2013. doi: 10.1109/FOCS.2013.53.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin (eds.), *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pp. 265–284. Springer, 2006. doi: 10.1007/11681878_14. URL https://doi.org/10.1007/11681878_14.

Oded Goldreich, Shafi Goldwasser, and Silvio Micali. How to construct random functions. *J. ACM*, 33(4):792–807, August 1986. ISSN 0004-5411. doi: 10.1145/6490.6503. URL https://doi.org/10.1145/6490.6503.

Abe Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. SemStamp: A semantic watermark with paraphrastic robustness for text generation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4067–4082, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.226. URL https://aclanthology.org/2024.naacl-long.226/.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code, 2024. URL https://arxiv.org/abs/2403.07974.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *CoRR*, abs/1407.1338, 2014. URL http://arxiv.org/abs/1407.1338.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning*

*Research*, pp. 17061–17084. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples, 2024. URL https://arxiv.org/abs/2307.11729.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 27469–27500. Curran Associates, Inc., 2023. doi: 10.52202/075280-1195. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/575c450013d0e99e4b0ecf82bd1afaa4-Paper-Conference.pdf.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models, 2024. URL https://arxiv.org/abs/2307.15593.

Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. MarkLLM: An open-source toolkit for LLM watermarking. In Delia Irazu Hernandez Farias, Tom Hope, and Manling Li (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 61–71, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-demo.7. URL https://aclanthology.org/2024.emnlp-demo.7/.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019. URL http://arxiv.org/abs/1910.10683.

Sebastien Roch. *Modern Discrete Probability: An Essential Toolkit*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2024. doi: 10.1017/9781009305129.

Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Locally differentially private protocols for frequency estimation. In *26th USENIX Security Symposium (USENIX Security 17)*, pp. 729–745, Vancouver, BC, August 2017. USENIX Association. ISBN 978-1-931971-40-9. URL https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/wang-tianhao.

Stanley L. Warner. Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60 309:63–6, 1965. URL https://api.semanticscholar.org/CorpusID:35435339.

Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models, 2025. URL https://arxiv.org/abs/2311.04378.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for ai-generated text, 2023. URL https://arxiv.org/abs/2306.17439.

# A Appendix

## A.1 Discussion on the definition of local DP watermarks

The likelihood ratio condition in Definition 1 says that the likelihoods of observing any text $y$ under the watermarked model and the base model are not too different, but does not yet apply to any arbitrary distribution independent of the secret key, such as the distribution of human-written text. Nevertheless, Definition 1 implies a similar likelihood ratio bound for any two keys $k$ and $k'$.

Denote $p(\cdot)$ as the prior distribution of keys, and $p(\cdot|y)$ the posterior distribution of keys given observed text $y$. Let $q(y \mid k)$ be the shorthand for $\Pr[\Gamma(k; p) = y \mid k]$. Definition 1 then implies:

$$\begin{aligned} p(k \mid y) &= \frac{q(y \mid k)p(k)}{\sum_{k'} q(y \mid k')p(k')} \leq \frac{e^{\varepsilon}p(y)p(k)}{\sum_{k'} e^{-\varepsilon}p(y)p(k')} \\ &= e^{2\varepsilon}p(k). \end{aligned}$$

Applying the same argument in the other direction gives $p(k \mid y) \geq e^{-2\varepsilon}p(k)$. Viewing the watermark as introducing an association of the text with the secret key, this condition limits the strength of such association that can be inferred by an observer with the key, and thus provides a natural notion of privacy for arbitrary distributions of non-watermarked text.

### A.2 RGS VIA REJECTION SAMPLING

**Algorithm 2** RGS via rejection sampling

**Require:** Prompt $x$, maximum number of attempts $M$, privacy budget $\varepsilon$, keyed hash $h_k$ with parameter $\pi_0$
**Ensure:** One sampled sequence $Y$.

```
Initialize counter C = 0
while C < M do
    Sample a sequence completion Y ~ LM(· | x) independently
    Increment the counter C ← C + 1
    if h_k(Y) = G then
        return Y.
    else
        With probability e^{-ε}, return Y
    end if
end while
return Y.
```

We show formally that our nominal privacy level $\varepsilon$ and expected samples are directly linked.

**Fact 1** (RGS-reject expected sample count)**.** *The expected number of samples needed for the rejection sampler in Algorithm 2 is at most* $1/(\Pr[h_k(Y) = G] + e^{-\varepsilon} \Pr[h_k(Y) = R])$.

The proof is given in A.5.9.

As this is a rejection sampler, there is a natural bound on the probability ratio between the original distribution and our sampler, giving a local DP guarantee.

**Proposition 4** (RGS-reject privacy)**.** *The rejection-sampler in Algorithm 2 is* $\varepsilon$*-DP.*

*Proof.* The proof follows directly from the choice of rejection weights, which can be found in A.5.2. □

### A.3 PRF

We construct a favored set with nominal probability $\pi_0 \in [0, 1]$ using keyed BLAKE2b with a 128-bit digest. Let $h_k(y) \in \{0, \ldots, 2^{128} - 1\}$ denote the integer representation of the digest of an encoded token sequence $y$. We define

$$G_k(\pi_0) = \{y : h_k(y) < T\}, \qquad T = \lfloor \pi_0 2^{128} \rfloor.$$

For analysis, we idealize the keyed hash as a uniformly random function. For any fixed sequence $y$ independent of this random function, exactly $T$ of the $2^{128}$ possible hash outputs imply membership, so

$$\Pr[y \in G_k(\pi_0)] = \frac{T}{2^{128}}, \qquad 0 \leq \pi_0 - \frac{T}{2^{128}} < 2^{-128}.$$

Then, we state the following assumption on the coloring properties of the hash function.

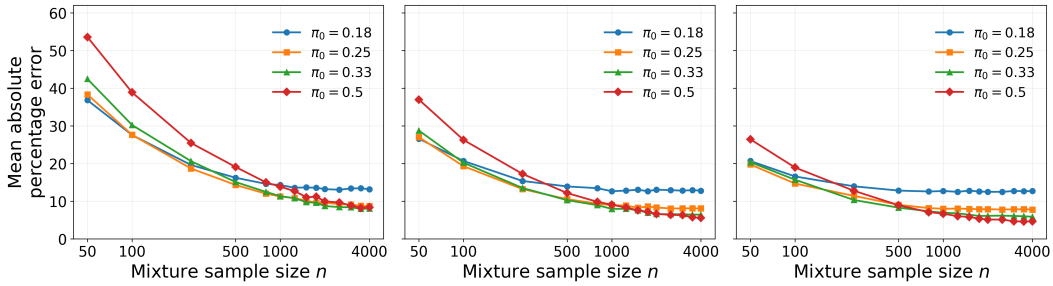


Figure 3: Mean absolute percentage errors ( $100\% \frac{|\hat{\lambda}-\lambda|}{\lambda}$ against sample sizes $n$ for various $\pi_0$ used. From left to right: $\lambda = 0.2, 0.3, 0.4$.

**Assumption 1** (Ideal coloring)**.** *For the keyed hash function $h_k$ with parameter $\pi_0$, we assume that for distinct sequences $y_1, \ldots, y_m$, the colors $h_k(y_1), \ldots, h_k(y_m)$ as independent and identically distribution* $\mathrm{Ber}(\pi_0)$ *random variables.*

In the formal PRF definition, these are only computationally indistinguishable (Goldreich et al., 1986), but we omit the negligible distinguishing advantage from our notation.

### A.4 Sample size scaling and bias-variance tradeoffs

We empirically test the effectiveness of the bias-mitigation strategy by tuning $\pi_0$ proposed in Section 5.3. We use the same C4 prompt mixture setup as 6.1, but sample from Llama-3.1-8B-Instruct with temperature=0.2, and truncate each completion to 20 tokens, reflecting low per-prompt entropy but high prompt diversity. We find that adjusting $\pi_0$ can significantly improve estimation error at large $n$ via the bias variance tradeoff. Figure 3 shows the mean absolute percentage error ($100\% \frac{|\hat{\lambda}-\lambda|}{\lambda}$) averaged over all mixtures and keys, for $\pi_0 \in \{0.18, 0.25, 0.33, 0.5\}$ and $\lambda \in \{0.2, 0.3, 0.4\}$, at $\varepsilon = 3$. The optimal choice of $\pi_0$ under perfect entropy is $\frac{1}{1+e^{-\varepsilon/2}} \approx 0.18$. We observe that at small sample sizes, smaller $\pi_0$ generally achieves lower error, but gradually get taken over by bigger values of $\pi_0$, and the largest $\pi_0 = 0.5$ eventually becomes the best across all three $\lambda$'s. In particular, for $\pi_0 = 0.18$, the errors plateau quickly across $\lambda$'s. This corroborates our theory that bias persists, and at large sample sizes where bias dominates the total error, a larger $\pi_0$ achieves a better tradeoff.

### A.5 Proofs and supporting technical lemmas

#### A.5.1 Proof: RGS Privacy

*Proof.* It suffices to show that $\forall y \in \mathcal{Y}$,

$$e^{-\varepsilon} \Pr_{\text{base}}[Y = y \mid k] \leq \Pr_{\text{RGS}}[Y = y \mid k] \leq e^{\varepsilon} \Pr_{\text{base}}[Y = y \mid k]$$

First, since $Q_R = 1 - Q_G$, we have $Q_R = \max\{1 - e^{\varepsilon} P_G, e^{-\varepsilon} P_R\}$, so $Q_R \geq e^{-\varepsilon} P_R$. Also, for $\varepsilon > 0$ we have $1 - e^{\varepsilon}(1 - P_R) = e^{\varepsilon} P_R - (e^{\varepsilon} - 1) \leq e^{\varepsilon} P_R$. Therefore, $Q_R \leq e^{\varepsilon} P_R$ and $Q_R \geq e^{-\varepsilon} P_R$. Similarly, $1 - e^{-\varepsilon} P_R = 1 - e^{-\varepsilon}(1 - P_G) \geq e^{-\varepsilon} P_G$.

Conditioned on the realized candidate set $Y^M = (Y_1, \ldots, Y_M)$, the probability of RGS outputting $y$ is

$$\begin{aligned}
\Pr_{\text{RGS}}[Y = y \mid Y^M, k] &= \mathbf{1}\{y \in Y^M\} \frac{c_y}{M} \left( \mathbf{1}\{h_k(y) = G\} \frac{Q_G}{P_G} + \mathbf{1}\{h_k(y) = R\} \frac{Q_R}{P_R} \right) \\
&\leq \frac{c_y}{M} \left( \mathbf{1}\{h_k(y) = G\} e^{\varepsilon} + \mathbf{1}\{h_k(y) = R\} e^{\varepsilon} \right) \\
&= e^{\varepsilon} \frac{c_y}{M}
\end{aligned}$$

For the other direction, we have

$$\begin{aligned}\Pr_{\text{RGS}}[Y = y \mid Y^M, k] &\geq \frac{c_y}{M}\left(\mathbf{1}\{h_k(y) = G\}e^{-\varepsilon} + \mathbf{1}\{h_k(y) = R\}e^{-\varepsilon}\right) \\ &= e^{-\varepsilon}\frac{c_y}{M}\end{aligned}$$

Taking expectation over candidate sets $Y^M$ we get

$$\begin{aligned}\Pr_{\text{RGS}}[Y = y \mid k] &= \mathbb{E}_{Y^M}\left[e^{\varepsilon}\frac{c_y}{M}\right] \\ &\leq e^{\varepsilon}\Pr_{\text{base}}[Y = y] \\ &= e^{\varepsilon}\Pr_{\text{base}}[Y = y \mid k]\end{aligned}$$

Similarly,

$$\Pr_{\text{RGS}}[Y = y \mid k] \geq e^{-\varepsilon}\Pr_{\text{base}}[Y = y \mid k]$$

□

### A.5.2 Proof for RGS rejection sampler privacy

*Proof.* We first assume there is no limit on the maximum candidate count. For any round in the rejection sampling procedure, let acc be the event that the sample $Y$ from the base LM is accepted by the rejection sampler and returned in the round. It suffices to show that $\forall y \in \mathcal{Y}$,

$$e^{-\varepsilon}\Pr[Y = y] \leq \Pr[Y = y \mid \mathsf{acc}, k] \leq e^{\varepsilon}\Pr[Y = y]$$

For simplicity, define

$$\begin{aligned}P_G &= \sum_{y\in\mathcal{Y}}\mathbf{1}\{h_k(y) = G\}\Pr[Y = y], \\ P_R &= \sum_{y\in\mathcal{Y}}\mathbf{1}\{h_k(y) = R\}\Pr[Y = y] \\ &= 1 - P_G\end{aligned}$$

The conditional probability of acceptance takes two values depending on the color of $y$:

$$\begin{aligned}\Pr[\mathsf{acc} \mid Y = y, k] &= \mathbf{1}\{h_k(y) = G\} + e^{-\varepsilon}\mathbf{1}\{h_k(y) = R\} \\ &\in \{1, e^{-\varepsilon}\}\end{aligned}$$

Then, the total probability of acceptance is

$$\begin{aligned}\Pr[\mathsf{acc} \mid k] &= \sum_{y\in\mathcal{Y}}\Pr[\mathsf{acc} \mid Y = y, k]\Pr[Y = y \mid k] \\ &= \sum_{y\in\mathcal{Y}}\Pr[\mathsf{acc} \mid Y = y, k]\Pr[Y = y] \\ &= \sum_{y\in\mathcal{Y}}\mathbf{1}\{h_k(y) = G\}\Pr[Y = y] + e^{-\varepsilon}\sum_{y\in\mathcal{Y}}\mathbf{1}\{h_k(y) = R\}\Pr[Y = y] \\ &= P_G + e^{-\varepsilon}P_R \\ &\in [e^{-\varepsilon}, 1].\end{aligned}$$

Therefore,

$$\begin{aligned}\Pr[Y = y \mid \mathsf{acc}, k] &= \frac{\Pr[\mathsf{acc} \mid Y = y, k]\cdot\Pr[Y = y \mid k]}{\Pr[\mathsf{acc} \mid k]} \\ &\leq \frac{1}{e^{-\varepsilon}}\Pr[Y = y] \\ &= e^{\varepsilon}\Pr[Y = y]\end{aligned}$$

Similarly for the other direction,

$$\begin{aligned}\Pr[Y = y \mid \mathsf{acc}, k] &\geq \frac{e^{-\varepsilon}}{1} \Pr[Y = y] \\ &= e^{-\varepsilon} \Pr[Y = y]\end{aligned}$$

Since reaching the finite $M$ maximum sample limit just falls back to a regular sample from the base LM, privacy is preserved. □

### A.5.3 PROOF: PRF IDENTITIES

*Proof.* Use the short hand $Z_y = \mathbf{1}\{h_k(y) = G\}$. Under the ideal random-coloring model, for distinct $y$'s, the $Z_y$'s are computationally indistinguishable with independent $\mathrm{Ber}(\pi_0)$ random variables. Since $g_i(k) = \sum_{y \in \mathcal{Y}} p_i(y) Z_y$, we have

$$\mathbb{E}_k[g_i(k)] = \sum_{y \in \mathcal{Y}} p_i(y) \mathbb{E}_k[Z_y] = \pi_0.$$

Furthermore, independence gives

$$\begin{aligned}\mathrm{Cov}_k[g_i(k), g_j(k)] &= \sum_{y, y' \in \mathcal{Y}} p_i(y) p_j(y') \mathrm{Cov}_k[Z_y, Z_{y'}] \\ &= \pi_0(1 - \pi_0) \sum_{y \in \mathcal{Y}} p_i(y) p_j(y) \\ &= \pi_0(1 - \pi_0) \langle p_i, p_j \rangle.\end{aligned}$$

□

### A.5.4 PROOF: CONDITIONAL MSE

*Proof.* With the unclipped moment estimator $\tilde{\lambda} = \frac{\frac{1}{n}\sum_{i=1}^n H_i - \pi_0}{\pi_1 - \pi_0}$, we have the decomposition

$$\tilde{\lambda} - \lambda = \frac{1}{\pi_1 - \pi_0} \left[ \frac{1}{n} \sum_{i=1}^n (H_i - q_i) + \frac{1}{n} \sum_{i \in I_1} (q_i - \pi_1) + \frac{1}{n} \sum_{i \in I_0} (g_i - \pi_0) \right].$$

Denote $V(k) := \frac{1}{n} \sum_{i=1}^n (H_i - q_i)$. Then

$$\tilde{\lambda} - \lambda = \frac{1}{\pi_1 - \pi_0} \Big[ V(k) + \lambda B(k) + (1 - \lambda) U(k) \Big]$$

Note that $B(k)$ and $U(k)$ are fixed given $k$. The only random quantity is $V(k)$, the sampling noise, with

$$\mathbb{E}[V(k) \mid k, W] = 0.$$

Then, the conditional MSE is

$$\mathbb{E}[(\tilde{\lambda} - \lambda)^2 \mid k, W] = \frac{1}{(\pi_1 - \pi_0)^2} \Big[ \mathbb{E}[V(k)^2 \mid k] + (\lambda B(k) + (1 - \lambda) U(k))^2 \Big].$$

By conditional independence, we have

$$\begin{aligned}\mathbb{E}[V(k)^2 \mid k, W] &= \mathrm{Var}[V(k) \mid k, W] \\ &= \mathrm{Var}\left[ \frac{1}{n} \sum_{i=1}^n H_i \mid k \right] \\ &= \frac{1}{n^2} \sum_{i=1}^n \mathrm{Var}[H_i \mid k] \\ &= \frac{1}{n^2} \left( \sum_{i \in I_1} q_i (1 - q_i) + \frac{1}{n^2} \sum_{i \in I_0} g_i (1 - g_i) \right) \\ &\leq \frac{1}{4n}.\end{aligned}$$

Since projection onto $[0,1]$ can only decrease the MSE, combining the above we have

$$\mathbb{E}[(\hat{\lambda}-\lambda)^2 \mid k, W] \le \frac{1}{(\pi_1-\pi_0)^2}\left[\frac{1}{4n} + (\lambda B(k) + (1-\lambda)U(k))^2\right]$$

□

### A.5.5 PROOF: PERSISTENT BIAS

*Proof.* Recall $f(x) = \frac{e^\varepsilon x}{e^\varepsilon x + 1 - x}$. We have the exact identity

$$f(x) - f(\pi_0) = f'(\pi_0)(x-\pi_0) - (e^\varepsilon - 1)f'(\pi_0)\frac{(x-\pi_0)^2}{1+(e^\varepsilon-1)x},$$

where $f'(\pi_0) = \frac{e^\varepsilon}{(1+(e^\varepsilon-1)\pi_0)^2}$ is decreasing in $\pi_0$.

Therefore,

$$\begin{aligned} B(k) &= \frac{1}{m}\sum_{i=1}^m f(g_i) - f(\pi_0) \\ &= f'(\pi_0)\left(\frac{1}{m}\sum_{i=1}^m g_i - \pi_0\right) - (e^\varepsilon-1)f'(\pi_0)\frac{1}{m}\sum_{i=1}^m \frac{(g_i-\pi_0)^2}{1+(e^\varepsilon-1)g_i} \end{aligned}$$

The first term is the counterpart of $U(k)$ for the watermarked texts, which vanishes with prompt diversity. The second term represents the low-entropy induced nonlinearity error, which does not simply vanish with sample size or prompt diversity.

$$\begin{aligned} \mathbb{E}_k[B(k)] &= f'(\pi_0)\left(\frac{1}{m}\sum_{i=1}^m \mathbb{E}_k[g_i] - \pi_0\right) - (e^\varepsilon-1)f'(\pi_0)\mathbb{E}_k\left[\frac{1}{m}\sum_{i=1}^m \frac{(g_i-\pi_0)^2}{1+(e^\varepsilon-1)g_i}\right] \\ &= -(e^\varepsilon-1)f'(\pi_0)\mathbb{E}_k\left[\frac{1}{m}\sum_{i=1}^m \frac{(g_i-\pi_0)^2}{1+(e^\varepsilon-1)g_i}\right] \end{aligned}$$

$$\begin{aligned} \mathbb{E}_k\left[\frac{1}{m}\sum_{i=1}^m \frac{(g_i-\pi_0)^2}{1+(e^\varepsilon-1)g_i}\right] &\ge \mathbb{E}_k\left[\frac{1}{m}\sum_{i=1}^m \frac{(g_i-\pi_0)^2}{1+(e^\varepsilon-1)}\right] \\ &= e^{-\varepsilon}\frac{1}{m}\sum_{i=1}^m \mathbb{E}_k[(g_i-\pi_0)^2] \\ &= e^{-\varepsilon}\pi_0(1-\pi_0)\frac{1}{m}\sum_{i=1}^m \|p_i\|_2^2 \end{aligned}$$

□

**Lemma 4** (Variance of $B(k)$)**.** *For the rejection sampler with $q_i = f(g_i)$ and $m = |I_1|$, under Assumption 1,*

$$\mathrm{Var}_k[B(k)] \le \frac{e^{2\varepsilon}\pi_0(1-\pi_0)}{m^2}\Big(\sum_{i\in I_1}\|p_i\|^2 + \sum_{i\ne j\in I_1}\langle p_i, p_j\rangle\Big).$$

The proof is given in Appendix A.5.5.

*Proof.* Use the shorthand $Z_y = \mathbf{1}\{h_k(y) = G\}$, which are i.i.d. $\mathrm{Ber}(\pi_0)$ under Assumption 1. We have that $B(k)$ is a function of $Z_y$'s. We have the decomposition

$$f(x) - f(x') = \frac{e^\varepsilon(x-x')}{(1+(e^\varepsilon-1)x)(1+(e^\varepsilon-1)x')}$$

, so replacing $Z_y$ by an i.i.d. copy $Z'_y$ changes each $g_i$ changes $B(k)$ by at most $c_y := \frac{e^\varepsilon}{m}\sum_{i\in I_1} p_i(y)$.

Note that this change is only nonzero only if $Z_y \neq Z'_y$, and since they are independent $\mathrm{Ber}(\pi_0)$ variables, this event has probability $2\pi_0(1-\pi_0)$.

By the Efron-Stein inequality (Boucheron et al., 2004), we have

$$\begin{aligned}\mathrm{Var}_k[B(k)] &\le \tfrac{1}{2}\sum_y 2\pi_0(1-\pi_0)\, c_y^2 \\ &= \frac{e^{2\varepsilon}\pi_0(1-\pi_0)}{m^2}\sum_{i,j\in I_1}\langle p_i, p_j\rangle.\end{aligned}$$

□

### A.5.6 PROOF: REJECTION SAMPLER BIAS MITIGATION

*Proof.*

$$f(x) - f(y) = \frac{e^\varepsilon(x-y)}{(1+(e^\varepsilon-1)x)(1+(e^\varepsilon-1)y)}$$

We can use the leave one out trick.

$$\begin{aligned}\mathbb{E}_k\left[f(g_i) - f(\pi_0)\right] &= \mathbb{E}_k\left[\frac{e^\varepsilon(g_i-\pi_0)}{(1+(e^\varepsilon-1)g_i)(1+(e^\varepsilon-1)\pi_0)}\right] \\ &= \mathbb{E}_k\left[\frac{e^\varepsilon(\sum_{y\in\mathcal{Y}} p_i(y)\mathbf{1}\{h_k(y)=1\}-\pi_0)}{(1+(e^\varepsilon-1)g_i)(1+(e^\varepsilon-1)\pi_0)}\right] \\ &= e^\varepsilon\sum_{y\in\mathcal{Y}} p_i(y)\mathbb{E}\left[\frac{\mathbf{1}\{h_k(y)=1\}-\pi_0}{(1+(e^\varepsilon-1)g_i)(1+(e^\varepsilon-1)\pi_0)}\right]\end{aligned}$$

Denote $g_i^{-y} = \sum_{z\neq y} p_i(z)\mathbf{1}\{h_k(z)=G\}$. Shorthand $Z_y = \mathbf{1}\{h_k(y)=G\}$. Then, $g_i = g_i^{-y} + p_i(y)Z_y$. Note that conditional on $g_i^{-y}$, $Z_y \sim \mathrm{Ber}(\pi_0)$.

$$\begin{aligned}\mathbb{E}\left[\frac{p_i(y)Z_y-\pi_0}{1+(e^\varepsilon-1)g_i}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{Z_y-\pi_0}{1+(e^\varepsilon-1)(g_i^{-y}+p_i(y)Z_y)} \mid g_i^{-y}\right]\right] \\ &= \mathbb{E}\left[\pi_0\frac{1-\pi_0}{1+(e^\varepsilon-1)(g_i^{-y}+p_i(y))} + (1-\pi_0)\frac{-\pi_0}{1+(e^\varepsilon-1)g_i^{-y}}\right] \\ &= -\pi_0(1-\pi_0)(e^\varepsilon-1)p_i(y)\mathbb{E}_k\left[\frac{1}{(1+(e^\varepsilon-1)(g_i^{-y}+p_i(y))(1+(e^\varepsilon-1)g_i^{-y})}\right]\end{aligned}$$

Consolidating,

$$\begin{aligned}\frac{\mathbb{E}_k[B(k)]}{f(\pi_0)-\pi_0} &= \frac{1}{f(\pi_0)-\pi_0}\frac{1}{m}\sum_{i=1}^m \mathbb{E}_k[f(g_i)-f(\pi_0)] \\ &= \frac{-\pi_0(1-\pi_0)e^\varepsilon(e^\varepsilon-1)}{m(f(\pi_0)-\pi_0)(1+(e^\varepsilon-1)\pi_0)}\sum_{i=1}^m\sum_{y\in\mathcal{Y}} p_i(y)^2\mathbb{E}_k\left[\frac{1}{r(g_i^{-y})}\right]\end{aligned}$$

Since $g_i^{-y} = \sum_{y'\neq y} Z_{y'}$ and each $Z_{y'} \overset{i.i.d.}{\sim} \mathrm{Ber}(\pi_0)$, by Corollary 4.2.6. of Roch (2024) we have that $\mathbb{E}_k[\frac{1}{r(g_i^{-y})}]$ is nonincreasing in $\pi_0$. Also, $\frac{-\pi_0(1-\pi_0)e^\varepsilon(e^\varepsilon-1)}{m(f(\pi_0)-\pi_0)(1+(e^\varepsilon-1)\pi_0)}$ is decreasing in $\pi_0$ so overall the positive gap $\frac{\mathbb{E}_k[-B(k)]}{f(\pi_0)-\pi_0}$ is decreasing in $\pi_0$. □

### A.5.7 PERSISTENT ESTIMATOR BIAS FOR THE WEIGHTED COUNTS SAMPLER

**Fact 2.** *For both samplers, the bias term $B(k) = \sum_{i \in I_1}(q_i - \pi_1)$ is non-positive in expectation over $k$.*

*Proof.* Rejection sampler: Let $f(x) = \frac{e^\varepsilon x}{e^\varepsilon x + 1 - x}$ denote the reweighting function.

$$B(k) = \frac{1}{|I_1|} \sum_{i \in I_1} f(g_i(k)) - f(\pi_0).$$

Since $f$ is concave, by Jensen's inequality we have

$$\pi_1 = f(\pi_0) = f(\mathbb{E}_k[g_i]) \geq \mathbb{E}_k \left[ \frac{1}{m} \sum_{i=1}^{m} f(g_i) \right].$$

For the parallel sampler: With prompt $x_i$ and its corresponding base distribution $p_i$, conditioned on the candidate set $Y^M$, the probability of RGS outputting green sequence is

$$q_i = \Pr[h_k(Y) = G \mid Y^M, k, W] = \min\{e^\varepsilon P_G^i, 1 - e^{-\varepsilon} P_R^i\}$$

We have that given $k$ and over the randomness in $Y^M$,

$$M \cdot P_G^i \sim \mathrm{Bin}(M, g_i).$$

Let $q(x) = \min\{e^\varepsilon x, 1 - e^{-\varepsilon}(1 - x)\}$. We note that $q$ is concave because it is the minimum of two affine functions. Then, applying Jensen's inequality twice, we have

$$\begin{aligned}
\mathbb{E}_k[q_i] &= \mathbb{E}_k \left[ \mathbb{E}_{Y^M} \left[ \min\{e^\varepsilon P_G^i, 1 - e^{-\varepsilon} P_R^i\} \mid k \right] \right] \\
&= \mathbb{E}_k \left[ \mathbb{E}_{Y^M} \left[ q(P_G^i) \mid k \right] \right] \\
&\leq \mathbb{E}_k \left[ q(\mathbb{E}_{Y^M}[P_G^i \mid k]) \right] \\
&= \mathbb{E}_k \left[ q(g_i(k)) \right] \\
&\leq q(\mathbb{E}_k[g_i(k)]) \\
&= q(\pi_0) \\
&= \pi_1.
\end{aligned}$$

Averaging the above over $i \in I_1$ gives the desired fact. □

### A.5.8 PROOF: KEY-AVERAGED MSE

*Proof.* We bound the last three terms of equation 5 Since $\mathbb{E}_k[U(k)] = 0$, we have

$$\mathbb{E}_k[U(k)^2] = \mathrm{Var}_k[U(k)] = \pi_0(1 - \pi_0)\|\bar{p}_0\|_2^2.$$

Then, we bound the cross term using Cauchy-Schwarz:

$$\begin{aligned}
\mathbb{E}_k[B(k)U(k)] &\leq |\mathbb{E}_k[B(k)U(k)]| \\
&\leq \sqrt{\mathbb{E}_k[B(k)^2]} \sqrt{\mathbb{E}_k[U(k)^2]} \\
&= \sqrt{\mathbb{E}_k[B(k)^2]} \sqrt{\pi_0(1 - \pi_0)\|\bar{p}_0\|_2^2} \\
&\leq \sqrt{\pi_0(1 - \pi_0)}\, \|\bar{p}_0\|_2,
\end{aligned}$$

Finally, we bound $\mathrm{Var}_k[B(k)]$ using Lemma 4:

$$\begin{aligned}
\mathbb{E}_k[B(k)^2] &= \mathrm{Var}_k[B(k)] + (\mathbb{E}_k[B(k)])^2 \\
&\leq e^{2\varepsilon} \pi_0(1 - \pi_0)\|\bar{p}_1\|_2^2 + (\mathbb{E}_k[B(k)])^2.
\end{aligned}$$

Substituting the above into equation 5 gives the desired inequality with remainder $R$. Every term of $R$ is proportional to $\|\bar{p}_0\|_2$, $\|\bar{p}_0\|_2^2$ or $\|\bar{p}_1\|_2^2$, so $R \to 0$ under the assumption. □

### A.5.9 PROOF: REJECTION SAMPLER EXPECTED SAMPLE COUNT

*Proof.* Fix any key $k$. In each round $m \in [M]$ of the rejection sampling procedure, the probability of accepting a completion $Y$ is

$$a := 1 \cdot \Pr[h_k(Y) = G] + e^{-\varepsilon} \Pr[h_k(Y) = R].$$

Since the $Y$'s are sampled independently from the base model across all $M$ rounds, the acceptance indicator are independent $\mathrm{Ber}(a)$ random variables. Let $X$ be the number of rounds until acceptance, then $X \sim \text{Geometric}(a)$. The sample count needed by the rejection sampler is thus $T = \min(M, X)$, with $\mathbb{E}[T] \leq \mathbb{E}[X] = \frac{1}{a}$. □