WATERMARKABLE MULTI-DRAFT SPECULATIVE SAMPLING VIA POISSON PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have achieved state-of-the-art performance across a wide range of tasks, motivating two important aspects of deployment: inference efficiency and output provenance, which can be tackled by speculative sampling and watermarking, respectively. However, recent works have shown that combining these two goals is highly nontrivial and can be potentially impossible. In this work, we develop a novel multi-draft speculative sampling algorithm based on Poisson processes that improves the frontier of this fundamental trade-off. The proposed algorithm has strong sampling efficiency on its own and, more interestingly, is naturally *watermarkable*: we can embed an unbiased watermark without degrading speculative acceptance. Moreover, our algorithm is based on an *exact* list-coupling-without-communication scheme and satisfies certain level of *drafter invariance* that provides robustness to drafter substitution. Our multi-draft construction jointly provides exact target sampling, stable target-keyed watermark strength as the number of drafts increases, and strong empirical sampling efficiency.

1 INTRODUCTION

Large language models (LLMs) have become a central component of modern generative AI, achieving strong performance in dialogue, reasoning, and open-ended content creation. Their probabilistic nature based on autoregressive sampling enables them to generate diverse responses (Brown et al., 2020; Singh et al., 2025). However, as LLM-generated text becomes increasingly fluent and difficult to distinguish from human-written content, concerns about accountability, provenance, misuse, and intellectual property have attracted increasing attention (Wu et al., 2025). For instance, ChatGPT-generated scientific abstracts are hard for human reviewers to distinguish from human-generated ones, yet they may contain fabricated data (Shumailov et al., 2024). Hence, watermarking has become an important, and potentially necessary, tool for tracing the provenance of LLM-generated text by embedding statistical signals into the sampling process while preserving output quality (Kirchenbauer et al., 2023; Aaronson and Kirchner, 2023). For instance, reflecting its importance, Anthropic recently announced that future Claude models will generate text with embedded watermarks (Anthropic, 2026).

Apart from provenance, the efficiency of text generation is another bottleneck in the system pipeline. *Speculative sampling* has emerged as a promising approach to accelerate content generation without hurting the target model distribution (Chen et al., 2023; Leviathan et al., 2023). However, it is not independent of watermarking. On the contrary, recent works (Hu and Huang, 2024; He et al., 2026) show that combining the two is highly nontrivial: by Hu and Huang (2024), a "no-go" theorem establishes a fundamental trade-off between watermark strength and speculative sampling efficiency. He et al. (2026) improves this trade-off by considering alternative watermark metrics.

In this work, we ask the following question: can we design a watermarkable speculative sampling scheme that further improves the trade-off (Hu and Huang, 2024; He et al., 2026) by increasing speculative sampling efficiency without compromising watermark detectability?

We answer this question in the **affirmative**: we design a novel multi-draft speculative sampling algorithm that improves speculative sampling efficiency while naturally embedding an unbiased watermark and, more importantly, preserves watermark strength as the number of drafts and sampling efficiency increase. Our contributions are two-fold:

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.

- **Multi-draft speculative sampling**: We design a novel multi-draft speculative sampling algorithm that generates and verifies multiple draft continuations. Our scheme uses desirable properties of Poisson processes, which yield a drafter invariance property (Daliri et al., 2025) and provides operational robustness when the drafter is changed. Compared with baseline methods, we prove a stronger acceptance guarantee and empirically obtain higher token rates consistently across all tested Qwen/Llama multi-draft settings.
- **Efficiency-preserving watermarking**: Besides desirable sampling efficiency, our algorithm is naturally *watermarkable* with the use of Poisson processes, which can embed a recoverable statistical signal while maintaining the distribution *unbiased*. By building the watermark directly into the sampling procedure through keyed randomness, we preserve the watermark strength while improving sampling efficiency along the additional multi-draft dimension.[1]

2 Related Works

LLM Decoding and Speculative Sampling. Standard autoregressive decoding methods, including top-k sampling (Radford et al., 2019; Fan et al., 2018), nucleus sampling (Holtzman et al., 2020), and the permute-and-flip decoder (Zhao et al., 2025a), generate tokens sequentially; and therefore, suffer from the inherent latency of one target-model call per token (Stern et al., 2018). Speculative sampling (Chen et al., 2023; Leviathan et al., 2023) addresses this bottleneck by using a smaller draft model to propose several future tokens, while the larger target model verifies the proposed continuations in parallel. The efficiency gain depends on how well the draft distribution matches the target distribution. Recent works further improve this paradigm through optimal-transport-based token selection (Sun et al., 2023), tree-based inference and verification (Miao et al., 2024), and multi-draft architectures (Khisti et al., 2025; Rowan et al., 2025). It can also be used to accelerate diffusion models (De Bortoli et al., 2025; Bullo et al., 2026).

Watermarking. LLM watermarking aims to embed detectable statistical signals into generated text while preserving fluency and utility. Early logit-bias methods pseudorandomly select green-list tokens and detect the resulting bias by hypothesis testing (Kirchenbauer et al., 2023), while later works improve robustness, accessibility, and probability balance (Wu et al., 2023; Giboulot and Furon, 2024; Park et al., 2026). A parallel line studies unbiased watermarking, where the sampler maintains the distribution while correlating the output with secret randomness (Kuditipudi et al., 2023; Hu et al., 2024; Christ et al., 2024). They connect watermarking to information theory (Moulin and O'Sullivan, 2003) and statistics, including detection power, false-positive control, optimal tests, and couplings (Li et al., 2025; He et al., 2025; Tsur et al., 2025). Related works include robust or multi-bit attribution (Zhao et al., 2023; Boroujeny et al., 2024; Qu et al., 2025), and efficient identification of watermarked segments (Zhao et al., 2025b). In recent works (Hu and Huang, 2024; He et al., 2026) an inevitable trade-off between speculative sampling and watermarking has been discussed.

Poisson Functional Representation (PFR). PFR is a sampling scheme introduced in (Li and El Gamal, 2018), with related Poisson process constructions also discussed by Maddison et al. (2014); Maddison (2016). It selects a sample from a target distribution through a Poisson process and a proposal distribution, and has been applied to information theory settings (Li and Anantharam, 2021; Khisti et al., 2024; Liu and Li, 2024; 2025; Liu et al., 2026) and machine learning (He et al., 2024; Liu et al., 2024; Zhou et al., 2026; Flamich et al., 2026). It can be extended to list-coupling (Li and Anantharam, 2021) by the mapping theorem (Last and Penrose, 2017) of Poisson process, and achieves the best known bound on coupling without communication (Daliri et al., 2025).

Notations. $x_{i:j}$ denotes the sequence $x_i, x_{i+1}, \ldots, x_j$, and we let $x_{:j} := x_{1:j}$. $P \ll Q$ denotes that probability measure P is absolutely continuous with respect to Q. The logarithm is on base 2.

3 Speculative Sampling

We first briefly review the technical background of speculative sampling (Chen et al., 2023; Leviathan et al., 2023; Sun et al., 2023; Rowan et al., 2025). Given volcabulary $\mathcal{V}$ and a context $x_{:t} :=$

[1] Our construction does not contradict the single-draft trade-off of Hu and Huang (2024); rather, it introduces the number of drafts as an additional design dimension.

$(x_1, x_2, \ldots, x_t)$, an autoregressive language model $\mathcal{M}_b$ generates the next token by sampling from the conditional distribution $\mathcal{M}_b(\cdot|x_{:t})$ under temperature sampling (Ackley et al., 1985; Ficler and Goldberg, 2017). *Speculative sampling* accelerates this process by introducing a smaller and cheaper *draft model* $\mathcal{M}_s$, which proposes several candidate tokens before the target model verifies them. Given a prefix $x_{:t}$, one iteration of the speculative sampling consists of the following three stages:

1. **Draft Stage.** The draft model $\mathcal{M}_s$ sequentially samples L candidate tokens $\tilde{x}_{t+1}, \ldots, \tilde{x}_{t+L}$. For each position $i = 1, \ldots, L$, we record the draft distribution $\mathcal{M}_s(\cdot|x_{:t}, \tilde{x}_{t+1:t+i-1})$.
2. **Computation Stage.** Conditioned on the drafted context, the target model $\mathcal{M}_b$ evaluates the corresponding conditional distributions, potentially in parallel:
$$\mathcal{M}_b(\cdot|x_{:t}), \mathcal{M}_b(\cdot|x_{:t}, \tilde{x}_{t+1}), \ldots, \mathcal{M}_b(\cdot|x_{:t}, \tilde{x}_{t+1:t+L}),$$
3. **Selection Stage.** By the distributions $\mathcal{M}_s$ and $\mathcal{M}_b$, we accept the longest prefix of the drafted tokens until the first rejection, and a *correction token* is sampled from a specific residual distribution. If all L drafts are accepted, one more token from the target is sampled.

The acceptance-correction procedure ensures that the generated tokens follow *exactly* the same distribution as direct autoregressive sampling, i.e., improving decoding efficiency does not change the target law. The exactness is achieved via a recursive token-level maximal coupling (Algorithm 1), where we use P, Q to denote the target and draft distributions, respectively, and achieve

$$\Pr(X = Y) = \sum\nolimits_{x \in \mathcal{V}} \min(P(x), Q(x)) = 1 - \|P - Q\|_{\mathrm{TV}},$$

i.e., the closer P and Q are, the higher the probability that the draft token will be accepted. Together with the extra sampled token, there could be at most $L + 1$ tokens generated in each iteration, and therefore, the speedup is up to $(L + 1)$ times if the decoding time of the draft model is negligible.

Kobus and Gündüz (2025) proposed a new sampling scheme (Algorithm 2) based on exponential random variables, which, together with the token probabilities, determine the scores assigned to the tokens. The token with the smallest score wins the race and follows the desired distribution.

Algorithm 1: Token-level sampling (Sun et al., 2023; Chen et al., 2023; Leviathan et al., 2023)

Input: Target and draft distribution P, Q; $X \sim Q$.
Compute the residual P_{res}: for any $x \in \mathcal{V}$,

$$P_{\mathrm{res}}(x) = \frac{P(x) - \min\{P(x), Q(x)\}}{1 - \sum_y \min\{P(y), Q(y)\}}.$$

Sample $U \sim \mathrm{Unif}(0, 1)$
if $U \leq \min(1, P(X)/Q(X))$ **:**
  return $Y = X$ `// Accept`
return $Y \sim P_{\mathrm{res}}$ `// Correction`

Algorithm 2: Token-level sampling by exponential race (Kobus and Gündüz, 2025)

Input: Target and draft distribution P, Q.

for $i \in \mathcal{V}$ **do**
  $E_i \sim \mathrm{Exp}(1)$
$X \leftarrow \arg\min_{i \in \mathcal{V}} E_i / Q(i)$
$Y \leftarrow \arg\min_{i \in \mathcal{V}} E_i / P(i)$
if $X = Y$ **:**
  return $Y = X$ `// Accept`
return Y `// Correction`

4 Watermarkable Speculative Sampling

We first present a single-draft version for illustration. Our use of Poisson processes naturally embeds the watermark into the sampling process via pseudorandomness. The formal multi-draft scheme, which further accelerates autoregressive generation without compromising watermark strength, will be presented in Section 5. Before proceeding, we first introduce two key components of our scheme.

4.1 Unbiased Watermark

A good watermark should be detectable from the generated text while preserving output quality. *Unbiased* schemes use pseudorandomness to couple tokens with secret keys, so that the watermark can be detected through statistical tests without significantly altering the marginal distribution.

Suppose $\tilde{P}$ is a base distribution over $\mathcal{V}$, an unbiased watermark is a decoding rule such that, for each seed ζ, it induces a watermarked distribution $\tilde{P}_\zeta$, which preserves $\tilde{P}$ after averaging over the seed,

$$\mathbf{E}_\zeta[\tilde{P}_\zeta(v)] = \tilde{P}(v), \qquad \forall v \in \mathcal{V}.$$

The Gumbel watermark (Aaronson and Kirchner, 2023), which uses the Gumbel-max trick (Gumbel, 1954) and lets the seed ζ determine the Gumbel noise, is unbiased, and we leverage a similar idea for our scheme. Let $u_t(\cdot|x, y_{t-1}) : \mathcal{V} \to \mathbb{R}$ be a real-valued function (a.k.a. *logits*) that encodes the model's preferences on words and T be the temperature, the softmax sampling is equivalent to

$$y_t = \text{argmax}_{y\in\mathcal{V}}\big(u_t(y)/\mathsf{T}\big) + G_t(y), \quad G_t(y) \sim \text{Gumbel}(0,1) \text{ i.i.d for each } t, y,$$

where the Gumbel noise can be generated by $\text{Gumbel}(0,1) \sim -\log(\log(1/\text{Unif}([0,1])))$.

A random $r_t \sim (\text{Unif}([0,1]))^{|\mathcal{V}|}$ can be represented by $G_t(y) = -\log(-log(r_t(y)))$. We can replace $\text{Unif}([0,1])$ by a pseudo-random function $r_t(y) = F_{y_{t-m:t-1},\mathsf{k}}(y)$ with prefix length m, and employ

$$y_t = \text{argmax}_{y\in\mathcal{V}}\big(u_t(y)/\mathsf{T}\big) - \log(-\log(r_t(y))). \tag{1}$$

Conditioned on a secret key k, $r_t(y)$ is a deterministic function, but over the distribution of k, $r_t(y)$ is computationally indistinguishable from samples drawn from a truly i.i.d. uniform distribution.

The detector has the key k and computes the watermark score $\sum_{t=m+1}^{n} -\log\big(1 - r_t(y_t)\big)$. If $y_{1:n}$ is unwatermarked, the expected score is $n - m$; otherwise, it is larger and the watermark is detected.

Our algorithm relies on Poisson processes and also incorporates the use of pseudo-random functions.

4.2 Poisson Functional Representation (PFR)

We leverage a scheme called PFR (Li and El Gamal, 2018); related works are discussed in Section 2.

For a finite vocabulary, PFR has a simple exponential-race interpretation: assign each token v an independent clock $E_v \sim \text{Exp}(1)$ and select $\arg\min_v E_v/P(v)$. Sharing the same keyed clocks between the target and draft couples their samples while preserving each marginal. In the single-draft discrete setting, this is the same shared-randomness construction as Gumbel-max/DeltaGumbel-MWS (Hu and Huang, 2024); the Poisson representation becomes useful for the multi-draft extension because its mapped order statistics provide an exact list of i.i.d. samples.

Definition 4.1 (PFR)**.** Let $(T_i)_i \sim \text{PP}(1)$ be a Poisson process of rate 1, independent of $Z_i \overset{\text{iid}}{\sim} \nu$ for $i = 1, 2, \ldots$. $(Z_i, T_i)_i$ is a Poisson process of intensity measure $Q \times \lambda_{[0,\infty)}$, where $\lambda_{[0,\infty)}$ is the Lebesgue measure over $[0, \infty)$. Fix distribution P over $\mathcal{Z}$ such that $P \ll Q$. PFR selects the point

$$Z = Z_K \quad \text{where} \quad K := \text{argmin}_i\big(T_i \cdot \big(\mathrm{d}P/\mathrm{d}Q(Z_i)\big)^{-1}\big). \tag{2}$$

Remark 4.2. *By drawing a sequence $(Z_i)_i$ from the* reference *distribution Q and a sequence of* times *$(T_i)_i$, the PFR selects a sample following the* target *distribution P. The sample Z_i with the smallest T_i only follows Q; to obtain a sample from P exactly, we inflate the time by factor $(\mathrm{d}P/\mathrm{d}Q(Z_i))^{-1}$.*

4.3 Watermarkable Speculative Sampling via Poisson Processes

Core idea. We unify watermarking and speculative sampling through a keyed Poisson coupling. Every emitted token is target-distributed, while the draft model only controls how many such samples are emitted. The watermarking is performed at the *coupling level*, rather than through a post-hoc reweighting of marginals. This allows our scheme to stay efficient when embedding watermark. [2]

Algorithm 3 is a sketch of our single-draft speculative sampling with watermark, see a detailed implementation in Algorithm 9. Fixing a Poisson process Π_k by key k; with context c, we denote the labeled Poisson process and the sample selected by PFR by $G(\mathsf{k}, c)$ and $\text{PFR}(\bar{P}(\cdot|c), \Pi)$, respectively.

For discrete LLM target distributions, as we prove in Appendix A.1, PFR reduces to the exponential race (Kobus and Gündüz, 2025) and is functionally equivalent to the shared Gumbel/DeltaGumbel

[2]The Maintain Watermark Strength scheme (Hu and Huang, 2024) shares the same spirit; see Figure 1. Nevertheless, our scheme can be extended to the multi-draft version in Section 5, achieving further acceleration.

Algorithm 3: Watermarkable Speculative Sampling via Poisson Processes (Simplified)

Given : lookahead K, output length N, initial prompt $w_{1:n}$, Poisson generator G, secret key k

while $n < N$ **do**

  $h \leftarrow w_{1:n}$; $L \leftarrow \min\{K,\, N-n\}$; $accepted \leftarrow$ **true**;

  for $s = 1$ **to** L **do**

    $c_s \leftarrow h \,\|\, \tilde{w}_{1:s-1}$; $\Pi_s \leftarrow G(\mathsf{k}, c_s)$ `// keyed Poisson for watermark`

    $\tilde{w}_s \leftarrow \mathrm{PFR}(Q(\cdot|c_s), \Pi_s)$

  for $s = 1$ **to** L **do** `// evaluation in parallel`

    $y_s \leftarrow \mathrm{PFR}(P(\cdot|c_s), \Pi_s)$; $a_s \leftarrow \mathbb{1}\{y_s = \tilde{w}_s\}$; $w_{n+1} \leftarrow y_s$; $n \leftarrow n+1$;

    if $a_s = 0$ **:**

      $accepted \leftarrow$ **false** and **break** `// later draft contexts are invalid`

  if $accepted$ ***and*** $n < N$ **:** `// all draft tokens accepted`

    $\Pi^\star \leftarrow G(\mathsf{k}, w_{1:n})$ `// keyed Poisson source`

    $w_{n+1} \leftarrow \mathrm{PFR}(P(\cdot|w_{1:n}), \Pi^\star)$; $n \leftarrow n+1$ `// bonus`

construction underlying MWS (Hu and Huang, 2024). We use it as the single-draft specialization and starting point for MPFR, rather than as a distinct single-draft coupling. Poisson processes become particularly useful for our multi-draft construction, where the mapping theorem (Last and Penrose, 2017) generates multiple exact samples from one shared process.

He et al. (2026) used a different watermark metric to improve the trade-off frontier. In comparison, our scheme is robust across different watermark measurements, as we consider various scores (Aaronson and Kirchner, 2023; Hu and Huang, 2024; Li et al., 2025; Lattimore, 2026) in Appendix E, which makes our scheme directly comparable to existing watermarking schemes. Our construction is not a black-box wrapper for arbitrary watermarking rules: it instantiates a distortion-free keyed watermark, while the detector scores in Appendix E are different tests of the same keyed pivot. We use Aaronson's score (Aaronson and Kirchner, 2023) in the main section, which is introduced as follows.

If the output of key k is $w_{1:N}$ and n_0 is the initial prompt length, the detector evaluates the scored positions $\mathcal{T} := \{n_0+1, \ldots, N\}$. For $t \in \mathcal{T}$, let $c_t := w_{:t-1}$ be the prefix before position t, and the detector regenerates $\Pi_t := G(\mathsf{k}, c_t)$. For each $v \in \mathcal{V}$, let $r_t(v) := \exp(-\mu(v)\tau_t(v))$ be the induced uniform value where $\tau_t(v) \sim \mathrm{Exp}(\mu(v))$ is the first arrival time of v in Π_t, and the detector computes

$$S_{\mathrm{A}} := \sum\nolimits_{t\in\mathcal{T}} -\log(1 - r_t(w_t)), \tag{3}$$

whose expectation is $\mathbf{E}[S_{\mathrm{A}}] \geq |\mathcal{T}| + (\pi^2/6 - 1)\sum_{t\in\mathcal{T}} \mathbf{E}[H(P(\cdot|w_{:t-1}))]$. If the sequence is not generated with k, the observed w_t is independent of Π_t, $r_t(w_t) \sim \mathrm{Unif}[0,1]$ and $\mathbf{E}[S_{\mathrm{A}}] = |\mathcal{T}|$.

At-a-Glance Experimental Validation. Figure 1 gives an at-a-glance comparison. The single-draft PFR curve is the Poisson realization of the shared Gumbel coupling described above; the new trade-off improvement comes from the additional draft-count dimension in MPFR, where sampling efficiency increases while the target-keyed watermark signal remains stable.

4.4 Drafter Invariance and Watermarking

Drafter-invariance asks that changing the draft model not introduce additional draft dependence into the emitted target-side sequence (Daliri et al., 2025). This is operationally useful when a drafter is updated or replaced. We use a stopping-time notion: the drafter may change the accepted block length, but conditional on that stopping time, the emitted prefix is governed by the target-side randomness.

Let $Y_{1:\tau}$ and $\widetilde{Y}_{1:\widetilde{\tau}}$ denote the output sequences obtained with two drafters in a block. We define:

Definition 4.3. The algorithm $\mathcal{A}$ is said to be *stopping-time drafter invariant* if, for every stopping time value τ_0 in the support of τ, every $1 \leq j \leq \tau_0$, and every output prefix $y_{1:j}$, we have

$$\mathbf{P}\{Y_{1:j} = y_{1:j} \,|\, \mathcal{R},\, x_{:t},\, \tau = \tau_0\} = \mathbf{P}\{\widetilde{Y}_{1:j} = y_{1:j} \,|\, \mathcal{R},\, x_{:t},\, \widetilde{\tau} = \tau_0\}.$$

Proposition 4.4. *Our Algorithm 3 is stopping-time drafter invariant.*

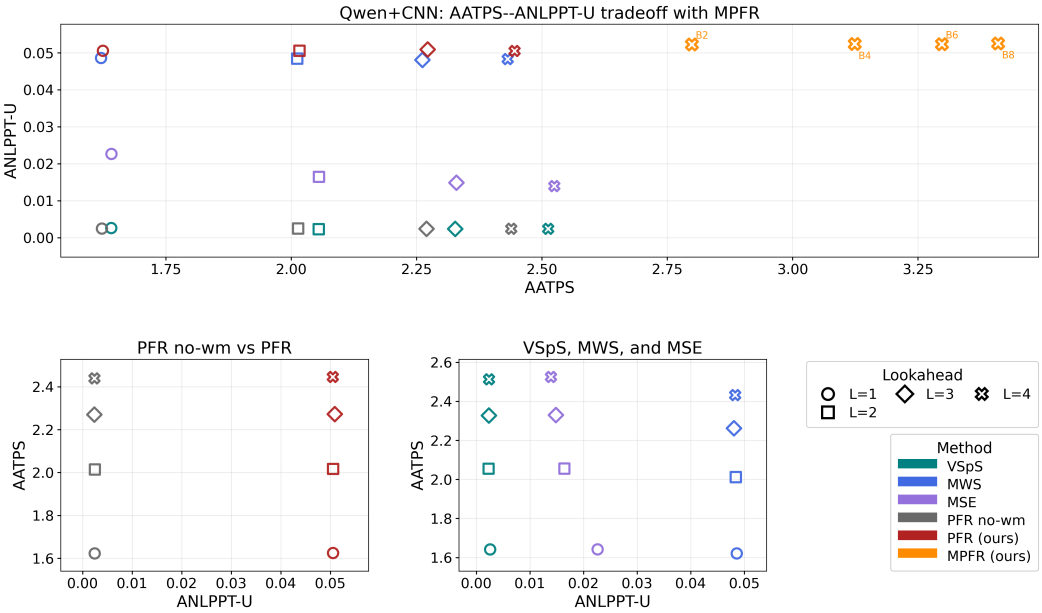


Figure 1: Sampling efficiency versus watermark strength for existing single-draft methods (Hu and Huang, 2024), single-draft PFR, and multi-draft MPFR (Section 5). MPFR increases AATPS while maintaining ANLPPT; see Table 1 for token rate improvements. This figure uses Qwen2.5-7B-Instruct and Qwen2.5-0.5B-Instruct as the target and draft, respectively, on CNN/DAILYMAIL. Experiments on more models and datasets are shown in Appendix D.

The proof is in Appendix C.1.1. This notion lies between strong drafter invariance (Daliri et al., 2025) and the conditional drafter invariance of Rowan et al. (2025). It is neither necessary for nor equivalent to strong watermark detection; here its role is operational robustness. Conditioning on τ separates the block length, which the drafter can affect, from the emitted content, which under our keyed construction does not acquire additional dependence on the realized draft trajectory.

There is nevertheless an efficiency tension. Maximal coupling can attain $\mathbf{P}(X = Y) = 1 - \|P - Q\|_{\mathrm{TV}}$ for $X \sim P, Y \sim Q$ (Thorisson, 2000; Den Hollander, 2012), whereas the best matching probability under coupling without communication satisfies

$$\mathbf{P}(X = Y) \leq (1 - \|P - Q\|_{\mathrm{TV}})/(1 + \|P - Q\|_{\mathrm{TV}}). \tag{4}$$

5 WATERMARKABLE MULTI-DRAFT SPECULATIVE SAMPLING

We now extend our Algorithm 3 to multi-draft speculative sampling, by a multi-sample generalization of PFR built on the mapping theorem for Poisson processes (Li and Anantharam, 2021).

We use $\mathcal{M}_t$ and $\mathcal{M}_d^{(b)}$ to denote the *target* and *draft* language models with $1 \leq b \leq B$. They induce conditional distributions $\mathcal{M}_t(\cdot|x_{:t})$ and $\mathcal{M}_d^{(b)}(\cdot|x_{:t})$, which give the probability of the next token given the current context $x_{:t}$, usually denoted by c. In multi-draft speculative sampling (Miao et al., 2024; Sun et al., 2023; Khisti et al., 2025; Rowan et al., 2025), at each context c, we generate a list of B proposal drafts, often i.i.d. in practice (Rowan et al., 2025; Sun et al., 2023) from $\mathcal{M}_d^{(1)}, \ldots, \mathcal{M}_d^{(B)}$. These proposal drafts induce a depth-L speculative tree rooted at the current prefix, which is then verified in parallel with $\mathcal{M}_t$. At each step, if at least one of the B candidate tokens is accepted, the first accepted token will be appended to the final output; otherwise, the verification procedure stops. The output is $Y_{1:\tau}$, where τ is the number of accepted tokens plus one.

The PFR (Definition 4.1) selects a sample $\tilde{Z}_P$ from $(Z_i)_i$ that has the smallest *score* $T_i \cdot \left(\frac{\mathrm{d}P}{\mathrm{d}Q}(Z_i)\right)^{-1}$, where $(T_i)_i \sim \mathrm{PP}(1)$, and it follows the target distribution P. To get B samples, multi-sample PFR

Algorithm 4: Watermarkable Multi-Draft Speculative Sampling (Simplified)

Given : lookahead L, number of homogeneous drafts B, output length N, target model P, draft model Q, initial prompt $w_{1:n}$, keyed source generator G, and secret key k

while $n < N$ **do**
- $h \leftarrow w_{1:n}$; $\ell \leftarrow \min\{L, N-n\}$ `// root context of current block`
- initialize root context h with multiplicity B
- **for** $s = 1$ **to** ℓ **do** `// build context tree up to depth` ℓ
  - **for** *each currently occupied context* c **do**
    - $\Pi(c) \leftarrow G(\mathsf{k}, c)$ `// keyed Poisson source for watermark`
    - generate $\nu(c)$ drafts by $Q(\cdot|c)$ using $\mathrm{MPFR}(Q(\cdot|c), \Pi(c), \cdot)$ `// Eq.(5)`
    - merge equal draft tokens into child contexts and update their multiplicities
- compute target-side sample $y(c) = \mathrm{MPFR}(P(\cdot|c), \Pi(c), 1)$ for every c `// in parallel`
- $c^\star \leftarrow h$; *accepted* $\leftarrow$ **true**
- **for** $s = 1$ **to** ℓ **do**
  - emit the target token $y(c^\star)$ and append it to the output
  - **if** $y(c^\star) \notin \mathcal{D}(c^\star)$ **:** `// $\mathcal{D}(c^\star)$ contains distinct draft tokens at $c$`
    - *accepted* $\leftarrow$ **false** and **break** `// first rejection ends the block`
  - **if** $n = N$ **:**
    - **break**
  - $c^\star \leftarrow c^\star \,\|\, y(c^\star)$ `// move to the next realized context`
- **if** *accepted* ***and*** $n < N$ **:**
  - emit one bonus token by $\mathrm{MPFR}(P(\cdot|c^\star), G(\mathsf{k}, c^\star), 1)$

(MPFR) selects samples with the smallest, second smallest, and so on up to the B-th smallest scores:

$$\tilde{Z}_P(1), \tilde{Z}_P(2), \ldots, \tilde{Z}_P(B) \overset{\text{iid}}{\sim} P, \tag{5}$$

where the exactness follows (Li and Anantharam, 2021); see Appendix A.3 for the formal definition and implementation.

For multi-draft speculative sampling, we use B i.i.d. drafts from a model $\mathcal{M}_d$, and at each sampling step we consider $P := \mathcal{M}_t(\cdot|x_{:t})$ as the target distribution and $Q := \mathcal{M}_d(\cdot|x_{:t})$ as the proposal distribution. Note that we consider *identical* proposals (Sun et al., 2023) in this section, but our framework can also be generalized to non-identically distributed proposals (Rowan et al., 2025), which is deferred to Appendix B.4.

Our scheme is sketched in Algorithm 4; see Algorithm 10 for details. The core idea is as follows. To implement a multi-draft algorithm, Algorithm 4 uses *context-indexed* mapped Poisson processes. Starting from the root prefix $h = w_{1:n}$, we build a depth-ℓ speculative tree, where each occupied context c is assigned a keyed source $\Pi(c)$ and generates $\nu(c)$ drafts from $Q(\cdot|c)$ via MPFR. We then compute the target-side sample $\mathrm{MPFR}(P, \Pi(c), 1)$ for every c and emit tokens by following the unique realized target path, continuing as long as the current target winner belongs to the draft set.

The multi-draft decoder is hence more than taking B order statistics independently at each position: the same context-indexed keyed source jointly couples the target winner and the B proposal samples over a branching autoregressive tree. The multiplicity bookkeeping and the unique target-realized path preserve target-distribution and stopping-time drafter invariance after earlier accept/reject decisions.

Importantly, the Poisson process is indexed by contexts rather than by drafter identities, and thus MPFR works in a way that makes the emitted tokens tightly tied to the prefix and the secret key, which helps watermark detection. In comparison, (Rowan et al., 2025) uses draft-indexed algorithms, keeps a set of currently viable draft indices, and chooses the next emitted token using only the active drafts. This is a natural implementation, but the active set makes the next emitted token depend on the realized draft sequences, causes the output to retain hidden dependence on draft-side trajectories, and hence only satisfies their conditional drafter invariance, but not Definition 4.3.

Theorem 5.1. *The acceptance probability of at least one token at the current step satisfies*

$$\mathbf{P}(\text{accept}) \geq 1 - \sum_{i \in \mathcal{V}} P(i)\big(P(i)/(P(i)+Q(i))\big)^B. \tag{6}$$

Table 1: Multi-draft results on Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct on CNN/DAILYMAIL and ELI5, $L = 4$. We report the average accepted tokens per step (AATPS) and token rate (TR). The better result between MPFR and GLS for each setting with the same B is in bold. Each cell is computed over the 1000 prompts and 3 different seeds. See more detailed data in Appendix D.4.

		Qwen-CNN/DM		Qwen-ELI5		Llama-CNN/DM		Llama-ELI5	
Decoder	B	AATPS	TR	AATPS	TR	AATPS	TR	AATPS	TR
MPFR	2	2.797	**25.11**	**2.563**	**23.57**	**3.459**	**37.44**	**3.103**	**34.29**
MPFR	4	**3.123**	**27.50**	**2.918**	**26.46**	**3.784**	**40.20**	**3.455**	**37.79**
MPFR	6	**3.294**	**28.50**	**3.109**	**27.81**	**3.935**	**41.71**	**3.635**	**39.59**
MPFR	8	**3.409**	**29.13**	**3.235**	**28.73**	**4.034**	**42.04**	**3.749**	**39.76**
GLS	2	**2.803**	24.78	2.523	22.73	3.431	36.79	3.080	33.72
GLS	4	3.109	27.12	2.850	25.60	3.741	39.31	3.408	37.11
GLS	6	3.272	27.84	3.039	27.01	3.894	40.51	3.588	39.12
GLS	8	3.384	28.62	3.156	28.09	3.987	39.87	3.700	39.37

The proof is in Appendix B, where we also give a generalized version that accounts for the number of active drafts, and demonstrate our advantages over the existing schemes (Rowan et al., 2025).

6 EXPERIMENTS

We evaluate whether our algorithms improve the empirical frontier between sampling efficiency and watermark detectability, while preserving generation quality. The headline configuration uses Qwen2.5-7B-Instruct as the target and Qwen2.5-0.5B-Instruct as the drafter on CNN/DAILYMAIL. Other model-dataset cells, full per-L and per-B tables, and ablation details are deferred to Appendix D. Here L is the maximum depth of a speculative block: the drafter proposes up to L future positions, and the target evaluates the corresponding contexts in parallel.

Baselines. We compare against four families of methods, all run in a common harness with the same prompts, decoding parameters, and watermark key. **(i) Watermark-only reference.** BASIC-UWM is autoregressive Gumbel-style unbiased watermarking without speculation (Aaronson and Kirchner, 2023); it sets the watermark-strength ceiling. **(ii) Speculation-only reference.** VSPS denotes vanilla speculative sampling without watermark (Leviathan et al., 2023), and the multi-draft GLS baseline (the drafter-invariant *InvariantMultiDraftStrategy*) of Rowan et al. (2025); these are speculation-only efficiency references. **(iii) Single-draft trade-off frontier.** MWS /MSE of Hu and Huang (2024) are the two endpoints of the single-draft no-go trade-off, and MSE-PSEUDO is the pseudorandom-r variant of He et al. (2026) that improves the trade-off via an alternative watermark metric. **(iv) Ours.** PFR (single-draft) and MPFR (multi-draft); PFR-NOWM is PFR with the keyed source replaced by an unkeyed Poisson source, isolating the cost of watermarking from the cost of Poisson coupling.

Metrics. Following (Hu and Huang, 2024; He et al., 2026) we separate three concerns. **Efficiency:** AATPS and token rate (TR); AATPS measures accepted output tokens per target-verification step, while TR is a hardware- and implementation-dependent throughput diagnostic. **Detectability:** ANLPPT, the average negative log p-value per token under three score variants U / Li / PL (Aaronson and Kirchner, 2023; Li et al., 2025; Lattimore, 2026) (see Appendix E); and TPR@1%FPR, calibrated on non-watermarked generations. **Quality:** per-token log-perplexity (LPPL) under the target model, used as an unbiasedness audit; we also report ROUGE-L when references are available.

Single-draft efficiency-watermark frontier. Figure 1 plotted every method in the (AATPS, ANLPPT-U) plane sweeping $L \in \{1, 2, 3, 4\}$. The prior speculative watermarking baselines (MWS, MSE, MSE-PSEUDO) exhibit the expected trade-off. As discussed in Section 4.3, the single-draft PFR curve should be interpreted as the Poisson realization of the same shared Gumbel/DeltaGumbel coupling underlying MWS, rather than as a separate way around the single-draft no-go result. The relevant ablation is that PFR and PFR-NOWM have nearly identical AATPS, showing that keying the shared source does not add an acceptance penalty within this coupling. Our new trade-off improvement is the $B > 1$ extension in Section 5. Other model/dataset results are tested in Appendix D.3.

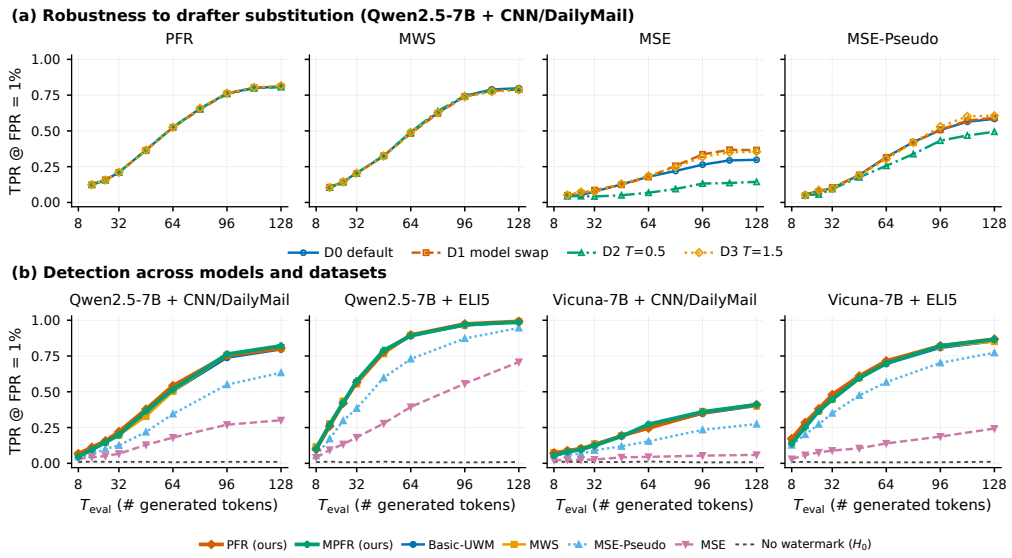


Figure 2: **Detection and robustness to drafter substitution.** **(a)** TPR@1%FPR under drafter substitution on Qwen2.5-7B + CNN/DAILYMAIL: the target model, watermark key, prompts, and target sampling parameters are fixed, and only the drafter is changed. PFR varies by <1 percentage point, whereas MSE varies by 12–22 points and MSE-PSEUDO by 6–11 points. **(b)** TPR@1%FPR versus detection token budget T_{eval} at the default drafter, across Qwen2.5-7B and Vicuna-7B on CNN/DAILYMAIL and ELI5. PFR and MPFR closely track the strong-watermark references BASIC-UWM and MWS, while MSE and MSE-PSEUDO are consistently weaker.

Multi-draft scaling. We compare MPFR with GLS (Rowan et al., 2025) at lookahead $L{=}4$ and $B \in \{2, 4, 6, 8\}$ across the four Qwen/Llama–dataset cells in Table 1. MPFR has higher measured token rate in all 16 matched settings and higher AATPS in 15 of 16; the only AATPS exception is Qwen–CNN/DailyMail at $B = 2$ (2.797 vs. 2.803). Meanwhile, the MPFR watermark signal is stable across B (ΔANLPPT ≤ 0.001), so the per-token detection signal does not dilute as additional drafts are added. Full per-cell numbers across the four (TARGET, DATASET) cells, all three ANLPPT variants, and the LPPL audit column are in Appendix D.4 (Tables 6–9).

Detection at a fixed false-positive rate. ANLPPT measures average evidence per token; TPR@1%FPR measures operational detectability under a calibrated false-positive budget. Figure 2(b) shows TPR@1%FPR versus token budget T_{eval} at the default drafter on Qwen2.5-7B and Vicuna-7B, each evaluated on CNN/DAILYMAIL and ELI5. PFR and MPFR consistently track the strong-watermark references BASIC-UWM and MWS, while MSE and MSE-PSEUDO are weaker across these four detection settings. MSE-PSEUDO improves over MSE but still falls below our methods, showing that the target-side keyed coupling retains strong detector-visible evidence.

Drafter-substitution ablation. We fix the target model, watermark key, prompts, and target sampling parameters, and vary only the drafter in model scale and temperature. As shown in Figure 2(a), PFR changes by less than 1 percentage point in TPR@1%FPR across the tested drafters, whereas MSE and MSE-PSEUDO vary by 12–22 and 6–11 points, respectively.

CONCLUDING REMARKS AND LIMITATIONS

We designed a drafter-invariant, multi-draft speculative sampling algorithm that embeds a distortion-free keyed watermark and improves the trade-off between watermark strength and sampling efficiency (Hu and Huang, 2024). In the reported Qwen/Llama 7–8B multi-draft settings, MPFR achieves higher measured token rate than GLS while maintaining strong watermark signal. Further reducing coupling, tree, and watermark-bookkeeping overhead is an important systems direction. Moreover, a fundamental characterization of the trade-off among target-draft communication, watermark strength, and speculative sampling efficiency remains not fully clear.

AI USE STATEMENT

Generative AI tools were used to assist with writing and language polishing, including grammar correction, as well as software code development and data cleaning. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content.

REFERENCES

Scott Aaronson and H Kirchner. Watermarking of large language models. In *Large Language Models and Transformers Workshop at Simons Institute for the Theory of Computing*, 2023.

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

Anthropic. How claude's text watermark works. `https://www.anthropic.com/news/claude-text-watermark`, August 2026.

Massieh Kordi Boroujeny, Ya Jiang, Kai Zeng, and Brian Mark. Multi-bit distortion-free watermarking for large language models. *arXiv preprint arXiv:2402.16578*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Marcello Bullo, Yanxiao Liu, Öykü Sıla Güner, Arpan Mukherjee, and Deniz Gündüz. Accelerating diffusion sampling via speculative draft trees. *arXiv preprint*, 2026. URL `https://arxiv.org/abs/2609.17691`.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1125–1139. PMLR, 2024.

Majid Daliri, Christopher Musco, and Ananda Theertha Suresh. Coupling without communication and drafter-invariant speculative decoding. In *2025 IEEE International Symposium on Information Theory (ISIT)*, pages 1–6. IEEE, 2025.

Valentin De Bortoli, Alexandre Galashov, Arthur Gretton, and Arnaud Doucet. Accelerated diffusion models via speculative sampling. *arXiv preprint arXiv:2501.05370*, 2025.

Frank Den Hollander. Probability theory: The coupling method. *Lecture notes available online (http://websites. math. leidenuniv. nl/probability/lecturenotes/CouplingLectures. pdf)*, 3, 2012.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 889–898, 2018.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. Eli5: Long form question answering. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3558–3567, 2019.

Jessica Ficler and Yoav Goldberg. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, 2017.

Gergely Flamich, Oykü Sıla Güner, Yanxiao Liu, and Deniz Gündüz. Scalable differentially private data compression via diffusion and stochastic codes. *arXiv preprint arXiv:2607.03392*, 2026.

Eva Giboulot and Teddy Furon. Watermax: breaking the llm watermark detectability-robustness-quality trade-off. *Advances in Neural Information Processing Systems*, 37:18848–18881, 2024.

Emil Julius Gumbel. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office, 1954.

Jiajun He, Gergely Flamich, and José Miguel Hernández-Lobato. Accelerating relative entropy coding with space partitioning. *Advances in Neural Information Processing Systems*, 37:75791–75828, 2024.

Weiqing He, Xiang Li, Tianqi Shang, Li Shen, Weijie Su, and Qi Long. On the empirical power of goodness-of-fit tests in watermark detection. *arXiv preprint arXiv:2510.03944*, 2025.

Weiqing He, Xiang Li, Li Shen, Weijie Su, and Qi Long. Improving the trade-off between watermark strength and speculative sampling efficiency for language models. In *International Conference on Learning Representations (ICLR)*, 2026.

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

Zhengmian Hu and Heng Huang. Inevitable trade-off between watermark strength and speculative sampling efficiency for language models. *Advances in Neural Information Processing Systems*, 37: 55370–55402, 2024.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Ashish Khisti, Arash Behboodi, Gabriele Cesa, and Pratik Kumar. Unequal message protection: One-shot analysis via poisson matching lemma. In *2024 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2024.

Ashish Khisti, M Reza Ebrahimi, Hassan Dbouk, Arash Behboodi, Roland Memisevic, and Christos Louizos. Multi-draft speculative sampling: Canonical decomposition and theoretical limits. In *International Conference on Learning Representations (ICLR)*, 2025.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International conference on machine learning*, pages 17061–17084. PMLR, 2023.

Szymon Kobus and Deniz Gündüz. Speculative sampling via exponential races. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18189–18204, 2025.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.

Günter Last and Mathew Penrose. *Lectures on the Poisson process*, volume 7. Cambridge University Press, 2017.

Tor Lattimore. Refined detection for gumbel watermarking. *arXiv preprint arXiv:2603.30017*, 2026.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning (ICML)*, pages 19274–19286. PMLR, 2023.

Cheuk Ting Li and Venkat Anantharam. A unified framework for one-shot achievability via the Poisson matching lemma. *IEEE Transactions on Information Theory*, 67(5):2624–2651, 2021.

Cheuk Ting Li and Abbas El Gamal. Strong functional representation lemma and applications to coding theorems. *IEEE Transactions on Information Theory*, 64(11):6967–6978, 2018.

Cheuk Ting Li et al. Channel simulation: Theory and applications to lossy compression and differential privacy. *Foundations and Trends® in Communications and Information Theory*, 21(6): 847–1106, 2024.

Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *The Annals of Statistics*, 53(1):322–351, 2025.

Yanxiao Liu and Cheuk Ting Li. One-shot information hiding. In *2024 IEEE Information Theory Workshop (ITW)*, pages 169–174. IEEE, 2024.

Yanxiao Liu and Cheuk Ting Li. One-shot coding over general noisy networks. *IEEE Transactions on Information Theory*, 71(11):8346–8357, 2025.

Yanxiao Liu, Wei-Ning Chen, Ayfer Özgür, and Cheuk Ting Li. Universal exact compression of differentially private mechanisms. *Advances in Neural Information Processing Systems*, 37: 91492–91531, 2024.

Yanxiao Liu, Sepehr Heidari Advary, and Cheuk Ting Li. Nonasymptotic oblivious relaying and variable-length noisy lossy source coding. *IEEE Transactions on Information Theory*, 72(6): 4555–4564, 2026.

Chris J Maddison. A Poisson process model for Monte Carlo. *Perturbation, Optimization, and Statistics*, pages 193–232, 2016.

Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. *Advances in Neural Information Processing Systems*, 27, 2014.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Zhengxin Zhang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, et al. Specinfer: Accelerating large language model serving with tree-based speculative inference and verification. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, pages 932–949, 2024.

Pierre Moulin and Joseph A O'Sullivan. Information-theoretic analysis of information hiding. *IEEE Transactions on information theory*, 49(3):563–593, 2003.

Shinwoo Park, Hyejin Park, Hyeseon Ahn, and Yo-Sub Han. Watermod: Modular token-rank partitioning for probability-balanced llm watermarking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32683–32691, 2026.

Wenjie Qu, Wengrui Zheng, Tianyang Tao, Dong Yin, Yanze Jiang, Zhihua Tian, Wei Zou, Jinyuan Jia, and Jiaheng Zhang. Provably robust multi-bit watermarking for {AI-generated} text. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 201–220, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Joseph Rowan, Buu Phan, and Ashish Khisti. List-level distribution coupling with applications to speculative decoding and lossy compression. *Advances in Neural Information Processing Systems*, 38, 2025.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. Blockwise parallel decoding for deep autoregressive models. *Advances in Neural Information Processing Systems*, 31, 2018.

Ziteng Sun, Ananda Theertha Suresh, Jae Hun Ro, Ahmad Beirami, Himanshu Jain, and Felix Yu. Spectr: Fast speculative decoding via optimal transport. *Advances in Neural Information Processing Systems*, 36:30222–30242, 2023.

Lucas Theis and Noureldin Y Ahmed. Algorithms for the communication of samples. In *International Conference on Machine Learning (ICML)*, pages 21308–21328. PMLR, 2022.

Hermann Thorisson. Coupling, stationarity, and regeneration. *Probability and its Applications*, 2000.

Dor Tsur, Carol Xuan Long, Claudio Mayrink Verdun, Hsiang Hsu, Haim Permuter, and Flavio P Calmon. Optimized couplings for watermarking large language models. *arXiv preprint arXiv:2505.08878*, 2025.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338, 2025.

Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. A resilient and accessible distribution-preserving watermark for large language models. *arXiv preprint arXiv:2310.07710*, 2023.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*, 2023.

Xuandong Zhao, Lei Li, and Yu-Xiang Wang. Permute-and-flip: An optimally stable and watermarkable decoder for LLMs. In *International Conference on Learning Representations, (ICLR)*, 2025a.

Xuandong Zhao, Chenwen Liao, Yu-Xiang Wang, and Lei Li. Efficiently identifying watermarked segments in mixed-source texts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6304–6316, 2025b.

Chuqin Zhou, Xiaoyue Ling, Yunuo Chen, Jincheng Dai, Guo Lu, and Wenjun Zhang. Dual-representation image compression at ultra-low bitrates via explicit semantics and implicit textures. *arXiv preprint arXiv:2602.05213*, 2026.

A More on PFR and Related Works

A.1 PFR for Discrete Distributions

It is known that the general PFR (Li and El Gamal, 2018; Li and Anantharam, 2021) reduces to the use of exponential random variables (Li and El Gamal, 2018; Liu and Li, 2025) in the case of simulating discrete distributions, which is similar to the Gumbel-max trick (Gumbel, 1954) used for speculative sampling (Kobus and Gündüz, 2025).

In (Li et al., 2024), it is discussed how to extend the discrete case to the general case. However, the other direction, namely reducing the general case to the discrete case, has not been discussed in the literature, to the best of the authors' knowledge. To make the paper self-contained, we provide formal statements and proofs for this direction.

The first define the exponential functional representation (EFR) as follows (also see (Li and El Gamal, 2018; Liu and Li, 2025)). Let $\mathcal{Z}$ be a finite or countably infinite set, and let P be a distribution on $\mathcal{Z}$. Let $(E_z)_{z\in\mathcal{Z}}$ be independent $\mathrm{Exp}(1)$ random variables. With the convention that $E_z/P(z) = +\infty$ whenever $P(z) = 0$, define

$$Z_{\mathrm{EFR}} := \arg\min_{z\in\mathcal{Z}} \frac{E_z}{P(z)}. \tag{7}$$

As proved in Appendix A.2, EFR is equivalent to the Gumbel-max trick. In the following, we prove that PFR reduces to EFR in the discrete case, i.e., when the target and proposal distributions are discrete.

Theorem A.1. *Let $\mathcal{Z}$ be a finite or countably infinite set. Let Q and P be distributions on $\mathcal{Z}$ such that $P \ll Q$. Write $q_z := Q(z), p_z := P(z), z \in \mathcal{Z}$. The PFR considers a Poisson process $(Z_i, T_i)_{i\geq 1}$ on $\mathcal{Z}\times[0,\infty)$ with intensity measure $Q\times\lambda_{[0,\infty)}$, and $\tilde{T}_i = T_i\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(Z_i)\right)^{-1}$. For each z with $q_z > 0$, define the first arrival time of symbol z by*

$$S_z := \inf\{T_i : Z_i = z\}, \tag{8}$$

and define

$$E_z := q_z S_z. \tag{9}$$

For z with $q_z = 0$, define E_z to be an arbitrary independent $\mathrm{Exp}(1)$ random variable, independent of everything else. Then the random variables $(E_z)_{z\in\mathcal{Z}}$ are independent $\mathrm{Exp}(1)$ random variables, and the PFR output satisfies

$$Z_{\mathrm{PFR}} = \arg\min_i \tilde{T}_i = \arg\min_{z\in\mathcal{Z}} \frac{E_z}{p_z} = Z_{\mathrm{EFR}} \qquad \textit{a.s.}, \tag{10}$$

where we use the convention that $E_z/p_z = +\infty$ if $p_z = 0$.

Proof. For $z \in \mathcal{Z}$ and $t \geq 0$, define

$$N_z(t) := \#\{i : T_i \leq t,\ Z_i = z\}.$$

Since $(Z_i, T_i)_{i\geq 1}$ is a Poisson process with intensity measure $Q\times\lambda_{[0,\infty)}$, for each z with $q_z > 0$, the process $(N_z(t))_{t\geq 0}$ is a Poisson process with rate q_z. Moreover, for distinct symbols $z_1, \ldots, z_m$, the processes

$$(N_{z_1}(t))_{t\geq 0}, \ldots, (N_{z_m}(t))_{t\geq 0}$$

are independent, because they count disjoint subsets of the underlying Poisson process.

Therefore, for z with $q_z > 0$, the first arrival time

$$S_z = \inf\{t : N_z(t) \geq 1\}$$

satisfies

$$\mathbf{P}(S_z > s) = \mathbf{P}(N_z(s) = 0) = e^{-q_z s}, \qquad s \geq 0.$$

Thus $S_z \sim \mathrm{Exp}(q_z)$. Hence

$$E_z = q_z S_z \sim \mathrm{Exp}(1).$$

The independence of the Poisson counting processes for different symbols implies that the random variables

$$(E_z)_{z:q_z>0}$$

are independent. For symbols with $q_z = 0$, we have $p_z = 0$ because $P \ll Q$, and these symbols can never be selected by PFR. Defining independent $\mathrm{Exp}(1)$ variables E_z for those symbols therefore does not change the output. Hence the full family $(E_z)_{z\in\mathcal{Z}}$ is independent and identically distributed as $\mathrm{Exp}(1)$.

It remains to show that the PFR selection rule reduces to the EFR selection rule. Since $P \ll Q$, for every z with $p_z > 0$, we also have $q_z > 0$, and

$$\frac{\mathrm{d}P}{\mathrm{d}Q}(z) = \frac{p_z}{q_z}.$$

Therefore, for every index i with $Z_i = z$ and $p_z > 0$,

$$\tilde{T}_i = T_i \left(\frac{p_z}{q_z}\right)^{-1} = T_i \frac{q_z}{p_z}.$$

Taking the minimum over all Poisson points with mark z, we get

$$\inf_{i:Z_i=z} \tilde{T}_i = \left(\inf_{i:Z_i=z} T_i\right) \frac{q_z}{p_z} = S_z \frac{q_z}{p_z} = \frac{E_z}{p_z}.$$

If $p_z = 0$, then $\mathrm{d}P/\mathrm{d}Q(z) = 0$ whenever $q_z > 0$, and hence

$$\tilde{T}_i = T_i \left(\frac{\mathrm{d}P}{\mathrm{d}Q}(z)\right)^{-1} = +\infty$$

for every Poisson point with mark z. This agrees with the EFR convention

$$\frac{E_z}{p_z} = +\infty.$$

If $q_z = 0$, then also $p_z = 0$, and no Poisson point has mark z almost surely, so such a symbol is irrelevant in both constructions.

Thus,

$$\inf_i \tilde{T}_i = \inf_{z\in\mathcal{Z}} \inf_{i:Z_i=z} \tilde{T}_i = \inf_{z\in\mathcal{Z}} \frac{E_z}{p_z}.$$

Moreover, the minimizer is unique almost surely. Indeed, for every z with $p_z > 0$,

$$\frac{E_z}{p_z} \sim \mathrm{Exp}(p_z),$$

and these variables are independent. Since exponential random variables are continuous, ties have probability zero. Therefore, almost surely,

$$Z_{\mathrm{PFR}} = Z_K = \arg\min_{z\in\mathcal{Z}} \frac{E_z}{p_z} = Z_{\mathrm{EFR}}.$$

□

A.2 Relation between PFR and Gumbel-Max Trick

As discussed in (Zhao et al., 2025a), the Gumbel Watermark (Aaronson and Kirchner, 2023) is based on

$$y_t = \mathrm{argmin}_{y\in\mathcal{V}} \frac{u_t(y)}{T} + G_t(y)$$

where $G(y) \sim \text{Gumbel}(0,1)$ i.i.d. for each t and y. Here, we show that it is equivalent to the PFR in the discrete case. Recall $\text{Gumbel}(0,1) \sim -\log\big(\log(1/\text{Unif}[0,1])\big)$. We calculate

$$\begin{aligned}
y_t &= \arg\max_{y\in\mathcal{V}} \left(\frac{u_t(y)}{T} + G_t(y)\right) \\
&= \arg\max_{y\in\mathcal{V}} \left(\frac{u_t(y)}{T} - \log Z_t(y)\right) \qquad \Big(Z_t(y) \overset{\text{iid}}{\sim} \text{Exp}(1),\ G_t(y) \overset{d}{=} -\log Z_t(y)\Big) \\
&= \arg\max_{y\in\mathcal{V}} \Big(-\log\big(Z_t(y)e^{-u_t(y)/T}\big)\Big) \\
&= \arg\min_{y\in\mathcal{V}} Z_t(y)e^{-u_t(y)/T} \\
&= \arg\min_{y\in\mathcal{V}} \frac{Z_t(y)}{\exp(u_t(y)/T)} \\
&= \arg\min_{y\in\mathcal{V}} \frac{Z_t(y)}{w_t(y)} \qquad \Big(w_t(y) := \exp(u_t(y)/T)\Big) \\
&= \arg\min_{y\in\mathcal{V}} \frac{Z_t(y)}{P_t(y)},
\end{aligned}$$

where the last equality is because $P_t(y) := \frac{w_t(y)}{\sum_{y'\in\mathcal{V}} w_t(y')}$ and scaling by a positive constant doesn't change the argmin.

A.3 Mapped Poisson Process

A.3.1 Definitions

Definition A.2 (Mapped Poisson Process (Li and Anantharam, 2021))**.** Let $(T_i)_i$ be a Poisson process of rate 1, independent of $Z_i \overset{\text{iid}}{\sim} Q$ for $i = 1, 2, \ldots$. Let $i_{P,1}, i_{P,2}, \ldots \in \mathbb{N}$ be a sequence of distinct integers such that $\bigcup_{j=1}^{\infty}\{i_{P,j}\} = \{i : \frac{\mathrm{d}P}{\mathrm{d}Q}(Z_i) > 0\}$ and $\{T_{i_{P,j}}(\frac{\mathrm{d}P}{\mathrm{d}Q}(Z_{i_{P,j}}))^{-1}\}_{j\in\mathbb{N}}$ is sorted in ascending order with arbitrary tie-breaking.[3] For $j \in \mathbb{N}$ and suppose P is the target distribution, we denote

$$\left\{\tilde{Z}_P(j), \tilde{T}(j)\right\}_{j\in\mathbb{N}} := \left\{Z_{i_{P,j}}, T_{i_{P,j}}(\frac{\mathrm{d}P}{\mathrm{d}Q}(Z_{i_{P,j}}))^{-1}\right\}_{j\in\mathbb{N}}. \tag{11}$$

Give a positive integer B, the mapped Poisson Process selects the first B samples in $\left\{\tilde{Z}_P(j), \tilde{T}(j)\right\}_{j\in\mathbb{N}}$.

Proposition A.3. *By the mapping theorem (Last and Penrose, 2017), $\left\{\tilde{Z}_P(j), \tilde{T}(j)\right\}_{j\in\mathbb{N}}$ is a Poisson process with intensity measure $P \times \lambda_{\mathbb{R}_{\geq 0}}$ where $\mathbb{R}_{\geq 0}$ to denote nonnegative real numbers, and therefore we have*

$$\tilde{Z}_P(1), \tilde{Z}_P(2), \ldots \overset{\text{iid}}{\sim} P. \tag{12}$$

A.3.2 Direct Implementations

We can implement the MPFR just like (Theis and Ahmed, 2022) for PFR, as follows:

[3] Note a tie occurs with probability 0.

Algorithm 5: Finite implementation of the B-sample mapped PFR

Given : target distribution P, proposal distribution Q, likelihood ratio $r = dP/dQ$, number of samples B, and a constant $\underline{w} > 0$ such that $1/r(u) \geq \underline{w}$ Q-a.s. on $\{r(u) > 0\}$

Return : B samples from P

$H \leftarrow \emptyset$ `// H stores the current B smallest-score candidates`

$t \leftarrow 0$; $i \leftarrow 1$

while $|H| < B$ ***or*** $\max_{(s,j,z)\in H} s > t\underline{w}$ **do**

  Sample $E_i \sim \mathrm{Exp}(1)$ and set $t \leftarrow t + E_i$ `// Poisson process`

  Sample $Z_i \sim Q$

  if $r(Z_i) > 0$ **:**

    $S_i \leftarrow t/r(Z_i)$

  else

    $S_i \leftarrow +\infty$

  if $|H| < B$ **:**

    Insert (S_i, i, Z_i) into H

  else if $S_i < \max_{(s,j,z)\in H} s$ **:**

    Remove an element $(s^\star, j^\star, z^\star) \in H$ satisfying $s^\star = \max_{(s,j,z)\in H} s$

    Insert (S_i, i, Z_i) into H

  $i \leftarrow i + 1$

Sort the elements of H as

$$(S_{i_1}, i_1, Z_{i_1}), \ldots, (S_{i_B}, i_B, Z_{i_B})$$

so that

$$S_{i_1} \leq S_{i_2} \leq \cdots \leq S_{i_B}.$$

return $(Z_{i_1}, \ldots, Z_{i_B})$

A.3.3 Proof of Correctness

We then prove that the implementation of Algorithm 5 is correct.

Theorem A.4. *Let $P \ll Q$ be probability measures on a Polish space $\mathcal{U}$, and let $r := dP/dQ$. Let $Z_1, Z_2, \ldots \overset{iid}{\sim} Q$ and let $0 < T_1 < T_2 < \cdots$ be the arrival times of a unit-rate Poisson process, independent of $(Z_i)_{i\geq 1}$. Define*

$$S_i := \begin{cases} T_i/r(Z_i), & r(Z_i) > 0, \\ +\infty, & r(Z_i) = 0. \end{cases}$$

Assume that there exists $\underline{w} > 0$ such that

$$1/r(u) \geq \underline{w} \qquad Q\text{-a.s. on } \{r(u) > 0\}.$$

Fix $B \in \mathbb{N}$. Consider the algorithm that scans the candidates in the order $i = 1, 2, \ldots$, keeps the B smallest observed scores, and stops at the first time n such that the current B-th smallest observed score is at most $T_n\underline{w}$. Then the algorithm terminates almost surely after finitely many iterations. Moreover, at termination, the retained candidates are exactly the B smallest-score candidates among the infinite sequence. Consequently, the returned marks are distributed as

$$(\widetilde{U}_P(1), \ldots, \widetilde{U}_P(B)),$$

the first B points of the mapped Poisson process with respect to P, and hence are i.i.d. according to P.

Proof. Let

$$\Pi_Q := \{(Z_i, T_i) : i \geq 1\}.$$

Since $0 < T_1 < T_2 < \cdots$ are the arrival times of a unit-rate Poisson process and $Z_1, Z_2, \ldots \overset{iid}{\sim} Q$ are independent marks, Π_Q is a Poisson point process on $\mathcal{U} \times \mathbb{R}_{\geq 0}$ with intensity measure $Q \times \lambda$, where λ denotes the Lebesgue measure on $\mathbb{R}_{\geq 0}$.

Consider the measurable mapping

$$\Phi : \mathcal{U} \times \mathbb{R}_{\geq 0} \to \mathcal{U} \times \mathbb{R}_{\geq 0}, \qquad \Phi(u, t) = \left(u, \frac{t}{r(u)}\right)$$

on the set $\{u : r(u) > 0\}$. Points with $r(u) = 0$ are assigned score $+\infty$ and hence never appear among the finite mapped arrival times. For any nonnegative measurable function g on $\mathcal{U} \times \mathbb{R}_{\geq 0}$, we have

$$\begin{aligned}\int_{\mathcal{U}} \int_0^\infty g\left(u, \frac{t}{r(u)}\right) dt\, Q(du) &= \int_{\{r>0\}} \int_0^\infty g\left(u, \frac{t}{r(u)}\right) dt\, Q(du) \\ &= \int_{\{r>0\}} \int_0^\infty g(u,s) r(u)\, ds\, Q(du) \\ &= \int_{\mathcal{U}} \int_0^\infty g(u,s)\, ds\, P(du).\end{aligned}$$

Therefore, by the mapping theorem for Poisson point processes, the mapped point process

$$\Pi_P := \{(Z_i, S_i) : i \geq 1,\ r(Z_i) > 0\}, \qquad S_i := \frac{T_i}{r(Z_i)},$$

is a Poisson point process on $\mathcal{U} \times \mathbb{R}_{\geq 0}$ with intensity measure $P \times \lambda$.

Let

$$S_{i_1} < S_{i_2} < \cdots$$

be the increasing ordering of the finite scores in Π_P. Since the time-marginal intensity of Π_P is λ, the ordered times $S_{i_1}, S_{i_2}, \ldots$ are the arrival times of a unit-rate Poisson process. Moreover, conditional on these arrival times, the associated marks are independent with common distribution P. Hence

$$(Z_{i_1}, Z_{i_2}, \ldots) \stackrel{d}{=} (U_1, U_2, \ldots),$$

where $U_1, U_2, \ldots \overset{iid}{\sim} P$. In particular,

$$(Z_{i_1}, \ldots, Z_{i_B})$$

are exactly the first B points of the mapped Poisson process with respect to P, and they are i.i.d. according to P.

It remains to show that the finite algorithm stops safely and terminates almost surely. After observing candidates $1, \ldots, n$, let

$$S_{n,(B)}$$

denote the B-th smallest value among $S_1, \ldots, S_n$, whenever at least B finite scores have been observed. By construction, the algorithm keeps the B smallest observed scores, so the largest score retained by the algorithm is exactly $S_{n,(B)}$.

Suppose the algorithm stops at time n. Then

$$S_{n,(B)} \leq T_n \underline{w}.$$

For every future candidate $m > n$, we have $T_m > T_n$. If $r(Z_m) = 0$, then $S_m = +\infty$, so such a candidate cannot improve the current list. If $r(Z_m) > 0$, then by the assumption $1/r(u) \geq \underline{w}$ Q-a.s.,

$$S_m = \frac{T_m}{r(Z_m)} = T_m \frac{1}{r(Z_m)} \geq T_m \underline{w} > T_n \underline{w} \geq S_{n,(B)}.$$

Thus every future score is strictly larger than the current B-th smallest observed score. Therefore no future candidate can enter the set of the global B smallest scores. Hence, when the algorithm stops, the retained candidates are exactly the global B smallest-score candidates among the infinite sequence.

Finally, we prove almost sure finite termination. Since Π_P is a Poisson point process with intensity $P \times \lambda$, its first B mapped arrival times are finite almost surely. Let

$$S_{i_B}$$

denote the B-th smallest mapped score, and let

$$N_0 := \max\{i_1, \ldots, i_B\}.$$

Then $N_0 < \infty$ almost surely. After the algorithm has processed candidates $1, \ldots, N_0$, it has already observed the true global first B mapped candidates, so the largest retained score is S_{i_B}. Since

$$T_n \to \infty \qquad \text{almost surely},$$

and since $\underline{w} > 0$, there exists an almost surely finite random index $N \geq N_0$ such that

$$T_N \underline{w} \geq S_{i_B}.$$

At time N, the algorithm therefore satisfies the stopping condition

$$S_{N,(B)} = S_{i_B} \leq T_N \underline{w}.$$

Thus the algorithm terminates almost surely after finitely many iterations.

Combining the safe stopping argument with the mapped Poisson process argument, the algorithm returns exactly

$$(Z_{i_1}, \ldots, Z_{i_B}),$$

which are the first B points of the mapped Poisson process with respect to P. Consequently, the returned samples are i.i.d. according to P. □

A.4 A BETTER FINITE-SUPPORT MPFR IMPLEMENTATION

The proposal-scanning implementation in Appendix A.3.2 can be slow in finite-vocabulary LLM decoding. Its stopping rule depends on a lower bound on the likelihood ratio, which can be very small when a rare token has small proposal probability but non-negligible target probability. In our experiments, the logits are processed by temperature and top-k truncation, so the resulting distribution has finite support of size at most K. In this setting, we can simulate the relevant Poisson clocks directly and avoid proposal scanning altogether.

Let P be a distribution on a finite vocabulary $\mathcal{U}$, and define

$$\mathcal{A}_P := \{u \in \mathcal{U} : P(u) > 0\}.$$

For each $u \in \mathcal{A}_P$, let

$$G_{u,j} := \sum_{m=1}^{j} E_{u,m}, \qquad j \geq 1,$$

where the $E_{u,m}$'s are independent $\mathrm{Exp}(1)$ random variables. We associate to (u, j) the mapped-Poisson score

$$S_{u,j} := \frac{G_{u,j}}{P(u)}.$$

The first B MPFR samples are the labels of the B smallest scores among

$$\{S_{u,j} : u \in \mathcal{A}_P,\ j \geq 1\}.$$

Since no token can contribute more than B arrivals among the first B total arrivals, it is enough to generate $j = 1, \ldots, B$ for every $u \in \mathcal{A}_P$.

Theorem A.5 (Exactness of direct finite-support MPFR)**.** *Let P be a distribution on a finite set $\mathcal{U}$. The output $(U_{(1)}, \ldots, U_{(B)})$ of Algorithm 6 consists of B independent samples from P. That is, for every $(u_1, \ldots, u_B) \in \mathcal{U}^B$,*

$$\mathbf{P}\{U_{(1)} = u_1, \ldots, U_{(B)} = u_B\} = \prod_{b=1}^{B} P(u_b).$$

Proof. For each $u \in \mathcal{A}_P$, the sequence $(G_{u,j})_{j \geq 1}$ is the arrival process of a unit-rate Poisson process. Therefore,

$$\left(\frac{G_{u,j}}{P(u)}\right)_{j \geq 1}$$

is a Poisson process on $\mathbb{R}_+$ with rate $P(u)$, since

$$\#\left\{j : \frac{G_{u,j}}{P(u)} \leq t\right\} = \#\{j : G_{u,j} \leq P(u)t\}$$

is Poisson with mean $P(u)t$. The processes corresponding to different tokens are independent. Their superposition is therefore a marked Poisson process on $\mathbb{R}_+ \times \mathcal{U}$ with intensity measure

$$dt \otimes P.$$

Algorithm 6: Direct finite-support implementation of B-sample MPFR

Given : finite distribution P on $\mathcal{U}$, number of samples B, and keyed source Π
Return : B samples from P
$\mathcal{A}_P \leftarrow \{u \in \mathcal{U} : P(u) > 0\}$
$H \leftarrow \emptyset$ `// candidate triples` (S, u, j)
for *each* $u \in \mathcal{A}_P$ **do**
  $G \leftarrow 0$
  for $j = 1$ **to** B **do**
    Generate $E_{u,j} \sim \mathrm{Exp}(1)$ from Π
    $G \leftarrow G + E_{u,j}$
    $S \leftarrow G/P(u)$
    Insert (S, u, j) into H
Let

$$(S_{(1)}, U_{(1)}, J_{(1)}), \ldots, (S_{(B)}, U_{(B)}, J_{(B)})$$

be the B elements of H with smallest scores, ordered by increasing score
return $(U_{(1)}, \ldots, U_{(B)})$

Equivalently, the total arrival rate is $\sum_{u \in \mathcal{A}_P} P(u) = 1$, and the marks of the ordered arrivals are independent with common distribution P.

Algorithm 6 constructs the first B candidate arrivals of each token clock. This is sufficient because among the first B arrivals of the superposed process, no individual token clock can contribute more than B arrivals. Hence the B smallest elements in H are exactly the first B arrivals of the superposed marked Poisson process, and their labels are B independent samples from P. □

We now describe the batched speculative decoding implementation. For a context c, let $P_k(\cdot|c)$ and $Q_k(\cdot|c)$ denote the processed target and draft distributions after applying the same logits processing, such as temperature and top-k truncation. The draft side builds a tree of candidate continuations using Algorithm 6 with Q_k. The target side then evaluates all contexts needed for verification in a batched manner and uses the same MPFR clocks with P_k to follow the realized target path. This batching changes only the order and grouping of model evaluations, not the sampling law.

Theorem A.6 (Exactness of batched-target direct-top-k MPFR decoding)**.** *Assume that for every context c, the keyed source $\Pi_{\mathsf{k}}(c)$ provides independent exponential clocks across tokens and local arrival indices, and that the sources $\{\Pi_{\mathsf{k}}(c)\}_c$ are independent across distinct contexts. Then Algorithm 7 generates exactly from the autoregressive model with transition kernel $P_k(\cdot|c)$. That is, for every finite sequence $u_1, \ldots, u_m$,*

$$\mathbf{P}\{W_{n+1} = u_1, \ldots, W_{n+m} = u_m\} = \prod_{t=1}^{m} P_k(u_t | w_{1:n}, u_1, \ldots, u_{t-1}).$$

Proof. Fix a context c. By Theorem A.5,

$$Y(c) = \mathrm{MPFR}(P_k(\cdot|c), \Pi_{\mathsf{k}}(c), 1)$$

has distribution $P_k(\cdot|c)$. The draft samples generated from $Q_k(\cdot|c)$ may use the same keyed clocks, and hence are coupled with $Y(c)$, but this does not change the marginal distribution of $Y(c)$.

Let

$$C_t := (w_{1:n}, W_{n+1}, \ldots, W_{n+t-1})$$

be the random context before generating W_{n+t}. The event $\{C_t = c\}$ is determined by the keyed sources associated with strict prefix contexts of c. By the assumed independence across contexts, $\Pi_{\mathsf{k}}(c)$ is independent of this event. Therefore, conditional on $C_t = c$,

$$W_{n+t} = Y(c) \sim P_k(\cdot|c),$$

and hence

$$\mathbf{P}\{W_{n+t} = u | C_t = c\} = P_k(u|c).$$

Algorithm 7: Batched-target direct-top-k MPFR speculative sampling

Given : lookahead L, number of drafts B, output length N, target model P, draft model Q, prompt $w_{1:n}$, key k

while $n < N$ **do**

$h \leftarrow w_{1:n}$, $\ell \leftarrow \min\{L, N-n\}$

$\mathcal{C}_0 \leftarrow \{h\}$ and $\nu(h) \leftarrow B$ `// context multiplicities`

`// Draft tree construction`

for $s = 1$ **to** ℓ **do**

$\mathcal{C}_s \leftarrow \emptyset$

for *each* $c \in \mathcal{C}_{s-1}$ **do**

Generate

$$(X_1(c), \ldots, X_{\nu(c)}(c)) \leftarrow \text{MPFR}(Q_k(\cdot|c), \Pi_{\mathsf{k}}(c), \nu(c)).$$

Let $\mathcal{D}(c)$ be the set of distinct tokens among these samples

for *each* $u \in \mathcal{D}(c)$ **do**

$m(c,u) \leftarrow |\{i : X_i(c) = u\}|$

Add $c\|u$ to $\mathcal{C}_s$

$\nu(c\|u) \leftarrow \nu(c\|u) + m(c,u)$

`// Batched target evaluation`

Compute the processed target distributions $P_k(\cdot|c)$ for all contexts

$$c \in \bigcup_{s=0}^{\ell} \mathcal{C}_s$$

using batched target-model forward passes

for *each such context* c **do**

Generate

$$Y(c) \leftarrow \text{MPFR}(P_k(\cdot|c), \Pi_{\mathsf{k}}(c), 1).$$

`// Follow the realized target path`

$c^\star \leftarrow h$, accepted $\leftarrow$ **true**

for $s = 1$ **to** ℓ **do**

Emit $Y(c^\star)$ and append it to the output

if $Y(c^\star) \notin \mathcal{D}(c^\star)$ **:**

accepted $\leftarrow$ **false**

break `// first rejection ends the block`

if $n = N$ **:**

break

$c^\star \leftarrow c^\star\|Y(c^\star)$

if accepted ***and*** $n < N$ **:**

Emit the precomputed bonus token $Y(c^\star)$

Applying the chain rule gives

$$\begin{aligned}&\mathbf{P}\{W_{n+1}=u_1,\dots,W_{n+m}=u_m\}\\&\quad=\prod_{t=1}^{m}\mathbf{P}\{W_{n+t}=u_t|W_{n+1}=u_1,\dots,W_{n+t-1}=u_{t-1}\}\\&\quad=\prod_{t=1}^{m}P_k(u_t|w_{1:n},u_1,\dots,u_{t-1}).\end{aligned}$$

Thus the generated sequence has exactly the desired autoregressive law. Finally, batched target evaluation only changes the computational order of deterministic model evaluations. It does not change the context-indexed MPFR clocks or the resulting samples $Y(c)$, so it does not change the output distribution. □

This implementation removes the proposal-scanning overhead by simulating the token-wise Poisson clocks directly. For support size at most K, producing B MPFR samples requires KB exponential variables followed by a top-B selection. The batched target implementation further reduces the number of target-model calls: instead of verifying the realized path sequentially with roughly one target forward pass per generated token, it evaluates the contexts in a speculative block together and extracts all logits needed for verification. If a block emits BE tokens on average, the target-forward cost per emitted token is reduced from approximately 1 to approximately $1/\mathrm{BE}$, while preserving the exact target marginal distribution.

B Theoretic Guarantees

B.1 Proof of Theorem 5.1

Proof. The proof is based on the Poisson matching lemma (Li and Anantharam, 2021). Let $p_i := P(i)$ and $q_i := Q(i)$ for $i \in \mathcal{V}$. Let $Y := \tilde{Z}_P(1)$ and $X^{(b)} := \tilde{Z}_Q(b)$ for $b = 1, \dots, B$ be generated from the same mapped Poisson process, and define the token-level acceptance event

$$\mathcal{A}_t := \left\{ Y \in \{X^{(1)}, \dots, X^{(B)}\} \right\},$$

which is the acceptance event. Apply the (Li and Anantharam, 2021, Lemma 3) with $j = 1$ and $k = B$. Conditioned on $Y = \tilde{Z}_P(1) = i$, the event $\mathcal{A}_t^c$ is exactly the event that the first mapped P-sample does not appear among the first B mapped Q-samples. Hence

$$\mathbf{P}(\mathcal{A}_t^c|Y=i) \le \left(1 - \left(1 + \frac{p_i}{q_i}\right)^{-1}\right)^B = \left(\frac{p_i}{p_i+q_i}\right)^B.$$

Taking complements gives

$$\mathbf{P}(\mathcal{A}_t|Y=i) \ge 1 - \left(\frac{p_i}{p_i+q_i}\right)^B.$$

Finally, average over $Y \sim P$:

$$\mathbf{P}(\mathcal{A}_t) = \sum_{i\in\mathcal{V}} p_i \mathbf{P}(\mathcal{A}_t|Y=i) \ge \sum_{i\in\mathcal{V}} p_i \left[1 - \left(\frac{p_i}{p_i+q_i}\right)^B\right].$$

For $B = 1$, this simplifies to

$$1 - \sum_{i\in\mathcal{V}} \frac{p_i^2}{p_i+q_i} = \sum_{i\in\mathcal{V}} \frac{p_i q_i}{p_i+q_i}.$$ □

B.2 Comparison with (Rowan et al., 2025) on Acceptance Probability

We compare our token-level acceptance probability with the token-level acceptance probability from (Rowan et al., 2025) in two levels, and we prove that in each level our algorithm's acceptance probability is strictly better.

B.2.1

We first prove a stronger version of Theorem 5.1, but not as clean as it.

Theorem B.1. *Let P be the target distribution and Q be the draft distribution on a finite vocabulary $\mathcal{V}$. Assume first that $P(i), Q(i) > 0$ for all $i \in \mathcal{V}$. For each $i \in \mathcal{V}$, define*

$$\alpha_i := P(i) \sum_{j \in \mathcal{V}} \left(\frac{Q(j)}{Q(i)} - \frac{P(j)}{P(i)} \right)_+ .$$

If the target token is generated as $\tilde{U}_P(1)$ and the B draft tokens are generated as $\tilde{U}_Q(1), \ldots, \tilde{U}_Q(B)$ from the same underlying Poisson process, then the acceptance probability of at least one token satisfies

$$\mathbf{P}(\text{accept}) \geq 1 - \sum_{i \in \mathcal{V}} P(i) \left(\frac{\alpha_i}{1 + \alpha_i} \right)^B . \tag{13}$$

Proof. Let the target token be

$$Y := \tilde{U}_P(1),$$

and let the B draft tokens be

$$X_b := \tilde{U}_Q(b), \qquad b = 1, \ldots, B.$$

The acceptance event is

$$\text{accept} = \{Y \in \{X_1, \ldots, X_B\}\}.$$

Conditioning on $Y = i$, the failure event is

$$\{Y \notin \{X_1, \ldots, X_B\}\}.$$

By the exact $j = 1$ tail formula in the generalized Poisson matching lemma,

$$\mathbf{P}(Y \notin \{X_1, \ldots, X_B\} \,|\, Y = i) \leq \left(\frac{\alpha_i}{1 + \alpha_i} \right)^B ,$$

where

$$\alpha_i = P(i) \sum_{j \in \mathcal{V}} \left(\frac{Q(j)}{Q(i)} - \frac{P(j)}{P(i)} \right)_+ .$$

Therefore,

$$\mathbf{P}(\text{accept} | Y = i) \geq 1 - \left(\frac{\alpha_i}{1 + \alpha_i} \right)^B .$$

Averaging over $Y \sim P$ gives

$$\begin{aligned} \mathbf{P}(\text{accept}) &= \sum_{i \in \mathcal{V}} P(i) \mathbf{P}(\text{accept} | Y = i) \\ &\geq \sum_{i \in \mathcal{V}} P(i) \left[1 - \left(\frac{\alpha_i}{1 + \alpha_i} \right)^B \right] \\ &= 1 - \sum_{i \in \mathcal{V}} P(i) \left(\frac{\alpha_i}{1 + \alpha_i} \right)^B . \end{aligned}$$

This proves (13). □

Moreover,

$$1 - \sum_{i \in \mathcal{V}} P(i) \left(\frac{\alpha_i}{1 + \alpha_i} \right)^B \geq 1 - \sum_{i \in \mathcal{V}} P(i) \left(\frac{P(i)}{P(i) + Q(i)} \right)^B . \tag{14}$$

Thus (13) strictly improves the weaker bound

$$\mathbf{P}(\text{accept}) \geq 1 - \sum_{i \in \mathcal{V}} P(i) \left(\frac{P(i)}{P(i) + Q(i)} \right)^B$$

whenever the inequality in (14) is strict.

We then show above Theorem B.1 is tighter than (Rowan et al., 2025, Proposition 2).

Theorem B.2. *Let P be the target distribution and Q be the draft distribution on a finite vocabulary $\mathcal{V}$. Write*

$$p_i := P(i), \qquad q_i := Q(i), \qquad i \in \mathcal{V},$$

and assume first that $p_i, q_i > 0$ for all $i \in \mathcal{V}$. For each $i \in \mathcal{V}$, define

$$\alpha_i := p_i \sum_{j \in \mathcal{V}} \left(\frac{q_j}{q_i} - \frac{p_j}{p_i} \right)_+ .$$

Let

$$L_{\mathrm{PML}} := 1 - \sum_{i \in \mathcal{V}} p_i \left(\frac{\alpha_i}{1 + \alpha_i} \right)^B$$

be the acceptance lower bound obtained from the exact $j = 1$ tail formula of the generalized Poisson matching lemma. Let

$$L_{\mathrm{LML}} := \sum_{i \in \mathcal{V}} \frac{B}{\sum_{j \in \mathcal{V}} \left[\max\left\{ \frac{p_j}{p_i}, \frac{q_j}{q_i} \right\} + (B-1)\frac{p_j}{p_i} \right]}$$

be the lower bound in Eq. (3) of the list matching lemma, written in the same target–draft notation. Then

$$L_{\mathrm{PML}} \geq L_{\mathrm{LML}}.$$

Moreover, if $B > 1$ and $P \neq Q$, then

$$L_{\mathrm{PML}} > L_{\mathrm{LML}}.$$

Proof. The proof is based on the Poisson matching lemma (Li and Anantharam, 2021). Let $Y := \tilde{Z}_P(1)$ and $X^{(b)} := \tilde{Z}_Q(b)$ for $b = 1, \ldots, B$ be generated from the same mapped Poisson process. By the mapping theorem, $\tilde{Z}_Q(1), \tilde{Z}_Q(2), \ldots$ are i.i.d. according to Q, and $\tilde{Z}_P(1) \sim P$. Define the token-level acceptance event

$$\mathcal{A}_t := \left\{ Y \in \{X^{(1)}, \ldots, X^{(B)}\} \right\}.$$

Apply the exact $j = 1$ tail formula of the generalized Poisson matching lemma with the first distribution equal to P and the second distribution equal to Q. Conditioned on $Y = \tilde{Z}_P(1) = i$, the event $\mathcal{A}_t^c$ is the event that the first mapped P-sample does not appear among the first B mapped Q-samples. Hence

$$\mathbf{P}(\mathcal{A}_t^c | Y = i) \leq \left(\frac{\alpha_i}{1 + \alpha_i} \right)^B,$$

where

$$\alpha_i = p_i \sum_{j \in \mathcal{V}} \left(\frac{q_j}{q_i} - \frac{p_j}{p_i} \right)_+ .$$

Taking complements gives

$$\mathbf{P}(\mathcal{A}_t | Y = i) \geq 1 - \left(\frac{\alpha_i}{1 + \alpha_i} \right)^B.$$

Finally, averaging over $Y \sim P$ gives

$$\begin{aligned} \mathbf{P}(\mathcal{A}_t) &= \sum_{i \in \mathcal{V}} p_i \mathbf{P}(\mathcal{A}_t | Y = i) \\ &\geq \sum_{i \in \mathcal{V}} p_i \left[1 - \left(\frac{\alpha_i}{1 + \alpha_i} \right)^B \right] \\ &= 1 - \sum_{i \in \mathcal{V}} p_i \left(\frac{\alpha_i}{1 + \alpha_i} \right)^B = L_{\mathrm{PML}}. \end{aligned}$$

We now rewrite the LML bound in terms of the same quantities α_i. For each fixed $i \in \mathcal{V}$,

$$\begin{aligned} p_i \sum_{j\in\mathcal{V}} \max\left\{\frac{p_j}{p_i}, \frac{q_j}{q_i}\right\} &= \sum_{j\in\mathcal{V}} \max\left\{p_j, \frac{p_i}{q_i} q_j\right\} \\ &= \sum_{j\in\mathcal{V}} \left[p_j + \left(\frac{p_i}{q_i} q_j - p_j\right)_+\right] \\ &= 1 + p_i \sum_{j\in\mathcal{V}} \left(\frac{q_j}{q_i} - \frac{p_j}{p_i}\right)_+ \\ &= 1 + \alpha_i. \end{aligned}$$

Therefore,

$$\begin{aligned} \sum_{j\in\mathcal{V}} \left[\max\left\{\frac{p_j}{p_i}, \frac{q_j}{q_i}\right\} + (B-1)\frac{p_j}{p_i}\right] &= \frac{1+\alpha_i}{p_i} + (B-1)\frac{1}{p_i} \\ &= \frac{B+\alpha_i}{p_i}. \end{aligned}$$

Thus Eq. (3) of the list matching lemma can be written as

$$L_{\mathrm{LML}} = \sum_{i\in\mathcal{V}} p_i \frac{B}{B+\alpha_i}.$$

It remains to compare the two expressions term by term. For every $\alpha \geq 0$, we claim that

$$1 - \left(\frac{\alpha}{1+\alpha}\right)^B \geq \frac{B}{B+\alpha}.$$

If $\alpha = 0$, both sides are equal to 1. If $\alpha > 0$, the inequality is equivalent to

$$\left(\frac{\alpha}{1+\alpha}\right)^B \leq \frac{\alpha}{B+\alpha},$$

or equivalently

$$\left(1 + \frac{1}{\alpha}\right)^B \geq 1 + \frac{B}{\alpha}.$$

This follows immediately from the binomial theorem:

$$\left(1 + \frac{1}{\alpha}\right)^B = 1 + \frac{B}{\alpha} + \sum_{m=2}^{B} \binom{B}{m} \alpha^{-m} \geq 1 + \frac{B}{\alpha}.$$

Hence, for each $i \in \mathcal{V}$,

$$p_i \left[1 - \left(\frac{\alpha_i}{1+\alpha_i}\right)^B\right] \geq p_i \frac{B}{B+\alpha_i}.$$

Summing over $i \in \mathcal{V}$ yields

$$L_{\mathrm{PML}} \geq L_{\mathrm{LML}}.$$

Finally, suppose $B > 1$ and $P \neq Q$. We show that the inequality is strict. If $\alpha_i = 0$ for every $i \in \mathcal{V}$, then for every $i, j \in \mathcal{V}$,

$$\frac{q_j}{q_i} \leq \frac{p_j}{p_i}.$$

Equivalently,

$$\frac{q_j}{p_j} \leq \frac{q_i}{p_i}.$$

Since this holds for every pair (i, j), all ratios q_i/p_i must be equal to a common constant. Because both P and Q are probability distributions, this constant must be 1, and hence $P = Q$, a contradiction. Therefore, when $P \neq Q$, there exists at least one i such that $\alpha_i > 0$.

For $B > 1$ and $\alpha > 0$, the binomial expansion above is strict:

$$\left(1 + \frac{1}{\alpha}\right)^B > 1 + \frac{B}{\alpha}.$$

Hence

$$1 - \left(\frac{\alpha}{1+\alpha}\right)^B > \frac{B}{B+\alpha}.$$

Applying this to an index i with $\alpha_i > 0$ gives a strict term in the sum. Therefore

$$L_{\mathrm{PML}} > L_{\mathrm{LML}}.$$

This completes the proof. □

B.2.2

We then show the clean version, Theorem 5.1, is already tighter than (Rowan et al., 2025, Eq. (4)).

Theorem B.3. *Let P be the target distribution and Q be the draft distribution on a finite vocabulary $\mathcal{V}$. Write*

$$p_i := P(i), \qquad q_i := Q(i), \qquad i \in \mathcal{V}.$$

Assume first that $p_i, q_i > 0$ for all $i \in \mathcal{V}$. Define

$$L_{\mathrm{PML,weak}} := 1 - \sum_{i \in \mathcal{V}} p_i \left(\frac{p_i}{p_i + q_i}\right)^B$$

and

$$L_{\mathrm{LML,rel}} := \sum_{i \in \mathcal{V}} p_i \left(1 + \frac{p_i}{Bq_i}\right)^{-1} = \sum_{i \in \mathcal{V}} p_i \frac{Bq_i}{p_i + Bq_i}.$$

Then

$$L_{\mathrm{PML,weak}} \geq L_{\mathrm{LML,rel}}.$$

Moreover, if $B > 1$, then the inequality is strict.

Proof. The proof is based on the Poisson matching lemma (Li and Anantharam, 2021). Let $Y := \tilde{Z}_P(1)$ and let

$$X^{(b)} := \tilde{Z}_Q(b), \qquad b = 1, \ldots, B,$$

be generated from the same mapped Poisson process. Define the token-level acceptance event

$$\mathcal{A}_t := \left\{ Y \in \{X^{(1)}, \ldots, X^{(B)}\} \right\}.$$

Applying the relaxed Poisson matching bound with $j = 1$ and $k = B$, conditioned on $Y = \tilde{Z}_P(1) = i$, gives

$$\mathbf{P}(\mathcal{A}_t^c | Y = i) \leq \left(1 - \left(1 + \frac{p_i}{q_i}\right)^{-1}\right)^B = \left(\frac{p_i}{p_i + q_i}\right)^B.$$

Taking complements yields

$$\mathbf{P}(\mathcal{A}_t | Y = i) \geq 1 - \left(\frac{p_i}{p_i + q_i}\right)^B.$$

Averaging over $Y \sim P$ gives

$$\begin{aligned}
\mathbf{P}(\mathcal{A}_t) &= \sum_{i \in \mathcal{V}} p_i \mathbf{P}(\mathcal{A}_t | Y = i) \\
&\geq \sum_{i \in \mathcal{V}} p_i \left[1 - \left(\frac{p_i}{p_i + q_i}\right)^B\right] \\
&= 1 - \sum_{i \in \mathcal{V}} p_i \left(\frac{p_i}{p_i + q_i}\right)^B = L_{\mathrm{PML,weak}}.
\end{aligned}$$

We now compare this lower bound with the relaxed version of the list matching lemma. In the notation of the list matching lemma paper, the proposal distribution is denoted by p and the target distribution by q. Since here we use P for the target and Q for the draft, their relaxed conditional bound, Eq. (4), becomes

$$\mathbf{P}(\mathcal{A}_t|Y=i) \geq \left(1+\frac{p_i}{Bq_i}\right)^{-1} = \frac{Bq_i}{p_i+Bq_i}.$$

Thus the corresponding unconditional relaxed LML bound is

$$L_{\mathrm{LML,rel}} = \sum_{i\in\mathcal{V}} p_i \frac{Bq_i}{p_i+Bq_i}.$$

It remains to compare the two conditional terms. Let

$$r_i := \frac{p_i}{q_i}.$$

Then

$$1-\left(\frac{p_i}{p_i+q_i}\right)^B = 1-\left(\frac{r_i}{1+r_i}\right)^B,$$

whereas

$$\frac{Bq_i}{p_i+Bq_i} = \frac{B}{B+r_i}.$$

Hence it suffices to prove that, for every $r>0$,

$$1-\left(\frac{r}{1+r}\right)^B \geq \frac{B}{B+r}.$$

Equivalently,

$$\left(\frac{r}{1+r}\right)^B \leq \frac{r}{B+r}.$$

Since $r>0$, this is equivalent to

$$\left(1+\frac{1}{r}\right)^B \geq 1+\frac{B}{r}.$$

This follows from the binomial theorem:

$$\left(1+\frac{1}{r}\right)^B = 1+\frac{B}{r}+\sum_{m=2}^{B}\binom{B}{m}r^{-m} \geq 1+\frac{B}{r}.$$

Therefore, for each $i\in\mathcal{V}$,

$$1-\left(\frac{p_i}{p_i+q_i}\right)^B \geq \frac{Bq_i}{p_i+Bq_i}.$$

Multiplying by p_i and summing over $i\in\mathcal{V}$ gives

$$L_{\mathrm{PML,weak}} \geq L_{\mathrm{LML,rel}}.$$

Finally, if $B>1$, then for every $r>0$,

$$\sum_{m=2}^{B}\binom{B}{m}r^{-m} > 0.$$

Thus the above inequality is strict for every i with $p_i, q_i > 0$. Under full support, all such terms have positive weight p_i, and hence

$$L_{\mathrm{PML,weak}} > L_{\mathrm{LML,rel}}.$$

This completes the proof. □

B.3 Generalized Theoretic Guarantee and Proof

Theorem B.4. *Consider the exact draft-indexed active-set version of multi-draft speculative sampling with total draft count B. Fix a current decoding step t, and let $S_t \subseteq \{1, \dots, B\}$ be the active set before sampling the next token. Write $J_t := |S_t|$. Let the current realized context be c_t, and similarly define $P_t := \mathcal{M}_t(\cdot|c_t), Q_t := \mathcal{M}_d(\cdot|c_t)$. For $i \in \mathcal{V}$, write $p_i^{(t)} := P_t(i)$ and $q_i^{(t)} := Q_t(i)$. Then the token-level acceptance probability at step t satisfies*

$$\mathbf{P}(\mathcal{A}_t|c_t, S_t) \geq \sum_{i\in\mathcal{V}} p_i^{(t)} \left[1 - \left(1 - \left(1 + \frac{p_i^{(t)}}{q_i^{(t)}}\right)^{-1}\right)^{J_t}\right] = 1 - \sum_{i\in\mathcal{V}} p_i^{(t)} \left(\frac{p_i^{(t)}}{p_i^{(t)} + q_i^{(t)}}\right)^{J_t}.$$

In particular, at the first step of each speculative block one has $J_t = B$, so

$$\mathbf{P}(\mathcal{A}_t|c_t) \geq 1 - \sum_{i\in\mathcal{V}} p_i^{(t)} \left(\frac{p_i^{(t)}}{p_i^{(t)} + q_i^{(t)}}\right)^{B}.$$

Proof. Conditioned on the realized context c_t and the active set S_t, the exact active-set algorithm uses only the J_t active draft streams for both target-side selection and acceptance. Inactive drafts do not participate in the current-step minimum and therefore do not affect the current-step coupling.

Hence, after conditioning on (c_t, S_t), the current step is identical to the setting of Theorem 5.1 with the number of drafts equal to J_t. Applying Theorem 5.1 with B replaced by J_t gives

$$\mathbf{P}(\mathcal{A}_t|c_t, S_t) \geq 1 - \sum_{i\in\mathcal{V}} p_i^{(t)} \left(\frac{p_i^{(t)}}{p_i^{(t)} + q_i^{(t)}}\right)^{J_t}.$$

At the first step of a speculative block, all drafts are active, so $J_t = B$, which yields the final claim. □

B.4 Extension of Theorem B.1 to non-identically distributed proposals

Proposition B.5. *Let P be the target distribution on a finite vocabulary $\mathcal{V}$, and let $Q_1, \dots, Q_B$ be B proposal distributions. Write*

$$p_i := P(i), \qquad q_{b,i} := Q_b(i), \qquad i \in \mathcal{V},\ b \in [B].$$

Assume first that $p_i, q_{b,i} > 0$ for all i and b. For each $i \in \mathcal{V}$ and $b \in [B]$, define

$$\alpha_i^{(b)} := p_i \sum_{j\in\mathcal{V}} \left(\frac{q_{b,j}}{q_{b,i}} - \frac{p_j}{p_i}\right)_+.$$

Then there exists a common-randomness coupling such that $Y \sim P$, $X^{(b)} \sim Q_b$ for each $b \in [B]$, the proposals $X^{(1)}, \dots, X^{(B)}$ are mutually independent, and

$$\Pr\Big[Y \in \{X^{(1)}, \dots, X^{(B)}\} \,\Big|\, Y = i\Big] \geq \sum_{b=1}^{B} \frac{1}{B + \alpha_i^{(b)}}. \tag{15}$$

Consequently,

$$\Pr\Big[Y \in \{X^{(1)}, \dots, X^{(B)}\}\Big] \geq \sum_{i\in\mathcal{V}} p_i \sum_{b=1}^{B} \frac{1}{B + \alpha_i^{(b)}}. \tag{16}$$

Proof. The proof is based on the Poisson matching lemma (Li and Anantharam, 2021). The main difference from the identical-proposal case is that there is no single proposal ordering when the proposal laws are $Q_1, \dots, Q_B$. We therefore augment the sample space by a block label. Define

$$\widetilde{\mathcal{V}} := \mathcal{V} \times [B],$$

and introduce the augmented target and proposal distributions

$$\widetilde{P}(i,b) := \frac{p_i}{B}, \qquad \widetilde{Q}_b(i,\ell) := \mathbf{1}\{\ell = b\} q_{b,i}.$$

Let

$$(Y, J) := \tilde{Z}_{\widetilde{P}}(1), \qquad (X^{(b)}, b) := \tilde{Z}_{\widetilde{Q}_b}(1), \quad b = 1, \dots, B,$$

be generated from the same underlying Poisson process on $\widetilde{\mathcal{V}} \times \mathbb{R}_{\geq 0}$. By the mapping theorem, $(Y, J) \sim \widetilde{P}$, and hence $Y \sim P$ and J is uniform on $[B]$ and independent of Y. Also, for each b, $(X^{(b)}, b) \sim \widetilde{Q}_b$, so $X^{(b)} \sim Q_b$. Since the laws $\widetilde{Q}_b$ are supported on disjoint blocks, the corresponding proposal samples are mutually independent.

Fix $i \in \mathcal{V}$. Define the acceptance event

$$\mathcal{A} := \left\{ Y \in \{X^{(1)}, \dots, X^{(B)}\} \right\}.$$

For each $b \in [B]$, let

$$\mathcal{A}_b := \left\{ (Y, J) = (X^{(b)}, b) \right\}.$$

Then $\mathcal{A} \supseteq \bigcup_{b=1}^{B} \mathcal{A}_b$, and the events $\mathcal{A}_1, \dots, \mathcal{A}_B$ are pairwise disjoint because they occur in different blocks. Therefore,

$$\begin{aligned}
\mathbf{P}(\mathcal{A}|Y = i) &\geq \sum_{b=1}^{B} \mathbf{P}(\mathcal{A}_b|Y = i) \\
&= \sum_{b=1}^{B} \mathbf{P}(J = b|Y = i)\mathbf{P}(\mathcal{A}_b|Y = i, J = b) \\
&= \frac{1}{B} \sum_{b=1}^{B} \Pr\left[\tilde{Z}_{\widetilde{P}}(1) = \tilde{Z}_{\widetilde{Q}_b}(1) \,\middle|\, \tilde{Z}_{\widetilde{P}}(1) = (i,b) \right]. \qquad (17)
\end{aligned}$$

We now apply the exact $j = 1$ tail formula in the generalized Poisson matching lemma to the pair $(\widetilde{P}, \widetilde{Q}_b)$. Conditioned on $\tilde{Z}_{\widetilde{P}}(1) = (i, b)$, the local PML parameter is

$$\begin{aligned}
\widetilde{\alpha}_i^{(b)} &:= \widetilde{P}(i,b) \sum_{\ell=1}^{B} \sum_{j \in \mathcal{V}} \left(\frac{\widetilde{Q}_b(j,\ell)}{\widetilde{Q}_b(i,b)} - \frac{\widetilde{P}(j,\ell)}{\widetilde{P}(i,b)} \right)_+ \\
&= \frac{p_i}{B} \sum_{j \in \mathcal{V}} \left(\frac{q_{b,j}}{q_{b,i}} - \frac{p_j}{p_i} \right)_+ \\
&= \frac{\alpha_i^{(b)}}{B}.
\end{aligned}$$

Hence

$$\Pr\left[\tilde{Z}_{\widetilde{P}}(1) = \tilde{Z}_{\widetilde{Q}_b}(1) \,\middle|\, \tilde{Z}_{\widetilde{P}}(1) = (i,b) \right] \geq 1 - \frac{\widetilde{\alpha}_i^{(b)}}{1 + \widetilde{\alpha}_i^{(b)}} = \frac{B}{B + \alpha_i^{(b)}}.$$

Substituting this into (17) gives

$$\mathbf{P}(\mathcal{A}|Y = i) \geq \frac{1}{B} \sum_{b=1}^{B} \frac{B}{B + \alpha_i^{(b)}} = \sum_{b=1}^{B} \frac{1}{B + \alpha_i^{(b)}}.$$

This proves (15). Averaging over $Y \sim P$ gives (16). □

C Drafter Invariance

C.1 Definitions

For a speculative sampling algorithm $\mathcal{A}$ with target model $\mathcal{M}_t$, draft model $\mathcal{M}_d$, shared randomness $\mathcal{R}$ and context $x_{:t}$, let $Y_{1:\tau}$ denote the emitted output sequence of $\mathcal{A}$ in the current speculative block.

Definition C.1 ((Daliri et al., 2025))**.** The algorithm $\mathcal{A}$ is said to be *strongly drafter invariant* if for every $1 \le j \le \tau$,

$$\mathbf{P}\big\{Y_{1:j} = y_{1:j} \,\big|\, \mathcal{R},\, x_{:t}\big\} = \mathbf{P}\big\{\widetilde{Y}_{1:j} = y_{1:j} \,\Big|\, \mathcal{R},\, x_{:t}\big\}.$$

Definition C.2 ((Rowan et al., 2025))**.** The algorithm $\mathcal{A}$ is said to be *conditional drafter invariant* if for every $1 \le j \le \tau$, letting $X_{1:L} := X_{1:L}(\mathcal{M}_d)$ denote the length-L draft sequence by $\mathcal{M}_d$, we have

$$\mathbf{P}\big\{Y_{1:j} = y_{1:j} \,\big|\, \mathcal{R},\, x_{:t},\, X_{1:L} = x_{1:L}\big\} = \mathbf{P}\big\{\widetilde{Y}_{1:j} = y_{1:j} \,\big|\, \mathcal{R},\, x_{:t},\, \widetilde{X}_{1:L} = x_{1:L}\big\}.$$

C.1.1 Proof of Proposition 4.4

We rephrase Proposition 4.4 to a more detailed version as follows.

Proposition C.3 (Stopping-time drafter invariance)**.** *Let τ denote the number of tokens emitted in one speculative block of Algorithm 3. Fix the initial prompt $h := w_{1:n}$ and the shared randomness*

$$\mathcal{R} := \{\, G(\mathsf{k}, c) : c \text{ is any context} \,\}.$$

Define the target-side PFR chain *recursively by*

$$Z_1 := \mathrm{PFR}\big(P(\cdot|h),\, G(\mathsf{k}, h)\big),$$

and for $s \ge 2$,

$$Z_s := \mathrm{PFR}\big(P(\cdot|h\|Z_{1:s-1}),\, G(\mathsf{k},\, h\|Z_{1:s-1})\big).$$

Then, for any stopping-time value τ_0 and every $1 \le j \le \tau_0$,

$$Y_{1:j} = Z_{1:j} \qquad \text{almost surely on the event } \{\tau = \tau_0\}.$$

Equivalently, for every output prefix $y_{1:j}$,

$$\mathbf{P}\big\{Y_{1:j} = y_{1:j} \,\big|\, \mathcal{R},\, h,\, \tau = \tau_0\big\} = \mathbb{1}\{Z_{1:j} = y_{1:j}\}.$$

In particular, for any two draft models Q and $\widetilde{Q}$,

$$\mathbf{P}\big\{Y_{1:j} = y_{1:j} \,\big|\, \mathcal{R},\, h,\, \tau = \tau_0\big\} = \mathbf{P}\big\{\widetilde{Y}_{1:j} = y_{1:j} \,\big|\, \mathcal{R},\, h,\, \widetilde{\tau} = \tau_0\big\},$$

so Algorithm 3 is stopping-time drafter invariant in the sense of Definition 4.3.

Proof. Fix $(\mathcal{R}, h)$ and a stopping-time value τ_0. We first show by induction on s that, on the event $\{\tau \ge s\}$,

$$Y_s = Z_s.$$

For $s = 1$, the first context is

$$c_1 = h.$$

Hence

$$Y_1 = \mathrm{PFR}\big(P(\cdot|c_1),\, G(\mathsf{k}, c_1)\big) = \mathrm{PFR}\big(P(\cdot|h),\, G(\mathsf{k}, h)\big) = Z_1.$$

Now suppose $s \ge 2$ and that $Y_r = Z_r$ holds on $\{\tau \ge r\}$ for all $r = 1, \dots, s-1$. On the event $\{\tau \ge s\}$, the block has reached step s, so the first $s-1$ draft tokens were accepted. Therefore

$$\tilde{w}_{1:s-1} = Y_{1:s-1}.$$

By the induction hypothesis, on $\{\tau \ge s\}$ we also have

$$Y_{1:s-1} = Z_{1:s-1},$$

and hence

$$c_s = h\|\tilde{w}_{1:s-1} = h\|Y_{1:s-1} = h\|Z_{1:s-1}.$$

Thus,

$$Y_s = \mathrm{PFR}\big(P(\cdot|c_s),\, G(\mathsf{k}, c_s)\big) = \mathrm{PFR}\big(P(\cdot|h\|Z_{1:s-1}),\, G(\mathsf{k},\, h\|Z_{1:s-1})\big) = Z_s.$$

This completes the induction.

Now let $1 \le j \le \tau_0$. On the event $\{\tau = \tau_0\}$ we have $\tau \ge j$, so the above induction gives

$$Y_{1:j} = Z_{1:j} \qquad \text{almost surely on } \{\tau = \tau_0\}.$$

Therefore,

$$\mathbf{P}\big\{Y_{1:j} = y_{1:j} \,\big|\, \mathcal{R},\, h,\, \tau = \tau_0\big\} = \mathbb{1}\{Z_{1:j} = y_{1:j}\}.$$

Since the right-hand side depends only on $(\mathcal{R}, h, \tau_0)$ and not on the draft model, the same conditional law holds for any other draft model $\widetilde{Q}$ conditioned on $\widetilde{\tau} = \tau_0$. This is exactly stopping-time drafter invariance. □

C.2 Multi-Draft Drafter invariance

For multi-draft speculative sampling scheme $\mathcal{A}$ with target model $\mathcal{M}_t$, draft models $\mathcal{M}_d^{(1)}, \ldots, \mathcal{M}_d^{(B)}$, context $x_{:t}$, shared randomness $\mathcal{R}$ and $b \in [B]$, let $X_{1:L}^{(b)} := X_{1:L}(\mathcal{M}_d^{(b)})$ denote the length-L draft sequence by $\mathcal{M}_d^{(b)}$ and $Y_{1:\tau}$ denote the emitted output sequence of $\mathcal{A}$ in the current speculative block.

Definition C.4 (Strong drafter invariance)**.** The algorithm $\mathcal{A}$ is *drafter invariant* if for every $1 \le j \le \tau$, every initial context $x_{:t}$, every realization of the shared randomness $\mathcal{R}$, and every output prefix $y_{1:j}$,

$$\mathbf{P}\Big\{Y_{1:j} = y_{1:j} \,\Big|\, \mathcal{R},\, x_{:t}\Big\} = \mathbf{P}\Big\{\widetilde{Y}_{1:j} = y_{1:j} \,\Big|\, \mathcal{R},\, x_{:t}\Big\},$$

for any two choices of draft models

$$\mathcal{M}_d^{(1)}, \ldots, \mathcal{M}_d^{(B)} \quad \text{and} \quad \widetilde{\mathcal{M}}_d^{(1)}, \ldots, \widetilde{\mathcal{M}}_d^{(B)}.$$

Definition C.5 (Conditional drafter invariance)**.** We say that the algorithm is *conditionally drafter invariant* if, for every $1 \le j \le \tau$, every initial context $x_{:t}$, every realization of the shared randomness $\mathcal{R}$, every draft sequences $x_{1:L}^{(1)}, \ldots, x_{1:L}^{(B)}$, and every output prefix $y_{1:j}$,

$$\begin{aligned}&\mathbf{P}\Big\{Y_{1:j} = y_{1:j} \,\Big|\, \mathcal{R},\, x_{:t},\, X_{1:L}^{(1)} = x_{1:L}^{(1)}, \ldots, X_{1:L}^{(B)} = x_{1:L}^{(B)}\Big\}\\ &= \mathbf{P}\Big\{\widetilde{Y}_{1:j} = y_{1:j} \,\Big|\, \mathcal{R},\, x_{:t},\, \widetilde{X}_{1:L}^{(1)} = x_{1:L}^{(1)}, \ldots, \widetilde{X}_{1:L}^{(B)} = x_{1:L}^{(B)}\Big\},\end{aligned}$$

for any two choices of draft models

$$\mathcal{M}_d^{(1)}, \ldots, \mathcal{M}_d^{(B)} \quad \text{and} \quad \widetilde{\mathcal{M}}_d^{(1)}, \ldots, \widetilde{\mathcal{M}}_d^{(B)}.$$

That is, conditioned on the shared randomness, the initial context, and the realized draft sequences, the conditional law of the emitted output prefix does not depend on which draft models generated those sequences.

Definition C.6 (stopping time drafter invariance)**.** Consider a speculative sampling algorithm with target model $\mathcal{M}_t$, draft models $\mathcal{M}_d^{(1)}, \ldots, \mathcal{M}_d^{(B)}$, initial context $x_{:t}$, shared randomness $\mathcal{R}$, and block output $Y_{1:\tau}$, where τ is the number of emitted tokens in the current speculative block.

We say that the algorithm is *stopping time drafter invariant* if, for every $1 \le j \le \tau$, every initial context $x_{:t}$, every realization of the shared randomness $\mathcal{R}$, every block length value τ_0, and every output prefix $y_{1:j}$,

$$\begin{aligned}&\mathbf{P}\Big\{Y_{1:j} = y_{1:j} \,\Big|\, \mathcal{R},\, x_{:t},\, \tau = \tau_0\Big\}\\ &= \mathbf{P}\Big\{\widetilde{Y}_{1:j} = y_{1:j} \,\Big|\, \mathcal{R},\, x_{:t},\, \widetilde{\tau} = \tau_0\Big\},\end{aligned}$$

for any two choices of draft models

$$\mathcal{M}_d^{(1)}, \ldots, \mathcal{M}_d^{(B)} \quad \text{and} \quad \widetilde{\mathcal{M}}_d^{(1)}, \ldots, \widetilde{\mathcal{M}}_d^{(B)}.$$

That is, conditioned on the shared randomness, the initial context, and the realized block length, the conditional law of the emitted output prefix does not depend on the draft models.

D Experiment Details and Full Results

This appendix expands the main-text evaluation along the two axes that mirror the contributions of Section 1: (i) the *single-draft regime* (B=1), where we certify that our keyed Poisson sampler preserves acceptance and quality while carrying watermark signal; and (ii) the *multi-draft regime* (B>1), where we extend the unbiasedness/strength guarantees to larger B and benchmark sampling efficiency against the list-coupling scheme of Rowan et al. (2025).

D.1 SETUP

Models. We use three target/drafter pairs, with different roles in the evaluation. The headline pair is `Qwen2.5-7B-Instruct/Qwen2.5-0.5B-Instruct`. The main multi-draft efficiency comparison additionally uses `Llama-3.1-8B-Instruct/Llama-3.2-1B-Instruct`. We retain `lmsys/vicuna-7b-v1.5/double7/vicuna-68m` only for the single-draft and detection generalization experiments; the FastChat v1.1 chat template is injected explicitly because the released Vicuna checkpoint does not ship a `chat_template` attribute. All models run in `float16` on a single NVIDIA H100 GPU.

Datasets. We evaluate on two tasks chosen to match the coverage of prior work: CNN/DAILYMAIL (summarization, prompted with the `System:/INPUT:/OUTPUT:` template of Hu and Huang (2024)) and ELI5 (open-ended explanation, Fan et al. (2019)). Each cell uses n=1000 prompts.

Sampling parameters. Unless stated otherwise we use top_k=50, top_p=1.0, temperature T=1.0, and `max_new_tokens`=128. The watermark key is fixed across all runs so that detection is deterministic.

Sweeps. For single-draft we sweep lookahead $L \in \{1, 2, 3, 4\}$, matching Hu and Huang (2024). For multi-draft we fix L=4 and sweep draft counts $B \in \{2, 4, 6, 8\}$, matching Rowan et al. (2025). Each (MODEL, DATASET, DECODER, L, B) configuration is run with multiple independent seeds; tables in Appendices D.3–D.4 report $\text{mean} \pm \text{std}$ across the 1000 prompts of the cell.

Baselines. We adopt the four-family taxonomy used in the main text (Section 6). For convenience we restate the role of each baseline:

- **Quality anchor** — BASIC-UWM: autoregressive Gumbel-style unbiased watermarking with no speculation. Provides the LPPL reference and the watermark-strength ceiling.
- **Speculation-only references** — VSPS (Leviathan et al., 2023) at B=1, and GLS (the drafter-invariant *InvariantMultiDraftStrategy*) of Rowan et al. (2025) at B>1. These provide speculation-only efficiency references.
- **Trade-off frontier** — MWS and MSE (Hu and Huang, 2024), plus MSE-PSEUDO, the dual-key pseudorandom-r variant of He et al. (2026). These provide the single-draft trade-off used for comparison.
- **Ours** — PFR (single-draft), MPFR (multi-draft), and the unwatermarked variant PFR-NOWM.

Metrics. We restate the three metric families used in the main text. *Efficiency.* **AATPS** (average accepted tokens per step) and **Token rate** in tokens/s. The latter is hardware-dependent and is reported only as a diagnostic. *Detectability.* **ANLPPT** — the average negative log p-value per token — under three score variants: **ANLPPT-U** (Aaronson and Kirchner, 2023; Hu and Huang, 2024), **ANLPPT-Li** (Li et al., 2025), and **ANLPPT-PL** (Lattimore, 2026). We additionally report **TPR@1%FPR** based on the Aaronson Gamma-tail test (Appendix D.5, "Detection statistic"). *Quality.* **LPPL** (per-token log-perplexity under the target), reported as an audit column. LPPL is only an empirical finite-sample audit; exact marginal preservation follows from the sampling construction, not from this statistic.

Reproducibility. The exact (seed, MODEL, DATASET, DECODER, L, B) configurations, software versions (`torch`, `transformers`, attention kernel), GPU model, and per-seed result files are released with the paper.

D.2 PROMPT TEMPLATES

For all instruction-tuned targets we apply the model's native chat template via the HuggingFace `apply_chat_template` interface. The Vicuna-v1.5 release ships without a `chat_template` attribute, so we explicitly inject the FastChat v1.1 template before tokenization.

CNN/DailyMail (summarization).

```
[
  {"role": "system",
   "content": "You are a helpful assistant."},
  {"role": "user",
   "content": "Summarize the following article in 3-5 sentences:\n\n"
              + article[:1500]}
]
```

The article is truncated to 1500 characters before tokenization.

ELI5 (open-ended QA).

```
[
  {"role": "system",
   "content": "You are a helpful assistant."},
  {"role": "user",
   "content": "Please explain like I'm five:\n\n" + question}
]
```

For each prompt the model generates up to 128 new tokens under the sampling parameters of Appendix D.1. The watermark key is fixed across all runs.

D.3 SINGLE-DRAFT EXPERIMENTS: FULL RESULTS

Goal. Hu and Huang (2024) characterize a single-draft trade-off between watermark strength and speculative acceptance. For a discrete vocabulary, our single-draft PFR implementation is functionally equivalent to the shared Gumbel/DeltaGumbel construction underlying MWS; therefore, we do not claim that PFR by itself bypasses the single-draft no-go result. We retain these experiments as a one-draft sanity check and as the base case for the MPFR extension.

Protocol. We follow Hu and Huang (2024): the same metric battery (AATPS, ANLPPT, LPPL) and the same lookahead range $L \in \{1, 2, 3, 4\}$. We evaluate on `Qwen2.5-7B-Instruct/0.5B` and on `vicuna-7b-v1.5/vicuna-68m`, on both CNN/DAILYMAIL and ELI5. On top of the Hu–Huang decoder set we add (i) our PFR sampler, (ii) the unwatermarked variant PFR-NOWM, (iii) BASIC-UWM as the unbiasedness anchor, and (iv) the pseudorandom-r variant MSE-PSEUDO of He et al. (2026).

Headline figure. Figure 1 (in Section 4) plots every method as a curve in the $(\text{AATPS}, \text{ANLPPT-}U)$ plane sweeping L on the headline cell. Figures 3–5 report the same plot on the remaining three cells. These plots verify strong detector-visible signal and isolate the effect of keying the shared source through the PFR/PFR-NOWM comparison; the single-draft curves are not claimed as a separate bypass of the Hu–Huang trade-off.

Unbiasedness audit. We report LPPL under the target model as an empirical audit of whether the keyed Poisson coupling introduces a measurable likelihood shift. PFR matches BASIC-UWM in LPPL within statistical noise, as shown in the final column of Tables 2–5. Together with the AATPS and ANLPPT results, this supports our empirical claim that coupling-level watermarking preserves both acceptance behavior and detector-visible signal; the exact marginal guarantee follows from the PFR construction.

Per-cell tables. Tables 2–5 report all metrics per $(L, \text{DECODER})$ for the four $(\text{TARGET}, \text{DATASET})$ cells. Each cell reports $\text{mean} \pm \text{std}$ across the 1000 prompts of the cell.

D.4 MULTI-DRAFT EXPERIMENTS: FULL PER-CELL RESULTS

Setup. The main-text multi-draft comparison uses the 2×2 matrix {Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct} $\times$ {CNN/DAILYMAIL, ELI5}, with tar-

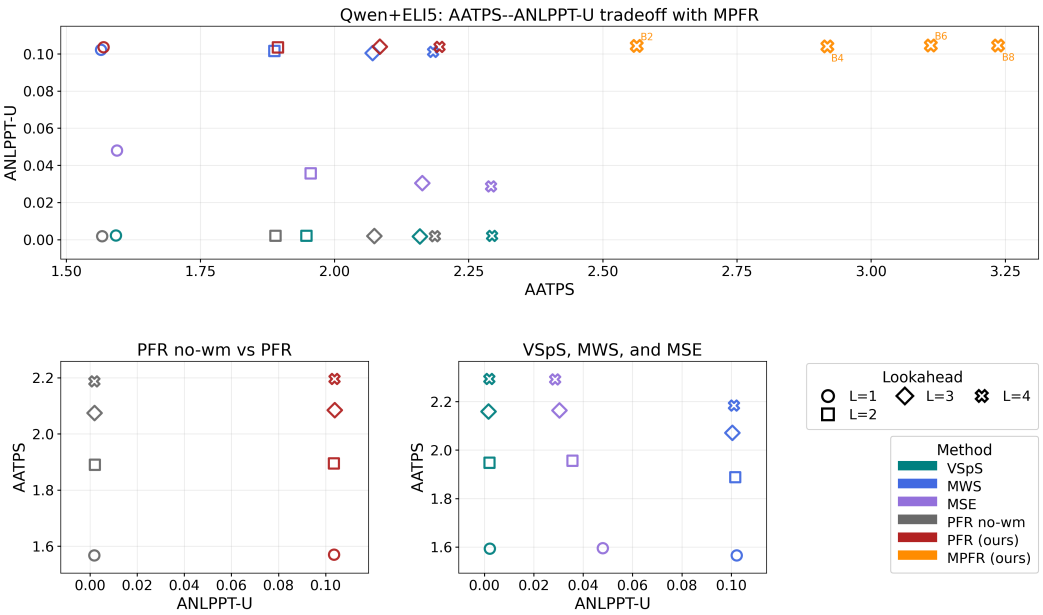


Figure 3: Single-draft (AATPS, ANLPPT-U) frontier on Qwen2.5-7B-Instruct / Qwen2.5-0.5B-Instruct × ELI5. The PFR/PFR-NoWM comparison isolates the effect of keying the shared source; these single-draft results are not claimed to bypass the single-draft trade-off.

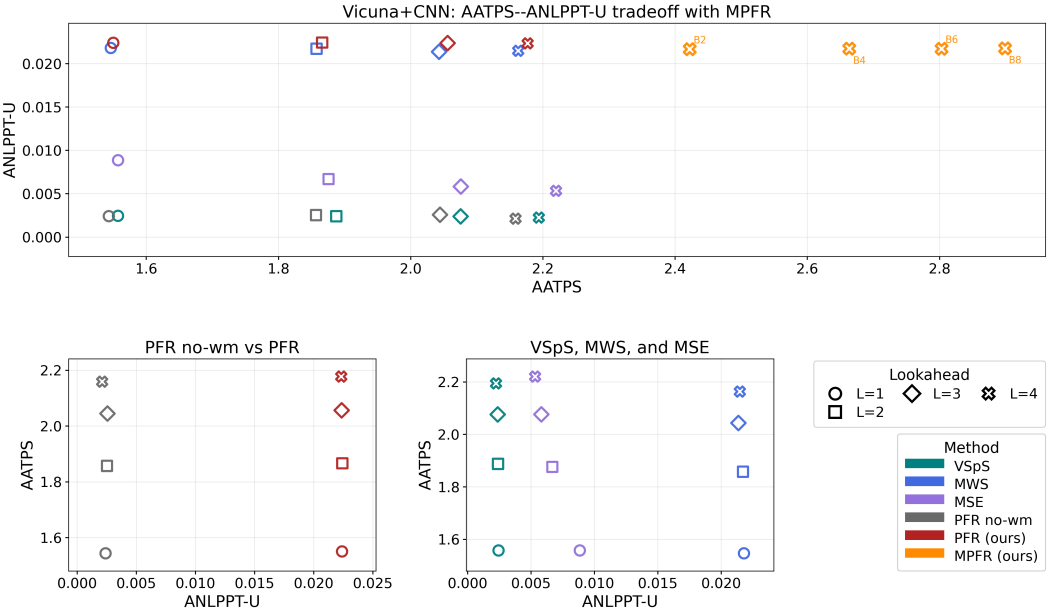


Figure 4: Single-draft (AATPS, ANLPPT-U) frontier on Vicuna-7B-v1.5 / Vicuna-68m × CNN/DailyMail. The PFR/PFR-NoWM comparison isolates the effect of keying the shared source; these single-draft results are not claimed to bypass the single-draft trade-off.

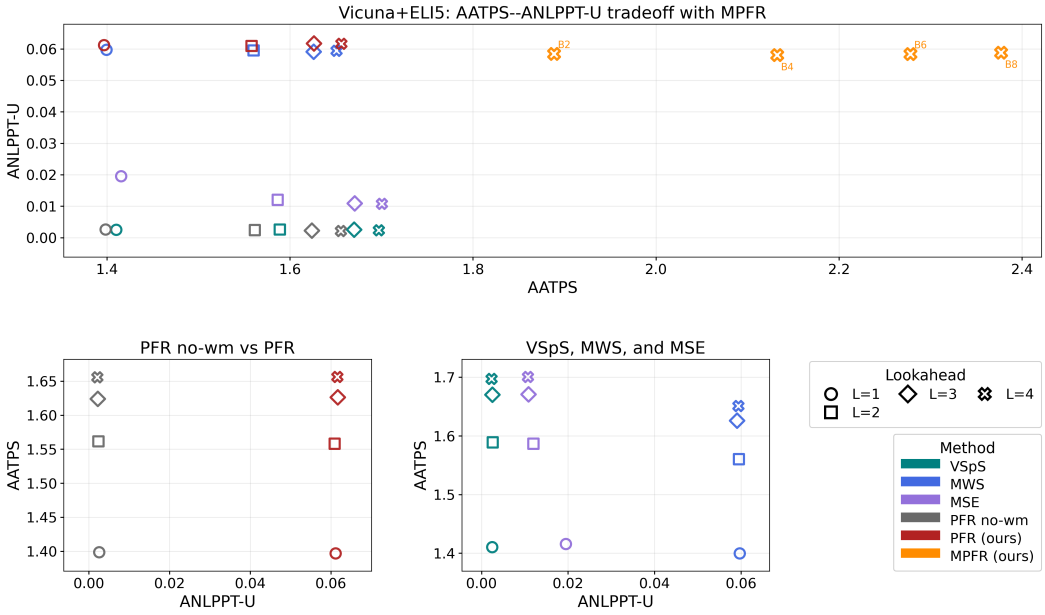


Figure 5: Single-draft (AATPS, ANLPPT-U) frontier on Vicuna-7B-v1.5 / Vicuna-68m × ELI5. The PFR/PFR-NoWM comparison isolates the effect of keying the shared source; these single-draft results are not claimed to bypass the single-draft trade-off.

Table 2: Single-draft results on Qwen2.5-7B-Instruct / Qwen2.5-0.5B-Instruct × CNN/DailyMail. Each cell: mean ± std across 1000 prompts.

L	Method	AATPS	Token rate	ANLPPT-U	ANLPPT-Li	ANLPPT-PL	LPPL
1	Basic-UWM	1.000 ± 0.000	33.353 ± 2.815	0.048 ± 0.035	0.035 ± 0.029	0.051 ± 0.036	0.473 ± 0.134
	VSpS	1.641 ± 0.070	33.229 ± 1.944	0.003 ± 0.005	0.003 ± 0.006	0.003 ± 0.006	0.477 ± 0.136
	PFR-NoWM	1.623 ± 0.075	32.147 ± 2.161	0.002 ± 0.006	0.003 ± 0.006	0.002 ± 0.006	0.474 ± 0.140
	PFR	1.625 ± 0.072	33.321 ± 1.706	0.051 ± 0.036	0.036 ± 0.030	0.053 ± 0.038	0.470 ± 0.143
	MSE	1.642 ± 0.070	20.578 ± 2.645	0.023 ± 0.021	0.016 ± 0.017	0.025 ± 0.023	0.472 ± 0.136
	MWS	1.621 ± 0.072	20.136 ± 2.952	0.049 ± 0.035	0.034 ± 0.029	0.052 ± 0.036	0.475 ± 0.133
	MSE-Pseudo	1.641 ± 0.072	20.703 ± 2.420	0.033 ± 0.028	0.023 ± 0.023	0.035 ± 0.029	0.475 ± 0.138
2	Basic-UWM	1.000 ± 0.000	33.353 ± 2.815	0.048 ± 0.035	0.035 ± 0.029	0.051 ± 0.036	0.473 ± 0.134
	VSpS	2.055 ± 0.147	29.399 ± 2.693	0.002 ± 0.006	0.002 ± 0.006	0.002 ± 0.006	0.482 ± 0.139
	PFR-NoWM	2.014 ± 0.151	30.119 ± 2.438	0.002 ± 0.005	0.003 ± 0.005	0.002 ± 0.005	0.474 ± 0.140
	PFR	2.016 ± 0.154	29.746 ± 2.374	0.051 ± 0.036	0.036 ± 0.030	0.053 ± 0.038	0.471 ± 0.143
	MSE	2.056 ± 0.149	16.954 ± 2.182	0.016 ± 0.017	0.012 ± 0.014	0.018 ± 0.019	0.475 ± 0.139
	MWS	2.012 ± 0.155	17.915 ± 2.679	0.048 ± 0.035	0.034 ± 0.028	0.052 ± 0.036	0.473 ± 0.134
	MSE-Pseudo	2.053 ± 0.150	17.768 ± 2.415	0.030 ± 0.026	0.022 ± 0.021	0.033 ± 0.028	0.481 ± 0.143
3	Basic-UWM	1.000 ± 0.000	33.353 ± 2.815	0.048 ± 0.035	0.035 ± 0.029	0.051 ± 0.036	0.473 ± 0.134
	VSpS	2.327 ± 0.228	26.272 ± 2.968	0.002 ± 0.006	0.002 ± 0.006	0.003 ± 0.006	0.481 ± 0.141
	PFR-NoWM	2.270 ± 0.228	26.709 ± 2.682	0.002 ± 0.006	0.002 ± 0.006	0.002 ± 0.005	0.474 ± 0.140
	PFR	2.272 ± 0.225	25.999 ± 2.742	0.051 ± 0.036	0.037 ± 0.030	0.053 ± 0.038	0.472 ± 0.144
	MSE	2.330 ± 0.228	13.453 ± 1.603	0.015 ± 0.017	0.011 ± 0.014	0.017 ± 0.018	0.476 ± 0.140
	MWS	2.262 ± 0.222	16.187 ± 2.509	0.048 ± 0.035	0.034 ± 0.028	0.051 ± 0.036	0.474 ± 0.134
	MSE-Pseudo	2.337 ± 0.232	14.933 ± 2.118	0.029 ± 0.026	0.021 ± 0.022	0.032 ± 0.027	0.480 ± 0.138
4	Basic-UWM	1.000 ± 0.000	33.353 ± 2.815	0.048 ± 0.035	0.035 ± 0.029	0.051 ± 0.036	0.473 ± 0.134
	VSpS	2.513 ± 0.288	23.229 ± 3.176	0.002 ± 0.006	0.002 ± 0.005	0.002 ± 0.006	0.475 ± 0.134
	PFR-NoWM	2.439 ± 0.292	23.285 ± 2.840	0.002 ± 0.005	0.002 ± 0.005	0.002 ± 0.006	0.474 ± 0.142
	PFR	2.446 ± 0.286	23.075 ± 2.755	0.050 ± 0.036	0.036 ± 0.030	0.053 ± 0.038	0.471 ± 0.144
	MSE	2.525 ± 0.298	12.623 ± 1.893	0.014 ± 0.016	0.010 ± 0.013	0.016 ± 0.017	0.479 ± 0.145
	MWS	2.432 ± 0.287	13.592 ± 2.213	0.048 ± 0.035	0.034 ± 0.028	0.052 ± 0.036	0.475 ± 0.132
	MSE-Pseudo	2.499 ± 0.287	13.494 ± 1.934	0.029 ± 0.026	0.021 ± 0.021	0.031 ± 0.027	0.481 ± 0.141

Table 3: Single-draft results on Qwen2.5-7B-Instruct / Qwen2.5-0.5B-Instruct × ELI5. Each cell: mean ± std across 1000 prompts.

L	Method	AATPS	Token rate	ANLPPT-U	ANLPPT-Li	ANLPPT-PL	LPPL
1	BASIC-UWM	1.000 ± 0.000	26.132 ± 1.924	0.100 ± 0.045	0.072 ± 0.039	0.104 ± 0.046	0.696 ± 0.141
	VSPS	1.593 ± 0.060	32.090 ± 1.889	0.002 ± 0.005	0.002 ± 0.005	0.002 ± 0.005	0.701 ± 0.133
	PFR-NOWM	1.567 ± 0.061	33.046 ± 1.405	0.002 ± 0.004	0.002 ± 0.004	0.002 ± 0.004	0.703 ± 0.139
	PFR	1.570 ± 0.062	33.213 ± 1.487	0.104 ± 0.048	0.075 ± 0.039	0.107 ± 0.049	0.699 ± 0.134
	MSE	1.595 ± 0.060	18.218 ± 2.022	0.048 ± 0.030	0.032 ± 0.024	0.053 ± 0.032	0.704 ± 0.135
	MWS	1.565 ± 0.063	20.177 ± 2.688	0.102 ± 0.047	0.075 ± 0.040	0.106 ± 0.048	0.700 ± 0.137
	MSE-PSEUDO	1.593 ± 0.062	19.553 ± 2.107	0.068 ± 0.038	0.049 ± 0.032	0.071 ± 0.039	0.703 ± 0.145
2	BASIC-UWM	1.000 ± 0.000	26.132 ± 1.924	0.100 ± 0.045	0.072 ± 0.039	0.104 ± 0.046	0.696 ± 0.141
	VSPS	1.948 ± 0.116	27.350 ± 2.434	0.002 ± 0.005	0.002 ± 0.005	0.002 ± 0.004	0.701 ± 0.136
	PFR-NOWM	1.890 ± 0.118	28.719 ± 1.782	0.002 ± 0.005	0.002 ± 0.004	0.002 ± 0.004	0.698 ± 0.137
	PFR	1.895 ± 0.125	28.851 ± 1.937	0.104 ± 0.048	0.075 ± 0.039	0.107 ± 0.049	0.699 ± 0.134
	MSE	1.956 ± 0.126	14.867 ± 1.179	0.036 ± 0.024	0.024 ± 0.020	0.040 ± 0.026	0.704 ± 0.140
	MWS	1.888 ± 0.121	16.985 ± 2.208	0.102 ± 0.046	0.074 ± 0.039	0.105 ± 0.047	0.700 ± 0.142
	MSE-PSEUDO	1.951 ± 0.118	24.767 ± 2.016	0.062 ± 0.034	0.043 ± 0.028	0.066 ± 0.036	0.706 ± 0.136
3	BASIC-UWM	1.000 ± 0.000	26.132 ± 1.924	0.100 ± 0.045	0.072 ± 0.039	0.104 ± 0.046	0.696 ± 0.141
	VSPS	2.159 ± 0.175	23.619 ± 2.707	0.002 ± 0.004	0.002 ± 0.004	0.002 ± 0.003	0.706 ± 0.139
	PFR-NOWM	2.074 ± 0.167	24.315 ± 2.065	0.002 ± 0.004	0.002 ± 0.005	0.002 ± 0.004	0.710 ± 0.139
	PFR	2.084 ± 0.174	24.675 ± 2.059	0.104 ± 0.049	0.075 ± 0.041	0.107 ± 0.049	0.697 ± 0.135
	MSE	2.164 ± 0.175	13.359 ± 2.005	0.030 ± 0.023	0.020 ± 0.019	0.035 ± 0.025	0.705 ± 0.135
	MWS	2.071 ± 0.169	14.243 ± 1.924	0.100 ± 0.046	0.073 ± 0.040	0.104 ± 0.047	0.698 ± 0.138
	MSE-PSEUDO	2.169 ± 0.172	21.339 ± 2.086	0.061 ± 0.034	0.043 ± 0.028	0.064 ± 0.035	0.709 ± 0.133
4	BASIC-UWM	1.000 ± 0.000	26.132 ± 1.924	0.100 ± 0.045	0.072 ± 0.039	0.104 ± 0.046	0.696 ± 0.141
	VSPS	2.294 ± 0.220	17.142 ± 3.485	0.002 ± 0.005	0.002 ± 0.005	0.002 ± 0.004	0.705 ± 0.137
	PFR-NOWM	2.187 ± 0.204	21.512 ± 2.145	0.002 ± 0.004	0.002 ± 0.005	0.002 ± 0.004	0.703 ± 0.138
	PFR	2.196 ± 0.214	21.319 ± 2.212	0.104 ± 0.048	0.075 ± 0.041	0.107 ± 0.049	0.698 ± 0.134
	MSE	2.292 ± 0.220	13.200 ± 2.021	0.029 ± 0.021	0.019 ± 0.017	0.033 ± 0.024	0.701 ± 0.136
	MWS	2.184 ± 0.208	11.570 ± 1.496	0.101 ± 0.046	0.073 ± 0.039	0.105 ± 0.047	0.696 ± 0.141
	MSE-PSEUDO	2.302 ± 0.214	18.298 ± 1.942	0.060 ± 0.033	0.042 ± 0.028	0.063 ± 0.035	0.705 ± 0.132

Table 4: Single-draft results on Vicuna-7B-v1.5 / Vicuna-68m × CNN/DAILYMAIL. Each cell: mean ± std across 1000 prompts.

L	Method	AATPS	Token rate	ANLPPT-U	ANLPPT-Li	ANLPPT-PL	LPPL
1	BASIC-UWM	1.000 ± 0.000	37.058 ± 0.848	0.021 ± 0.024	0.016 ± 0.019	0.022 ± 0.025	0.274 ± 0.190
	VSPS	1.557 ± 0.086	48.890 ± 3.439	0.002 ± 0.006	0.003 ± 0.006	0.002 ± 0.006	0.275 ± 0.167
	PFR-NOWM	1.544 ± 0.088	37.027 ± 5.064	0.002 ± 0.006	0.003 ± 0.006	0.002 ± 0.005	0.269 ± 0.121
	PFR	1.550 ± 0.090	38.256 ± 6.092	0.022 ± 0.024	0.017 ± 0.019	0.023 ± 0.025	0.272 ± 0.128
	MSE	1.557 ± 0.088	45.513 ± 4.490	0.009 ± 0.013	0.007 ± 0.011	0.010 ± 0.015	0.269 ± 0.123
	MWS	1.547 ± 0.091	30.814 ± 4.078	0.022 ± 0.024	0.017 ± 0.020	0.023 ± 0.026	0.276 ± 0.192
	MSE-PSEUDO	1.555 ± 0.088	27.597 ± 2.405	0.015 ± 0.018	0.012 ± 0.016	0.016 ± 0.018	0.279 ± 0.118
2	BASIC-UWM	1.000 ± 0.000	37.058 ± 0.848	0.021 ± 0.024	0.016 ± 0.019	0.022 ± 0.025	0.274 ± 0.190
	VSPS	1.888 ± 0.179	53.865 ± 5.770	0.002 ± 0.006	0.002 ± 0.006	0.002 ± 0.006	0.268 ± 0.129
	PFR-NOWM	1.857 ± 0.178	40.154 ± 5.953	0.003 ± 0.005	0.003 ± 0.006	0.003 ± 0.005	0.273 ± 0.123
	PFR	1.866 ± 0.183	42.639 ± 7.272	0.022 ± 0.024	0.017 ± 0.019	0.023 ± 0.025	0.272 ± 0.127
	MSE	1.876 ± 0.186	46.276 ± 7.794	0.007 ± 0.010	0.006 ± 0.009	0.007 ± 0.011	0.268 ± 0.131
	MWS	1.858 ± 0.181	31.588 ± 5.453	0.022 ± 0.024	0.016 ± 0.020	0.023 ± 0.025	0.278 ± 0.193
	MSE-PSEUDO	1.877 ± 0.182	29.023 ± 4.118	0.015 ± 0.018	0.012 ± 0.015	0.015 ± 0.020	0.273 ± 0.127
3	BASIC-UWM	1.000 ± 0.000	37.058 ± 0.848	0.021 ± 0.024	0.016 ± 0.019	0.022 ± 0.025	0.274 ± 0.190
	VSPS	2.076 ± 0.254	53.835 ± 7.463	0.002 ± 0.006	0.003 ± 0.006	0.002 ± 0.005	0.276 ± 0.212
	PFR-NOWM	2.044 ± 0.259	42.513 ± 7.326	0.003 ± 0.006	0.003 ± 0.006	0.002 ± 0.006	0.274 ± 0.128
	PFR	2.056 ± 0.263	41.813 ± 7.586	0.022 ± 0.024	0.017 ± 0.019	0.023 ± 0.025	0.272 ± 0.127
	MSE	2.076 ± 0.260	38.713 ± 8.338	0.006 ± 0.009	0.005 ± 0.008	0.006 ± 0.010	0.266 ± 0.118
	MWS	2.043 ± 0.258	28.960 ± 4.759	0.021 ± 0.024	0.016 ± 0.020	0.022 ± 0.026	0.277 ± 0.191
	MSE-PSEUDO	2.062 ± 0.257	32.256 ± 6.945	0.015 ± 0.019	0.012 ± 0.016	0.015 ± 0.020	0.277 ± 0.125
4	BASIC-UWM	1.000 ± 0.000	37.058 ± 0.848	0.021 ± 0.024	0.016 ± 0.019	0.022 ± 0.025	0.274 ± 0.190
	VSPS	2.194 ± 0.324	53.940 ± 8.237	0.002 ± 0.005	0.002 ± 0.005	0.002 ± 0.005	0.271 ± 0.120
	PFR-NOWM	2.159 ± 0.326	40.083 ± 7.877	0.002 ± 0.005	0.002 ± 0.005	0.002 ± 0.005	0.268 ± 0.117
	PFR	2.177 ± 0.323	41.800 ± 9.355	0.022 ± 0.023	0.017 ± 0.019	0.023 ± 0.025	0.272 ± 0.127
	MSE	2.220 ± 0.330	32.730 ± 6.784	0.005 ± 0.008	0.005 ± 0.008	0.006 ± 0.009	0.264 ± 0.119
	MWS	2.163 ± 0.319	27.253 ± 4.410	0.021 ± 0.024	0.016 ± 0.019	0.023 ± 0.025	0.277 ± 0.191
	MSE-PSEUDO	2.183 ± 0.317	33.007 ± 6.806	0.014 ± 0.018	0.011 ± 0.015	0.014 ± 0.018	0.277 ± 0.121

Table 5: Single-draft results on Vicuna-7B-v1.5 / Vicuna-68m × ELI5. Each cell: mean ± std across 1000 prompts.

L	Method	AATPS	Token rate	ANLPPT-U	ANLPPT-Li	ANLPPT-PL	LPPL
1	BASIC-UWM	1.000 ± 0.000	38.536 ± 1.896	0.060 ± 0.040	0.044 ± 0.033	0.062 ± 0.042	0.507 ± 0.152
	VSPS	1.410 ± 0.063	46.306 ± 3.178	0.002 ± 0.005	0.003 ± 0.005	0.002 ± 0.005	0.511 ± 0.150
	PFR-NOWM	1.399 ± 0.063	33.636 ± 4.757	0.003 ± 0.006	0.003 ± 0.007	0.002 ± 0.005	0.509 ± 0.152
	PFR	1.397 ± 0.065	35.231 ± 5.397	0.061 ± 0.043	0.045 ± 0.038	0.063 ± 0.042	0.513 ± 0.162
	MSE	1.415 ± 0.062	44.719 ± 3.192	0.020 ± 0.020	0.014 ± 0.017	0.022 ± 0.022	0.509 ± 0.147
	MWS	1.399 ± 0.064	26.309 ± 3.816	0.060 ± 0.041	0.044 ± 0.034	0.062 ± 0.042	0.507 ± 0.151
	MSE-PSEUDO	1.416 ± 0.063	28.987 ± 4.863	0.040 ± 0.030	0.029 ± 0.025	0.042 ± 0.031	0.504 ± 0.152
2	BASIC-UWM	1.000 ± 0.000	38.536 ± 1.896	0.060 ± 0.040	0.044 ± 0.033	0.062 ± 0.042	0.507 ± 0.152
	VSPS	1.589 ± 0.104	47.569 ± 3.837	0.003 ± 0.006	0.003 ± 0.006	0.003 ± 0.006	0.509 ± 0.146
	PFR-NOWM	1.562 ± 0.103	34.678 ± 4.452	0.002 ± 0.005	0.002 ± 0.005	0.002 ± 0.006	0.505 ± 0.147
	PFR	1.558 ± 0.106	34.425 ± 4.996	0.061 ± 0.043	0.044 ± 0.038	0.063 ± 0.042	0.513 ± 0.162
	MSE	1.587 ± 0.105	33.648 ± 4.994	0.012 ± 0.014	0.009 ± 0.012	0.014 ± 0.016	0.509 ± 0.147
	MWS	1.560 ± 0.109	23.481 ± 2.014	0.060 ± 0.040	0.043 ± 0.034	0.062 ± 0.042	0.510 ± 0.157
	MSE-PSEUDO	1.593 ± 0.110	28.595 ± 4.733	0.040 ± 0.032	0.029 ± 0.027	0.041 ± 0.032	0.509 ± 0.152
3	BASIC-UWM	1.000 ± 0.000	38.536 ± 1.896	0.060 ± 0.040	0.044 ± 0.033	0.062 ± 0.042	0.507 ± 0.152
	VSPS	1.670 ± 0.134	45.003 ± 5.451	0.003 ± 0.006	0.002 ± 0.005	0.002 ± 0.006	0.513 ± 0.157
	PFR-NOWM	1.624 ± 0.132	31.214 ± 4.662	0.002 ± 0.005	0.002 ± 0.005	0.002 ± 0.005	0.506 ± 0.149
	PFR	1.626 ± 0.130	35.694 ± 5.835	0.062 ± 0.043	0.045 ± 0.038	0.063 ± 0.042	0.515 ± 0.162
	MSE	1.671 ± 0.136	26.134 ± 3.956	0.011 ± 0.014	0.008 ± 0.012	0.012 ± 0.015	0.499 ± 0.145
	MWS	1.626 ± 0.136	22.536 ± 2.127	0.059 ± 0.040	0.043 ± 0.033	0.061 ± 0.041	0.508 ± 0.152
	MSE-PSEUDO	1.666 ± 0.134	28.611 ± 4.748	0.037 ± 0.028	0.027 ± 0.023	0.039 ± 0.029	0.505 ± 0.148
4	BASIC-UWM	1.000 ± 0.000	38.536 ± 1.896	0.060 ± 0.040	0.044 ± 0.033	0.062 ± 0.042	0.507 ± 0.152
	VSPS	1.697 ± 0.152	43.287 ± 5.333	0.002 ± 0.005	0.002 ± 0.006	0.002 ± 0.005	0.510 ± 0.152
	PFR-NOWM	1.656 ± 0.145	28.914 ± 4.147	0.002 ± 0.005	0.002 ± 0.005	0.002 ± 0.005	0.503 ± 0.149
	PFR	1.656 ± 0.146	32.316 ± 5.921	0.062 ± 0.043	0.045 ± 0.038	0.063 ± 0.042	0.515 ± 0.162
	MSE	1.700 ± 0.154	23.993 ± 3.469	0.011 ± 0.014	0.008 ± 0.012	0.012 ± 0.015	0.514 ± 0.150
	MWS	1.651 ± 0.149	22.198 ± 3.679	0.059 ± 0.039	0.043 ± 0.033	0.062 ± 0.041	0.509 ± 0.153
	MSE-PSEUDO	1.703 ± 0.155	26.004 ± 4.603	0.037 ± 0.030	0.027 ± 0.024	0.039 ± 0.031	0.508 ± 0.147

get/drafter pairs `Qwen2.5-7B-Instruct`/`Qwen2.5-0.5B-Instruct` and `Llama-3.1-8B-Instruct`/`Llama-3.2-1B-Instruct`. Each cell uses n=1000 prompts, lookahead L=4, and draft counts $B \in \{2, 4, 6, 8\}$, with the sampling parameters of Appendix D.1.

Decoders. MPFR is our keyed multi-draft Poisson sampler; GLS denotes the drafter-invariant *InvariantMultiDraftStrategy* of Rowan et al. (2025), run in the same harness on the same prompt set.

Per-cell tables. Tables 6–9 report all quality and watermark-strength metrics per (decoder, B). Columns are AATPS, token rate in tokens/s (TR), ANLPPT-{U, Li, PL}, and LPPL.

Table 6: Multi-draft results on Qwen2.5-7B-Instruct × CNN/DAILYMAIL, L=4.

Decoder	B	AATPS	TR	ANLPPT-U	ANLPPT-Li	ANLPPT-PL	LPPL
MPFR	2	2.797 ± 0.001	25.11 ± 0.22	0.051 ± 0.001	0.036 ± 0.001	0.053 ± 0.001	0.474 ± 0.002
MPFR	4	3.123 ± 0.002	27.50 ± 0.31	0.051 ± 0.001	0.036 ± 0.001	0.053 ± 0.001	0.474 ± 0.002
MPFR	6	3.294 ± 0.003	28.50 ± 0.49	0.050 ± 0.001	0.036 ± 0.001	0.053 ± 0.001	0.474 ± 0.001
MPFR	8	3.409 ± 0.002	29.13 ± 0.59	0.051 ± 0.002	0.036 ± 0.001	0.053 ± 0.002	0.474 ± 0.002
GLS	2	2.803 ± 0.007	24.78 ± 0.14	0.002 ± 0.000	0.003 ± 0.000	0.002 ± 0.000	0.483 ± 0.005
GLS	4	3.109 ± 0.008	27.12 ± 0.20	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.484 ± 0.003
GLS	6	3.272 ± 0.010	27.84 ± 0.60	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.481 ± 0.000
GLS	8	3.384 ± 0.012	28.62 ± 0.54	0.002 ± 0.000	0.003 ± 0.000	0.002 ± 0.000	0.476 ± 0.004

Speculative acceptance and throughput. Across Tables 6–9, MPFR has higher measured TR than GLS in all 16 matched settings and higher AATPS in 15 of 16. The only AATPS exception is Qwen–CNN/DailyMail at $B = 2$, where the values are 2.797 and 2.803, respectively. AATPS captures target-verification efficiency, whereas TR is hardware- and implementation-dependent.

Stability of the watermark signal across B. ANLPPT-U, ANLPPT-Li, and ANLPPT-PL for MPFR vary by at most 0.001 across $B \in \{2, 4, 6, 8\}$ on every cell of Tables 6–9. The per-token detection signal therefore does not dilute as the number of drafts increases.

Table 7: Multi-draft results on Qwen2.5-7B-Instruct × ELI5, L=4.

Decoder	B	AATPS	TR	ANLPPT-U	ANLPPT-Li	ANLPPT-PL	LPPL
MPFR	2	2.563 ± 0.007	23.57 ± 0.15	0.104 ± 0.001	0.075 ± 0.001	0.108 ± 0.001	0.705 ± 0.001
MPFR	4	2.918 ± 0.008	26.46 ± 0.25	0.104 ± 0.001	0.076 ± 0.001	0.108 ± 0.001	0.706 ± 0.000
MPFR	6	3.109 ± 0.005	27.81 ± 0.35	0.103 ± 0.001	0.075 ± 0.001	0.107 ± 0.001	0.705 ± 0.000
MPFR	8	3.235 ± 0.010	28.73 ± 0.49	0.104 ± 0.002	0.075 ± 0.002	0.108 ± 0.001	0.706 ± 0.001
GLS	2	2.523 ± 0.004	22.73 ± 0.22	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.747 ± 0.005
GLS	4	2.850 ± 0.010	25.60 ± 0.26	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.736 ± 0.002
GLS	6	3.039 ± 0.004	27.01 ± 0.32	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.728 ± 0.002
GLS	8	3.156 ± 0.009	28.09 ± 0.36	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.728 ± 0.002

Table 8: Multi-draft results on Llama-3.1-8B-Instruct × CNN/DAILYMAIL, $L = 4$. Mean ± std over 3 random seeds, 1000 prompts each. The drafter is Llama-3.2-1B-Instruct, the official distillation of the target sharing the same tokenizer.

Decoder	B	AATPS	TR	ANLPPT-U	ANLPPT-Li	ANLPPT-PL	LPPL
MPFR	2	3.459 ± 0.015	37.44 ± 0.57	0.058 ± 0.001	0.041 ± 0.000	0.061 ± 0.001	0.514 ± 0.002
MPFR	4	3.784 ± 0.013	40.20 ± 0.28	0.058 ± 0.000	0.041 ± 0.000	0.062 ± 0.001	0.514 ± 0.003
MPFR	6	3.935 ± 0.005	41.71 ± 0.30	0.058 ± 0.000	0.041 ± 0.000	0.061 ± 0.001	0.514 ± 0.003
MPFR	8	4.034 ± 0.005	42.04 ± 0.81	0.058 ± 0.001	0.041 ± 0.000	0.061 ± 0.001	0.514 ± 0.003
GLS	2	3.431 ± 0.005	36.79 ± 0.37	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.514 ± 0.003
GLS	4	3.741 ± 0.007	39.31 ± 0.29	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.508 ± 0.001
GLS	6	3.894 ± 0.007	40.51 ± 0.25	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.507 ± 0.001
GLS	8	3.987 ± 0.010	39.87 ± 0.66	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.505 ± 0.003

Distortion-free audit. Across the four cells, the largest absolute difference between the reported mean LPPL values of MPFR and GLS is 0.042 nats/token (Qwen–ELI5, $B = 2$). We use LPPL only as a finite-sample audit; exact marginal preservation follows from the MPFR construction rather than from this empirical statistic.

D.5 DRAFTER-SUBSTITUTION ABLATION: FULL RESULTS

Drafter conditions. D_0: Qwen2.5-0.5B-Instruct at T=1.0 (default). D_1: Qwen2.5-1.5B-Instruct at T=1.0 (3× scale, model swap). D_2: Qwen2.5-0.5B-Instruct at T=0.5 (sharper drafter). D_3: Qwen2.5-0.5B-Instruct at T=1.5 (more diffuse drafter). The target temperature, watermark key, and all other parameters are held fixed across D_0–D_3.

Detection statistic. For the Aaronson Gamma-tail test, the per-token score is $s_t = -\log(1 - U_t)$, where U_t is the per-token uniform recovered from the watermark code. Under the no-watermark null, the cumulative score $S_T = \sum_{t \le T} s_t$ is $\mathrm{Gamma}(T, 1)$; we declare detection at FPR= α when $\Pr_{H_0}[\mathrm{Gamma}(T, 1) > S_T] \le \alpha$, computed via the regularised upper incomplete gamma. We report α=1% throughout.

Pairwise ROUGE-L under drafter substitution. For each prompt p and decoder d we have one realised output token sequence per drafter condition. We compute ROUGE-L F1 over each pair (D_i, D_j) of drafter conditions on these token sequences, then report the per-prompt minimum over the six pairs (worst pair) averaged across prompts (Table 11). An unbiased verify rule that depends only on the target logits and the watermark key produces identical sequences across all drafters, so this quantity should approach 1.

E MORE ON WATERMARK

In this section, we discuss different watermarks that are related to our scheme.

Table 9: Multi-draft results on Llama-3.1-8B-Instruct × ELI5, $L = 4$. Mean ± std over 3 random seeds, 1000 prompts each.

Decoder	B	AATPS	TR	ANLPPT-U	ANLPPT-Li	ANLPPT-PL	LPPL
MPFR	2	3.103 ± 0.006	34.29 ± 0.35	0.135 ± 0.001	0.099 ± 0.001	0.139 ± 0.001	0.811 ± 0.004
MPFR	4	3.455 ± 0.010	37.79 ± 0.38	0.135 ± 0.002	0.099 ± 0.001	0.139 ± 0.002	0.812 ± 0.003
MPFR	6	3.635 ± 0.001	39.59 ± 0.43	0.135 ± 0.002	0.099 ± 0.001	0.139 ± 0.002	0.812 ± 0.005
MPFR	8	3.749 ± 0.002	39.76 ± 0.80	0.135 ± 0.001	0.098 ± 0.000	0.139 ± 0.001	0.811 ± 0.004
GLS	2	3.080 ± 0.011	33.72 ± 0.60	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.803 ± 0.003
GLS	4	3.408 ± 0.006	37.11 ± 0.12	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.794 ± 0.004
GLS	6	3.588 ± 0.008	39.12 ± 0.31	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.795 ± 0.008
GLS	8	3.700 ± 0.008	39.37 ± 0.58	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000	0.790 ± 0.003

Table 10: **Output quality.** LPPL↓ under target. ROUGE-L↑ vs. gold `highlights`. AATPS↑ accepted target tokens per step.

	Single-draft baselines							*Multi-draft (B=4)*	
Metric	VSPS	PFR-NOWM	BASIC-UWM	MSE	MWS	MSE-PSEUDO	**PFR**	GLS	**MPFR**
LPPL↓	0.474	0.469	0.472	0.470	0.475	0.477	**0.465**	0.484	**0.474**
ROUGE-L↑	0.219	0.222	0.223	0.221	0.223	0.219	**0.223**	0.222	**0.222**
AATPS↑	2.51	2.46	1.00	2.53	2.43	2.52	**2.46**	3.12	**3.13**

E.1 GUMBEL WATERMARK

We use a common notation for the scored positions throughout this subsection. Let $\mathcal{T} := \{m + 1, \dots, n\}$ and $M := |\mathcal{T}| = n - m$. For each $t \in \mathcal{T}$, let $r_t : \mathcal{V} \to (0, 1)$ be the keyed random function, and define

$$U_t := r_t(w_t), \qquad A_t := -\log(1 - U_t).$$

Aaronson's Gumbel watermark score (Aaronson and Kirchner, 2023) is

$$S_{\mathrm{A}} := \sum_{t \in \mathcal{T}} A_t = \sum_{t \in \mathcal{T}} -\log(1 - r_t(w_t)).$$

Under the null hypothesis, i.e., when $y_{1:n}$ is not generated from the watermarked model, the variables $(U_t)_{t \in \mathcal{T}}$ are independent $\mathrm{Unif}(0, 1)$ random variables. Hence $(A_t)_{t \in \mathcal{T}}$ are independent $\mathrm{Exp}(1)$ random variables, and

$$\mathbf{E}\left[S_{\mathrm{A}}\right] = M.$$

If $y_{1:n}$ is generated from the Gumbel-watermarked model, then

$$\mathbf{E}\left[S_{\mathrm{A}}\right] \geq M + \left(\frac{\pi^2}{6} - 1\right) \sum_{t \in \mathcal{T}} \mathbf{E}\left[\mathrm{Entropy}\left[p_t(\cdot)\right]\right],$$

where $p_t = \mathrm{Softmax}(u_t/T)$ and the entropy is measured in nats.

Since $S_{\mathrm{A}} \sim \mathrm{Gamma}(M, 1)$ under the null hypothesis, the corresponding p-value is

$$p_{\mathrm{A}} = \mathbf{P}\big(\mathrm{Gamma}(M, 1) \geq S_{\mathrm{A}}\big) = \frac{\Gamma(M, S_{\mathrm{A}})}{\Gamma(M)},$$

where $\Gamma(M)$ is the gamma function and $\Gamma(M, S_{\mathrm{A}})$ is the upper incomplete gamma function. The Average Negative Log P-value Per Token (ANLPPT) for Aaronson's score is

$$\mathrm{ANLPPT}_{\mathrm{A}} := -\frac{1}{M} \log p_{\mathrm{A}}.$$

Following the U-score framework (Hu and Huang, 2024), one may instead aggregate the keyed uniforms directly:

$$S_{\mathrm{U}} := \sum_{t \in \mathcal{T}} U_t.$$

Table 11: **Output stability under drafter substitution.** Pairwise ROUGE-L F1 between the realised token sequences across each (D_i, D_j) pair, per-prompt minimum averaged over prompts. An unbiased verify rule produces identical sequences across all drafters, so ROUGE-L $\to 1$.

Decoder	min-pair mean ROUGE-L
BASIC-UWM (no drafter, vacuous bound)	1.000
PFR (ours)	**0.990**
MWS	0.922
PFR-NOWM (no wm)	0.436
MSE-PSEUDO	0.426
MSE	0.406
VSPS (no wm)	0.394

Under the null hypothesis, $U_t \sim \mathrm{Unif}(0,1)$. Since for $U \sim \mathrm{Unif}(0,1)$,

$$\mathbf{E}\big[e^{\lambda U}\big] = \frac{e^{\lambda}-1}{\lambda}, \qquad \lambda \geq 0,$$

we have

$$\mathbf{E}\big[e^{\lambda S_{\mathrm{U}}}\big] = \left(\frac{e^{\lambda}-1}{\lambda}\right)^{M}.$$

Chernoff's bound gives

$$p_{\mathrm{U}} \leq \inf_{\lambda \geq 0} \exp\left(M \log \frac{e^{\lambda}-1}{\lambda} - \lambda S_{\mathrm{U}}\right),$$

and the corresponding ANLPPT is

$$\mathrm{ANLPPT}_{\mathrm{U}} := -\frac{1}{M} \log p_{\mathrm{U}}.$$

E.2 CONNECTION TO DELTAGUMBEL REWEIGHTING

We now relate the above notation to the DeltaGumbel reweighting rule used in (Hu and Huang, 2024). For each scored position $t \in \mathcal{T}$, let

$$P_t := P(\cdot \mid w_{:t-1})$$

be the target next-token distribution. Define

$$g_t(v) := -\log\big(-\log r_t(v)\big), \qquad v \in \mathcal{V}.$$

Then $(g_t(v))_{v\in\mathcal{V}}$ are independent standard Gumbel random variables. The DeltaGumbel rule selects

$$a_t^{\star} := \arg\max_{v \in \mathcal{V}} \{\log P_t(v) + g_t(v)\}.$$

Equivalently, for a fixed key, the reweighted distribution at position t is the point mass

$$Q_t := R_{g_t}(P_t) = \delta_{a_t^{\star}}.$$

The likelihood-agnostic U-score of (Hu and Huang, 2024) is

$$U_t := \exp\big(-\exp(-g_t(w_t))\big).$$

By the definition of g_t, this is exactly

$$U_t = r_t(w_t).$$

Therefore, Aaronson's per-token score is a deterministic transform of the DeltaGumbel U-score:

$$A_t = -\log(1 - r_t(w_t)) = -\log(1 - U_t).$$

Thus

$$S_{\mathrm{A}} = \sum_{t \in \mathcal{T}} A_t = \sum_{t \in \mathcal{T}} -\log(1 - U_t).$$

In particular, if the full vector $(U_t)_{t\in\mathcal{T}}$ is retained, then Aaronson's score can be computed exactly from the U-scores, and conversely

$$U_t = 1 - \exp(-A_t).$$

However, the aggregate statistics

$$\sum_{t\in\mathcal{T}} U_t \qquad \text{and} \qquad \sum_{t\in\mathcal{T}} -\log(1-U_t)$$

are not equivalent in general, since they apply different nonlinear transformations before aggregation.

E.3 Watermark Score by (Li et al., 2025)

The score of Li et al. (2025) is also computed from the same pivot $r_t(w_t)$. For a parameter Δ, define

$$k_\Delta := \left\lfloor \frac{1}{1-\Delta} \right\rfloor, \qquad q_\Delta := 1 - k_\Delta(1-\Delta).$$

Their per-token score is

$$h^\star_{\mathrm{gum},\Delta}(r) = \log\left(k_\Delta r^{\frac{\Delta}{1-\Delta}} + \mathbb{1}\{q_\Delta > 0\} r^{\frac{1-q_\Delta}{q_\Delta}}\right),$$

and the corresponding statistic is

$$S_{\mathrm{Li},\Delta} := \sum_{t\in\mathcal{T}} h^\star_{\mathrm{gum},\Delta}(r_t(w_t)).$$

We now discuss another watermark score for Aaronson's Gumbel Watermark, proposed by (Li et al., 2025), which is a statistically optimal score function for the Gumbel watermark under a class-dependent efficiency criterion. We introduce their result using our notation as follows. Their Gumbel pivot is exactly

$$Y_t^{\mathrm{gum}} := r_t(w_t).$$

Under the null hypothesis,

$$Y_t^{\mathrm{gum}} \sim \mathrm{Unif}[0,1].$$

Under the Gumbel-watermarked alternative, conditional on

$$P_t := P(\cdot|w_{:t-1}),$$

the pivot satisfies

$$\mathbf{P}_{H_1}(Y_t^{\mathrm{gum}} \le r|P_t) = \sum_{v\in\mathcal{V}} P_t(v) r^{1/P_t(v)}, \qquad r\in[0,1],$$

where terms with $P_t(v) = 0$ are omitted.

Aaronson's original score corresponds to the per-token score function

$$h_{\mathrm{A}}(r) := -\log(1-r),$$

so that

$$S_{\mathrm{A}} = \sum_{t\in\mathcal{T}} h_{\mathrm{A}}(r_t(w_t)).$$

In contrast, Li et al. (Li et al., 2025) consider the distribution class

$$\mathcal{P}_\Delta := \left\{P : \max_{v\in\mathcal{V}} P(v) \le 1-\Delta\right\}$$

and derive the optimal Gumbel score

$$h^\star_{\mathrm{gum},\Delta}(r) = \log\left(k_\Delta r^{\frac{\Delta}{1-\Delta}} + \mathbb{1}\{q_\Delta > 0\} r^{\frac{1-q_\Delta}{q_\Delta}}\right),$$

where

$$k_\Delta := \left\lfloor \frac{1}{1-\Delta} \right\rfloor, \qquad q_\Delta := 1 - k_\Delta(1-\Delta).$$

The corresponding statistic in our notation is

$$S_{\mathrm{Li},\Delta} := \sum_{t\in\mathcal{T}} h^{\star}_{\mathrm{gum},\Delta}(r_t(w_t)).$$

This score is the log-likelihood ratio associated with the least-favorable distribution

$$P^{\star}_{\Delta} = (\underbrace{1-\Delta,\ldots,1-\Delta}_{k_\Delta \text{ times}}, q_\Delta, 0, \ldots, 0),$$

with the q_Δ term omitted when $q_\Delta = 0$. Therefore, $h^{\star}_{\mathrm{gum},\Delta}$ is not the same as Aaronson's original score h_{A}. Indeed,

$$h_{\mathrm{A}}(r) \to \infty \qquad \text{as } r \uparrow 1,$$

whereas

$$h^{\star}_{\mathrm{gum},\Delta}(r) \to \log\left(k_\Delta + \mathbb{1}\{q_\Delta > 0\}\right) \qquad \text{as } r \uparrow 1.$$

Thus, Li et al.'s score and Aaronson's original score are based on the same Gumbel pivot $r_t(w_t)$, but they lead to different cumulative detection statistics.

E.4 Watermark Score by (Lattimore, 2026)

Lattimore (2026) proposed a refined detector for the Gumbel watermark based on a truncated power-law score. In our notation, the same Gumbel pivot is

$$r_t(w_t), \qquad t \in \mathcal{T}.$$

Instead of Aaronson's original per-token score

$$h_{\mathrm{A}}(r) := -\log(1-r),$$

Lattimore (Lattimore, 2026) considers

$$h_{\mathrm{PL},\epsilon}(r) = \min\left\{\epsilon^{-1/2}, (1-r)^{-1/2}\right\} - (2-\sqrt{\epsilon}),$$

and the corresponding statistic is

$$S_{\mathrm{PL},\epsilon} := \sum_{t\in\mathcal{T}} h_{\mathrm{PL},\epsilon}(r_t(w_t)).$$

Here the centering term $2-\sqrt{\epsilon}$ is chosen so that, under the null hypothesis,

$$\mathbf{E}[h_{\mathrm{PL},\epsilon}(U)] = 0, \qquad U \sim \mathrm{Unif}[0,1].$$

Thus, under the null and assuming independence across the scored positions,

$$\mathbf{E}[S_{\mathrm{PL},\epsilon}] = 0.$$

Both $S_{\mathrm{PL},\epsilon}$ and Aaronson's statistic

$$S_{\mathrm{A}} = \sum_{t\in\mathcal{T}} h_{\mathrm{A}}(r_t(w_t))$$

are model-agnostic statistics based on the same Gumbel pivot $r_t(w_t)$. However, they are not equivalent. In particular,

$$h_{\mathrm{A}}(r) \to \infty \qquad \text{as } r \uparrow 1,$$

whereas

$$h_{\mathrm{PL},\epsilon}(r) \le \epsilon^{-1/2} - (2-\sqrt{\epsilon})$$

is bounded because of the truncation. Therefore, Lattimore's truncated power-law score and Aaronson's original score use the same keyed randomness, but lead to different cumulative detection statistics.

F ALGORITHM IMPLEMENTATION

Algorithm 8: Compute Aaronson score, exact null p-value, and ANLPPT for keyed PFR watermark

Input : output sequence $w_{1:N}$, prompt length n_0, labeler Λ, secret key k, reference measure μ
Output : Aaronson score S_{A}, p-value p_{A}, ANLPPT $\mathrm{ANLPPT}_{\mathrm{A}}$ and $M = |\mathcal{T}|$

$S_{\mathrm{A}} \leftarrow 0$
$M \leftarrow 0$
for $t = n_0 + 1$ **to** N **do**
  $c_t \leftarrow w_{:t-1}$
  $\ell_t \leftarrow \Lambda(c_t)$
  $\Pi_t \leftarrow G(\mathsf{k}, \ell_t)$
  `// recover keyed uniform value for the observed token` w_t
  $\tau_t(w_t) \leftarrow \textsc{FirstArrival}(w_t, \Pi_t, \mu)$
  $r_t \leftarrow \exp(-\mu(w_t)\tau_t(w_t))$
  $S_{\mathrm{A}} \leftarrow S_{\mathrm{A}} - \log(1 - r_t)$
  $M \leftarrow M + 1$
end
`// Under` H_0, $S_{\mathrm{A}} \sim \mathrm{Gamma}(M, 1)$
$p_{\mathrm{A}} \leftarrow \Gamma(M, S_{\mathrm{A}})/\Gamma(M)$
$\mathrm{ANLPPT}_{\mathrm{A}} \leftarrow -\frac{1}{M} \log p_{\mathrm{A}}$

Algorithm 9: Watermarkable Speculative Sampling via Poisson Processes

Input : lookahead K, output length N, target model family P, draft model family Q, initial prefix $w_{1:n}$, common reference measure μ on $\mathcal{Z}$ such that $P(\cdot|c), Q(\cdot|c) \ll \mu$ for all contexts c, labeler Λ, secret key k

while $n < N$ **do**
  $h \leftarrow w_{1:n}$ // freeze block-start prefix
  $L \leftarrow \min\{K,\, N - n\}$
  for $s = 1$ **to** L // draft L tokens from shared PFR sources **do**
    $c_s \leftarrow h \,\|\, \tilde{w}_{1:s-1}$
    $\ell_s \leftarrow \Lambda(c_s)$ // includes repeated-context masking if used
    $\Pi_s \leftarrow G(\mathsf{k}, \ell_s)$
    $\tilde{w}_s \leftarrow \mathrm{PFR}(Q(\cdot|c_s), \Pi_s, \mu)$
  end
  if $n + L < N$ **:**
    $c_{L+1} \leftarrow h \,\|\, \tilde{w}_{1:L}$
  end
  // Prepare target-side evaluations for $c_1, \ldots, c_L$ and, if needed, c_{L+1} in parallel
  $accepted \leftarrow$ **true**
  // Verify each draft token using the same shared Poisson source
  for $s = 1$ **to** L **do**
    $y_s \leftarrow \mathrm{PFR}(P(\cdot|c_s), \Pi_s, \mu)$
    $w_{n+1} \leftarrow y_s$; $n \leftarrow n + 1$ // emit the target PFR winner
    if $y_s \neq \tilde{w}_s$ **:**
      $accepted \leftarrow$ **false**
      break // correction: later draft contexts are invalid
    end
  end
  if $accepted$ ***and*** $n < N$ **:**
    $\ell_{L+1} \leftarrow \Lambda(c_{L+1})$
    $\Pi_{L+1} \leftarrow G(\mathsf{k}, \ell_{L+1})$
    $w_{n+1} \leftarrow \mathrm{PFR}(P(\cdot|c_{L+1}), \Pi_{L+1}, \mu)$
    $n \leftarrow n + 1$ // bonus token
  end
end

Algorithm 10: Watermarkable Multi-Draft Speculative Sampling (Formal)

Given : lookahead L, number of drafts B, output length N, target model P, draft model Q, initial prompt $w_{1:n}$, keyed Poisson source generator G, and secret key k

while $n < N$ **do**
  $h \leftarrow w_{1:n}$; $\ell \leftarrow \min\{L,\, N-n\}$
  $\mathcal{C}_0 \leftarrow \{h\}$; $\nu(h) \leftarrow B$ // root context; all B homogeneous drafts start here
  for $s = 1$ **to** ℓ **do**
    $\mathcal{C}_s \leftarrow \emptyset$
    for $c \in \mathcal{C}_{s-1}$ **do**
      $\Pi(c) \leftarrow G(\mathsf{k}, c)$ // context-indexed keyed source shared by draft and target at context c
      for $i = 1$ **to** $\nu(c)$ **do**
        $\tilde{w}(c,i) \leftarrow \mathrm{MPFR}(Q(\cdot|c), \Pi(c), i)$
      $\mathcal{D}(c) \leftarrow \{\tilde{w}(c,1), \ldots, \tilde{w}(c,\nu(c))\}$ // distinct next-token proposals from context c
      for $u \in \mathcal{D}(c)$ **do**
        $\nu(c \,\|\, u) \leftarrow \big|\{\, i \in \{1,\ldots,\nu(c)\} : \tilde{w}(c,i) = u \,\}\big|$ // how many drafts move to child context $c\|u$
        $\mathcal{C}_s \leftarrow \mathcal{C}_s \cup \{\, c \,\|\, u \,\}$
  for $c \in \mathcal{C}_0 \cup \cdots \cup \mathcal{C}_{\ell-1}$ **do** // compute target-side winner at every internal context
    $y(c) \leftarrow \mathrm{MPFR}(P(\cdot|c), \Pi(c), 1)$ // target-side mapped-Poisson winner at context c
  $c^\star \leftarrow h$; *accepted* $\leftarrow$ **true** // start following the unique realized target path from the root
  for $s = 1$ **to** ℓ **do**
    $a_s \leftarrow \mathbb{1}\{y(c^\star) \in \mathcal{D}(c^\star)\}$ // accept if the target winner is among the drafts at the current context
    $w_{n+1} \leftarrow y(c^\star)$; $n \leftarrow n+1$ // emit the target-side token, not a draft token
    if $a_s = 0$ **:**
      accepted $\leftarrow$ **false** and **break** // first rejection ends the speculative block
    if $n = N$ **:**
      break
    $c^\star \leftarrow c^\star \,\|\, y(c^\star)$ // move to the next realized context
  if *accepted* ***and*** $n < N$ **:** // if all ℓ steps were accepted, emit one extra target token
    $\Pi^\star \leftarrow G(\mathsf{k}, c^\star)$
    $w_{n+1} \leftarrow \mathrm{MPFR}(P(\cdot|c^\star), \Pi^\star, 1)$
    $n \leftarrow n+1$ // bonus token