A Statistical Framework of Watermarks for Large Language Models: Pivot, Detection Efficiency and Optimal Rules

Xiang Li[1] Feng Ruan[2] Huiyuan Wang[1] Qi Long[1] Weijie Su[1]

[1] University of Pennsylvania, [2] Northwestern University.

Is it possible to (reliably) detect LLM-generated text?

Applications:

- Fostering original work in education and maintaining academic integrity
- Preventing fraud and deception
- Preserving the quality of data for training future AI models

Potential methods: Ad hoc methods leverage context, linguistic patterns, and other markers, which are often inaccurate, unreliable, and biased.

However: Worse, as AI models evolve, LLM-generated text increasingly resembles human-written text!

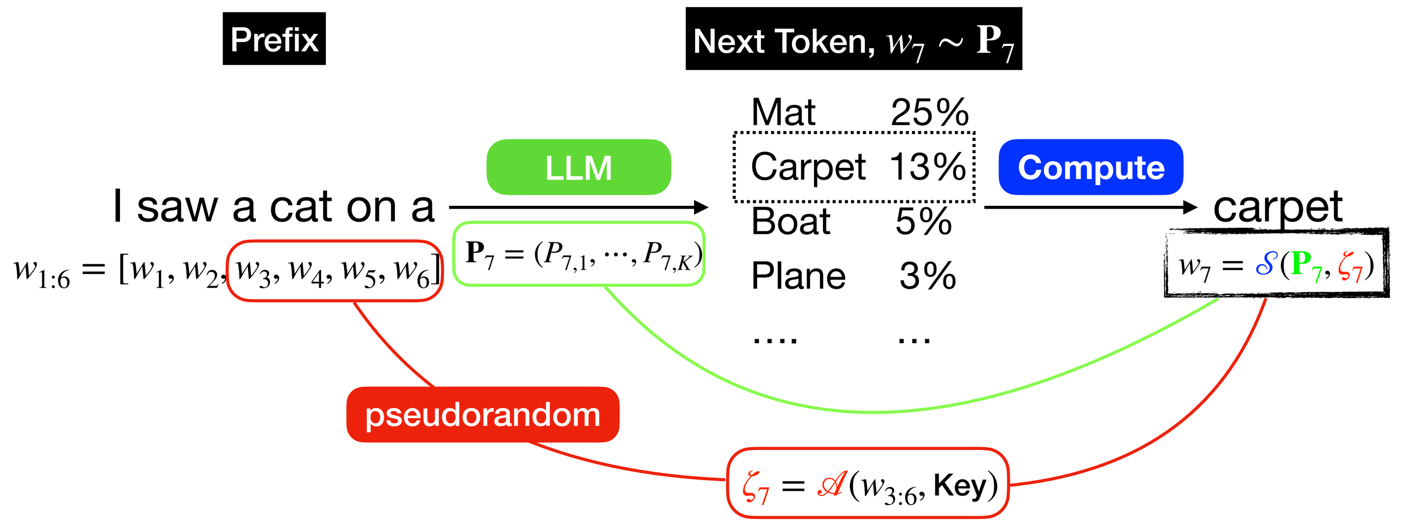
A principled approach: watermarking LLM-generated text

- Watermarking embeds subtle statistical signals into LLM-generated text
- These signal patterns are unlikely to appear in human-written text
- Watermarking enables provable detection of LLM-generated text

Watermark embedding

- The vocabulary $\mathcal{W} = \{1, \ldots, K\}$, a token w_t, and a text $w_{1:(t-1)} := w_1 \cdots w_{t-1}$.
- **Autoregresiveness**: An LLM generates each token sequentially by sampling from a probability distribution conditioned on previous tokens:

  $w_t \sim \boldsymbol{P}_t$ where $\boldsymbol{P}_t = \mathrm{LLM}(w_{1:(t-1)})$ is next-token prediction (NTP) dist. on $\mathcal{W}$.
- **Practical constraint**: $\boldsymbol{P}_t$ is unknown due to (i) limited access to LLMs and (ii) unknown system/user prompts.

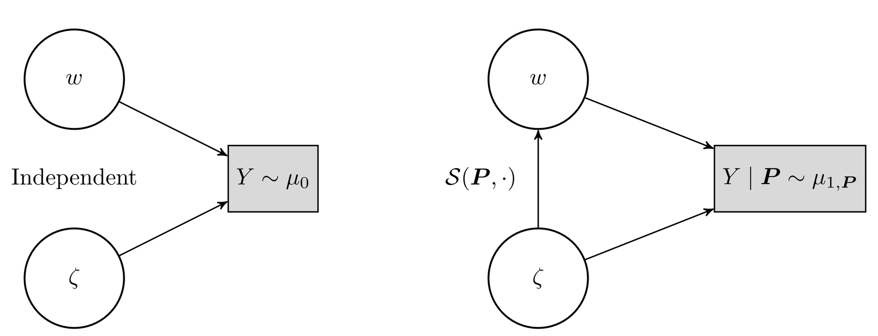


- Mathematical speaking: $w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t)$ where $\zeta_t = \mathcal{A}(w_{1:(t-1)}, \mathrm{Key})$.
- $\zeta_{1:t} := \zeta_1 \cdots \zeta_t$ is theoretically i.i.d. and practically recoverable.
- A watermark is defined by $(\mathcal{A}, \mathcal{S}, \mathrm{Key})$.
- Watermark signal is the dependence of each w_t on ζ_t.

Watermark detection

Pivotal statistic: Find a scalar function $Y_t = Y(w_t, \zeta_t)$ so that

- H_0: $Y_t \sim \mu_0$, regardless of $\boldsymbol{P}_t$
- H_1: $Y_t \sim Y(\mathcal{S}(\zeta_t, \boldsymbol{P}_t), \zeta_t) \overset{d}{=} \mu_{1,\boldsymbol{P}_t}$



Problem formulation: $H_0 : Y_t \overset{iid}{\sim} \mu_0, \ \forall t$ vs $H_1 : Y_t | \boldsymbol{P}_t \sim \mu_{1,\boldsymbol{P}_t}, \ \forall t$.

Detection rule: Test score $T_h = \sum_{t=1}^n h(Y_t)$ for some score function h. Reject H_0 if T_h is larger than a threshold.

Two considered watermarking schemes

- A watermark corresponds to sampling from the NTP distribution.
- **Gumbel-max watermark**: $\zeta = (U_1, U_2, \ldots U_K)$ contains iid $\mathcal{U}(0,1)$,

$$\mathcal{S}^{\mathrm{gum}}(\boldsymbol{P}, \zeta) = \arg\max_{w \in \mathcal{W}} \frac{\log U_w}{P_w}, \quad Y^{\mathrm{ars}}(w, \zeta) = U_w.$$

  Two scores: $h_{\mathrm{ars}}(y) = -\log(1-y) > h_{\log}(y) = \log(y)$.
- **Inverse transform watermark**: Let $\zeta = (\pi, U)$ with $U \sim \mathcal{U}(0,1)$ and π being sampled uniformly at random from all permutations on $\mathcal{W}$.

$$\mathcal{S}_{\mathrm{inv}}(\boldsymbol{P}, \xi) := \pi^{-1}(F^{-1}(U; \pi)), \quad Y^{\mathrm{dif}}(w, \zeta) = |U - \eta(\pi(w))|,$$

  where $\eta(w) := \frac{w-1}{|\mathcal{W}|-1}$ and $F(x; \pi) = \sum_{w'} P_{w'} \cdot \mathbf{1}_{\{\pi(w') \le x\}}$. One uses $h_{\mathrm{neg}}(y) = -y$.

Class-dependent detection efficiency

Questions: (i) How to theoretically compare different score functions and (ii) What is the "optimal" score function?

Class-dependent efficiency: (i) Select a class $\mathcal{P}$ that is believed to contain all $\boldsymbol{P}_{\le n}$, and (ii) Evaluate efficiency by the least-favorable power attained over $\mathcal{P}$.

Prior class: From empirical study, we choose Δ-regular set: $\mathcal{P}_\Delta := \{\boldsymbol{P} = (P_1, \cdots, P_K) : \max_w P_w \le 1 - \Delta\}$.

Formal definition: Fixing Type I error in $(0,1)$, the pivot-based test statistic $T_h = \sum h(Y_t)$ satisfies

$$\lim_{n \to \infty} \sup \sup_{\text{all} \boldsymbol{P}_t \in \mathcal{P}} [\text{Type II error}]^{\frac{1}{n}} \le \exp(-R_{\mathcal{P}}(h)),$$

where $\mathcal{P}$-efficiency rate $R_{\mathcal{P}}(h)$ is defined as

$$R_{\mathcal{P}}(h) = -\inf_{\theta \ge 0} \left\{ \theta \mathbb{E}_0[h(Y)] + \sup_{\boldsymbol{P} \in \mathcal{P}} \mathbb{E}_{1,\boldsymbol{P}} \log \left([\mathrm{e}^{-\theta h(Y)}] \right) \right\}.$$

Optimal score functions

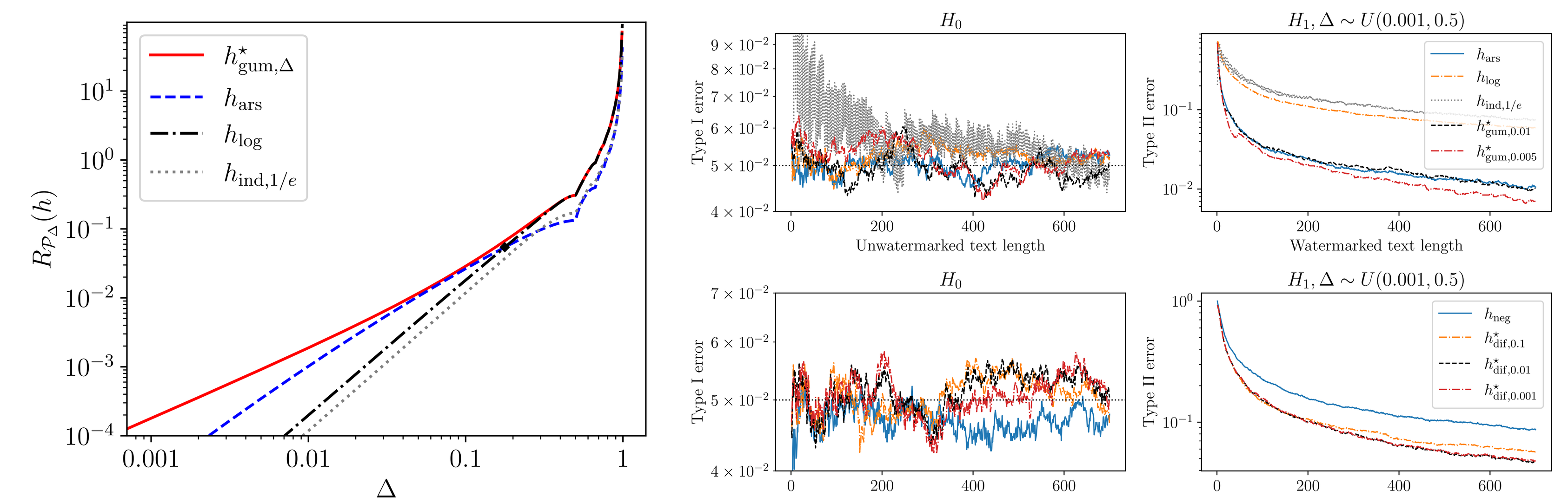
Finding the optimal score $h^\star = \arg\max_h R_{\mathcal{P}}(h)$ reduces to a minimax problem: $\min_h \max_{\boldsymbol{P} \in \mathcal{P}} L(h, \boldsymbol{P})$ where $L(h, \boldsymbol{P}) := \mathbb{E}_0[h(Y)] + \log \left(\mathbb{E}_{1,\boldsymbol{P}} \, \mathrm{e}^{-h(Y)} \right)$.

- The minimax problem is generally not convex-concave. Case-by-case analysis.
- If there exists an $\boldsymbol{P}^\star \in \mathcal{P}$ and a score function class $\mathcal{H}$ such that for all $h \in \mathcal{H}$,

$$h^\star := \log \frac{\mathrm{d}\mu_{1,\boldsymbol{P}^\star}}{\mathrm{d}\mu_0} \in \mathcal{H}, \qquad \sup_{\boldsymbol{P} \in \mathcal{P}} \mathbb{E}_1[\mathrm{e}^{-h(Y)} | \boldsymbol{P}] = \mathbb{E}_1[\mathrm{e}^{-h(Y)} | \boldsymbol{P}^\star],$$

  we then have $\max_h R_{\mathcal{P}}(h) = L(h^\star, \boldsymbol{P}^\star) = D_{\mathrm{KL}}(\mu_0, \mu_{1,\boldsymbol{P}^\star})$, where the maximum is obtained at $h^\star$.
- **Main results**: (i) For Gumbel-max, $h_{\mathrm{gum},\Delta}(y) = \log \frac{\mathrm{d}\mu_{1,\boldsymbol{P}^\star_\Delta}}{\mathrm{d}\mu_0}(y)$ with $\boldsymbol{P}^\star_\Delta = \left(1-\Delta, \ldots, 1-\Delta, 1-(1-\Delta) \cdot \left\lfloor \frac{1}{1-\Delta} \right\rfloor, 0, \ldots\right)$ and (ii) For inverse transform, $h^\star_{\mathrm{dif},\Delta}(r) = \log \frac{f_{\mathrm{dif},\Delta}(r)}{f_{\mathrm{dif},0}(r)}$ where $f_{\mathrm{dif},\Delta}(r) = \frac{2}{1-\Delta} \cdot \max\left\{1 - \frac{r}{1-\Delta}, 0\right\}$ when $K \to \infty$.

Simulation experiments



LLM experiments