Universally Optimal Watermarking Schemes for LLMs: from Theory to Practice

Haiyun He[1,*] (hh743@cornell.edu) Yepeng Liu [2,*] Ziqiao Wang[3] Yongyi Mao[4] Yuheng Bu[2]

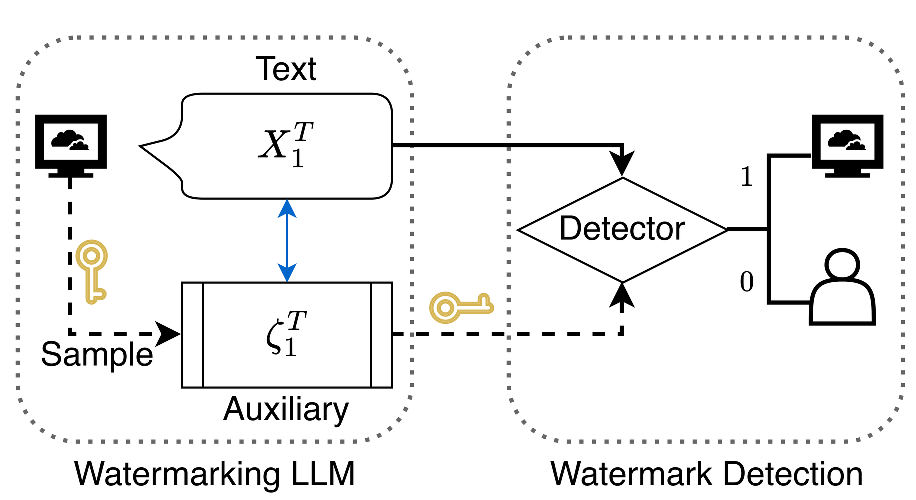
[1]Cornell University [2]University of Florida [3]Tongji University [4]University of Ottawa

Key Takeways

Jointly optimize the **watermarking** scheme and the **detector**:

- **Universally minimum** Type-II error $\Leftrightarrow$ fundamental trade-off between **detectability, distortion, and robustness**
- Theory to Practice: **practical token-level watermarking scheme** $\Rightarrow$ small Type-II error, worst-case Type-I error $\le \alpha$, robust, **model-agnostic**, computationally **efficient**
- **Experiments** (Llama2-13B, Mistral-8×7B) on multiple datasets $\Rightarrow$ **High detection accuracy** and robust to token replacement
- Universal optimal watermarking with **robustness against semantic-invariant attacks** $\Rightarrow$ guideline for future design

Watermarking LLM



Motivation: Risk of spreading disinformation, plagiarism $\Rightarrow$ distinguish AI-generated text from human-written one.

- Human text: $X_t \sim Q_{X_t|x_1^{t-1}}$ (NTP distribution)
- Watermarked text: $X_t \sim P_{X_t|x_1^{t-1},\zeta_t}$, dependent on auxiliary ζ_t
- Secret key (shared with detector) $\xrightarrow{\text{sample}} \zeta_1^T$

  e.g. $\zeta_t \leftarrow$ Random(seed = hash(x_{t-1}, key))
- **Watermarking scheme**: joint distrib. $P_{X_1^T,\zeta_1^T}$
- **ϵ-distorted**: distortion between text distrib.

$$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$$

$\Rightarrow \epsilon = 0$: **distortion-free** (ideal)

e.g. Gumbel-Max (Aaronson, 2023), EXP-edit (Kuditipudi et al., 2023)

Watermark Detection

Receive shared key and $X_1^T \xrightarrow{\text{recover}} \zeta_1^T$:

Watermarked text $X_1^T \not\perp\!\!\!\perp$ auxiliary ζ_1^T

v.s. Human text $X_1^T \perp\!\!\!\perp$ auxiliary ζ_1^T

$\Rightarrow$ Watermark detection = Hypothesis testing:

- H_0: human generated, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$;
- H_1: watermarked LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T,\zeta_1^T}$.

Any model-agnostic detector $\gamma : \mathcal{V}^T \times \mathcal{Z}^T \to \{0,1\}$ $\to$ performance metrics:

Type-I: $\beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) := (Q_{X_1^T} \otimes P_{\zeta_1^T})(\gamma(X_1^T,\zeta_1^T) \neq 0)$,

Type-II: $\beta_1(\gamma, P_{X_1^T,\zeta_1^T}) := P_{X_1^T,\zeta_1^T}(\gamma(X_1^T,\zeta_1^T) \neq 1)$.

Goal: jointly optimize watermark and detection

$$\inf_{\gamma, P_{X_1^T,\zeta_1^T}} \beta_1(\gamma, P_{X_1^T,\zeta_1^T})$$

$$\text{s.t. } \underbrace{\sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha}_{\text{guarantee worst-case Type-I}}, \quad \underbrace{\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon}_{\epsilon\text{-distorted}}.$$

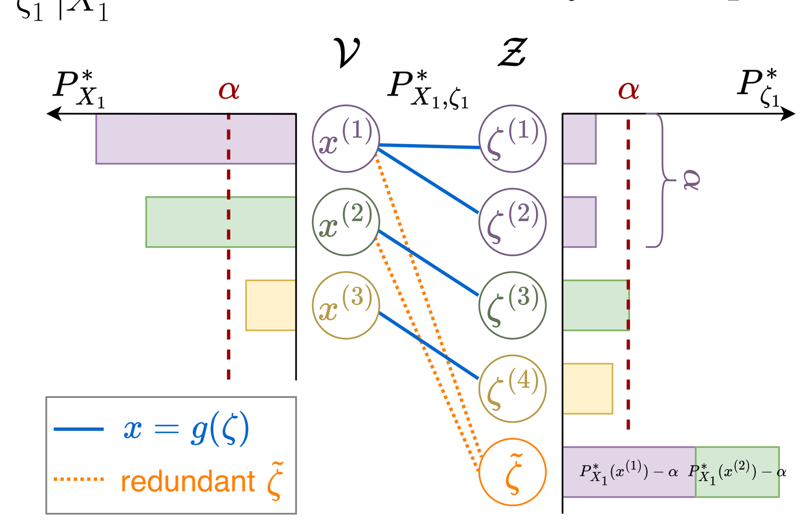
$\xrightarrow{\text{Result}}$ *universally minimum Type-II error* β_1^*

Universally Optimal Watermarking and Detection

- Optimal detector: $\gamma^*(X_1^T, \zeta_1^T) = \mathbb{1}\{X_1^T = g(\zeta_1^T)\}$, ($g$ surjective)
- Optimal watermarking scheme:

$$P^*_{X_1^T} = \mathop{\arg\min}_{P_{X_1^T}:\mathsf{D}(P_{X_1^T},Q_{X_1^T})\le\epsilon} \sum_{x_1^T}(P_{X_1^T}(x_1^T) - \alpha)_+$$

and $P^*_{\zeta_1^T|X_1^T}$ illustrated in the toy example:



1. Perform well on low-entropy text.
2. Distortion level controllable.

Theorem 1 (Universally minimum Type-II error)

$$\beta_1^*(Q_{X_1^T}, \alpha, \epsilon) = \min_{P_{X_1^T}:\mathsf{D}(P_{X_1^T},Q_{X_1^T})\le\epsilon} \sum_{x_1^T}(P_{X_1^T}(x_1^T) - \alpha)_+ \Rightarrow \text{Trade-off: distortion} \uparrow, \text{detection error} \downarrow$$

Experiments

Table: Watermark detection performance across different LLMs and datasets.

LLMs	Methods	C4 ROC-AUC	C4 TPR@1% FPR	C4 TPR@10% FPR	ELI5 (lower entropy) ROC-AUC	ELI5 (lower entropy) TPR@1% FPR	ELI5 (lower entropy) TPR@10% FPR
Llama-13B	KGW-1	0.995	0.991	1.000	0.989	0.974	0.986
	EXP-edit	0.986	0.968	0.996	0.983	0.960	0.995
	Gumbel-Max	0.996	0.993	0.994	0.999	0.991	0.994
	Ours	0.999	0.998	1.000	0.998	0.997	1.000
Mistral-8 × 7B	KGW-1	0.997	0.995	1.000	0.993	0.983	0.994
	EXP-edit	0.993	0.970	0.997	0.994	0.972	0.996
	Gumbel-Max	0.994	0.989	0.999	0.987	0.970	0.990
	Ours	0.999	0.998	1.000	0.999	0.999	1.000

Table: Watermark detection performance under token replacement attack.

LLMs	Methods	C4 ROC-AUC	C4 TPR@1% FPR	C4 TPR@10% FPR	ELI5 (lower entropy) ROC-AUC	ELI5 (lower entropy) TPR@1% FPR	ELI5 (lower entropy) TPR@10% FPR
Llama-13B	KGW-1	0.965	0.833	0.952	0.973	0.892	0.973
	EXP-edit	0.973	0.857	0.978	0.967	0.889	0.975
	Gumbel-Max	0.776	0.396	0.551	0.733	0.326	0.556
	Ours	0.989	0.860	0.976	0.995	0.969	0.994
Mistral-8 × 7B	EXP-edit	0.980	0.861	0.975	0.983	0.932	0.988
	Ours	0.990	0.881	0.966	0.993	0.991	0.995

Practical Token-Level Watermarking Scheme

Detector: $\gamma(X_1^T, \zeta_1^T) = \mathbb{1}\Big\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}\{h_{\text{key}}(X_t) = \zeta_t\} \ge \lambda\Big\}$

Algorithm Watermarked Text Generation

Require: LLM Q, Vocabulary $\mathcal{V}$, Prompt u, Secret key, Token-level false alarm $\eta \in (0, \min\{1, (\alpha/\binom{T}{\lceil T\lambda\rceil})^{\frac{1}{\lceil T\lambda\rceil}}\}]$.

1: $\mathcal{Z} = \{h_{\text{key}}(x)\}_{x\in\mathcal{V}} \cup \tilde{\zeta}$
2: **for** $t = 1, \ldots, T$ **do**
3: Construct $P_{\zeta_t|x_1^{t-1},u}(\zeta)$ using $(Q, \eta, \mathcal{Z})$
4: $(G_{t,\zeta})_{\zeta\in\mathcal{Z}} \leftarrow$ Gumbel(seed=hash(x_{t-n}^{t-1}, key)).
5: $\zeta_t \leftarrow \arg\max_{\zeta\in\mathcal{Z}} \log(P_{\zeta_t|x_1^{t-1},u}(\zeta)) + G_{t,\zeta}$.
6: **if** $\zeta_t \neq \tilde{\zeta}$ **then** $x_t \leftarrow h^{-1}_{\text{key}}(\zeta_t)$
7: **else** Sample $x_t \sim \left(\frac{(Q_{X_t|x_1^{t-1},u}(x)-\eta)_+}{\sum_{x\in\mathcal{V}}(Q_{X_t|x_1^{t-1},u}(x)-\eta)_+}\right)_{x\in\mathcal{V}}$
8: **end for**

Ensure: Watermarked text $x_1^T = (x_1, \ldots, x_T)$.

Algorithm Watermarked Text Detection

Require: SLM $\tilde{Q}$, Vocabulary $\mathcal{V}$, Text x_1^T, Secret key, Threshold λ, Token-level false alarm η.

1: score = 0, $\mathcal{Z} = \{h_{\text{key}}(x)\}_{x\in\mathcal{V}} \cup \tilde{\zeta}$
2: **for** $t = 1, \ldots, T$ **do**
3: Construct $P_{\zeta_t|x_1^{t-1},u}(\zeta)$ using $(\tilde{Q}, \eta, \mathcal{Z})$
4: $(G_{t,\zeta})_{\zeta\in\mathcal{Z}} \leftarrow$ Gumbel(seed=hash(x_{t-n}^{t-1}, key)).
5: $\zeta_t \leftarrow \arg\max_{\zeta\in\mathcal{Z}} \log(P_{\zeta_t|x_1^{t-1}}(\zeta)) + G_{t,\zeta}$.
6: score $\leftarrow$ score $+\mathbb{1}\{h_{\text{key}}(x_t) = \zeta_t\}$
7: **end for**
8: **if** score $> T\lambda$ **then**
9: **return** 1 {Input text is watermarked}
10: **else**
11: **return** 0 {Input text is unwatermarked}
12: **end if**

- Surrogate language model (SLM) is a much **smaller** language model $\xrightarrow{\text{obtain}} \tilde{Q}$ without prompt.
- Type-II error decays exponentially under certain condition, worst-case Type-I error $\le \alpha$