HeavyWater and SimplexWater: Distortion-Free LLM Watermarks for Low-Entropy Distributions

Dor Tsur*[2], Carol Long*[1], Claudio Verdun[1], Hsiang Hsu[3], Richard Chen[3], Haim Permuter[2], Sajani Vithana[1], Flavio P. Calmon[1]

[1]Harvard University [2]Ben Gurion University [3]JPMorgan Chase Bank (*Equal Contributions)

NEURAL INFORMATION PROCESSING SYSTEMS

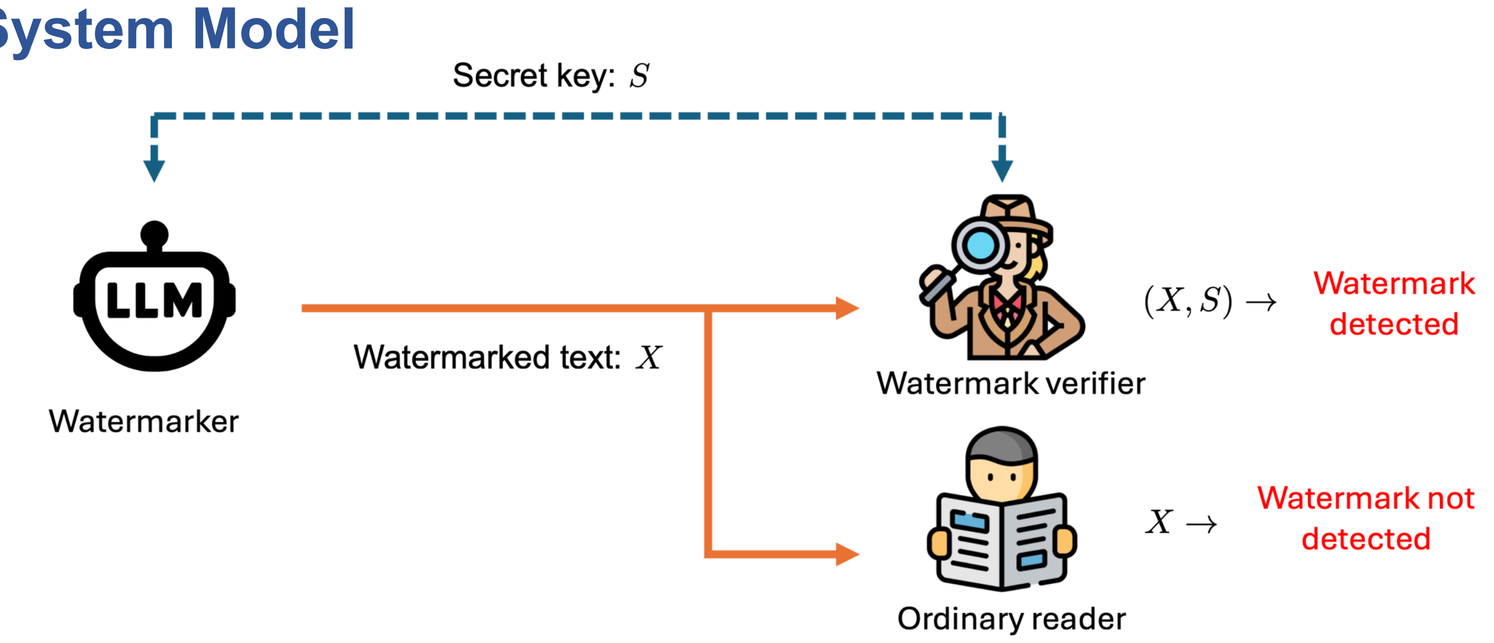
Harvard John A. Paulson School of Engineering and Applied Sciences

JPMorgan Chase & Co.

Background: Watermarking LLMs

Autoregressive token generation → Token-Level watermark

System Model



Watermarking Framework

Watermark Distribution

Given LLM distribution P_X, shared randomness $s \in \mathcal{S}$

Watermark: design *watermarked distribution* $P_{X|S}$

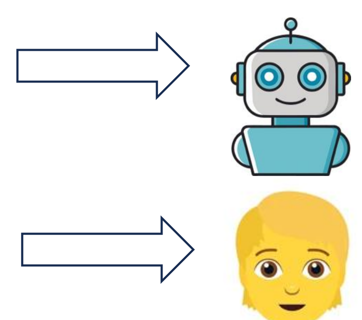
Distortion Free: $\mathbb{E}_S[P_{X|S}] = P_X$

Score Function

Detect: apply a statistical test via a **score function**

$f(x, s) : \mathcal{X} \times \mathcal{S} \rightarrow \mathbb{R}$

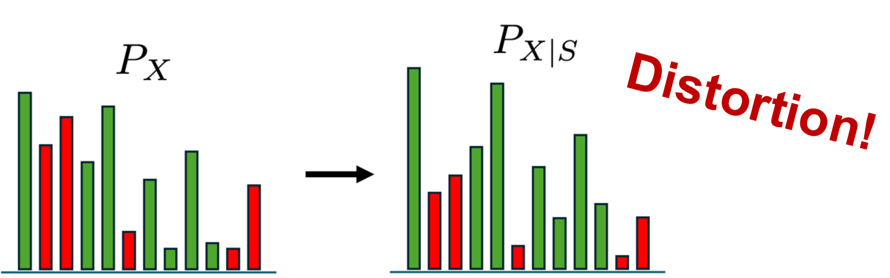
$$\frac{1}{n}\sum_{t=1}^{n} f(x_t, s_t) \quad \begin{cases} \geq \tau \\ < \tau \end{cases}$$

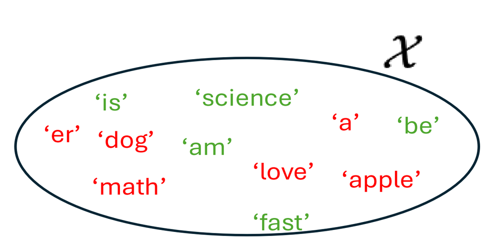


Example: Red-Green Watermark (Kirchenbauer et. al 23')

Upweight Green tokens

Shared Randomness: Random Binary partition





Score function – green list membership

$$f(x, s) = \mathbf{1}\{x \in \text{GreenList}(s)\}$$

TL;DR: Distortion-free watermarks via optimal transport and coding theory

Maximize Detection for Worst-Case Token Distribution

Maximize detection ⟷ **Maximize gap in expected score**

$$D_{\text{gap}} = \underbrace{\mathbb{E}_{P_S P_{X|S}}[f(X, S)]}_{\text{Watermarked}} - \underbrace{\mathbb{E}_{P_S P_X}[f(X, S)]}_{\text{Unwatermarked}}$$

Inner optimization is an **Optimal transport** problem!

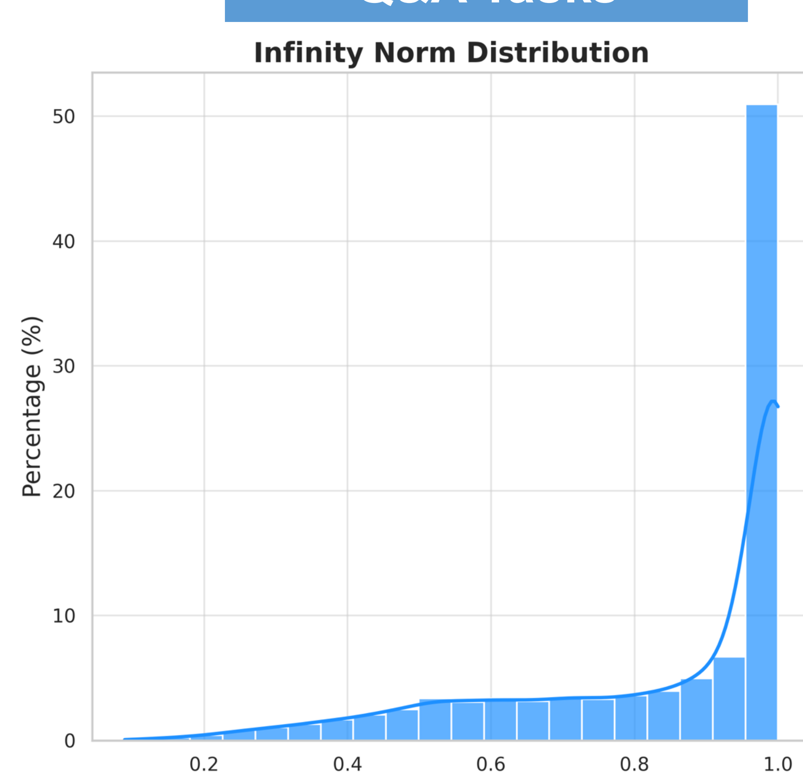
$$\max_{f \in \mathcal{F}} \ \min_{P_X \in \mathcal{P}_\lambda} \ \max_{\substack{P_{X,S} \\ \mathbb{E}_S[P_{X|S}] = P_X}} \ \mathbb{E}_{P_S P_{X|S}}[f(X, S)] - \mathbb{E}_{P_S P_X}[f(X, S)]$$

Score function | Worst-case LLM dist. | Watermarked Distribution

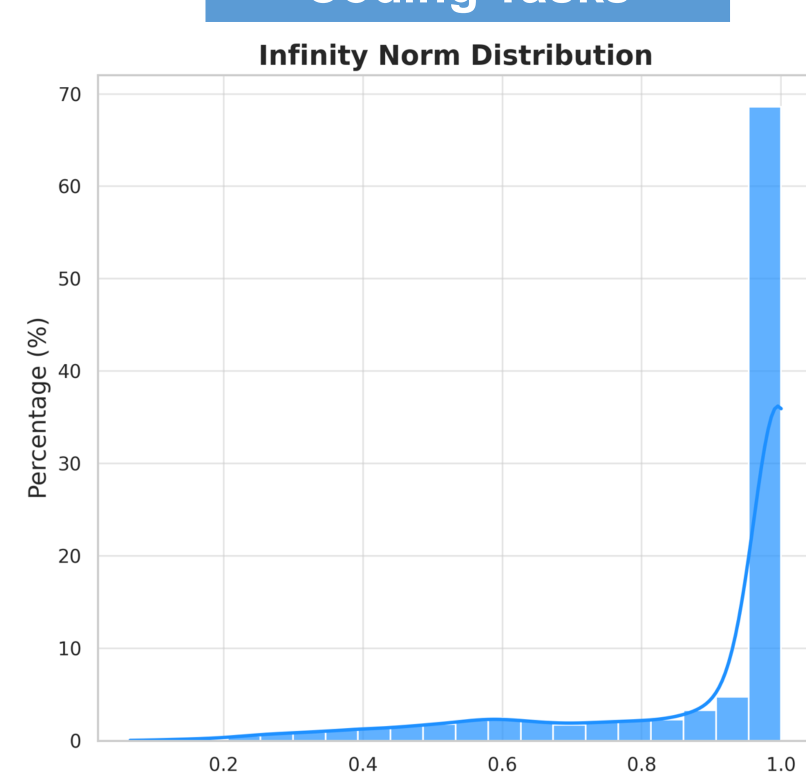
Min-Entropy Constraint $\mathcal{P}_\lambda = \left\{P \in \Delta_m : \max_x P(x) \leq \lambda\right\}, \lambda \in \left[\frac{1}{2}, 1\right]$

Worst-Case Token Distribution?

Q&A Tasks



Coding Tasks



Over 90% of LLM output distributions are dominated by a single token with probability above 0.5!

SimplexWater: Optimal Among All Binary-Score Watermarks

Optimal Transport: For a given s, perturb P_X to get $P_{X|S=s}$ that favors larger scores.

	s_1	s_2	...	s_k
x_1	1	0	0	1
x_2	0	0	1	1
x_3	0	0	1	0
...	1	1	0	0
x_n	0	1	0	1

Equivalence to coding theory:

Maximizing detection gap ⟷ **Designing a distance-maximizing code**

$$D_{\text{gap}} \leq \frac{|\mathcal{S}|(1-\lambda)}{2(|\mathcal{S}|-1)}$$

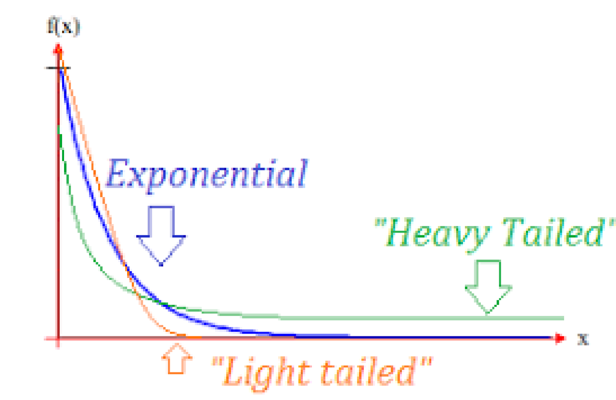
Proof uses Plotkin Bound (1960)! Simplex code **achieves** the bound.

HeavyWater: Better Detection with Continuous Scores

Scores sampled i.i.d. from P_F

Gumbel watermark is a special case of this setting!

	s_1	s_2	s_3	s_4
x_1	0.1	2.1	0.6	10
x_2	6.5	3.7	1.1	0.5
x_3	0.7	11	19	0.1
x_4	1.2	3.5	92	0.7
x_5	0.9	1.1	1.4	11



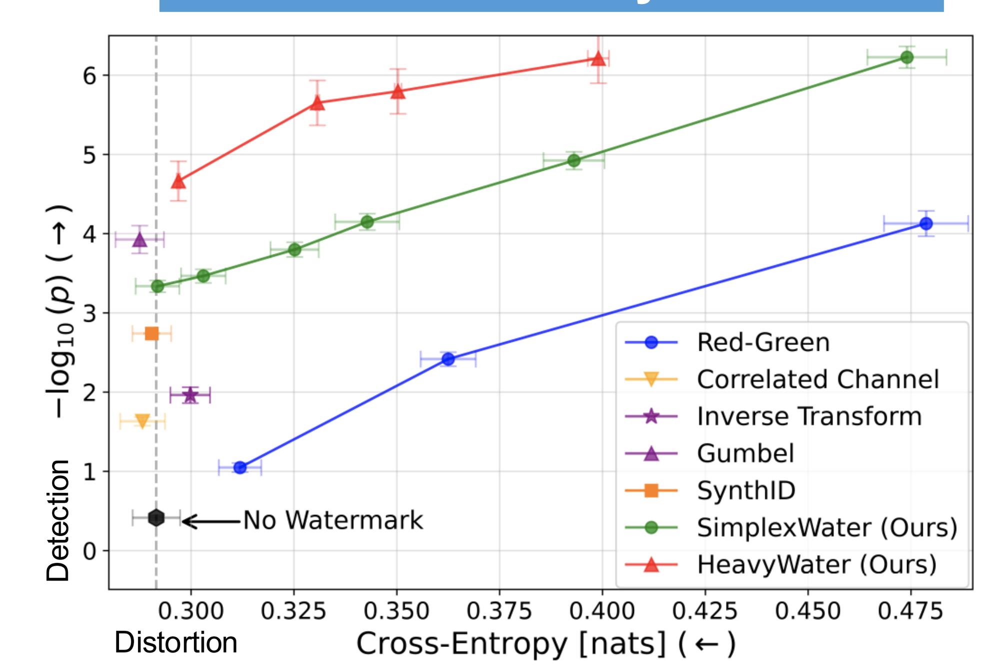
Heavier Tail of P_F ⟷ Higher detection gap

$$\lim_{|\mathcal{S}| \rightarrow \infty} D_{\text{gap}}(P_F, |\mathcal{S}|, \lambda) = \int_{1-\lambda}^{1} Q(u)du$$

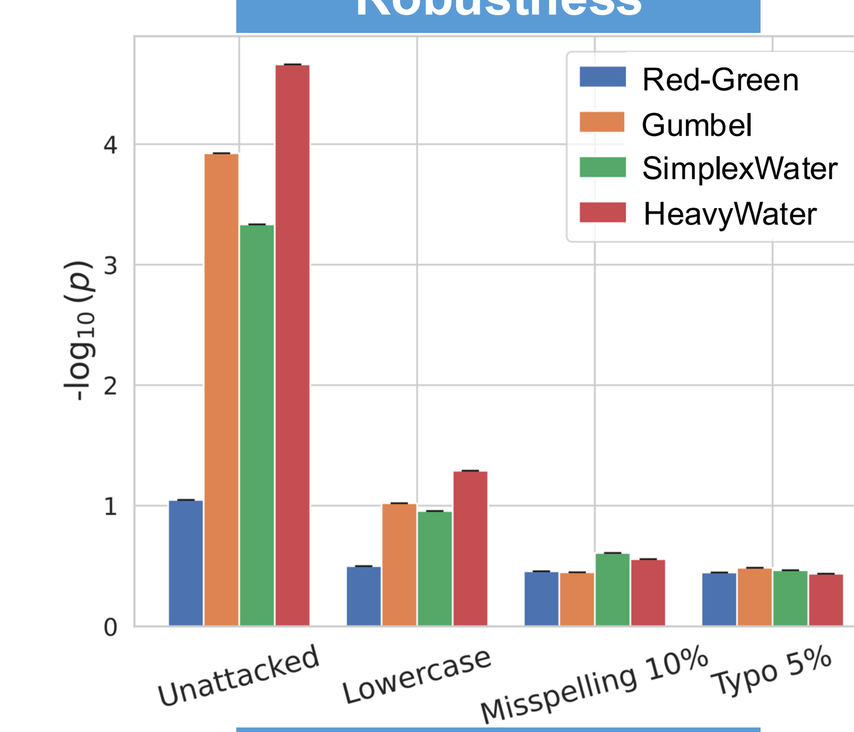
The tail of the quantile function of the i.i.d. difference distribution under P_F

Empirical Results: Superior Detection-Quality Tradeoff

Detection- Quality Tradeoff



Robustness



Watermark	Coding pass@5 (%)	Coding pass@10 (%)
None	20.7	28.1
HeavyWater	**18.9**	**27.8**
SimplexWater	**22.7**	**25.3**
Gumbel	20.5	25.6
Red/Green	19.5	23.2

Overhead