Language Model Watermarking

Detecting and Designing Statistical Dependence

Xiang Li
University of Pennsylvania

Seminar on Statistics and Data Science
HKUST Department of Mathematics · 9 November 2026

Why the origin of text matters

EDUCATION

UK universities warned to 'stress-test' assessments as 92% of students use AI

The Guardian · 26 February 2025

RESEARCH AND PEER REVIEW

Is ChatGPT corrupting peer review? Telltale words hint at AI use

Nature · 2024

NEWS AND MEDIA

Chicago Sun-Times prints summer reading list full of fake books

Ars Technica · May 2025

ONLINE REVIEWS

AI can write a more believable restaurant review than a human can

Yale Insights · 28 May 2024

EU AI Act: marking AI-generated text

ARTICLE 50(2) · TRANSPARENCY OBLIGATIONS

"... outputs of the AI system are marked in a **machine-readable format** and **detectable** as artificially generated or manipulated."

Regulation (EU) 2024/1689, Article 50(2)

European Commission

COVERS

Synthetic **text**, images, audio, and video

APPLIES FROM

2 August 2026

CAVEATS

Feasibility limits, exceptions

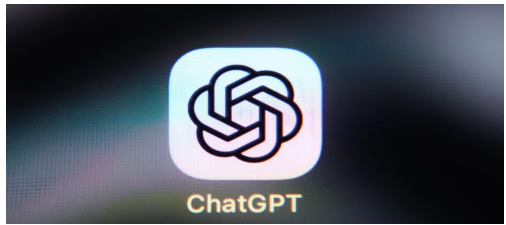
Text watermarking in industry



GOOGLE DEEPMIND · SYNTHID

Watermarking AI-generated text and video with SynthID

Google DeepMind blog · 14 May 2024

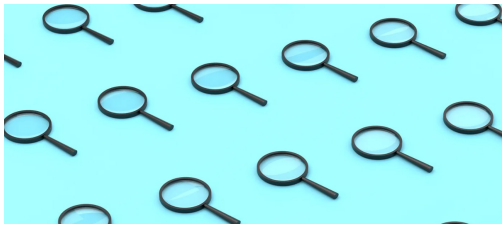


OPENAI · TEXTGRAIN

OpenAI will watermark ChatGPT outputs by default—but only in the EU

Ars Technica · 6 October 2026

Text watermarking in industry



ANTHROPIC · CLAUDE

Anthropic shares more details about how Claude's new watermarks will work

TechCrunch · 15 August 2026

Meta

META · TEXTSEAL FAMILY

Meta's TextSeal2 improves detection of AI-generated text with new watermarking technique

Crypto Briefing · 2 November 2026

Two lines of research

PART I

Statistical inference

Given a watermark

- Detection
- Fraction estimation
- Localization

PART II

Watermark design

Choose a watermark

- Detectability and diversity
- Entropy-calibrated coupling
- textGrain via optimal transport

PART I

Statistical inference for a given watermark

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

Annals of Statistics, 2025

Robust Detection of Watermarks for Large Language Models under Human Edits

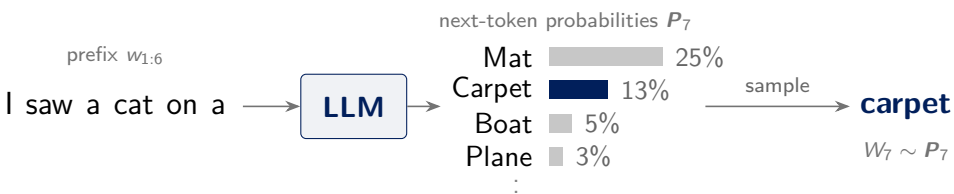
Journal of the Royal Statistical Society, Series B, 2026

With Feng Ruan, Huiyuan Wang, Qi Long, and Weijie Su

Autoregressive generation

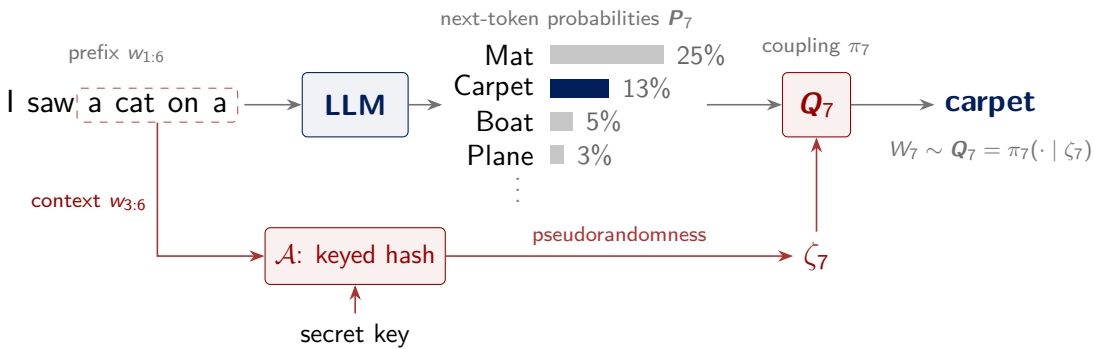
Vocabulary $\mathcal{V}$ of size $K = |\mathcal{V}|$, and W_t is the next token.

At step t: prefix $w_{1:t-1} \longrightarrow$ model probabilities $\boldsymbol{P}_t \longrightarrow$ sample W_t.



Autoregressive generation with a watermark

Keyed pseudorandomness changes the sampling step.



Unbiased watermarking is a coupling

At a fixed prefix, $W_t \in \mathcal{V}$ is the next token and $\boldsymbol{P}_t$ its distribution (Li et al.2025a).

$$\zeta_t \sim \mu, \qquad W_t \mid \zeta_t \sim \boldsymbol{Q}_t = \pi_t(\cdot \mid \zeta_t).$$

Here ζ_t is keyed pseudorandomness and $\boldsymbol{Q}_t$ is the (modified) sampling law.

Definition (Unbiasedness)

$$\mathbb{E}_{\zeta_t}[\boldsymbol{Q}_t] = \mathbb{E}_{\zeta_t}[\pi_t(\cdot \mid \zeta_t)] = \boldsymbol{P}_t.$$

- The joint law $\pi_t = \mathcal{L}(W_t, \zeta_t)$ is a **coupling** of $\boldsymbol{P}_t$ and μ.
- In essence, a coupling is a **sampling method**: draw $\zeta_t \sim \mu$, then $W_t \sim \boldsymbol{Q}_t$.
- **The signal is dependence between the token and the pseudorandomness.**

A concrete example: the Gumbel-max watermark (Aaronson2023)

Generate one keyed uniform value for each candidate token:

$$\zeta_t = (U_{t,w})_{w \in \mathcal{V}}, \qquad U_{t,w} \overset{\text{iid}}{\sim} \text{Unif}(0,1).$$

Gumbel-max sampling trick (Gumbel1948; Maddison et al.2014)

$$W_t = \arg\max_{w: P_{t,w} > 0} \frac{\log U_{t,w}}{P_{t,w}}.$$

- **Unbiased:** $\mathbb{P}_{\zeta_t}(W_t = w) = P_{t,w}$, averaging over ζ_t.
- **Dependent:** the selected value U_{t,W_t} tends to be large.
- **Fixed key and context:** the next token is deterministic.

Comments on pseudorandom numbers

$$\zeta_t = \underbrace{\mathcal{A}}_{\substack{\text{keyed hash /} \\ \text{pseudorandom generator}}} \Big(\underbrace{w_{t-\ell:t-1}}_{\substack{\text{public input} \\ \text{last } \ell \text{ tokens}}}, \quad \underbrace{\text{Key}}_{\substack{\text{secret input} \\ \text{watermark key}}} \Big).$$

- **Same context + same key $\Longrightarrow$ same pseudorandomness.**
- **Perfect pseudorandomness:** over a random key, distinct public inputs give independent outputs: $w_{t-\ell:t-1} \neq w_{s-\ell:s-1} \Rightarrow \zeta_t, \zeta_s \overset{\text{iid}}{\sim} \mu$.
- A standard assumption in the watermark literature. In essence, it requires **no repeated text windows**.
- What if it fails? Detection with pseudorandom collision (Cai et al.2025).

A pivot removes the unknown token probabilities

The detector has the **text and key**, but not $\boldsymbol{P}_t$.

$$H_0: \; W_t \perp \zeta_t \;\text{ or }\; (W_t, \zeta_t) \sim \boldsymbol{P}_t \otimes \mu \qquad H_1: \; (W_t, \zeta_t) \sim \pi_t \neq \boldsymbol{P}_t \otimes \mu.$$

Definition (Pivotal statistic) (Li et al.2025a)

Choose an upper-tail pivot $Y_t = Y(W_t, \zeta_t)$:

$$H_0: \quad Y_t \sim \mu_0 \qquad \text{(known law, for every } \boldsymbol{P}_t\text{)},$$

$$H_1: \quad Y_t \sim \mu_{1,\boldsymbol{P}_t} \qquad \text{(watermark alternative, stochastically larger)}.$$

The Gumbel pivot has a uniform null law

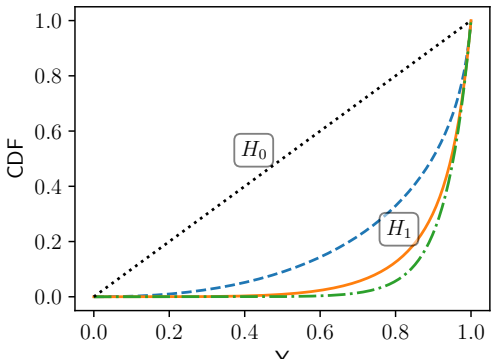
Pivot: $Y_t = U_{t,W_t}$

- Under H_0: uniform on $(0, 1)$.
- Under H_1: more mass near 1.

For $0 \le y \le 1$, the CDFs are

$$F_0(y) = y,$$

$$F_{1,\boldsymbol{P}_t}(y) = \sum_{w:P_{t,w}>0} P_{t,w} y^{1/P_{t,w}}.$$



From pivots to a test

For n scored positions and a score function h, use

$$T_h = \sum_{t=1}^{n} h(Y_t), \qquad \text{reject } H_0 \text{ if } T_h > c_{n,\alpha}(h).$$

The threshold $c_{n,\alpha}(h)$ controls false alarms at level $\alpha \in (0, 1)$.

Type I error false alarm under H_0

Type II error missed watermark under H_1

Two common choices:

- Ars score: $h_{\mathrm{ars}}(y) = -\log(1 - y)$ (Aaronson2023).
- Log score: $h_{\log}(y) = \log y$ (Kuditipudi et al.2024; Fernandez et al.2023).

Which detector is better in the worst case?

Fix the false-alarm level α and a token-uncertainty class (Li et al.2025a):

$$\mathcal{P}_\Delta = \{\boldsymbol{P} : \max_w P_w \le 1 - \Delta\}, \qquad 0 < \Delta \le 1 - |\mathcal{V}|^{-1}.$$

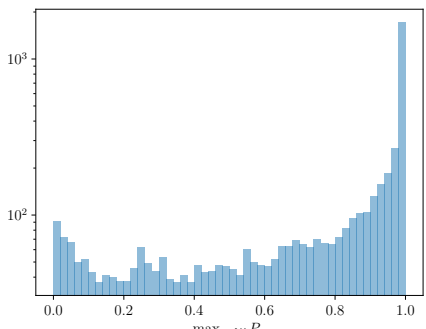
Definition (Efficiency rate)

$b_n(h)$: worst Type II error over sequences in $\mathcal{P}_\Delta$.

$$R_\Delta(h) = \liminf_{n\to\infty} -\frac{1}{n} \log b_n(h).$$

Larger rate: faster decay of missed detections.

ChatGPT-3.5-turbo



Optimal detection as a minimax problem (Li et al.2025a)

Finding $h^\star \in \arg\max_h R_\Delta(h)$ reduces to

Choose the score against the hardest token distribution

$$\inf_h \sup_{\boldsymbol{P}\in\mathcal{P}_\Delta} L(h, \boldsymbol{P}), \quad \text{where} \quad L(h, \boldsymbol{P}) = \mathbb{E}_0[h(Y)] + \log \mathbb{E}_{1,\boldsymbol{P}}[e^{-h(Y)}].$$

where $\mathbb{E}_0$ and $\mathbb{E}_{1,\boldsymbol{P}}$ are expectations under H_0 and $H_{1,\boldsymbol{P}}$.

- Generally **not convex–concave**.
- For Gumbel-max, exploit the class structure to find a **saddle point**.

Optimal score for the Gumbel-max watermark

Least-favorable likelihood-ratio score (Li et al.2025a)

$$h^\star_\Delta(y) = \log \frac{f_{1,\boldsymbol{P}^\star_\Delta}(y)}{f_0(y)},$$

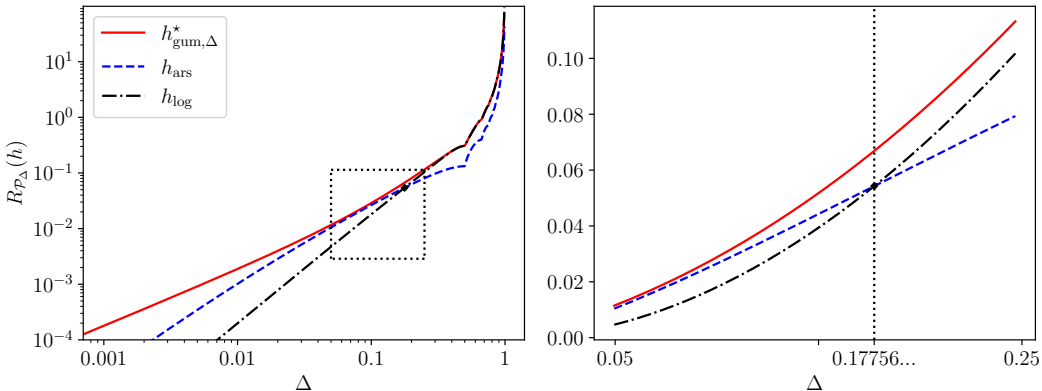
where $\boldsymbol{P}^\star_\Delta$ is least favorable in $\mathcal{P}_\Delta$ and $f_0(y) = 1$.

For $0 < \Delta < 1/2$,

$$\boldsymbol{P}^\star_\Delta = (1-\Delta, \Delta, 0, \ldots), \qquad h^\star_\Delta(y) = \log\left(y^{\Delta/(1-\Delta)} + y^{(1-\Delta)/\Delta}\right).$$

The score uses the class parameter Δ, not the unknown $\boldsymbol{P}_t$.

Detection efficiency



Editing turns inference into a mixture problem

Example: replace "cat" by "dog", with a one-token context ($\ell = 1$).

Generated	The	cat	sat	on	the	mat
Edited	The	**dog**	sat	on	the	mat
		token edited	context changed			

Definition (Surviving watermark evidence) (Li et al.2026a)

$$Y_t \mid \boldsymbol{P}_t \sim \underbrace{(1-\varepsilon)\mu_0}_{\text{erased signal: known null law}} + \underbrace{\varepsilon\mu_{1,\boldsymbol{P}_t}}_{\text{surviving signal: watermark law}} .$$

$\varepsilon \in [0,1]$ is the surviving fraction, and $\mu_{1,\boldsymbol{P}_t}$ is unknown.

Three questions under the mixture model

Surviving watermark evidence: $Y_t \mid \boldsymbol{P}_t \sim (1-\varepsilon)\mu_0 + \varepsilon\mu_{1,\boldsymbol{P}_t}$.

1	2	3
Global detection	**Fraction estimation**	**Localization**
Is any signal present?	How much remains?	Where does it remain?
Tr-GoF test (Li et al.2026a)	Moments (Li et al.2025b)	SPOT (Blanchet et al.2026)

evidence aggregated across the text — evidence at individual positions

Global detection: test uniformity

Convert the pivots into p-values:

$$p_t = 1 - F_0(Y_t), \qquad H_0 : \; p_t \sim \mathrm{Unif}(0,1).$$

A watermark creates an **excess of small p-values**.

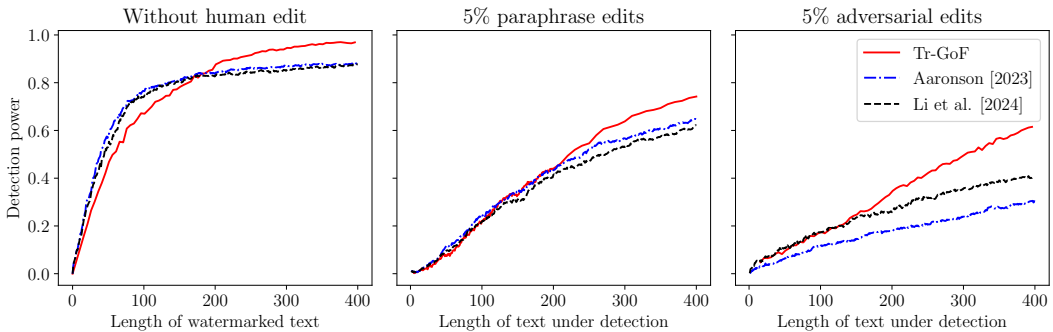
At a threshold $r \in (0,1)$, compare

$$\underbrace{\widehat{F}_n(r) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{p_t \le r\}}_{\text{observed tail mass}} \qquad \text{with} \qquad \underbrace{\mathbb{P}_0(p_t \le r) = r}_{\text{null tail mass}}.$$

- **Tr-GoF** (Li et al.2026a): scan thresholds and test deviations from uniformity, in the spirit of higher criticism (Donoho and Jin2004; Jager and Wellner2007).

Robust detection after edits

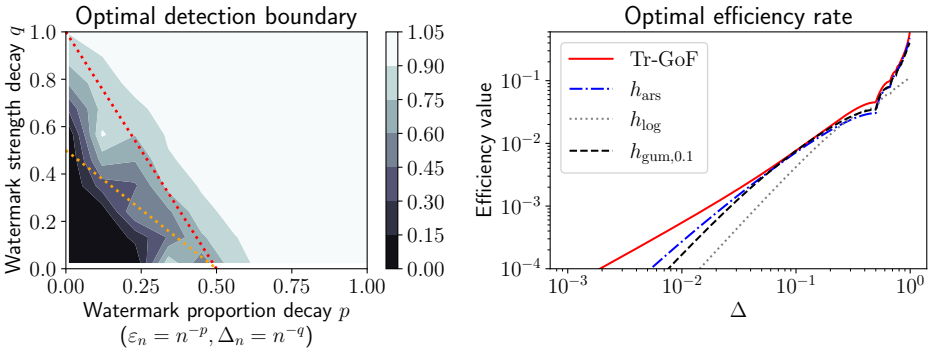
Gumbel-max watermark, OPT-1.3B (Zhang et al.2022), C4 news-like text (Raffel et al.2020).



- Other goodness-of-fit tests also bring power and robustness to edits (He et al.2025).

Why does Tr-GoF perform so well? (Li et al.2026a)

Surviving fraction $\varepsilon_n = n^{-p}$ and token uncertainty $\Delta_n = n^{-q}$, with $p, q \geq 0$.



Optimal boundary $2p + q < 1$, adaptive to unknown p and q.

Fraction estimation: start from a moment equation

Choose a weight $v : \mathbb{R} \to \mathbb{R}$ and take a moment of the mixture (Li et al.2025b).

$$m_v = (1-\varepsilon)\, m_{0,v} + \varepsilon\, m_{1,v}$$

- m_v: $\mathbb{E}_{\bar{F}}[v(Y)]$ — mixture: estimated from text
- $m_{0,v}$: $\mathbb{E}_{F_0}[v(Y)]$ — null: known
- $m_{1,v}$: $\mathbb{E}_{\bar{F}_1}[v(Y)]$ — signal: unknown

Match the sample moment to the mixture

$$\varepsilon = \frac{m_{0,v} - m_v}{m_{0,v} - m_{1,v}} \quad \Longrightarrow \quad \widehat{\varepsilon}_v = \frac{m_{0,v} - \widehat{m}_v}{m_{0,v} - \widehat{m}_{1,v}}, \qquad \widehat{m}_v = \frac{1}{n}\sum_{t=1}^{n} v(Y_t).$$

Key choice: approximate $m_{1,v}$ and choose the weight v.

Three estimators from the same equation (Li et al.2025b)

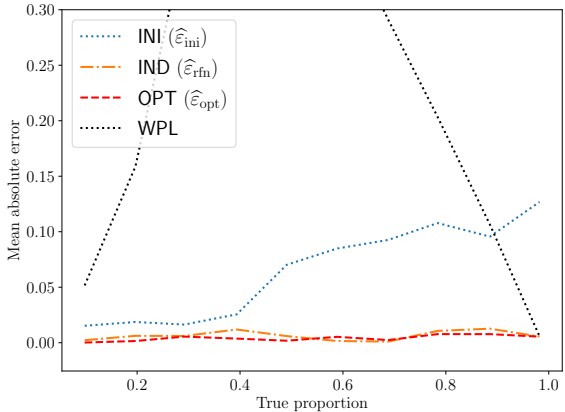
	Main idea	Estimator / weight
INI	Small-pivot tail: ignore the signal contribution	$1 - \dfrac{\widehat{F}(\delta)}{F_0(\delta)}$
IND	Correct the tail using watermarked reference text	$\dfrac{F_0(\delta) - \widehat{F}(\delta)}{F_0(\delta) - \widehat{F}_1(\delta)}$
OPT	Use variance-optimal weights from a plug-in fixed point	$v_\varepsilon^\star(y) = \dfrac{1 - g(y)}{1 - \varepsilon + \varepsilon g(y)}$

$\delta > 0$ is a tail cutoff, $\widehat{F}$ the empirical mixture CDF, and $\widehat{F}_1$ the reference signal CDF.

$g = d\bar{F}_1/dF_0$: signal-to-null density ratio, estimated from reference text.

Fraction estimation: performance

Gumbel-max watermark, OPT-1.3B, C4 news-like text (Li et al.2025b).



Localization: can we make a reliable discovery?

Let $\theta_t = 1$ mark surviving dependence. Select positions $S \subseteq \{1, \dots, n\}$.

Definition (Discoverable) (Blanchet et al.2026)

Local rules $\varphi_t : (0,1) \to \{0,1\}$ select $S = \{t : \varphi_t(Y_t) = 1\}$ with marginal FDR

$$\mathrm{mFDR} = \frac{\mathbb{E}[\sum_{t \in S}(1 - \theta_t)]}{\mathbb{E}|S|} \to 0, \qquad \mathbb{P}(|S| \geq 1) \to 1.$$

- **Global inference:** weak signals add up.
- **Discovery:** each pick must stand out.

SPOT: adaptive discovery (Blanchet et al.2026)

1. **Scan** thresholds τ with a nonempty selected tail.
2. **Estimate** the null share in each selected tail:

$$\frac{\text{estimated null tail mass}}{\text{observed tail mass}} \approx \frac{(1-\widehat{\varepsilon})(1-\tau)}{n^{-1}\sum_{t=1}^{n}\mathbf{1}\{Y_t > \tau\}}.$$

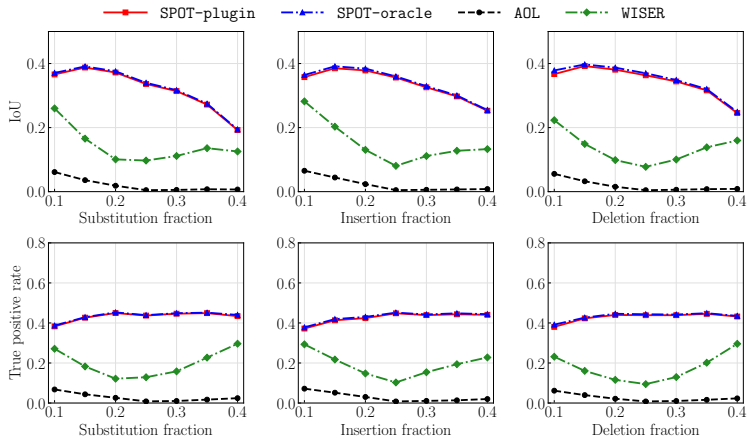
3. **Select** the largest tail passing the false-discovery check.

Adaptive boundary optimality

SPOT attains the discovery boundary without knowing the sparsity or signal strength.

Localization under random edits

OPT-1.3B, C4 news-like text, 400 tokens (Blanchet et al.2026).



PART II

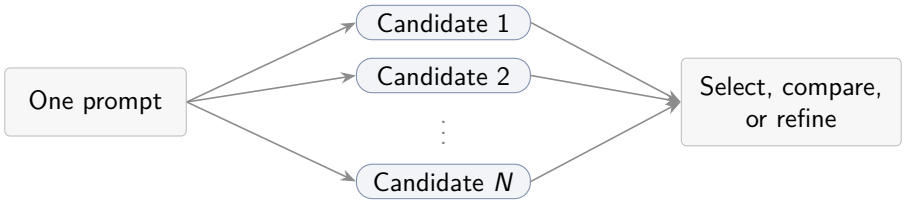
Designing the coupling

textGrain: Entropy-Calibrated Watermarking for Language Model Text

Technical report accompanying the OpenAI blog, 2026

With Garrett Wen, Xiaohong Chen, Qi Long, Arzav Jain,
Florent Joly, Mike Lam, Qingquan Song, and Weijie Su

A fixed key might lose diversity



Best-of-N (Cobbe et al.2021), test-time inference, repeated answers.

The fixed-key problem (Kuditipudi et al.2024; He et al.2025)

Same prompt + same key $\Longrightarrow$ same Gumbel-max continuation.

Unbiasedness alone does not preserve diversity (Li et al.2026b)

At one context, $\boldsymbol{P} = (1/2, 1/2)$ and $J \sim \mathrm{Unif}\{1, 2\}$ represents ζ.

Independent coupling

	$J = 1$	$J = 2$
$W = a$	$1/4$	$1/4$
$W = b$	$1/4$	$1/4$

Either key: still sample a or b.

Deterministic coupling

	$J = 1$	$J = 2$
$W = a$	$1/2$	0
$W = b$	0	$1/2$

Fixed key: always the same token.

Both preserve the same token and key marginals.

Dependence is an entropy cost

At a fixed context, let π couple token $W \sim \boldsymbol{P}$ with keyed randomness $\zeta \sim \mu$.

$$H(\boldsymbol{P}) = -\sum_{w} P_w \log P_w, \qquad H_\pi(W \mid \zeta) = \mathbb{E}_\zeta H(\boldsymbol{Q}), \qquad \boldsymbol{Q} = \pi(\cdot \mid \zeta).$$

The identity behind textGrain (Li et al.2026b)

$$\underbrace{H(\boldsymbol{P}) - H_\pi(W \mid \zeta)}_{\text{average entropy loss}} = \underbrace{I_\pi(W;\zeta)}_{\text{mutual information}} = \underbrace{D_{\mathrm{KL}}(\pi \| \boldsymbol{P} \otimes \mu)}_{\text{KL divergence from independence}} .$$

- More dependence uses more of the model's sampling randomness.
- The loss is averaged over the keyed randomness.

Detection requires positive dependence (Li et al.2026b)

In the sequential coupling model, let

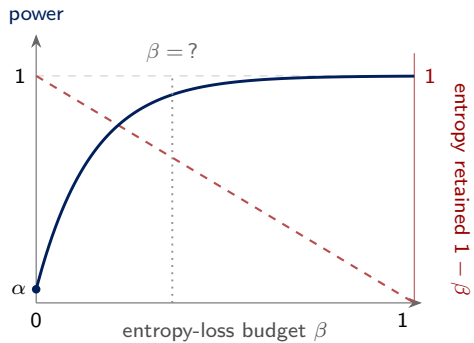
$$\mathcal{I}_n = \sum_{t=1}^{n} \mathbb{E}_1[I_{\pi_t}(W_t; \zeta_t)]$$

be the **total expected information**.

A bound for every level-α detector

$$\text{Power} \;\le\; \min\left\{1,\, \alpha + \sqrt{\mathcal{I}_n/2}\right\}.$$

$$\mathcal{I}_n = 0 \quad\Longrightarrow\quad \text{Power} \le \alpha.$$



We need dependence.
How much should we spend?

Optimal transport with an entropy budget

$\Pi(\boldsymbol{P}, \mu)$: couplings with token marginal $\boldsymbol{P}$ and key marginal μ (Peyré and Cuturi2019).

Design the coupling (Li et al.2026b)

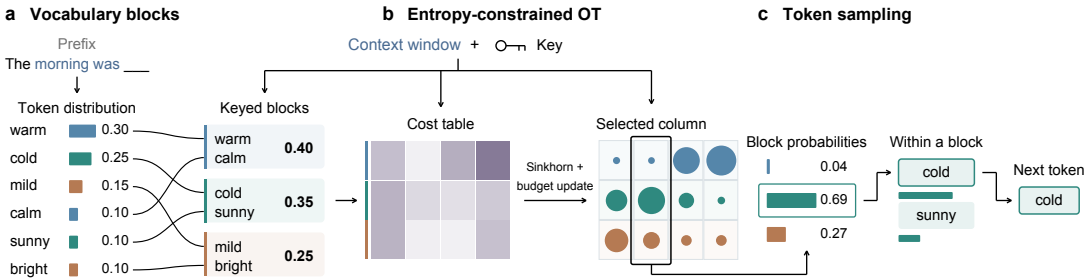
With a cost $c(w, \zeta)$ that favors detectable pairs, solve

$$\min_{\pi \in \Pi(\boldsymbol{P}, \mu)} \quad \mathbb{E}_\pi[c(W, \zeta)]$$

$$\text{subject to} \quad D_{\mathrm{KL}}(\pi \| \boldsymbol{P} \otimes \mu) \leq \underbrace{\beta H(\boldsymbol{P})}_{\text{entropy-loss budget}}, \qquad \beta \in [0, 1].$$

- Fixed marginals preserve **unbiasedness**.
- Average entropy retained $\geq (1 - \beta) H(\boldsymbol{P})$.

Block optimal transport (Li et al.2026b)



1. Group tokens into B keyed blocks.
2. Couple the blocks to m equally likely columns using Gumbel costs.
3. Select the keyed column, then sample a block and a token.

Solving the block problem

At block level, solve

$$\min_{\pi \in \Pi(\boldsymbol{P}^{\mathrm{blk}}, \boldsymbol{u}_m)} \sum_{b=1}^{B} \sum_{j=1}^{m} \pi(b,j) c(b,j) \quad \text{subject to} \quad D_{\mathrm{KL}}(\pi \| \boldsymbol{P}^{\mathrm{blk}} \otimes \boldsymbol{u}_m) \le \beta H(\boldsymbol{P}).$$

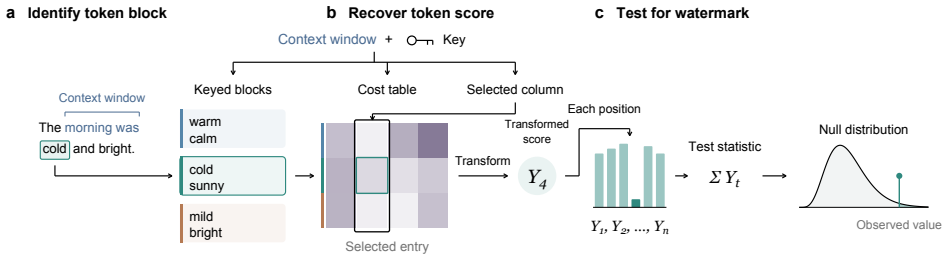
Here $\boldsymbol{P}^{\mathrm{blk}}$ gives block probabilities and $\boldsymbol{u}_m = (1/m, \ldots, 1/m)$ column probabilities.

1. **Sinkhorn:** rescale the rows and columns (Cuturi2013; Altschuler et al.2017).
2. **Budget update:** adjust the KL penalty $\lambda \ge 0$ toward β.
3. **Marginal correction:** preserve unbiasedness before sampling.

Typical configuration: $(B, m, N_{\mathrm{it}}, \beta) = (32, 64, 5, 0.36)$.

N_{it}: solver iteration limit.

Detection with a pivotal statistic (Li et al.2026b)



- **Text + key** $\rightarrow$ Gumbel value Z_t, pivot $V_t = F_G(Z_t)$.
  With $F_G(z) = e^{-e^{-z}}$, the null gives $V_t \sim \text{Unif}(0, 1)$.
- Reject for large $S_n = \sum_{t=1}^{n} -\log(1 - V_t)$.

Evaluation: detection and diversity

Model: GPT-5.6 Sol (OpenAI2026). **Data:** 1,000 KILT ELI5 questions (Fan et al.2019). First 300 tokens of each eligible response.

- **Detection:** power at nominal 1% false-positive rate.
- **Sampling diversity:** empirical entropy retention

$$\widehat{R} = \frac{\sum H(\boldsymbol{Q}_t)}{\sum H(\boldsymbol{P}_t)}.$$

  $\boldsymbol{Q}_t = \pi_t(\cdot \mid \zeta_t)$ is the fixed-key sampling law, summed over evaluated positions.

- **Response diversity:** Self-BLEU-4 (Zhu et al.2018) across repeated answers.

  20 generation attempts per question. Lower overlap means more diversity.

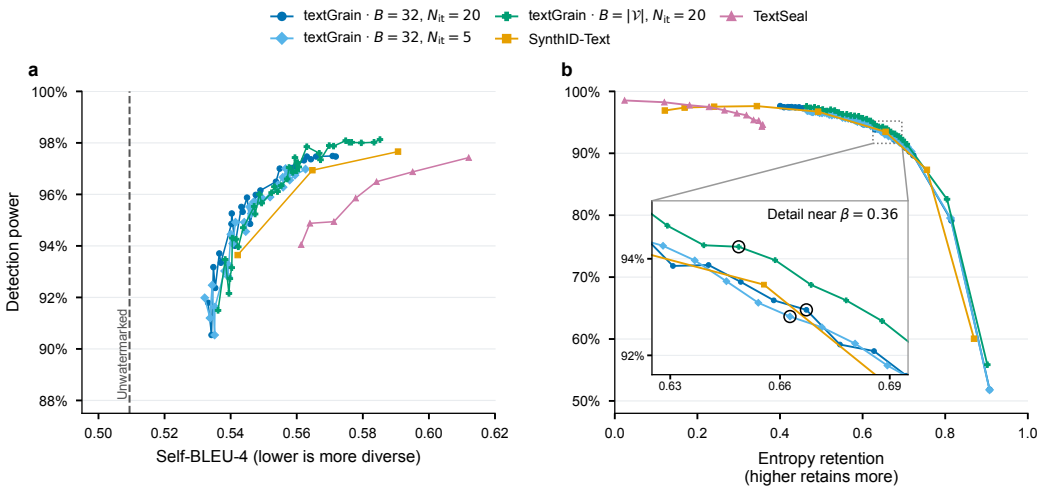
Similar Self-BLEU can hide different entropy costs

Qwen3-8B (Qwen Team2025) **on ELI5**, autoregressive sampling.

	SynthID-Text 6 layers (Dathathri et al.2024)	TextSeal mixing 0.4 (Sander et al.2026)
Self-BLEU-4 lexical overlap among responses	0.486	0.487
Entropy retained randomness left in token sampling	60.66%	39.16%

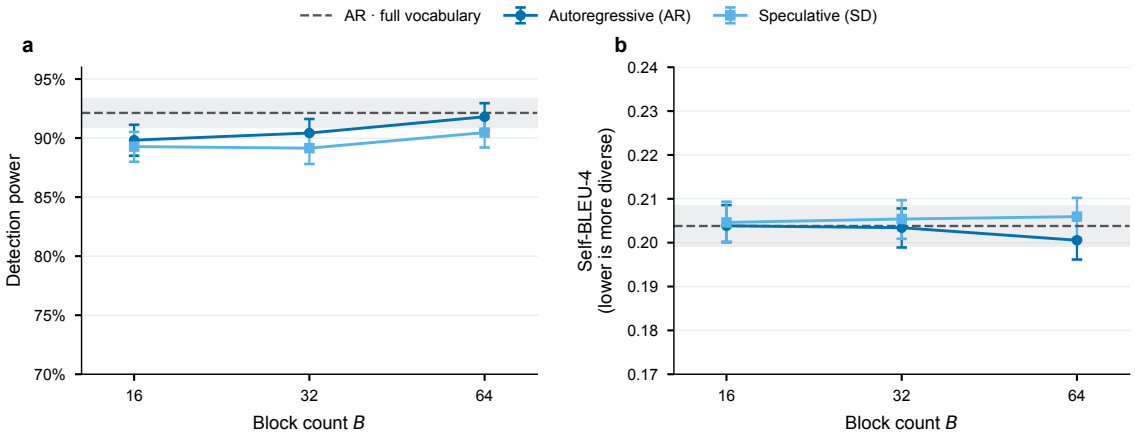
- Higher power need not mean a better design. It may come from **spending more entropy**.
- Compare watermarks at **comparable entropy retention**.

Detection–diversity trade-off on GPT-5.6 Sol



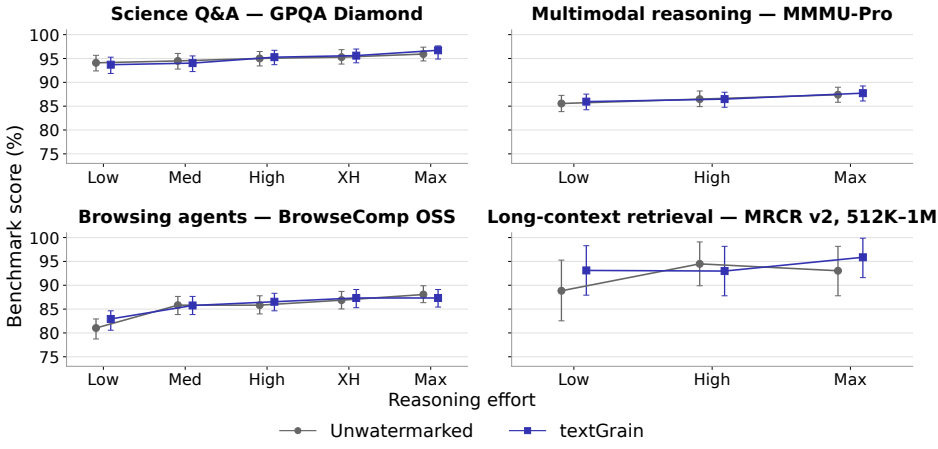
textGrain supports speculative sampling

textGrain stays unbiased under speculative sampling (Chen et al.2023).



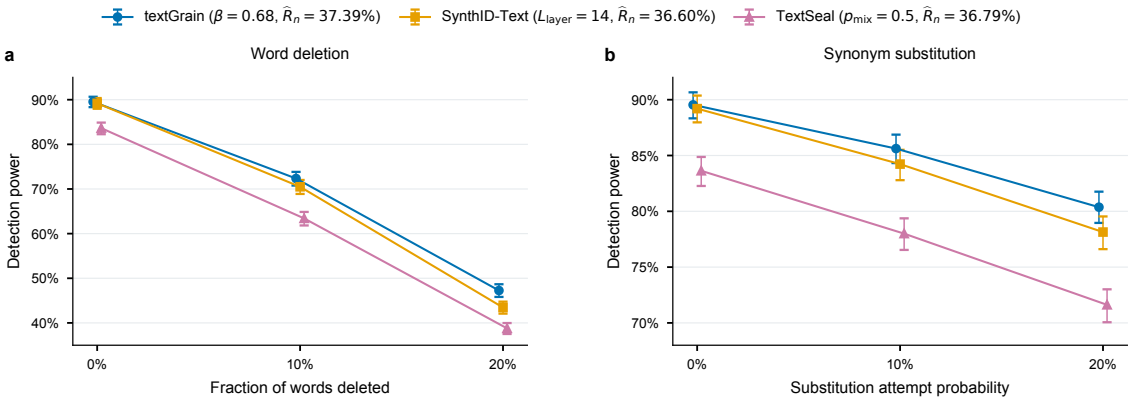
Task performance across four capabilities on GPT-6 Astra

Benchmarks (Rein et al.2023; Yue et al.2025; Wei et al.2025; OpenAI2025).



Robustness to editing on GPT-5.4

textGrain is more robust here, but editing still reduces detection power.



Summary: A coupling perspective

PART I

Detecting the coupling

Given a watermark

- Pivots reveal dependence.
- Minimax guides detection.
- Estimate how much survives and locate the signal.

PART II

Designing the coupling

Choose a watermark

- Dependence costs entropy.
- OT controls that cost.
- textGrain retains sampling diversity under a fixed key.

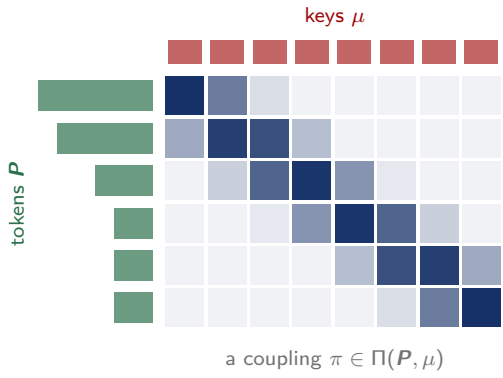
One object links both parts: the coupling between tokens and keyed randomness.

Thank you!

Questions are welcome.

Detecting the coupling: pivots, minimax tests, fraction estimation, localization

Designing the coupling: textGrain



Backup Slides

Block coupling and token sampling

- **Blocks.** A keyed map $g : \mathcal{V} \to \{1, \dots, B\}$ gives $P_b^{\mathrm{blk}} = \sum_{w:g(w)=b} P_w$.
- **Coupling.** π has row sums $\boldsymbol{P}^{\mathrm{blk}}$ and column sums $\boldsymbol{u}_m = (1/m, \dots, 1/m)$.
- **Sampling.** The key picks $J \sim \mathrm{Unif}\{1, \dots, m\}$, and the token law is

$$Q_w = m\,\pi(g(w), J)\,\frac{P_w}{P^{\mathrm{blk}}_{g(w)}}, \qquad \boldsymbol{Q} = (Q_w)_w = \text{law of } W \text{ given } J.$$

- **Unbiased:** $\mathbb{E}_J Q_w = P_w$.
- **Entropy cost:** $H(\boldsymbol{P}) - \mathbb{E}_J H(\boldsymbol{Q}) = D_{\mathrm{KL}}(\pi \| \boldsymbol{P}^{\mathrm{blk}} \otimes \boldsymbol{u}_m)$.

Zero-mass blocks are omitted.

Calibrating with random keys

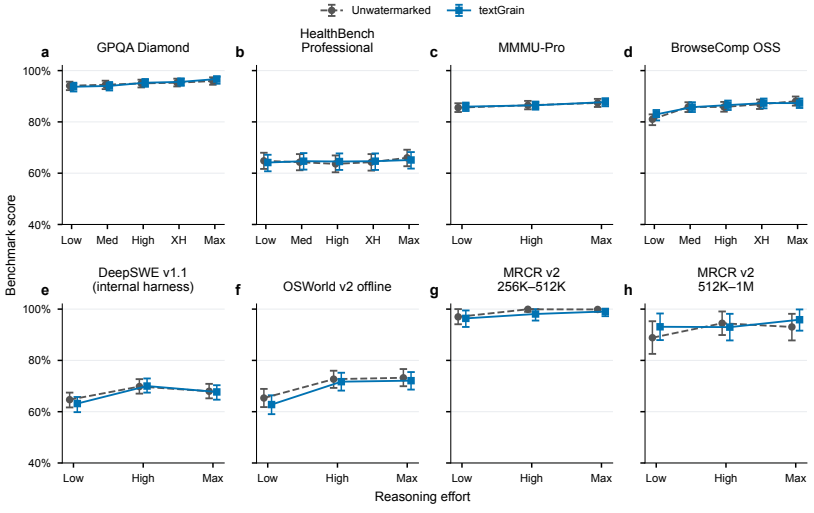
Keep the observed text x fixed. Draw M independent reference keys.

Compute $S_b = S(x, \mathrm{Key}_b)$ for $b = 0, \ldots, M$, where key 0 is the tested key.

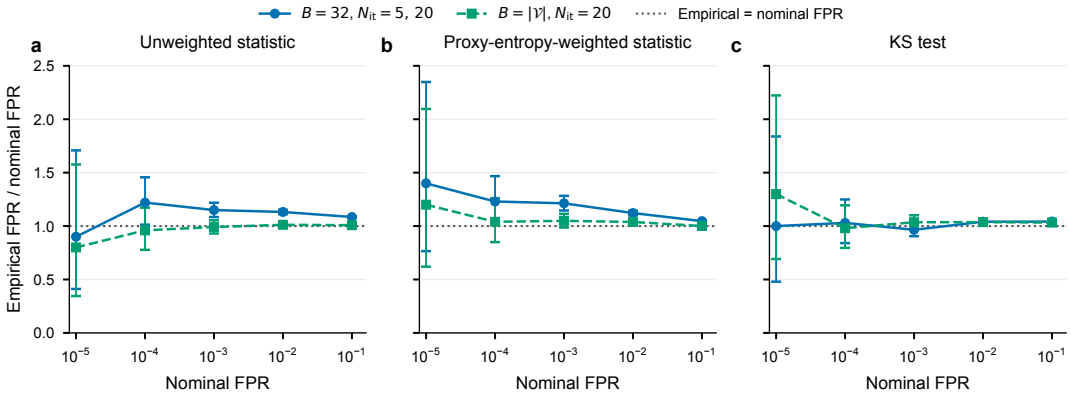
$$\widehat{p} = \frac{1 + \sum_{b=1}^{M} \mathbf{1}\{S_b \geq S_0\}}{M + 1}.$$

- Under key–text independence, scores are exchangeable across keys.
- Reject when $\widehat{p} \leq \alpha$.
- For every key, score only the **first occurrence of each (context window, block) pair**, so repeated windows are not counted twice.

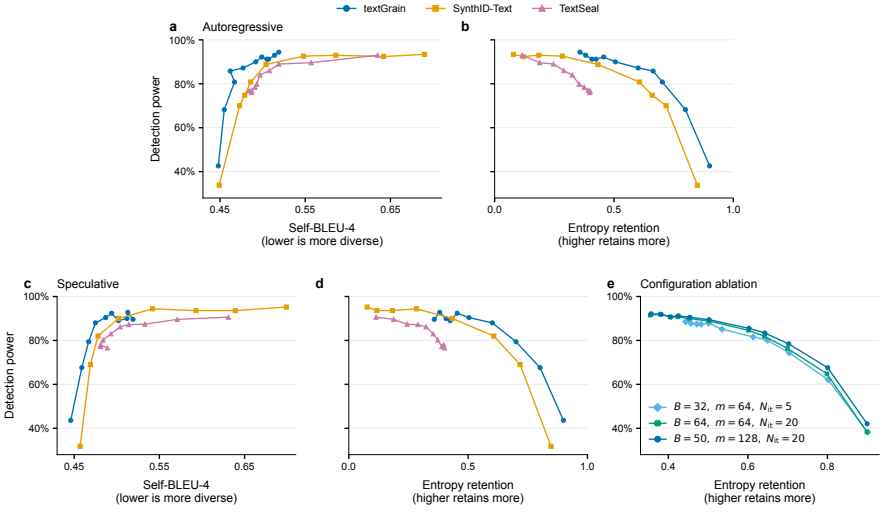
Task performance: all eight benchmarks



Null calibration



Trade-off on Qwen3-8B



References I

S. Aaronson. Watermarking of large language models, August 2023.

J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

J. H. Blanchet, T. T. Cai, X. Li, H. Liu, Q. Long, and W. J. Su. Optimal watermark localization in mixed-source large language model texts. *arXiv preprint arXiv:2608.14906*, 2026.

T. T. Cai, X. Li, Q. Long, W. J. Su, and G. G. Wen. Optimal detection for language watermarks with pseudorandom collision. *arXiv preprint arXiv:2510.22007*, 2025.

C. Chen, S. Borgeaud, G. Irving, J.-B. Lespiau, L. Sifre, and J. Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

S. Dathathri, A. See, S. Ghaisas, P.-S. Huang, R. McAdam, J. Welbl, V. Bachani, A. Kaskasoli, R. Stanforth, T. Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634 (8035):818–823, 2024.

References II

D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32 (3):962–994, 2004.

A. Fan, Y. Jernite, E. Perez, D. Grangier, J. Weston, and M. Auli. ELI5: Long Form Question Answering. *arXiv preprint arXiv:1907.09190*, 2019.

P. Fernandez, A. Chaffin, K. Tit, V. Chappelier, and T. Furon. Three bricks to consolidate watermarks for large language models. In *IEEE International Workshop on Information Forensics and Security*, pages 91–99. IEEE, 2023.

E. J. Gumbel. *Statistical theory of extreme values and some practical applications: A series of lectures*, volume 33. US Government Printing Office, 1948.

W. He, X. Li, T. Shang, L. Shen, W. J. Su, and Q. Long. On the empirical power of goodness-of-fit tests in watermark detection. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

L. Jager and J. A. Wellner. Goodness-of-fit tests via phi-divergences. *Annals of Statistics*, 35(5):2018–2053, 2007.

R. Kuditipudi, J. Thickstun, T. Hashimoto, and P. Liang. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

X. Li, F. Ruan, H. Wang, Q. Long, and W. 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, 2025a.

X. Li, G. G. Wen, W. He, J. Wu, Q. Long, and W. J. Su. Optimal estimation of watermark proportions in hybrid AI-human texts. *arXiv preprint arXiv:2506.22343*, 2025b.

References III

X. Li, F. Ruan, H. Wang, Q. Long, and W. J. Su. Robust detection of watermarks for large language models under human edits. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(2):491–515, 2026a.

X. Li, G. Wen, X. Chen, Q. Long, A. Jain, F. Joly, M. Lam, Q. Song, and W. J. Su. textgrain: Entropy-calibrated watermarking for language model text. Technical report, OpenAI, 2026b.

C. J. Maddison, D. Tarlow, and T. Minka. A* sampling. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

OpenAI. OpenAI MRCR: Long context multiple needle in a haystack benchmark. Dataset, 2025.

OpenAI. GPT-5.6 system card. Technical report, OpenAI, July 2026.

G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

Qwen Team. Qwen3-8B and Qwen3-0.6B model cards, 2025.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21 (140):1–67, 2020.

D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

References IV

T. Sander, H. Chang, S.-A. Rebuffi, T. Souček, T. Tran, V. Lacatusu, A. Mourachko, S. Parimi, C. Ropers, R. Moritz, V. Stark, H. Elsahar, and P. Fernandez. TextSeal: A localized LLM watermark for provenance & distillation protection. *arXiv preprint arXiv:2605.12456v2*, 2026.

J. Wei, Z. Sun, S. Papay, S. McKinney, J. Han, I. Fulford, H. W. Chung, A. T. Passos, W. Fedus, and A. Glaese. BrowseComp: A simple yet challenging benchmark for browsing agents. Technical report, OpenAI, 2025.

X. Yue, T. Zheng, Y. Ni, Y. Wang, K. Zhang, S. Tong, Y. Sun, B. Yu, G. Zhang, H. Sun, Y. Su, W. Chen, and G. Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, 2025.

S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu. Texygen: A benchmarking platform for text generation models. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1097–1100. ACM, 2018.

References: policy and industry

- European Union (2024). Regulation (EU) 2024/1689, Article 50(2).
- Google DeepMind (2024). Watermarking AI-generated text and video with SynthID. 14 May.
- Axon, S. (2026). OpenAI will watermark ChatGPT outputs.... *Ars Technica*, 6 October.
- Ha, A. (2026). Anthropic shares more details.... *TechCrunch*, 15 August.
- Crypto Briefing (2026). Meta's TextSeal2.... 2 November.
- The Guardian (2025). Student AI survey. 26 February.
- Nature (2024). AI use in peer review. 628, 483–484.
- Ars Technica (2025). AI-generated reading-list errors. May.
- Yale Insights (2024). Generated restaurant reviews. 28 May.