Watermarking Discrete Diffusion Language Models

Avi Bagchi Akhil Bhimaraju Moulik Choraria Daniel Alabi Lav Varshney

UPenn & UIUC

November 19, 2025

Summary

1. Problem Motivation
2. Large Language Diffusion Models (Nie et al., 2025)
3. Watermarking Schemes
4. Theoretical Results
5. Empirical Results

Why watermark language models?

The widespread deployment of AI agents motivate methods to distinguish AI-generated text from human-written content.

- Authenticity
- Traceability

Why watermark language models?

- As models improve, post-hoc “GPT-detectors” suffer
- Watermarking embeds a detectable, invisible signal *at generation time*

Watermark properties

Ideally, our watermark satisfies the following properties:

- **Soundness**: Detector reliably identifies unwatermarked content as unwatermarked.
- **Completeness**: Detector reliably identifies watermarked content as watermarked.
- **Distortion-Freeness**: Does not significantly reduce the quality of the text.
- **Robustness**: Still detectable following bounded modifications to the text.

Existing Watermarks for Generative Models

- Diffusion models for image generation
  - “Tree-Ring Watermark” (Wen et al., 2023)
- Autoregressive Large Language Models
  - “Red-Green List Watermark” (Kirchenbauer et al., 2024)

What about discrete diffusion language models?

- Discrete diffusion models generate tokens in parallel
  - Faster inference (Wang et al., 2025)
  - Greater controllability (Schiff et al., 2025)
  - Enhanced comprehension of global patterns (Hu et al., 2021)
- Rapid growth in both research and commercial use
  - Google (Gemini Diffusion), Inception Labs (Mercury) etc.
- No existing watermark!

Our Contribution

- We implement our watermark on the state-of-the-art Language Diffusion Model LLaDA (Nie et al., 2025)
  - Achieve high completeness and soundness while still preserving benchmark scores and perplexity
- We analytically prove:
  - False detection probability decays exponentially with the number of generated tokens
  - Our watermark leaves the token sampling distribution unchanged

Large Language Diffusion Models (LLaDA)

- Vocabulary $\mathcal{V}$, sequence length d, special `[MASK]` token.
- Start fully masked; iteratively unmask until $t = 0$.
- Model $p_\theta(\cdot \mid x_t)$ outputs per-position token distributions.

$$p_{t-\Delta t|t}(x_{t-\Delta t} \mid x_t) = \prod_{i=1}^{d} \left[p_\theta(x_{t-\Delta t}^i \mid x_t)\right]_i.$$

Large Language Diffusion Models (LLaDA)

x_0	the	Fed	raised	rates				
					Predictions			
x_1	MASK	MASK	MASK	MASK	letter	human	trail	quartz
$x_{0.75}$	the	MASK	MASK	MASK	the	city	next	rates
$x_{0.5}$	the	MASK	raised	MASK	the	state	up	home
$x_{0.25}$	the	Fed	raised	MASK	the	Fed	raised	rates

Watermarking Scheme: Attempt 1 (Tree-Ring)

- Wen et al. (2023) (originally for image diffusion) proposes embedding a signal in the initial noise vector and then reversing the sampling process (i.e. DDIM inversion) to recover the watermark
- Can we adapt the scheme for discrete diffusion?

Watermarking Scheme: Attempt 1 (Tree-Ring)

- Wen et al. (2023) (originally for image diffusion) proposes embedding a signal in the initial noise vector and then reversing the sampling process (i.e. DDIM inversion) to recover the watermark
- Can we adapt the scheme for discrete diffusion?
  - No, the sampling at every step in discrete diffusion makes a reversal difficult, so the watermark is essentially undetectable

Watermarking Scheme: Attempt 2 (Red/Green List)

- Kirchenbauer et al. (2024) (originally for autoregressive LLMs) proposes partitioning the vocabulary into a red and green list while applying a bias in sampling to favor the latter.
  - The partition is seeded by the previously generated token to enable detectability
- Can we adapt the scheme for discrete diffusion?

Watermarking Scheme: Attempt 2 (Red/Green List)

- Kirchenbauer et al. (2024) (originally for autoregressive LLMs) proposes partitioning the vocabulary into a red and green list while applying a bias in sampling to favor the latter.
  - The partition is seeded by the previously generated token to enable detectability
- Can we adapt the scheme for discrete diffusion?
- Algorithm (modifications in bold):
  - Partition vocab into green set G of size $\gamma|\mathcal{V}|$, **seeding by position in the sequence**
  - Bias sampling toward G via a logit boost $\delta > 0$:

$$p'(x) = \mathrm{Softmax}\big(\ell(x) + \delta \cdot \mathbf{1}\{x \in G\}\big).$$

  - **Repeat** across a subset of the sampling steps $S_W \subseteq S$
  - Regenerate G using the sequence position to enable detection (i.e. calculate z-score)

Watermarking Scheme: Attempt 2 (Red/Green List)

- **Limitation:** Increasing δ or S_W increases detectability but further distorts the text
- The "optimal" hyperparameters are task and model dependent

Watermarking Scheme: Attempt 2 (Red/Green List)

- **Limitation:** Increasing δ or S_W increases detectability but further distorts the text
- The "optimal" hyperparameters are task and model dependent

Physical fitness has numerous benefits for both physical and mental health. It can help improve cardiovascular health, increase muscle strength and endurance, and improve bone density. Additionally, it can help reduce stress, anxiety, and depression, and improve sleep quality. Regular physical activity can also help reduce the risk of chronic diseases such as diabetes, heart disease, and certain types of cancer. It can also help improve mood, concentration, and cognitive function. Finally, physical fitness can help improve overall quality of life and increase longevity.

$\delta = 0 \quad z = 1.53$

Physical fitness is a crucial aspect of maintaining a healthy lifestyle. It can help reduce the risk of chronic diseases, improve mental health, and improve overall wellbeing. It can also help strengthen muscles, increase flexibility, and improve digestion. Additionally, it can help boost energy, reduce stress, and improve sleep quality. Overall, the benefits of physical fitness are numerous.

$\delta = 4 \quad z = 8.38$

There are many benefits of physical fitness, including improving overall health and wellbeing, reducing stress and anxiety, boosting self-esteem, and reducing the risk of chronic diseases such as type 2 diabetes and heart disease. Regular physical activity can also improve digestion and circulation, improve sleep quality and duration, and help prevent and improve depression. Physical activity can also help prevent and improve osteoporosis by improving bone density. Regular physical activity can also help prevent and maintain a healthy weight and improve a person's bone density.

$\delta = 6 \quad z = 14.32$

There are a number of benefits to physical fitness. Here are a few brief examples: improve digestion

$\delta = 8 \quad z = 37.81$

Watermarking Scheme: Attempt 3 (Gumbel-max Trick)

- Idea: What if our scheme was unbiased at every step?
  - Effectively eliminates the need to tune δ, S_W, greatly simplifying our objective
  - We follow Aaronson and Kirchner (2022) (originally for autoregressive LLMs)

Watermarking Scheme: Attempt 3 (Gumbel-max Trick)

- Idea: What if our scheme was unbiased at every step?
  - Effectively eliminates the need to tune δ, S_W, greatly simplifying our objective
  - We follow Aaronson and Kirchner (2022) (originally for autoregressive LLMs)
- Algorithm (modifications in bold):
  - Generate $r_i \sim \text{Unif}[0, 1]$ at each position i
    - **Seed the RNG that generates r_i with the position in the sequence**
    - Sample token with maximum value of $r_i^{\frac{1}{p}}$ such that p is the probability of that token
  - **Repeat** across all sampling steps

Watermarking Scheme: Attempt 3 (Gumbel-max Trick)

- **Detection:** Regenerate r_i at each position i
  - Compute score
    $$\frac{1}{L}\sum_{i=1}^{L}\ln\left(\frac{1}{1-r_i}\right)$$
    where L is the length of the generated sequence.
  - Watermarked text: $\text{score} > \tau$ such that $\tau > 1$
  - Unwatermarked text: $\text{score} \approx 1$

Watermarking Scheme: Attempt 3 (Gumbel-max Trick)

- **Detection:** Regenerate r_i at each position i
  - Compute score
    $$\frac{1}{L}\sum_{i=1}^{L}\ln\left(\frac{1}{1-r_i}\right)$$
    where L is the length of the generated sequence.
  - Watermarked text: $\text{score} > \tau$ such that $\tau > 1$
  - Unwatermarked text: $\text{score} \approx 1$
- How is this detectable?
  - **Intuition:** Generated tokens x_i correspond to higher random values r_i on average, since the Gumbel-max rule favors tokens linked to larger r_i.
  - This correlation makes $\ln(1/(1-r_i))$ systematically larger for watermarked text, pushing the average score above 1.

Proof of distortion-freeness

WTS that $\arg\max_y \frac{\ln R_y}{p_y}$ has same distribution as p_y for $y \in \{1, 2, \ldots, |\mathcal{V}|\}$.

Proof of distortion-freeness

WTS that $\arg\max_y \frac{\ln R_y}{p_y}$ has same distribution as p_y for $y \in \{1, 2, \ldots, |\mathcal{V}|\}$.

$$\begin{aligned} \mathbb{P}(Y = y) &= \mathbb{P}\left(\frac{\ln R_y}{p_y} \geq \frac{\ln R_z}{p_z} \ \forall\ z \neq y\right) \\ &= \int_0^1 \prod_{z \neq y} r_y^{p_z/p_y} dr_y && (\{R_z\} \text{ independent}) \\ &= \int_0^1 r_y^{\frac{1-p_y}{p_y}} dr_y = p_y, \end{aligned}$$

which concludes the proof.

Proof of distortion-freeness

WTS that $\arg\max_y \frac{\ln R_y}{p_y}$ has same distribution as p_y for $y \in \{1, 2, \ldots, |\mathcal{V}|\}$.

$$\begin{aligned}
\mathbb{P}(Y = y) &= \mathbb{P}\left(\frac{\ln R_y}{p_y} \geq \frac{\ln R_z}{p_z} \ \forall \ z \neq y\right) \\
&= \int_0^1 \prod_{z \neq y} r_y^{p_z/p_y} dr_y & (\{R_z\} \text{ independent}) \\
&= \int_0^1 r_y^{\frac{1-p_y}{p_y}} dr_y = p_y,
\end{aligned}$$

which concludes the proof.

- We can watermark **every** sampling step without concern of distortion

Proof of false detection probability

WTS for $\tau = 1 + \zeta$:

$$\mathbb{P}(\text{Watermark detection} \mid \text{Unwatermarked}) \leq m \exp[-L(\zeta - \ln(1 + \zeta))] \,,$$

where L is the sequence length and m the number of RNG seeds.

- False detection probability decays exponentially in L
- Proof included in our paper

Aside: robustness

We follow Kuditipudi et al. (2024) to thwart prefix deletions.

Aside: robustness

We follow Kuditipudi et al. (2024) to thwart prefix deletions.

- Instead of simply seeding our RNG at position i, we can seed by $i \bmod m$ for some parameter m
- In detection, we iterate through $s \in \{0, 1, \ldots, m-1\}$ and compute $(i + s) \bmod m$
- Choose the offset with the greatest detection score

Empirical results

- We use the red/green list scheme as a baseline
- Compare GSM8k (math problems) and BBH (challenging logic) performance for unwatermarked vs watermarked text
- Assess detectability of open-ended prompts

Red / Green list results (baseline): benchmarks

Table: Comparison of Green-List Watermarking Results on GSM8K and BBH Benchmarks (100 prompts each).

Model (Benchmark)	Hyperparameters	Correctness (%)	Detectability (%)
Llama (GSM8K)	δ=0, γ=0.25	54	19
	δ=2, γ=0.25	32	90
LLaDA (GSM8K)	δ=0, γ=0.025	71	2
	δ=6, γ=0.025, S_W=$\{S_1 \ldots S_{200}\}$	21	92
Llama (BBH)	δ=0, γ=0.25	84	0
	δ=2, γ=0.25	67	46
LLaDA (BBH)	δ=0, γ=0.025	90	0
	δ=6, γ=0.025, S_W=$\{S_1 \ldots S_{200}\}$	75	3

Red / Green list results (baseline): benchmarks

Table: Comparison of Green-List Watermarking Results on GSM8K and BBH Benchmarks (100 prompts each).

Model (Benchmark)	Hyperparameters	Correctness (%)	Detectability (%)
Llama (GSM8K)	δ=0, γ=0.25	54	19
	δ=2, γ=0.25	32	90
LLaDA (GSM8K)	δ=0, γ=0.025	71	2
	δ=6, γ=0.025, S_W=$\{S_1 \ldots S_{200}\}$	21	92
Llama (BBH)	δ=0, γ=0.25	84	0
	δ=2, γ=0.25	67	46
LLaDA (BBH)	δ=0, γ=0.025	90	0
	δ=6, γ=0.025, S_W=$\{S_1 \ldots S_{200}\}$	75	3

- Takeaway 1: Benchmark performance significantly drops for watermarked text
- Takeaway 2: Again, the choice of parameters to yield a detectable watermark vary by task

Our watermark results: benchmarks

Table: Testing our watermarking scheme on GSM8K and BBH benchmarks ($\tau = 1.015$).

Model (Benchmark)	Watermark	Correctness (%)	Detectability (%)
LLaDA (GSM8K)	No	63	39
	Yes	**71**	86
LLaDA (BBH)	No	89	43
	Yes	**89**	47

Aside: How do we choose the detection threshold τ?

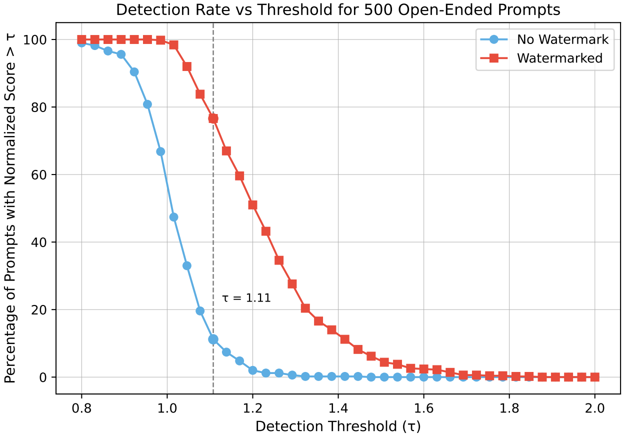

Figure: Percentage of open-ended prompts that exceed threshold τ, for different values of τ. We show results for unwatermarked and watermarked text, illustrating the tradeoff between soundness and completeness.

Our watermark results: open-ended generation

Table: Testing our watermarking scheme on open-ended generation (temp $= 1$, $\tau^* = 1.11$).

Model	Watermark	Perplexity	Detectability (%)
LLaDA	No	5.715	11
	Yes	**5.070**	**77**

Our watermark results: open-ended generation

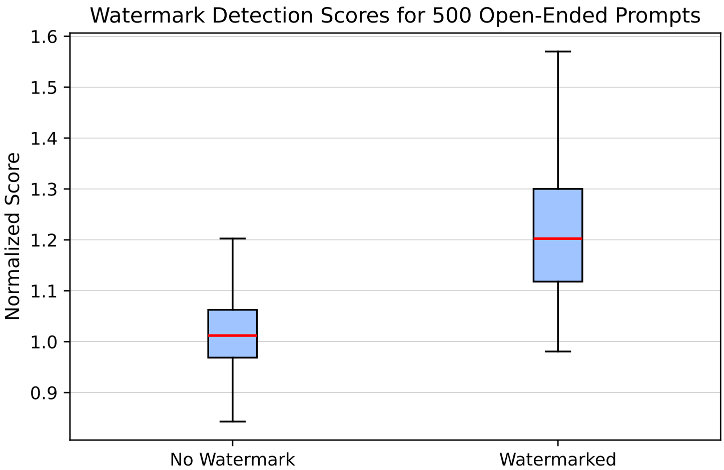

Figure: Distribution of normalized detection scores for unwatermarked vs watermarked text using our modified Gumbel-max scheme. We use 500 open-ended prompts.

Conclusion

- We introduced the first watermark for discrete diffusion language models
- We demonstrate its completeness empirically, and its soundness and distortion-freeness both theoretically and empirically

Conclusion

- We introduced the first watermark for discrete diffusion language models
- We demonstrate its completeness empirically, and its soundness and distortion-freeness both theoretically and empirically
- Future work can:
  - Implement our framework for additional models other than LLaDA
  - Improve robustness guarantees beyond prefix deletions

References I

Scott Aaronson and Hendrik Kirchner. Watermarking gpt outputs. Lecture slides https://www.scottaaronson.com/talks/watermark.ppt, 2022. Accessed: 2025-10-13.

Minghui Hu, Yujie Wang, Tat-Jen Cham, Jianfei Yang, and P. N. Suganthan. Global context with discrete diffusion in vector quantised modelling for image generation, 2021. URL https://arxiv.org/abs/2112.01799.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models, 2024. URL https://arxiv.org/abs/2301.10226.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models, 2024. URL https://arxiv.org/abs/2307.15593.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025. URL https://arxiv.org/abs/2502.09992.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P. de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models, 2025. URL https://arxiv.org/abs/2412.10193.

Xu Wang, Chenkai Xu, Yijie Jin, Jiachun Jin, Hao Zhang, and Zhijie Deng. Diffusion llms can do faster-than-ar inference via discrete diffusion forcing, 2025. URL https://arxiv.org/abs/2508.09192.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust, 2023. URL https://arxiv.org/abs/2305.20030.