GaussMark:
A Practical Approach for Structural Watermarking of Language Models

Adam Block, Alexander Rakhlin, Ayush Sekhari

Gahyun baek

GaussMark: A Practical Approach for
Structural Watermarking of Language Models

Adam Block[1], Alexander Rakhlin[2], and Ayush Sekhari[2]

[1]Microsoft Research
[2]MIT

Source: Block et al., ICML 2025

LLM Watermarking

- Watermark ?

watermark

/ˈwôdərˌmärk/

A **watermark** is a faint design, logo, or text embedded into paper or digital media to show ownership, prove authenticity, or mark a file's status. Dropbox +3

- The key requirements: **Invisible signal in text**

Source: Block et al., ICML 2025

Why do we need LLM Watermarking ?

- **AI-generated content is widespread**

1. Human-written vs. AI-generated text
2. Claim provenance, copyright

Can you guess which one was generated by AI ?

I attended the Red Teaming Challenge yesterday. At first, I solved the problems one by one. However, during the challenge, we noticed that other teams were solving the problems with automatic agents.
I was frustrated and almost lost motivation to solve the problems. I think the challenge rules should be changed to make it more fair.

Yesterday I went to a red teaming challenge.
We were solving every problem manually, one by one. But most other teams were using agents and automating everything. They could try way more prompts than us, so we were at a pretty big disadvantage.
It was a little frustrating, but next time I think we definitely need to prepare some automation too.

Source: Block et al., ICML 2025

Why do we need LLM Watermarking ?

- **AI-generated content is widespread**

1. Human-written vs. AI-generated text
2. Claim provenance, copyright

Can you guess which one was generated by AI ?

I attended the Red Teaming Challenge yesterday. At first, I solved the problems one by one. However, during the challenge, we noticed that other teams were solving the problems with automatic agents.
I was frustrated and almost lost motivation to solve the problems. I think the challenge rules should be changed to make it more fair.

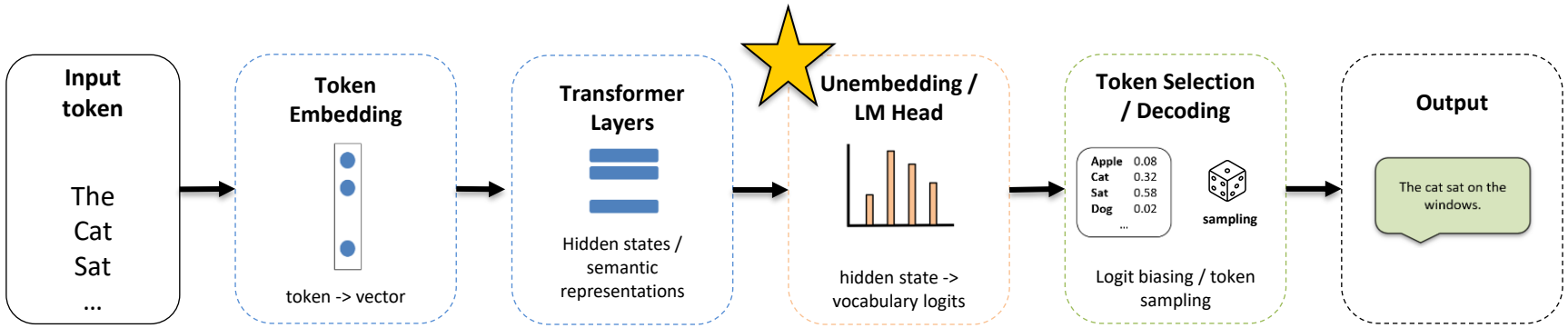
Gahyun

Yesterday I went to a red teaming challenge.
We were solving every problem manually, one by one.
But most other teams were using agents and automating everything. They could try way more prompts than us, so we were at a pretty big disadvantage.
It was a little frustrating, but next time I think we definitely need to prepare some automation too.

ChatGPT

Source: Block et al., ICML 2025

From Token-Level Watermarking to Structural Watermarking



Token sampling	Model weights
• Watermark -> logits -> sample • Appplied at every single token • The sampling pipeline must be rewritten • Latency grows with sequence length.	• Weights + ξ-> ordinary generation • Applied once, before generation • The sampling pipeline is untouched • No additional generation latency

Source: Block et al., ICML 2025

Perturbation in some weights

THE TRUTH IS IN THERE: IMPROVING REASONING IN LANGUAGE MODELS WITH LAYER-SELECTIVE RANK REDUCTION

Pratyusha Sharma[1] **Jordan T. Ash**[2*] **Dipendra Misra**[2*]
[1]Massachusetts Institute of Technology [2]Microsoft Research NYC

ABSTRACT

Transformer-based Large Language Models (LLMs) have become a fixture in modern machine learning. Correspondingly, significant resources are allocated towards research that aims to further advance this technology, typically resulting in models of increasing size that are trained on increasing amounts of data. This work, however, demonstrates the surprising result that it is often possible to significantly improve the performance of LLMs by selectively removing higher-order components[1] of their weight matrices. This simple intervention, which we call LAyer-SElective Rank reduction (LASER), can be done on a model after training has completed, and requires minimal additional parameters and data. We show extensive experiments demonstrating the generality of this finding across language models and datasets, and provide in-depth analyses offering insights into both when LASER is effective and the mechanism by which it operates[2].

=> Low components act as noise
=> Modifying selected weight components does not necessarily destroy model performance.

A small additive perturbation of model weights can instead be used as a watermark.

THE TRUTH IS IN THERE: IMPROVING REASONING IN LANGUAGE MODELS WITH LAYER-SELECTIVE RANK REDUCTION, Sharma et al., ICLR, 2024

Source: Block et al., ICML 2025

Hypothesis testing problem

- **Null hypothesis**
- This text is not related to the watermarking key.

$$(\xi', y') \sim \nu \otimes q \text{ with } q \in \Delta(\mathcal{Y}).$$

- **Alternative hypothesis**
- This text is exactly from the weight perturbed model.

$$\text{key } \xi', \text{ i.e., } y' \sim p_{\theta(\xi')}(\cdot \mid x).$$

y: candidate text, ξ: secret key

Source: Block et al., ICML 2025

Log-Likelihood Ratio

: A natural way to compare the null and alternative hypotheses

$$\log \frac{p_{H_A}(y)}{p_{H_0}(y)}$$

=> The larger the value is, the H_A is more available

[LLR]

- Compares the two competing hypotheses
- A higher ratio favours the watermark

Source: Block et al., ICML 2025

GaussMark - generate

- Structural inherent in model weights
- Modifies model weights
- Adds Gaussian perturbation
- Generates normally afterward
- No token-level biasing

Algorithm 1 GaussMark.Generate

1: **Input** Language Model $p_\theta : \mathcal{V}^\star \to \Delta(\mathcal{Y})$ with parameters $\theta \in \Theta$, prompt x, variance $\sigma^2 > 0$.
2: **Sample** watermark key $\xi \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$.
3: **Watermark** model by setting $\theta(\xi) = \theta + \xi$.
4: **Generate** using the perturbed model: $y \sim p_{\theta(\xi)}(\cdot \mid x)$.

θ: base model's weights for a layer
ξ: secret key
θ'=θ+ξ

$$\theta = \begin{bmatrix} 0.32 \\ -0.51 \\ 0.14 \end{bmatrix} \quad \xi = \begin{bmatrix} 0.003 \\ -0.001 \\ 0.002 \end{bmatrix}$$

$$\theta + \xi = \begin{bmatrix} 0.323 \\ -0.511 \\ 0.142 \end{bmatrix}$$

Source: Block et al., ICML 2025

GaussMark - detect

y: candidate text, ξ: secret key

Algorithm 2 GaussMark.Detect

1: **Input** Language Model $p_\theta : \mathcal{V}^\star \to \Delta(\mathcal{Y})$ with parameters $\theta \in \Theta$, prompt x, variance $\sigma^2 > 0$, watermark key ξ, candidate text $y \in \mathcal{Y}$, level $\alpha \in (0, 1)$.

2: **Evaluate** the test statistic

$$\psi(y, \xi \mid x) = \frac{\langle \xi, \nabla_\theta \log p_\theta(y \mid x) \rangle}{\sigma \left\| \nabla_\theta \log p_\theta(y \mid x) \right\|}. \tag{1}$$

3: **Return** 'Watermarked' **if** $\psi(y, \xi \mid x) \geq \Phi^{-1}(1 - \alpha)$; **else** fail to reject null.

Source: Block et al., ICML 2025

GaussMark - detect

$$\psi(y, \xi \mid x) = \frac{\langle \xi, \nabla_\theta \log p_\theta(y \mid x) \rangle}{\sigma \left\| \nabla_\theta \log p_\theta(y \mid x) \right\|}.$$

y: candidate text, ξ: secret key

1. Calculate text log-likelihood with base model

$$\log p_\theta(y|x)$$

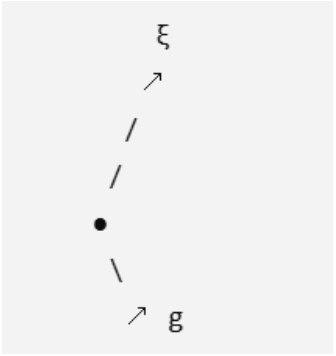
2. Backpropagation to obtain gradient **g**

$$g = \nabla_\theta \log p_\theta(y|x)$$

3. Measure key–gradient alignment

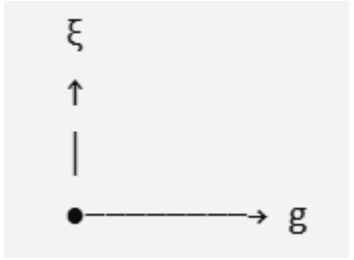
$$\xi \quad \text{vs.} \quad g$$

$$\psi = \frac{1}{\sigma} \left\langle \xi, \frac{g}{\|g\|} \right\rangle$$



watermarked

Stronger alignment



unwatermarked

No systematic alignment

Generation perturbs the model in direction ξ, and detection checks whether the generated text still contains evidence of that direction.

Source: Block et al., ICML 2025

The model's performance is preserved

Table 1. Performance of GaussMark on various models.

Model	SuperGLUE (**Avg**)		GSM–8K (**Acc**)		AlpacaEval–2.0 (**win rate**)	
	Unwatermarked	GaussMark	Unwatermarked	GaussMark	Unwatermarked	GaussMark
Llama3.1–8B	0.7012 ± 0.0336	0.7094 ± 0.0327	0.6300 ± 0.0133	0.6111 ± 0.0134	(0.46, 0.54)	0.45
Mistral–7B	0.6836 ± 0.0335	0.7033 ± 0.0332	0.4291 ± 0.0136	0.4321 ± 0.0136	(0.49, 0.51)	0.50
Phi3.5–Mini	0.6451 ± 0.0254	0.6489 ± 0.0258	0.8423 ± 0.0100	0.8552 ± 0.0097	(0.46, 0.54)	0.49

- GaussMark is **not distortion-free**, so model quality has to be evaluated empirically: quality has to be measured rather than guaranteed.

Source: Block et al., ICML 2025

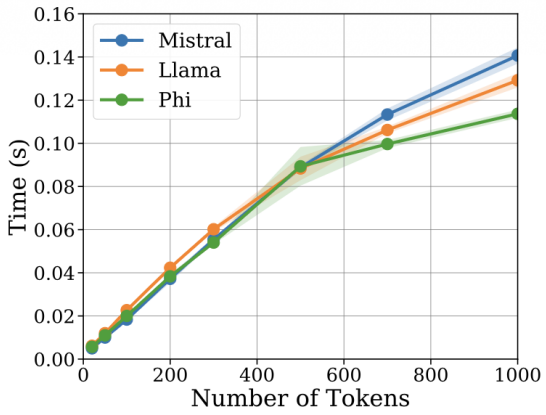
The selected layer is different

Model	Watermarked Layer	Watermarked Weight	Variance (σ^2)	Dimension (d)
Mistral–7B	20	up_proj	1e-05	58M
Llama3.1–8B	28	up_proj	3e-04	58M
Phi3.5–Mini	20	down_proj	1e-03	25M

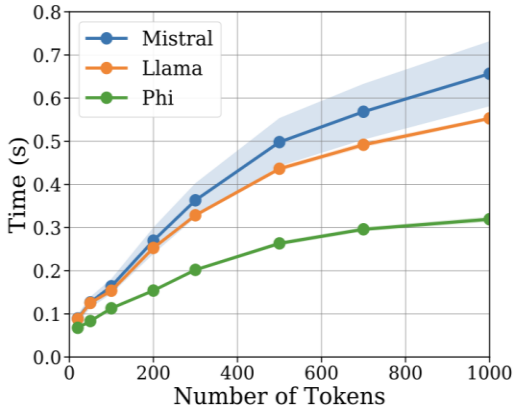
- Balancing detectability-text quality
- Different layers have different sensitivity to perturbations

Source: Block et al., ICML 2025

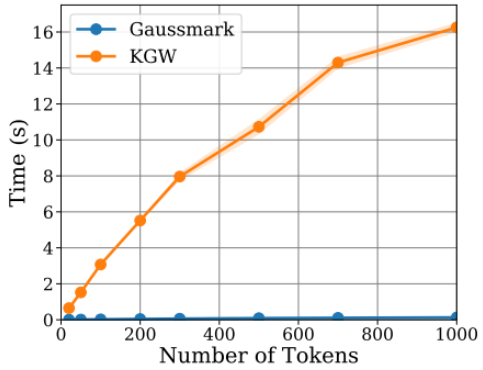
Generation Adds No Watermarking Latency



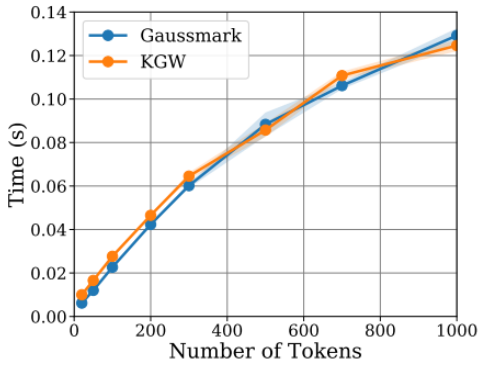
(a) Generation time



(b) Detection time



(a) Generation time (Llama3.1–8B).



(b) Detection time (Llama3.1–8B).

- Generation cost equals the base model
- 1K tokens detected in under a second
- Orders of magnitude faster than KGW

Only the weights change -> no latency degradation

Source: Block et al., ICML 2025

Limitation and Future Work

- **Limitation**
  - Not multi-bit watermarking scheme
  - Prompt is required when detection
  - Base model is required
  - No distortion free guarantee
  - White-box access
  - Requires more tokens

- **Personal thoughts**
  - Not token-level watermarking.
  - Can this extend to multi-bit watermarking?
  - The first try modifying weights to embed watermarking signals.

Source: Block et al., ICML 2025

Thank you

Source: Block et al., ICML 2025