

Reproduction of jwkirchenbauer/lm-watermarking: watermark detection with OPT-1.3B

K-Veritas Team

Submitted by K-Veritas Team

kv:2610.00039v1 · Published 2026-10-11

Independent reproduction · Artifact evaluation · Security and cryptography

Abstract

Ran the repository's demo_watermark.py at the pinned commit with facebook/opt-1.3b and the default watermark settings (gamma 0.25, delta 2.0, simple_1 seeding, multinomial sampling at temperature 0.7, 200 new tokens, z-score threshold 4.0), once for each of five generation seeds (123, the default, then 1 to 4). Each run continues the script's built-in prompt with and without the watermark and runs the detector on both outputs, reporting the z-score and greenlist fraction. The experiments/, watermark_reliability_release/ and hf_hub_space_demo/ folders are withheld as hash-only entries since this run does not use them. PyTorch 2.11 (CUDA 12.8), transformers 5.19. One NVIDIA RTX 5060 Ti.

About this record

Title, authors and abstract were provided by the authors and reviewed for completeness, not for scientific soundness. Every number in this record is copied from the signed data of the sealed reports [2], which remain authoritative.

This cover page is not signed. Check it against the SHA-256 listed at <https://kveritas.org/records/2610.00039v1>, and verify each report with its code bundle: `kveritas verify <report.pdf> --bundle <bundle.zip>`.

1 Sealed reports

The experiments were sealed with K-Veritas [1], which binds each result to the code, hardware and time that produced it.

Report	Label	Sealed	Runs	Invocations	Signer
r1 [2]	OPT-1.3B watermark demo, 5 ...	2026-10-10	1	1	Signed by the K-Veritas server

2 Results

r1: OPT-1.3B watermark demo, 5 seeds

Run	Metric	Final value	Step
1	plain_z_score	0.707	4
1	plain_green_fraction_pct	28.1	4
1	watermarked_z_score	10.8	4
1	watermarked_green_fraction_pct	58.3	4

Run	Inline claim	Value	Phase
1	plain_z_score_seed_123	0.205	
1	watermarked_z_score_seed_123	11.3	
1	plain_z_score_seed_1	-2.25	
1	watermarked_z_score_seed_1	8.55	

Run	Inline claim	Value	Phase
1	plain_z_score_seed_2	0.532	
1	watermarked_z_score_seed_2	9.37	
1	plain_z_score_seed_3	-0.123	
1	watermarked_z_score_seed_3	9.04	
1	plain_z_score_seed_4	0.707	
1	watermarked_z_score_seed_4	10.8	

3 Setup

Report	Item	Value
r1	Operating system	linux / amd64
r1	CPU	Intel(R) Xeon(R) W-2145 CPU @ 3.70GHz, 16 cores
r1	GPU	NVIDIA GeForce RTX 5060 Ti, 16311 MiB
r1	Memory	31.0 GB
r1	Total run time	64.1 s

4 Files

Each report and code bundle must hash to these values. A bundle belongs to the report whose signed data names its hash.

File	SHA-256	Size
r1.pdf	8320420999fc0b4c2b31f32206cc02750fe21a6036f12d80510952892e564f7c	755 KB
r1.kvbundle.zip	0f22f086a877ed4650cba6fdadc3830c36a48fd30fb7e09394c9f140e497d765	121 KB

5 Links and licence

Paper: <https://arxiv.org/abs/2301.10226>

Reproduction of:

<https://github.com/jwkirchenbauer/lm-watermarking/tree/82922516930c02f8aa322765defdb5863d07a00e>

Record: CC-BY 4.0. Code: Apache-2.0, as declared by the authors. Tags: watermarking, llm, detection, opt.

Listed, not endorsed. Review checks the artifact, never the science. Reports are never edited; changing them makes a new version. Other fields are corrected in place, and every correction is logged in metadata.yaml.

References

[1] M. K. Keita and C. Homan. Computer Science Conferences Should Require Nonrepudiable Experimental Results. NeurIPS 2026 Position Paper Track. arXiv:2605.08586.

[2] Sealed experiment report r1 (OPT-1.3B watermark demo, 5 seeds), kv:2610.00039v1/r1. Data hash ac03b8e5033482f06efa21de3c8e0ceb7f32d4968df7a3a05837d1e49e879464.

<https://kveritas.org/records/2610.00039v1/r1.pdf>