

Reproduction of locuslab/wanda: pruning OPT-1.3B with Wanda and magnitude

K-Veritas Team

Submitted by K-Veritas Team

kv:2610.00037v1 · Published 2026-10-11

Independent reproduction · Artifact evaluation · Machine learning

Abstract

Pruned OPT-1.3B with the repository's main_opt.py at the pinned commit and measured WikiText-2 perplexity for three settings: Wanda at 50% unstructured sparsity, Wanda with 2:4 semi-structured sparsity, and the magnitude-pruning baseline at 50% unstructured, each with 128 C4 calibration samples. Changes: transformers 4.40 and PyTorch 2.11 instead of the pinned 4.28 and 1.10, and lib/data.py loads C4 without the retired 'allenai--c4' config name (same data files). 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/2610.00037v1>, 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: Wanda 50%, Wan...	2026-10-10	1	1	Signed by the K-Veritas server

2 Results

r1: OPT-1.3B: Wanda 50%, Wanda 2:4, magnitude 50%

Run	Metric	Final value	Step
1	wanda_unstructured_50_actual_sparsity	0.5	
1	wanda_unstructured_50_wikitext2_ppl	18.4108	
1	wanda_2of4_actual_sparsity	0.5	
1	wanda_2of4_wikitext2_ppl	28.2521	
1	magnitude_unstructured_50_actual_sparsity	0.5001	
1	magnitude_unstructured_50_wikitext2_ppl	1713.59	

Run	Inline claim	Value	Phase
1	wanda_unstructured_50_wikitext2_ppl	18.4108	
1	wanda_2of4_wikitext2_ppl	28.2521	
1	magnitude_unstructured_50_wikitext2_ppl	1713.59	

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	459.3 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	cc04c3467266f10122da40eaf79193cc58a7305b4bcaad23b482bbb0ca299f6c	4.1 MB
r1.kvbundle.zip	7945ac5100d3fc6c4caef643c5992d0b2c233d2c3ce468af4eab55a77aa23858	27 KB

5 Links and licence

Paper: <https://arxiv.org/abs/2306.11695>

Reproduction of: <https://github.com/locuslab/wanda/tree/8e8fc87b4a2f9955baa7e76e64d5fce7fa8724a6>

Record: CC-BY 4.0. Code: MIT, as declared by the authors. Tags: pruning, sparsity, llm-compression, 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: Wanda 50%, Wanda 2:4, magnitude 50%), kv:2610.00037v1/r1. Data hash ebe63d9fbaa8d39c38ecb640cf95a443f663d88873ca18a3c8c1902a8850f49f.

<https://kveritas.org/records/2610.00037v1/r1.pdf>