AI Course Project · CSCI-B-M315 · Advanced AI Topics

Wanda

Pruning Large Language Models with Weights AND Activations

Zineb ABERCHA

Omar Alfarouq BOUHADI

Supervised by Prof. Hamza KEURTI

12 January 2026

The Problem: Deploying LLMs is Expensive

LLMs require massive resources

- Billions of parameters
- 14GB+ GPU memory for 7B model
- High inference cost

Solution: Neural Network Pruning

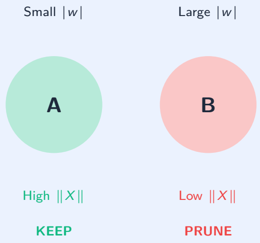
Remove redundant weights while preserving accuracy.

Traditional approach: Magnitude Pruning

- Remove weights with smallest $|w|$
- Simple but fundamentally flawed

Key Insight

"Weight magnitude alone does not determine importance."



Small $|w|$ | Large $|w|$

A | B

High $||X||$ | Low $||X||$

KEEP | PRUNE

The Approach: Wanda (Weights AND Activations)

Magnitude Pruning

$$S_{\text{mag}} = |w_{ij}|$$

Only considers weight size

× Ignores input patterns

× Fails at high sparsity

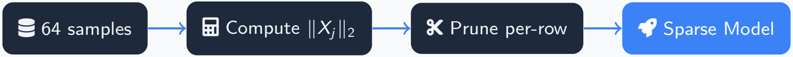
Wanda Pruning

$$S_{\text{wanda}} = |w_{ij}| \cdot \|X_j\|_2$$

Weight × activation norm

✓ Preserves critical paths

✓ No retraining needed



64 samples → Compute $\|X_j\|_2$ → Prune per-row → Sparse Model

One-shot pruning · Layer-by-layer · Per-row sparsity

Implementation & Setup

Our Implementation

- PyTorch + HuggingFace Transformers
- Layer-by-layer pruning with hooks
- Per-row sparsity allocation
- 64 calibration samples (WikiText-2)

Models Tested

LLaMA-2-7B	6.7B params
LLaMA-3.1-8B	8.0B params

Evaluation

Perplexity (WikiText-2 test)

- Lower = better language modeling

Zero-Shot (5 benchmarks)

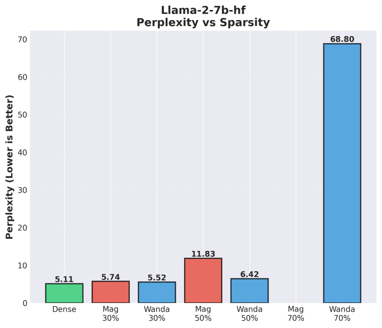
- PIQA, HellaSwag, ARC-E, BoolQ, RTE

Sparsity Levels

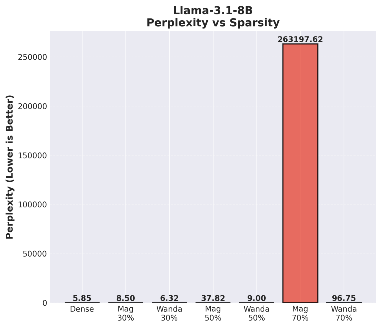
30% **50%** **70%**

NVIDIA L40S (48GB)

Results: Perplexity on WikiText-2



LLaMA-2-7B

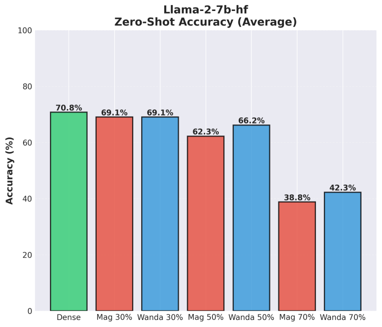


LLaMA-3.1-8B

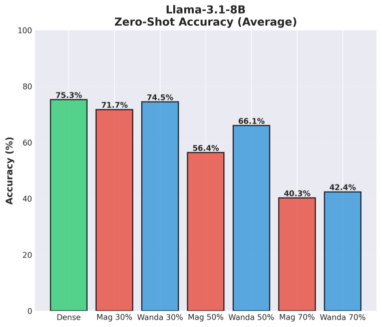
Model	Magnitude	Wanda	Improvement
LLaMA-2 @ 50%	11.83	**6.42**	**45.8%↓**
LLaMA-3 @ 50%	37.82	**9.00**	**76.2%↓**

At 70%: Magnitude → NaN while Wanda remains functional

Results: Zero-Shot Task Accuracy



LLaMA-2-7B



LLaMA-3.1-8B

Model	Dense	Mag 50%	Wanda 50%	Retained
LLaMA-2	70.8%	62.3%	66.2%	93.5%
LLaMA-3	75.3%	56.4%	66.1%	87.8%

Wanda: **+4pp** (LLaMA-2) and **+10pp** (LLaMA-3) more accuracy

Comparison with Original Paper

WikiText-2 Perplexity at 50% Unstructured Sparsity

Model	Dense	Magnitude	Wanda	Source
LLaMA-7B	5.68	17.29	7.26	Original Paper
LLaMA-2-7B	5.11	11.83	**6.42**	Our Reproduction

Reproduction Success

- ✓ Same trend: Wanda **58%** better
- ✓ Similar improvement ratio
- ✓ Confirms paper's claims

Differences Explained

- LLaMA-2 vs LLaMA-1
- 64 vs 128 calibration samples
- Different checkpoints

✓ **Wanda's effectiveness is reproducible**

Comparison to Theory & Conclusions

Theory Validated

- ✓ Wanda outperforms magnitude at all sparsity
- ✓ Advantage scales: 3.7%→76.2%
- ✓ Stable where magnitude fails

Why It Works

- Emergent high-activation features
- $\|X_j\|_2$ captures input importance

Limitations

- ⚠ 70%+ degrades all methods
- ⚠ Needs specialized HW for speedup

Key Takeaways

1. **50% sparsity** = optimal
2. 2× time worth 45-76% better PPL
3. One-shot, **no retraining**

Wanda = Simple + Effective + Practical

Thank You

Questions?

Reference

Sun, M., Liu, Z., Bair, A., & Kolter, J. Z. (2024)

A Simple and Effective Pruning Approach for LLMs

ICLR 2024