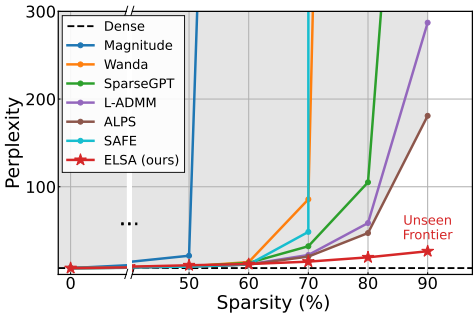
The Unseen Frontier: Pushing the Limits of LLM Sparsity with Surrogate-free ADMM

Kwanhee Lee[1] Hyeondo Jang[1] Dongyeop Lee[1] Dan Alistarh[2] Namhoon Lee[1]

ICLR 2026

[1]Pohang University of Science and Technology (POSTECH)
[2]Institute of Science and Technology Austria (ISTA)

Motivation: The Sparsity Wall



- LLM pruning can drastically cut costs (Frantar and Alistarh 2023; Sun et al. 2024)
- But existing methods hit a **"sparsity wall"** at 50–60%
- ELSA pushes LLM sparsity to **80–90%** while staying stable

Sparsity Wall: Limitations of Layer-wise REM

Prior methods minimize a **surrogate** objective layer by layer (Shin et al. 2024; Huang et al. 2025):

$$x^* = \left\{ \arg \min_{\|x_i\|_0 \le k_i} \mathbb{E}_{\mathcal{D}} \left[\|\tilde{x}_i^T g(x_{i-1}; \mathcal{D}) - x_i^T g(x_{i-1}; \mathcal{D})\|^2 \right] \text{ for } i = 1, \ldots, L \right\}$$

1. **Compounding errors** — approximate layer-wise solutions accumulate errors across layers
2. **Suboptimality** — sequential, independent layers are never jointly optimized
3. **Surrogate $\neq$ true objective** — layer-wise reconstruction error $\tilde{f}$ is not the LLM loss f

The wall is an artifact of the formulation, not an inherent limit.

ELSA: Surrogate-free LLM Pruning via ADMM

Optimize the **true** LLM objective under a sparsity constraint:

$$\min_{x,z\in\mathbb{R}^d} f(x) + I_{\mathcal{S}}(z) \quad \text{subject to} \quad x = z$$

ADMM decomposes this into three tractable subproblems:

$$x_{k+1} = \operatorname*{argmin}_x f(x) + \tfrac{\lambda}{2}\|x - z_k + u_k\|_2^2 \qquad \text{(gradient descent on true loss)}$$

$$z_{k+1} = \mathrm{proj}_{\|\cdot\|_0 \le k}(x_{k+1} + u_k) \qquad \text{(sparsity projection)}$$

$$u_{k+1} = u_k + x_{k+1} - z_{k+1} \qquad \text{(dual update)}$$

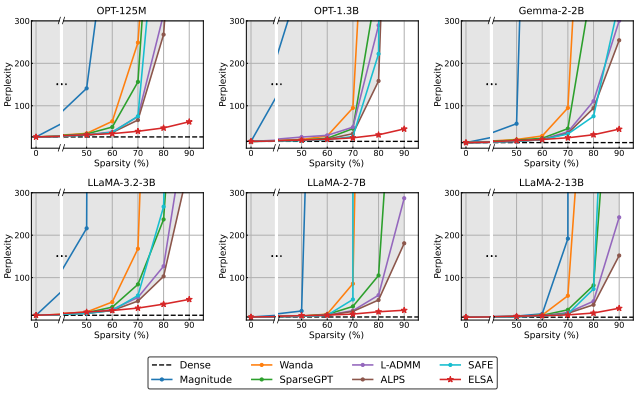
More details

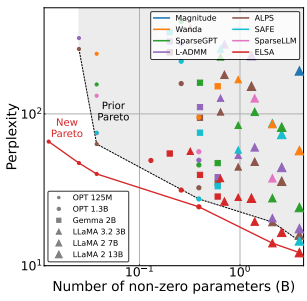
- **Objective-aware projection**: use Hessian (empirical Fisher) instead of magnitude

$$z^{t+1} = \operatorname*{argmin}_{z \in \mathcal{S}} \sum_{i \leq d} \hat{\mathbf{F}}_{ii} \left(z_i - (x_i^{t+1} + u_i^t) \right)^2$$

- **Elsa-L**: low-precision ADMM states for scaling to 27B parameters
- **Convergence guarantees**: provably converges to a stationary point, even with quantized variant ELSA-L

Main Results: Perplexity vs. Sparsity





7.8× less perplexity than best baseline on LLaMA-2-7B at **90%** sparsity

Practical Benefits

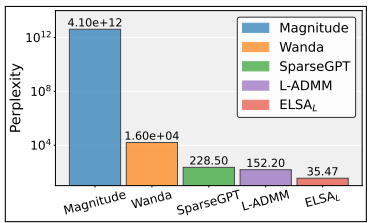
Inference speedup & memory (LLaMA-2-7B)

	Dense	90%	95%
Speedup	1×	2.56×	**3.98**×
Memory	13.6 GB	2.9 GB	1.7 GB

Perplexity (LLaMA-2-7B, C4)

Method	90%	95%
Wanda + LoRA	65.56	143.0
Wanda + Full	34.87	53.62
Elsa	**23.14**	**28.39**

- ELSA stays stable at 95% sparsity: 3.98× speedup, 7.80× memory reduction
- ELSA-L scales to **Gemma-2-27B** at 90% sparsity



Conclusion

- The "sparsity wall" was an artifact — ELSA breaks through to **80–90%**
- Principled ADMM optimization on the **true LLM objective**
- Practical: real speedup & memory savings on commodity hardware

Paper

Code

References I

- Frantar, Elias and Dan Alistarh (2023). "SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot". In: *ICML*.
- Huang, Weizhong et al. (2025). "Determining layer-wise sparsity for large language models through a theoretical perspective". In: *arXiv preprint arXiv:2502.14770*.
- Shin, Sungbin et al. (2024). "Rethinking Pruning Large Language Models: Benefits and Pitfalls of Reconstruction Error Minimization". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1182–1191.
- Sun, Mingjie et al. (2024). "A Simple and Effective Pruning Approach for Large Language Models". In: *ICLR*.