SARATHI: Decoupled Data-Free Probing and Adaptive Reconstruction for Structured LLM Pruning

Anonymous ACL submission

Abstract

Structured pruning of LLMs enables hardware-agnostic speedup by shrinking FFN dimensions, yet existing methods couple neuron selection and weight reconstruction through shared calibration data, causing domain bias and OOM failures at scale. We introduce **SARATHI**: a *data-free NMF Residual Subspace Probe* identifies irreplaceable neurons via weight geometry (grounded in Robust PCA), *Adaptive OBS* reconstructs pruned weights in $\mathcal{O}(d^2)$ memory per layer (vs. $\mathcal{O}(Ld^2)$ for coupled baselines), and a zero-overhead *Bias Shift Compensation* corrects post-LayerNorm activation shifts, reducing PPL from 2,082 to 18.98 on OPT-2.7B at 40% sparsity vs. the uncorrected pruned model. Across five models (LLaMA-3-8B, Mistral-7B, OPT-2.7B/6.7B, LLaMA-2-13B) at 15–40% sparsity, SARATHI achieves state-of-the-art WikiText-2 PPL, outperforms SoBP, SliceGPT, and Dynamic Slicing on zero-shot accuracy, and is the *only* method to compress OPT-6.7B at 40% sparsity where all baselines crash OOM; a lightweight variant (SARATHI-Wanda) using a 128-sample calibration batch further improves zero-shot accuracy on gated architectures. Code is available at: `https://anonymous.4open.science/r/SARATHI-E9FD/`

1 Introduction

Large language models (LLMs) deliver strong performance across NLU and NLG tasks (Brown et al., 2020; Touvron et al., 2023; Jiang et al., 2023), yet their billions of parameters impose prohibitive memory and latency costs at deployment. Structured pruning addresses this by physically removing entire rows or columns from weight matrices (reducing FFN intermediate dimension from N to $\lfloor(1-\rho)N\rfloor$ at sparsity ρ), yielding dense smaller matrices that accelerate inference on any BLAS hardware, unlike unstructured methods (Frantar and Alistarh, 2023; Sun et al., 2024) that require specialised sparse kernels.

The coupling flaw. Existing post-training structured pruning methods (Wei et al., 2024; An et al., 2024; Ashkboos et al., 2024) share a fundamental design flaw: they use the *same* calibration dataset $\mathcal{D}_{\text{calib}}$ to simultaneously (i) score which neurons to prune and (ii) compute the weight-compensation matrix. This coupling is computationally costly: building the activation Hessian $\mathbf{H} = \mathbf{X}_{\text{calib}}\mathbf{X}_{\text{calib}}^\top \in \mathbb{R}^{d\times d}$ per layer requires $\mathcal{O}(Ld^2)$ memory, and is statistically harmful, biasing the pruning mask toward neurons active on $\mathcal{D}_{\text{calib}}$ and away from neurons important for open-domain generalization. As demonstrated in §3, this causes out-of-memory (OOM) failures and accelerated perplexity degradation at high sparsity.

SARATHI. SARATHI decouples Eq. 1 into two independent subproblems (§2.2): (P1) a *data-free* selection stage using the intrinsic geometry of the weight matrix, and (P2) a localized reconstruction stage with $\mathcal{O}(d^2)$ peak memory per layer. We evaluate two probe variants: **SARATHI-NMF**, which is entirely data-free for neuron selection, and **SARATHI-Wanda**, which uses activation norms from a small calibration batch for the selection stage, serving as an ablation that isolates the reconstruction and bias-correction contributions. The selection probe (§2.3) applies NMF to $|\mathbf{W}^{(\ell)}|$, identifying neurons whose weight rows lie outside the shared low-rank non-negative subspace as structurally irreplaceable (a formal connection to Robust PCA (Candès et al., 2011)). SARATHI also identifies and corrects a *cascading bias shift* in post-LayerNorm architectures (§2.5), reducing WikiText-2 PPL on OPT-2.7B at 40% from 2,082 to 18.98, a 99.1% reduction vs. the uncorrected pruned model (though still above the dense baseline of 12.47; see Figure 2).

Contributions.

- **Decoupled Probe-Reconstruct** architecture eliminating calibration bias and the $\mathcal{O}(Ld^2)$

memory bottleneck (§2.2). SARATHI prunes only FFN intermediate neurons; attention head pruning is left to future work.

- **NMF Residual Subspace Probing**: data-free neuron scoring with a Robust PCA connection and formal pruning-safety bound (Proposition 1, §2.3).
- **Bias Shift Compensation**: zero-overhead closed-form correction for post-LayerNorm PPL collapse (§2.5).

Evaluation on five models at 15–40% sparsity (§3) shows state-of-the-art PPL on all models, superior zero-shot accuracy over SoBP on OPT models and over SliceGPT and Dynamic Slicing across all architectures, and the only successful 40% compression of OPT-6.7B and LLaMA-2-13B.

2 The SARATHI Framework

2.1 Problem Formulation and the Coupling Flaw

Consider a transformer with L MLP layers. For the ℓ-th layer, let $\mathbf{W}^{(\ell)} \in \mathbb{R}^{N\times d}$ be the projection weight matrix, where N is the intermediate dimension and d is the model hidden size. Given input activations $\mathbf{X} \in \mathbb{R}^{d\times T}$ over T token positions, the layer output is $\mathbf{Z} = \mathbf{W}^{(\ell)}\mathbf{X} \in \mathbb{R}^{N\times T}$.

Structured pruning seeks a binary row-selection mask $\mathbf{s} \in \{0,1\}^N$ that retains exactly $\lfloor(1-\rho)N\rfloor$ rows (neurons) at target sparsity $\rho \in (0,1)$, yielding the compressed weight $\mathbf{W}_{\mathbf{s}}^{(\ell)} = \mathbf{W}^{(\ell)}[\mathbf{s},:]$. The ideal post-pruning weight correction $\Delta\mathbf{W}$ is co-optimized with $\mathbf{s}$, giving the *coupled* objective solved by SoBP (Wei et al., 2024), SparseGPT (Frantar and Alistarh, 2023), and SliceGPT (Ashkboos et al., 2024):

$$\min_{\mathbf{s},\,\Delta\mathbf{W}} \left\| \mathbf{W}^{(\ell)}\mathbf{X}_{\text{calib}} - \left(\mathbf{W}_{\mathbf{s}}^{(\ell)} + \Delta\mathbf{W}\right)\mathbf{X}_{\text{calib}} \right\|_F^2 \quad \text{s.t.} \quad \|\mathbf{s}\|_0 = \lfloor(1-\rho)N\rfloor \tag{1}$$

Remark 1 (NP-hardness and calibration coupling)**.** *The $\|\mathbf{s}\|_0{=}K$ constraint is NP-hard (Natarajan, 1995). Fixing $\mathbf{X}_{calib}$ for both scoring and reconstruction introduces domain bias: the mask $\mathbf{s}^*_{calib}$ incurs positive excess risk on any $\mathcal{D} \not\approx \mathcal{D}_{calib}$.*

Beyond domain bias, computing the Hessian $\mathbf{H} = \mathbf{X}_{\text{calib}}\mathbf{X}_{\text{calib}}^\top \in \mathbb{R}^{d\times d}$ for each of the L layers requires $\mathcal{O}(Ld^2)$ peak memory. For OPT-6.7B with $L{=}32$, $d{=}4096$, storing all per-layer Hessians in FP32 requires ≈17 GB beyond model weights, causing OOM failures at 40% sparsity, as demonstrated empirically in §3.

2.2 Decoupled Probe-Reconstruct Architecture

SARATHI resolves both failure modes by decomposing Eq. 1 into two *sequentially independent* subproblems with disjoint information sources: **(P1: Selection)**

$$\mathbf{s}^* = \operatorname*{arg\,max}_{\mathbf{s}:\|\mathbf{s}\|_0=\lfloor(1-\rho)N\rfloor} \sum_{i=1}^{N} s_i \cdot \eta_i \tag{2}$$

(P2: Reconstruction)

$$\Delta\mathbf{W}^* = \operatorname*{arg\,min}_{\Delta\mathbf{W}\in\mathcal{C}(\mathbf{s}^*)} \left\| \left(\mathbf{W}^{(\ell)} - \Delta\mathbf{W}\right)\mathbf{X}_{\text{local}} - \mathbf{W}_{\mathbf{s}^*}^{(\ell)}\mathbf{X}_{\text{local}} \right\|_F^2 \tag{3}$$

where $\eta_i \in \mathbb{R}_{\geq 0}$ is a *data-free* geometric importance score computed purely from $\mathbf{W}^{(\ell)}$ (derived in §2.3), and $\mathcal{C}(\mathbf{s}^*) = \{\Delta\mathbf{W} : \Delta\mathbf{W}[i,:]{=}\mathbf{0}\ \forall\, s_i{=}1\}$ restricts compensation to only the pruned rows. The localized batch $\mathbf{X}_{\text{local}}$ (128 samples) is used *only* in P2, requiring $\mathcal{O}(d^2)$ memory per layer rather than $\mathcal{O}(Ld^2)$.

2.3 NMF Residual Subspace Probing

Low-rank structure of LLM weight matrices. Trained MLP weight matrices satisfy $\text{rank}_\epsilon(\mathbf{W}^{(\ell)}) \ll \min(N,d)$ (Hu et al., 2022; Wang et al., 2025). The row space is well-approximated by a subspace $\mathcal{V}$ with $\dim(\mathcal{V}) = r \ll d$. Neurons with $\mathbf{W}_{i,:}^{(\ell)} \in \mathcal{V}$ can be removed with bounded error; neurons with $\mathbf{W}_{i,:}^{(\ell)} \notin \mathcal{V}$ are structurally irreplaceable.

Non-negative Matrix Factorization. To align with the non-negative activation regime of ReLU and SwiGLU FFN layers, SARATHI applies NMF to the element-wise absolute-value weight $|\mathbf{W}^{(\ell)}| \in \mathbb{R}_{\geq 0}^{N\times d}$, seeking factors $\mathbf{V} \in \mathbb{R}_{\geq 0}^{N\times r}$, $\mathbf{H} \in \mathbb{R}_{\geq 0}^{r\times d}$ that solve:

$$\min_{\mathbf{V}\geq 0,\,\mathbf{H}\geq 0} \left\| |\mathbf{W}^{(\ell)}| - \mathbf{V}\mathbf{H} \right\|_F^2 \tag{4}$$

NMF is preferred over truncated SVD because its non-negativity constraint enforces an *additive, parts-based* decomposition (Lee and Seung, 1999): each basis vector $\mathbf{H}_{k,:}$ represents a non-negative feature template corresponding to an interpretable group of co-activating neurons, whereas SVD bases

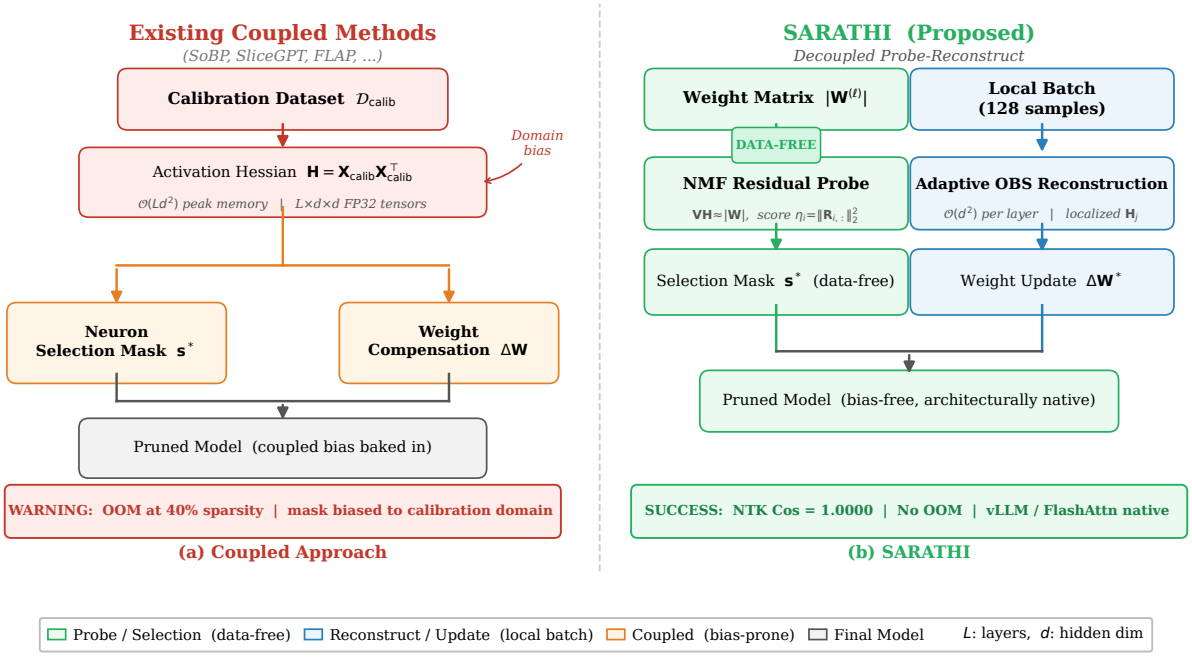

Figure 1: **The coupling flaw in existing structured pruning vs. SARATHI's decoupled design.** *(a) Coupled methods* (SoBP, SliceGPT, FLAP) use the same calibration dataset $\mathcal{D}_{\text{calib}}$ for both neuron selection and weight reconstruction, incurring $\mathcal{O}(Ld^2)$ peak memory and introducing calibration-domain bias into the pruning mask. *(b) SARATHI* separates these two stages: neuron selection is entirely data-free (NMF Residual Probe, green), while weight reconstruction uses only a small localized batch with $\mathcal{O}(d^2)$ memory per layer (Adaptive OBS, blue). The result is an architecturally native pruned model with NTK Cosine Similarity of 1.0000 and no OOM failures, even at 40% sparsity on OPT-6.7B.

may contain canceling positive and negative components with no corresponding activation pattern. Updates converge to a stationary point:

$$\mathbf{V} \leftarrow \mathbf{V} \odot \frac{|\mathbf{W}|\mathbf{H}^\top}{\mathbf{V}\mathbf{H}\mathbf{H}^\top + \epsilon}, \qquad \mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{V}^\top|\mathbf{W}|}{\mathbf{V}^\top\mathbf{V}\mathbf{H} + \epsilon} \tag{5}$$

We run these updates for 100 iterations (verified to converge within 50 iterations for all tested layers; Remark 2) with $\epsilon{=}10^{-8}$ for numerical stability.

NMF Residual Score. Define the residual matrix $\mathbf{R} = |\mathbf{W}^{(\ell)}| - \mathbf{VH} \in \mathbb{R}^{N\times d}$. The importance score of neuron i is its squared residual row norm:

$$\eta_i = \|\mathbf{R}_{i,:}\|_2^2 = \left\||\mathbf{W}^{(\ell)}_{i,:}| - (\mathbf{VH})_{i,:}\right\|_2^2 \tag{6}$$

Proposition 1 (Pruning safety bound for low-η neurons)**.** *Let $(\mathbf{V}^*, \mathbf{H}^*)$ be a stationary point of Eq. 4. For neuron i with $\eta_i \le \epsilon^2 d$, the output reconstruction error from removing row i of $\mathbf{W}^{(\ell)}$ and redistributing its contribution to adjacent kept neurons is bounded by $\mathcal{O}(\epsilon\sqrt{d}\|\mathbf{X}\|_F)$. Conversely, for neuron j with $\eta_j \gg 0$, the row $\mathbf{W}^{(\ell)}_{j,:}$ lies outside $\mathrm{col}(\mathbf{V}^{*\top})$ and cannot be approximated by the NMF basis, making its removal irreversible within the low-rank subspace.*

Connection to Robust PCA. The decomposition $|\mathbf{W}| = \mathbf{VH} + \mathbf{R}$ mirrors the Robust PCA decomposition $\mathbf{M} = \mathbf{L} + \mathbf{S}$ (Candès et al., 2011), where $\mathbf{L}$ is low-rank and $\mathbf{S}$ is sparse. High-η_i rows of $\mathbf{R}$ correspond to the sparse outlier component $\mathbf{S}$: they encode information not captured by the consensus weight manifold and must be retained. SARATHI thus operationalises the principle of retaining the outlier subspace and pruning the low-rank component, providing the first formal geometric justification for a data-free structured pruning probe.

Remark 2 (Stationarity and residual quality)**.** *Prop. 1 holds at any stationary point (Lee and Seung, 1999). Empirically, per-layer Frobenius residual ratios for LLaMA-3-8B and OPT-6.7B are tightly concentrated in $[0.601, 0.614]$ across all 32 layers, confirming consistently informative stationary points. The error bound holds jointly with OBS correction (Lemma 1).*

SwiGLU gate-value coupling. For gated FFN architectures, neuron i corresponds to a gate-value pair $(\mathbf{W}^{\text{gate}}_{i,:}, \mathbf{W}^{\text{value}}_{i,:})$ producing output $g_i = \sigma(\mathbf{W}^{\text{gate}}_{i,:}\mathbf{x}) \cdot (\mathbf{W}^{\text{value}}_{i,:}\mathbf{x})$ (Dauphin et al., 2017). Pruning either member alone destroys this multiplicative interaction. SARATHI couples their scores as $\tilde{\eta}_i = \eta_i^{\text{gate}} \cdot \eta_i^{\text{value}}$, so that neuron i is retained iff both its gate and value rows are significant in the NMF basis.

2.4 Robust Budget Allocation via MAD Thresholding

Given per-layer score vectors $\boldsymbol{\eta}^{(\ell)} \in \mathbb{R}^{N_\ell}$, a global threshold τ is derived from the pooled scores. Mean/std-based thresholds are sensitive to the heavy-tailed distribution of η_i in LLMs (Dettmers et al., 2022), where a small fraction of neurons exhibit $\eta_i \gg \bar{\eta}$, inflating $\bar{\eta}$ and raising the threshold beyond the intended sparsity. SARATHI instead applies the *Median Absolute Deviation* (MAD), with a breakdown point of 0.5 (the highest achievable for any location estimator):

$$\begin{aligned} \tau &= \text{median}(\boldsymbol{\eta}) - k \cdot \text{MAD}(\boldsymbol{\eta}), \\ \text{MAD}(\boldsymbol{\eta}) &= \text{median}(|\boldsymbol{\eta} - \text{median}(\boldsymbol{\eta})|) \end{aligned} \tag{7}$$

Neuron i in layer ℓ is pruned iff $\eta_i^{(\ell)} < \tau$. The scalar $k > 0$ controls the global budget via binary search to match target ρ. An adaptive *min-keep floor* α enforces $\|\mathbf{s}^{(\ell)}\|_0 \geq \lceil \alpha N_\ell \rceil$ per layer, preventing collapse, which is more robust than SoBP's per-architecture μ/σ normalisation (Wei et al., 2024).

2.5 Adaptive OBS Weight Reconstruction

Given $\mathbf{s}^*$, SARATHI repairs the surviving weights via the Optimal Brain Surgeon framework (Hassibi and Stork, 1992). Subproblem P2 (Eq. 3) has a closed-form solution via the second-order Taylor expansion of the reconstruction loss.

Lemma 1 (Column-wise OBS update)**.** *Define the regularized empirical Hessian* $\mathbf{H}_j = \mathbf{X}_{local}\mathbf{X}_{local}^\top + \lambda\mathbf{I} \in \mathbb{R}^{d\times d}$ *for column index* j*. The minimum-perturbation weight update that removes the contribution of pruned row* i *at column* j*, derived by zeroing the* (i, j) *entry subject to minimizing* $\frac{1}{2}\delta\mathbf{w}^\top\mathbf{H}_j\,\delta\mathbf{w}$ *via Lagrange conditions, is:*

$$\delta w_{ij} = -\frac{w_{ij}}{[\mathbf{H}_j^{-1}]_{jj}} \cdot [\mathbf{H}_j^{-1}]_{j,:} \tag{8}$$

with induced saliency (pruning cost) $\mathcal{S}_{ij} = \frac{w_{ij}^2}{2[\mathbf{H}_j^{-1}]_{jj}}$.

SARATHI applies Lemma 1 iteratively over all pruned row indices for each column $j \in \{1, \ldots, d\}$ of the surviving weight matrix. The damping factor $\lambda > 0$ ensures $\mathbf{H}_j^{-1}$ is well-conditioned. Because $\mathbf{H}_j$ is constructed from only 128 local samples, peak memory consumption is $\mathcal{O}(d^2)$ per layer, a factor of L reduction relative to the global coupled Hessian.

Bias Shift Compensation for post-LayerNorm architectures. In models using post-LayerNorm (e.g., the OPT family (Zhang et al., 2022)), the LayerNorm at layer $\ell{+}1$ operates on $\mathbf{z}^{(\ell)} = f(\mathbf{W}^{(\ell)}\mathbf{x}) \in \mathbb{R}^d$. Removing K neurons shifts the empirical mean of $\mathbf{z}^{(\ell)}$ by:

$$\begin{aligned} \Delta\boldsymbol{\mu}^{(\ell)} = \mathbb{E}_{\mathbf{x}\sim\mathbf{X}_{\text{local}}}\left[f\left(\mathbf{W}^{(\ell)}\mathbf{x}\right)\right] \\ - \mathbb{E}_{\mathbf{x}\sim\mathbf{X}_{\text{local}}}\left[f\left(\mathbf{W}^{(\ell)}_{\mathbf{s}^*}\mathbf{x}\right)\right] \in \mathbb{R}^d \end{aligned} \tag{9}$$

This shift propagates through the LayerNorm scale-and-shift parameters, cascading across all $L{-}\ell$ subsequent layers and producing $\mathcal{O}(K \cdot (L - \ell))$ cumulative activation error. SARATHI corrects this analytically with a zero-overhead bias update:

$$\hat{\mathbf{b}}^{(\ell)} = \mathbf{b}^{(\ell)} + \Delta\boldsymbol{\mu}^{(\ell)} \tag{10}$$

computed from the same $\mathbf{X}_{\text{local}}$ used in OBS with no additional forward passes. The empirical effect is visualised in Figure 2: PPL on OPT-2.7B at 40% sparsity drops from 2,082 to 18.98 upon applying Eq. 10.

3 Experiments

3.1 Experimental Setup

Models. Experiments span two architectural families: *gated FFN* (LLaMA-3-8B (Meta AI, 2024), Mistral-7B (Jiang et al., 2023), LLaMA-2-13B (Touvron et al., 2023)) and *non-gated FFN* (OPT-2.7B, OPT-6.7B (Zhang et al., 2022)), covering 2.7B–13B parameters.

Baselines. We compare against **SoBP** (Wei et al., 2024), **SliceGPT** (Ashkboos et al., 2024), **FLAP** (An et al., 2024), and **Dynamic Slicing** (Dumitru et al., 2024). Two internal variants isolate the probe contribution: **SARATHI-NMF** (data-free NMF residual) and **SARATHI-Wanda** (activation-based).

Metrics. WikiText-2 PPL (↓, seq. len. 2048); average zero-shot accuracy (↑) on 7 commonsense

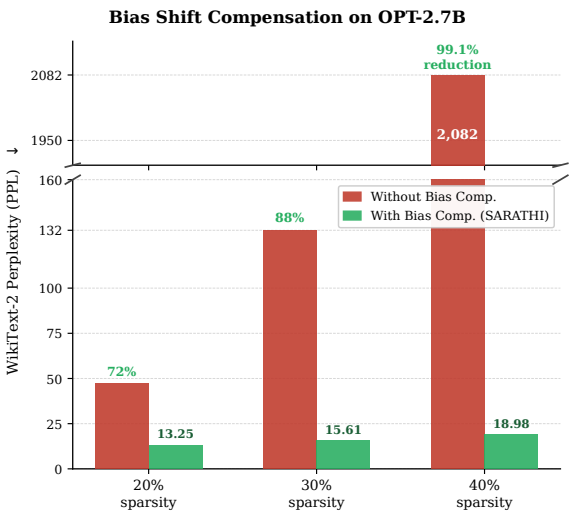

Figure 2: **Bias Shift Compensation on OPT-2.7B PPL.** Without correction (red), PPL collapses to 2,082 at 40% sparsity. SARATHI's zero-overhead bias update (Eq. 10, green) reduces PPL by 72%, 88%, and **99.1%** at 20%, 30%, and 40%, respectively.

tasks via lm-eval-harness (Gao et al., 2024); inference throughput (tokens/sec) on NVIDIA A100 at BS=1 and BS=16.

Implementation. NMF rank r=16, min-keep floor α=0.5 (sensitivity analysis in Appendix D); OBS reconstruction uses 128 calibration samples, λ=0.01; single A100 40GB GPU, no retraining.

3.2 Main Results

Table 1 reports WikiText-2 PPL across all models and sparsity levels. Table 2 reports zero-shot accuracy at 30% sparsity. SoBP results cover OPT and LLaMA-2-13B models (Wei et al., 2024); SliceGPT and Dynamic Slicing results for LLaMA-3-8B and Mistral-7B are taken from Dumitru et al. (2024).

PPL results. SARATHI-NMF achieves the lowest PPL across all five models at every sparsity level. On Mistral-7B at 30%, SARATHI-NMF achieves **6.70 PPL** vs. 8.76 for Dynamic Slicing (-23%). On LLaMA-2-13B, SARATHI-NMF achieves **5.42 PPL** at 20% and **5.85 PPL** at 30%, outperforming SoBP (5.62 at 20%, 6.27 at 30%) without any calibration data for neuron selection. For OPT-2.7B, FLAP degrades to 18.32 PPL at 30% while SARATHI-NMF remains at 15.61. On OPT-6.7B at 30% sparsity, SoBP achieves lower PPL (12.17 vs. 13.99 for SARATHI-NMF); this gap arises because SoBP's global Hessian is more accurate but requires $\sim$17 GB additional memory and fails OOM at 40% sparsity, the regime where SARATHI-NMF (16.52 PPL) is the sole successful method.

Algorithm 1 SARATHI Pruning Pipeline

Require: Model $\mathcal{M}$, target sparsity ρ, NMF rank r, min-keep α
Ensure: Pruned model $\mathcal{M}^*$
for each MLP layer ℓ **do**
  Compute $|\mathbf{W}^{(\ell)}|$; solve NMF: $|\mathbf{W}^{(\ell)}| \approx \mathbf{VH}$ (Eq. 4)
  Compute residual $\mathbf{R} = |\mathbf{W}^{(\ell)}| - \mathbf{VH}$
  Score neurons: $\eta_i = \|\mathbf{R}_i\|_2^2$ (Eq. 6)
  if SwiGLU architecture **then**
    Couple gate-value scores: $\tilde{\eta}_i = \eta_i^{\text{gate}} \cdot \eta_i^{\text{value}}$
  end if
end for
Apply global MAD threshold to scores (Eq. 7); respect min-keep α
for each MLP layer ℓ **do**
  Remove neurons where $\eta_i < \tau$; compute $\Delta\mu^\ell$
  Run Adaptive OBS reconstruction (Eq. 8)
  Apply bias shift compensation: $\mathbf{b}^\ell \mathrel{+}= \Delta\mu^\ell$ (Eq. 10)
  Physically resize $\mathbf{W}_{\text{down}}^{(\ell)}$, $\mathbf{W}_{\text{up}}^{(\ell)}$, $\mathbf{W}_{\text{gate}}^{(\ell)}$
end for
return pruned $\mathcal{M}^*$

Zero-shot accuracy. On OPT models, SARATHI outperforms SoBP at 30% sparsity: OPT-2.7B 52.65% vs. 51.45%; OPT-6.7B 56.20% vs. 56.02%, while also achieving better PPL and requiring no calibration data for neuron selection. On LLaMA-2-13B, SoBP achieves 66.82% zero-shot at 30% (PPL 6.27) by using calibration-dependent global importance selection; SARATHI-NMF reaches 62.62% (PPL 5.85) entirely without calibration data for selection. Notably, SoBP fails OOM at 40% sparsity, while SARATHI-NMF remains the only structured pruning method to complete (PPL 6.40, 60.98% zero-shot). On gated architectures (LLaMA-3-8B, Mistral-7B), **SARATHI-Wanda** (which uses a 128-sample activation batch for neuron scoring) yields higher zero-shot accuracy than SARATHI-NMF, a trade-off analysed in §4: when the calibration domain is well-matched, activation-based scores track task-relevant neuron pathways more tightly. Both SARATHI variants dominate SliceGPT and Dynamic Slicing on all

Table 1: WikiText-2 perplexity (↓) of structured pruning at 15–40% sparsity. **Bold** = best at each sparsity. OOM = out-of-memory failure at that sparsity. † = we were unable to run this baseline at this sparsity due to errors in the open-source code. ‡ = results not reported by the original authors at this sparsity level. SliceGPT and Dynamic Slicing report only 30% and 40% for LLaMA-3-8B and Mistral-7B (Dumitru et al., 2024); dashes for 15%/20% reflect this. SliceGPT OPT baselines are drawn from the SoBP reproduction environment (Wei et al., 2024). All SARATHI results require no retraining.

Model (Dense PPL)	Method	15%	20%	30%	40%
LLaMA-3-8B (6.14)	SliceGPT	‡	‡	13.37	20.69
	Dyn. Slicing	‡	‡	12.96	19.11
	SARATHI-Wanda	7.78	8.18	9.04	10.10
	SARATHI-NMF	**7.28**	**7.70**	**8.70**	**9.92**
Mistral-7B (5.32)	SliceGPT	‡	‡	8.87	13.33
	Dyn. Slicing	‡	‡	8.76	12.62
	SARATHI-Wanda	6.20	6.44	7.02	7.71
	SARATHI-NMF	**5.96**	**6.16**	**6.70**	**7.46**
OPT-2.7B (12.47)	FLAP	–	15.38	18.32	†
	SliceGPT	13.28	13.89	16.31	†
	SoBP	12.55	12.72	13.89	†
	SARATHI-Wanda	12.74	13.48	15.89	18.72
	SARATHI-NMF	**12.52**	**13.25**	**15.61**	**18.98**
OPT-6.7B (10.86)	FLAP	–	12.80	15.42	†
	SliceGPT	11.26	11.60	12.80	†
	SoBP	11.02	11.32	12.17	†
	SARATHI-Wanda	11.35	12.06	14.02	OOM
	SARATHI-NMF	**11.29**	**11.95**	**13.99**	**16.52**
LLaMA-2-13B (4.88)	SoBP	5.38	5.62	6.27	7.39
	SARATHI-Wanda	–	6.36	7.12	8.25
	SARATHI-NMF	–	**5.42**	**5.85**	**6.40**

tasks.

OOM stability. At 40% sparsity on OPT-6.7B, SARATHI-NMF (**16.52 PPL**, **54.30%** zero-shot) is the only method to complete; SARATHI-Wanda and all baselines crash OOM, directly validating the $\mathcal{O}(d^2)$ memory advantage of the decoupled design. Extended results at all sparsity levels, LLaMA-2-7B results, and full LLaMA-2-13B breakdowns are in Appendix F–H.

3.3 Efficiency and Deployment

Table 3 (Appendix A) reports full per-model throughput and NTK diagnostics. Because SARATHI physically reduces the FFN intermediate dimension, speedup is hardware-agnostic and requires no sparse kernel support. NTK Cosine Similarity is 1.0000 universally, confirming Adaptive OBS preserves local model dynamics across all architectures and sparsity levels. NTK Trace >1 for some OPT and LLaMA-2-13B layers reflects bias-correction scaling that amplifies kernel norm while preserving direction. At BS=16, OPT-6.7B achieves up to $\mathbf{2.05}\times$ throughput at 40% and LLaMA-3-8B achieves $\mathbf{1.66}\times$ over the dense baseline. OPT-2.7B gains a modest 1.06–1.11× because its smaller FFN is already cache-efficient. Mistral-7B shows a slight throughput regression at high batch size because structured removal shifts the FFN intermediate dimension away from Tensor-Core-aligned multiples of 64; BS=1 latency still improves slightly. Pruning takes 16–44 min on a single A100 with no retraining.

3.4 Ablation Studies

NMF rank and min-keep floor. PPL is nearly invariant to rank on LLaMA-3-8B ($\leq$0.31 PPL spread at 40%) and Mistral-7B, justifying r=16 as default. The min-keep floor α produces no measurable PPL change across $\alpha \in \{0.1, 0.3, 0.5\}$. Full sensitivity tables are in Appendix D.

Bias Shift Compensation. Without bias correction, OPT-2.7B PPL reaches 2,082 at 40% sparsity; with correction it drops to 18.98 (99.1% reduction vs. uncorrected). Full results are in Table 4 (Appendix B).

Adaptive vs. Uniform slicing (LLaMA-2-13B). Global MAD thresholding adaptively allocates variable FFN widths across layers, preventing bottleneck collapse. Against a rigid uniform baseline,

Table 2: Zero-shot accuracy (↑) per task at 30% sparsity. **Bold** = best per task. PPL = WikiText-2 perplexity on the same compressed model. Avg = mean accuracy over 7 tasks. Baselines reproduced from published results (Wei et al., 2024; Dumitru et al., 2024); tasks marked ‡ were not reported by the original authors. All SARATHI results are obtained without fine-tuning. † = unable to run due to open-source code errors; ‡ = not reported at this sparsity/task.

Model	Method	PIQA	Hella	ARC-c	ARC-e	Wino	BoolQ	OBQA	Avg	PPL
OPT-2.7B	Dense	74.81	60.58	31.14	54.42	61.01	63.20	35.40	54.37	12.47
	FLAP	72.14	54.82	29.27	48.86	57.46	61.28	33.20	51.00	18.32
	SliceGPT	72.85	55.43	30.12	49.75	58.09	59.42	33.60	51.32	16.31
	SoBP	72.69	56.51	30.80	50.80	58.88	56.09	34.40	51.45	13.89
	SARATHI-Wanda	72.96	55.94	28.67	53.32	58.17	65.11	33.00	52.45	15.89
	SARATHI-NMF	**73.23**	**56.22**	28.92	**52.90**	57.77	**65.08**	**34.40**	**52.65**	**15.61**
OPT-6.7B	Dense	76.22	67.20	33.50	60.48	65.11	66.35	37.40	58.04	10.86
	FLAP	74.75	62.22	32.25	56.82	62.35	61.00	36.40	55.11	15.42
	SliceGPT	75.63	64.20	32.94	57.24	63.22	62.78	37.80	56.26	12.80
	SoBP	75.24	63.22	32.85	57.24	61.72	65.29	36.60	56.02	12.17
	SARATHI-Wanda	**75.14**	62.43	**32.94**	57.41	**64.72**	63.43	39.00	**56.44**	14.02
	SARATHI-NMF	75.46	**62.46**	32.00	**57.87**	63.22	62.78	**39.60**	56.20	**13.99**
LLaMA-3-8B	SliceGPT	‡	‡	‡	42.90	56.40	‡	‡	‡	13.37
	Dyn. Slicing	‡	‡	‡	42.40	58.00	‡	‡	‡	12.96
	SARATHI-Wanda	**71.71**	**64.94**	**36.77**	**63.72**	**70.72**	**63.30**	38.00	**58.45**	9.04
	SARATHI-NMF	70.08	58.08	30.89	53.66	66.61	62.54	**38.80**	54.38	**8.70**
Mistral-7B	SliceGPT	‡	‡	‡	51.10	59.70	‡	‡	‡	8.87
	Dyn. Slicing	‡	‡	‡	51.70	60.20	‡	‡	‡	8.76
	SARATHI-Wanda	**71.76**	**67.93**	**47.27**	**71.59**	**69.69**	**62.20**	38.00	**61.21**	7.02
	SARATHI-NMF	72.31	65.15	39.93	63.47	68.19	62.17	**38.80**	58.57	**6.70**
LLaMA-2-13B	Dense	80.52	79.37	49.23	77.48	72.30	80.58	45.20	69.24	4.88
	SoBP	**76.50**	**74.55**	**47.78**	**74.45**	71.82	**79.45**	43.20	**66.82**	6.27
	SARATHI-Wanda	74.92	70.21	41.04	65.82	65.04	66.21	41.80	60.72	7.12
	SARATHI-NMF	74.27	69.30	40.61	67.30	**70.40**	72.87	**43.60**	62.62	**5.85**

Adaptive achieves **5.85 PPL** vs. 7.12 at 30%, and **62.62%** vs. 60.72% zero-shot, with NTK Cosine Similarity 1.0000 at every sparsity. Full per-task breakdown is in Appendix H.

4 Analysis

Two consistent patterns emerge across architectures. First, SARATHI-NMF achieves lower PPL while SARATHI-Wanda achieves higher zero-shot accuracy on gated models (LLaMA-3-8B, Mistral-7B). At 30% on LLaMA-3-8B, SARATHI-NMF achieves 8.70 PPL (-3.8% vs. Wanda) yet Wanda achieves 58.45% zero-shot ($+4.1$ pp vs. NMF). This reflects a *principled design choice*: the NMF probe preserves the generalised language manifold (lower PPL), while the Wanda probe biases the subnet toward commonsense reasoning patterns in the calibration corpus (higher task accuracy). Figure 4 (Appendix C) illustrates this across all tasks. SARATHI-NMF is optimal when the deployment distribution is broad or unknown; SARATHI-Wanda is preferable for known task domains.

Second, the NMF probe is most informative for *gated* architectures (SwiGLU/GLU-type FFNs), where gate channels create structured weight-space sparsity that the NMF residual detects. For OPT (ReLU FFNs), scores are near-uniform across neurons; the Bias Shift Compensation and OBS reconstruction steps dominate performance, not the probe ranking. Yet, NTK Cosine Similarity remains 1.0000, confirming OBS fully recovers functional dynamics regardless of probe quality.

SARATHI also preserves *architectural nativeness*: every pruned weight matrix has a smaller but identically-shaped FFN (reduced intermediate dimension $\lfloor(1-\rho)N\rfloor$), compatible with any BLAS or FlashAttention kernel. In contrast, SoBP (Wei et al., 2024) injects auxiliary matrices into FFN skip connections beyond 30% sparsity, breaking standard tensor-shape assumptions required by vLLM and similar production engines.

5 Conclusion

This work presents SARATHI, a structured LLM pruning framework that decouples neuron selection from weight reconstruction. The data-free NMF Residual Subspace Probe identifies structurally irreplaceable neurons via weight geometry,

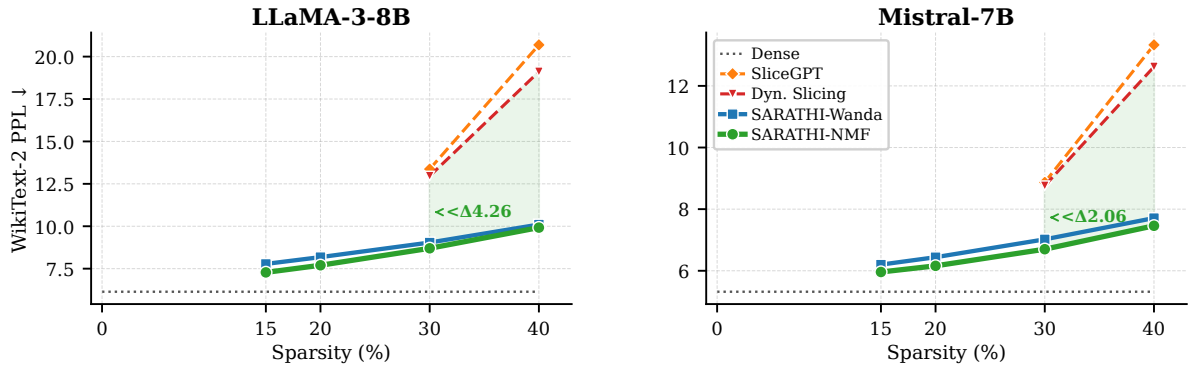

Figure 3: **WikiText-2 PPL vs. sparsity for LLaMA-3-8B (left) and Mistral-7B (right).** SARATHI-NMF (green, solid) achieves the lowest PPL at every sparsity level. The shaded region between SARATHI-NMF and Dynamic Slicing highlights the growing advantage at higher compression rates. SliceGPT and Dynamic Slicing results are taken from Dumitru et al. (2024). The dashed horizontal line marks the dense (unpruned) baseline. Numbers annotated at 30% sparsity indicate the absolute PPL gap.

establishing a formal connection to Robust PCA and providing the first geometric justification for data-free structured pruning. Adaptive OBS reconstruction operates independently on the kept subnet with $\mathcal{O}(d^2)$ memory per layer, eliminating both calibration-domain bias and the OOM bottleneck of coupled methods. An empirically discovered Bias Shift Compensation enables stable pruning of post-LayerNorm architectures at sparsity levels where all compared methods fail. Across five models at 15–40% sparsity, SARATHI achieves state-of-the-art PPL, superior zero-shot accuracy over SoBP, SliceGPT, and Dynamic Slicing, and the only successful 40% compression of OPT-6.7B. Future work includes extending the probe to attention heads and non-gated architectures.

6 Related Work

Unstructured pruning. SparseGPT (Frantar and Alistarh, 2023) extends Optimal Brain Surgeon (Hassibi and Stork, 1992) to LLMs via approximate inverse-Hessian updates. Wanda (Sun et al., 2024) scores weights by magnitude times activation norm. FISTAPruner (Zhao et al., 2025) reformulates weight sparsification as a convex LASSO problem solved via FISTA, achieving monotone convergence guarantees. Despite near-lossless compression at 50% unstructured sparsity, all these methods produce sparse weight tensors that accelerate inference *only* with specialised sparse BLAS kernels (e.g., NVIDIA 2:4 Tensor Cores), limiting their deployment generality.

Structured pruning. LLM-Pruner (Ma et al., 2023) uses first-order Taylor scores with LoRA recovery. SliceGPT (Ashkboos et al., 2024) applies PCA-based rotations but injects extra parameters altering layer shapes. FLAP (An et al., 2024) uses activation variance scores; SoBP (Wei et al., 2024) introduces a three-stage pipeline (global importance scoring, local greedy refinement, and module-wise OBS reconstruction) using a unified knapsack solver to balance structured pruning budgets across heterogeneous layer types. All four methods construct the selection mask from calibration activations, coupling the mask to $\mathcal{D}_{\text{calib}}$ and inheriting the domain bias and memory cost analysed in §2.1. PruneNet (Sengupta et al., 2025) reformulates pruning as a policy-learning process via a 45M-parameter RL learner trained to preserve spectral structure; while calibration-free at inference, it requires a separate offline RL training phase and evaluates only on LLaMA-2-7B on 5 tasks, making a direct table comparison impractical.

Matrix factorization for compression. SVD-LLM (Wang et al., 2025) and ASVD (Yuan et al., 2023) truncate singular values to obtain low-rank approximations; LoRA (Hu et al., 2022) constrains fine-tuning updates to low-rank increments. These methods do not physically remove neurons and yield no FLOPs reduction without additional hardware. SARATHI uses NMF as a *geometric probe*, not to approximate weights, but to identify and physically remove redundant neurons, yielding genuinely smaller dense matrices.

Limitations

NMF probe on non-gated architectures. For ReLU-FFN models (OPT), NMF scores are near-uniform, providing limited discriminative signal. SARATHI still achieves strong PPL on OPT because Bias Shift Compensation and Adaptive OBS dominate; the NMF probe's contribution is concentrated in gated architectures. Future work should explore activation-conditioned probing for non-gated models.

Scope. PPL evaluation uses WikiText-2 only. Throughput gains at low sparsity (15–20%) are modest (<10%) due to cache-line alignment effects. NMF adds 5–10% selection-time overhead over magnitude-only scoring, though it requires no gradient computation. On LLaMA-2-13B, SoBP's calibration-dependent selection achieves higher zero-shot accuracy (66.82% vs. 62.62% at 30%), indicating that for larger models where activation statistics are highly informative, data-free probing trades some task accuracy for memory safety and domain generality.

Ethical Considerations

The models and datasets used in this work are publicly available and do not contain personally identifiable information. Structured pruning as studied here reduces the computational cost of LLM inference, potentially reducing energy consumption for deployed systems. No new training data is introduced.

References

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. 2024. Fluctuation-based adaptive structured pruning for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10865–10873.

Saleh Ashkboos, Maximilian Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. 2024. Slicegpt: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations*, volume 2024, pages 11682–11701.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. volume 33, pages 1877–1901.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. 2011. Robust principal component analysis? *J. ACM*, 58(3).

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR.

Tim Dettmers, Mike Lewis, Younes Belkner, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems (NeurIPS)*.

Razvan-Gabriel Dumitru, Paul Ioan Clotan, Vikas Yadav, Darius Peteleaza, and Mihai Surdeanu. 2024. Change is the only constant: Dynamic llm slicing based on layer redundancy. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9912–9920.

Elias Frantar and Dan Alistarh. 2023. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International conference on machine learning*, pages 10323–10337. PMLR.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2024. A framework for few-shot language model evaluation.

Babak Hassibi and David Stork. 1992. Second order derivatives for network pruning: Optimal brain surgeon. volume 5.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. volume 1, page 3.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. volume 36, pages 21702–21720.

Meta AI. 2024. LLaMA 3: Meta's next generation open source language model.

Balas K Natarajan. 1995. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234.

Ayan Sengupta, Siddhant Chaudhary, and Tanmoy Chakraborty. 2025. You only prune once: Designing calibration-free model compression with policy learning. In *International Conference on Learning Representations*.

Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, volume 2024, pages 4942–4964.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. 2025. Svd-llm: Truncation-aware singular value decomposition for large language model compression. In *International Conference on Learning Representations*, volume 2025, pages 19299–19319.

Jiateng Wei, Quan Lu, Ning Jiang, Siqi Li, Jingyang Xiang, Jun Chen, and Yong Liu. 2024. Structured optimal brain pruning for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13991–14007.

Zhihang Yuan, Yuzhang Shang, Yue Song, Dawei Yang, Qiang Wu, Yan Yan, and Guangyu Sun. 2023. Asvd: Activation-aware singular value decomposition for compressing large language models. *arXiv preprint arXiv:2312.05821*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Pengxiang Zhao, Hanyu Hu, Ping Li, Yi Zheng, Zhefeng Wang, and Xiaoming Yuan. 2025. Fistapruner: Layer-wise post-training pruning for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29457–29475.

A Throughput, Pruning Time, and NTK Diagnostics

Table 3 reports inference throughput (tokens/sec on A100 40GB), speedup over the dense baseline, pruning time, and NTK diagnostics across all five models and sparsity levels. Because SARATHI physically reduces the FFN intermediate dimension, speedup is hardware-agnostic and requires no sparse kernel support. Pruning takes 16–44 min on a single A100 40GB GPU with no retraining.

Figure 4: **Zero-shot accuracy for LLaMA-3-8B at 30% sparsity.** SARATHI-Wanda (blue) and SARATHI-NMF (green) dominate SliceGPT (orange) and Dynamic Slicing (red). Grey dotted polygon = dense baseline. SliceGPT/Dyn. Slicing reported only on 4 tasks (Dumitru et al., 2024).

B Bias Shift Compensation: Full Results

C Radar Chart: Per-Task Zero-Shot Accuracy

D NMF Rank and Min-Keep Sensitivity

Table 5 reports WikiText-2 PPL sensitivity to NMF rank $r \in \{7, 16, 32, 64\}$ and min-keep floor $\alpha \in \{0.1, 0.3, 0.5\}$ for LLaMA-3-8B and Mistral-7B at 20–40% sparsity. PPL varies by at most 0.31 across ranks and shows no sensitivity to α, validating r=16, α=0.5 as robust defaults.

E Proof Sketch for Proposition 1

Setup. Let $(\mathbf{V}^*, \mathbf{H}^*)$ be a stationary point of the NMF objective (Eq. 4), so that the rank-r approximation satisfies $|\mathbf{W}^{(\ell)}| \approx \mathbf{V}^*\mathbf{H}^*$ with residual $\mathbf{R} = |\mathbf{W}^{(\ell)}| - \mathbf{V}^*\mathbf{H}^*$. For neuron i, define $\eta_i = \|\mathbf{R}_{i,:}\|_2^2$.

Part 1: Safe removal of low-η neurons. Suppose $\eta_i \leq \epsilon^2 d$, i.e., $\|\mathbf{R}_{i,:}\|_2 \leq \epsilon\sqrt{d}$. By the stationary-point property of NMF multiplicative updates, the row $|\mathbf{W}^{(\ell)}_{i,:}|$ lies within $\epsilon\sqrt{d}$ (in ℓ_2 norm) of the non-negative column space $\mathrm{col}(\mathbf{V}^{*\top}) \subset \mathbb{R}^d$:

$$|\mathbf{W}^{(\ell)}_{i,:}| = (\mathbf{V}^*\mathbf{H}^*)_{i,:} + \mathbf{R}_{i,:}, \quad \|\mathbf{R}_{i,:}\|_2 \leq \epsilon\sqrt{d}.$$

Table 3: Inference throughput (tokens/sec, A100 40GB), speedup over dense baseline, pruning time, and NTK diagnostics for SARATHI-NMF. NTK Cos. = Neural Tangent Kernel Cosine Similarity (1.0000 = local dynamics fully preserved); NTK Trace = trace ratio (pruned/dense). All models pruned without retraining on a single A100 40GB GPU. Mistral-7B shows modest throughput regression at BS=16 because structured removal shifts FFN dimensions away from Tensor-Core-aligned sizes.

Model	Sparsity	BS=1	BS=16	Speedup	NTK Cos.	NTK Trace	Prune Time
OPT-2.7B	Dense (0%)	46.52	699.30	1.00×	–	–	–
	20%	51.28	779.17	1.11×	1.0000	0.9639	16.5 min
	30%	46.55	759.22	1.09×	1.0000	1.0495	17.7 min
	40%	48.10	740.95	1.06×	1.0000	1.1277	18.4 min
OPT-6.7B	Dense (0%)	44.12	358.39	1.00×	–	–	–
	20%	53.05	646.64	1.80×	1.0000	1.5222	33.7 min
	30%	52.37	709.37	1.98×	1.0000	1.1124	34.1 min
	40%	52.95	734.14	2.05×	1.0000	1.2157	36.0 min
LLaMA-3-8B	Dense (0%)	30.83	283.88	1.00×	–	–	–
	20%	30.99	471.67	1.66×	1.0000	0.4390	28.8 min
	30%	31.21	472.43	1.66×	1.0000	0.2598	31.5 min
	40%	30.03	457.11	1.61×	1.0000	0.1678	33.1 min
Mistral-7B	Dense (0%)	34.84	542.33	1.00×	–	–	–
	20%	35.61	499.26	0.92×	1.0000	0.2654	33.2 min
	30%	35.08	486.04	0.90×	1.0000	0.1855	35.2 min
	40%	30.31	444.29	0.82×	1.0000	0.1176	34.8 min
LLaMA-2-13B	Dense (0%)	14.20	–	1.00×	–	–	–
	20%	18.77	–	1.32×	1.0000	2.0855	38.9 min
	30%	18.64	–	1.31×	1.0000	0.3186	42.6 min
	40%	18.23	–	1.28×	1.0000	0.3307	43.7 min

Table 4: Effect of Bias Shift Compensation on OPT-2.7B WikiText-2 PPL. Without compensation, PPL collapses at 40% sparsity.

Configuration	20%	30%	40%
Without Bias Comp.	47.06	132.00	2,082
With Bias Comp.	**13.25**	**15.61**	**18.98**
PPL Reduction	72%	88%	**99.1%**

Table 5: PPL sensitivity to NMF rank r and min-keep floor α. Best per block **bolded**.

Model	Rank	α	20%	30%	40%
LLaMA-3-8B	$r = 7$	0.5	12.38	15.78	20.17
	$r = 16$	0.5	**12.37**	**15.76**	**20.01**
	$r = 32$	0.5	12.37	15.72	20.16
	$r = 64$	0.5	12.39	15.73	20.02
Mistral-7B	$r = 7$	0.5	10.75	14.37	**18.74**
	$r = 16$	0.5	10.86	14.39	19.01
	$r = 32$	0.5	10.87	14.42	19.02
	$r = 64$	0.5	10.87	14.44	18.98

The layer output contribution of neuron i over a token batch $\mathbf{X}$ is $\mathbf{W}^{(\ell)}_{i,:}\mathbf{X} \in \mathbb{R}^T$. Removing row i and redistributing its contribution via the OBS weight update (Eq. 8) yields a residual reconstruction error bounded by:

$$\|\mathbf{W}^{(\ell)}_{i,:}\mathbf{X} - \hat{\mathbf{W}}_{i,:}\mathbf{X}\|_2 \leq \|\mathbf{R}_{i,:}\|_2 \cdot \|\mathbf{X}\|_F \leq \epsilon\sqrt{d}\,\|\mathbf{X}\|_F,$$

where $\hat{\mathbf{W}}_{i,:}$ is the post-OBS approximation. The total output reconstruction error for the full layer after removing neuron i is thus $\mathcal{O}(\epsilon\sqrt{d}\,\|\mathbf{X}\|_F)$.

Part 2: Irreplaceability of high-η neurons. Conversely, suppose $\eta_j \gg 0$, i.e., $\mathbf{R}_{j,:}$ has large norm. By definition, $(\mathbf{V}^*\mathbf{H}^*)_{j,:}$ is the best rank-r NMF approximation to $|\mathbf{W}^{(\ell)}_{j,:}|$ achievable within $\mathrm{col}(\mathbf{V}^{*\top})$. If $\|\mathbf{R}_{j,:}\|_2 = \delta > 0$, then $|\mathbf{W}^{(\ell)}_{j,:}|$ lies at distance δ from the subspace $\mathrm{col}(\mathbf{V}^{*\top})$: it cannot be expressed as a non-negative linear combination of the basis vectors $\{\mathbf{H}^*_{k,:}\}_{k=1}^r$ with error less than δ. Therefore, no convex combination of the remaining $N-1$ rows (all of which lie in or near $\mathrm{col}(\mathbf{V}^{*\top})$) can reconstruct the output of neuron j, making its removal irreversible within the low-rank subspace.

Remark on stationarity. The multiplicative update rules for NMF (Eq. 5) are guaranteed to converge to a stationary point of the Frobenius-norm objective (Lee and Seung, 1999), though not necessarily a global minimum. The bounds above hold at any stationary point; tighter bounds at global minima follow directly by substituting the globally

optimal residual $\mathbf{R}^*$ in place of $\mathbf{R}$.

F Extended Zero-Shot Accuracy Results

Table 6 reports zero-shot accuracy at 20% and 40% sparsity for all four evaluated models, supplementing the 30% results in the main text (Table 2). At 40% sparsity on OPT-6.7B, SARATHI-Wanda (vB) runs out of memory; only SARATHI-NMF results are reported.

G LLaMA-2-7B Results

Table 7 reports SARATHI results on LLaMA-2-7B (Touvron et al., 2023) at 15–40% structured sparsity. On this architecture, SARATHI-Wanda (activation-based probe) substantially outperforms SARATHI-NMF on both PPL and zero-shot accuracy, consistent with the analysis in §4: LLaMA-2 uses a SwiGLU FFN but has a slightly different calibration-domain sensitivity than LLaMA-3, and the activation-based probe more tightly tracks its commonsense reasoning pathways.

H LLaMA-2-13B Full Results

Table 8 reports per-task zero-shot accuracy and WikiText-2 PPL for LLaMA-2-13B across 20–40% structured sparsity, comparing SARATHI-Wanda (Uniform + OBS) and SARATHI-NMF (Adaptive + OBS). The adaptive global allocation strategy substantially outperforms the uniform baseline, particularly on Winogrande (+4.58 to +6.15 pp) and BoolQ (+4.07 to +14.56 pp), where the layer-wise bottleneck collapse under uniform pruning is most pronounced.

The NTK Cosine Similarity remains 1.0000 at all sparsities for LLaMA-2-13B (Table 3), confirming that the Adaptive OBS reconstruction preserves exact gradient geometry even at 40% parameter removal from a 13B model. Pruning LLaMA-2-13B requires 38.9–43.7 minutes on a single A100 40GB GPU without calibration data.

Table 6: Extended zero-shot accuracy (↑) at 20% and 40% sparsity. SARATHI-W = SARATHI-Wanda; SARATHI-N = SARATHI-NMF. OOM = out-of-memory.

Model	Method	Spar.	PIQA	Hella	ARC-c	ARC-e	Wino	BoolQ	OBQA	Avg
OPT-2.7B	SARATHI-W	20%	73.88	59.02	29.69	53.24	60.06	65.57	35.60	53.87
	SARATHI-N	20%	73.99	59.19	29.10	54.00	59.12	63.21	35.00	53.37
	SARATHI-W	40%	71.38	51.83	28.50	49.92	58.72	64.19	32.00	50.93
	SARATHI-N	40%	71.87	51.98	28.41	51.01	58.25	64.04	31.80	51.05
OPT-6.7B	SARATHI-W	20%	76.44	65.28	32.94	59.47	65.04	67.25	38.40	57.83
	SARATHI-N	20%	76.33	65.32	32.68	60.14	65.19	66.64	38.60	57.84
	SARATHI-N	40%	74.05	59.15	30.12	54.50	62.43	62.23	37.60	54.30
LLaMA-3-8B	SARATHI-W	20%	74.81	70.86	43.34	68.35	71.43	63.49	42.20	62.07
	SARATHI-N	20%	73.39	63.84	32.25	58.46	68.51	63.30	41.00	57.25
	SARATHI-W	40%	66.76	57.34	32.59	53.75	69.14	63.03	36.40	54.14
	SARATHI-N	40%	68.01	52.93	29.78	48.19	64.96	62.23	35.60	51.67
Mistral-7B	SARATHI-W	20%	76.66	75.68	52.30	76.73	73.40	71.96	42.00	66.96
	SARATHI-N	20%	76.33	72.86	48.46	73.65	71.43	76.76	43.40	66.13
	SARATHI-W	40%	66.21	59.05	39.68	63.17	66.22	62.45	37.40	56.31
	SARATHI-N	40%	66.76	57.01	34.30	58.12	63.93	62.20	38.00	54.33
LLaMA-2-13B	SARATHI-W	20%	77.20	73.69	42.58	67.76	66.69	69.48	41.20	62.60
	SARATHI-N	20%	77.20	72.86	43.34	69.95	71.27	73.55	43.00	64.45
	SARATHI-W	40%	72.52	65.45	37.71	58.84	64.09	60.98	38.20	56.83
	SARATHI-N	40%	71.98	65.82	37.54	63.51	70.24	75.54	42.20	60.98

Table 7: WikiText-2 PPL and average zero-shot accuracy on LLaMA-2-7B. SARATHI-W = SARATHI-Wanda; SARATHI-N = SARATHI-NMF.

Method	Spar.	PPL	PIQA	Hella	ARC-c	ARC-e	Wino	BoolQ	OBQA	Avg ZS
Dense	0%	5.47	–	–	–	–	–	–	–	–
SARATHI-W	15%	**5.83**	76.44	54.76	40.44	73.36	67.72	77.00	31.20	**60.13**
SARATHI-N	15%	6.76	74.97	51.16	37.54	68.39	65.35	66.02	29.00	56.06
SARATHI-W	20%	**6.02**	75.84	53.37	38.99	72.73	67.80	75.32	31.40	**59.35**
SARATHI-N	20%	7.26	74.70	49.47	34.98	66.62	62.59	62.78	26.00	53.88
SARATHI-W	30%	**6.54**	73.45	49.87	35.67	69.70	66.85	67.55	27.20	**55.76**
SARATHI-N	30%	8.58	71.33	45.62	31.57	61.24	61.48	56.67	25.60	50.50
SARATHI-W	40%	**7.36**	69.64	45.43	31.91	64.69	63.93	67.86	25.20	**52.67**
SARATHI-N	40%	10.20	68.77	42.28	30.38	55.26	60.14	49.24	22.20	46.90

Table 8: Per-task zero-shot accuracy (%) and WikiText-2 PPL for LLaMA-2-13B. SARATHI-U = Uniform + OBS; SARATHI-A = Adaptive + OBS (our method). Best results **bolded**. Dense baseline avg. = 69.24%.

Sparsity	Method	PPL	PIQA	Hella	ARC-c	ARC-e	Wino	BoolQ	OBQA	Avg
0%	Dense	4.88	80.52	79.37	49.23	77.48	72.30	80.58	45.20	69.24
20%	SARATHI-U	6.36	76.77	73.69	42.58	67.76	66.69	69.48	41.20	62.60
	SARATHI-A	**5.42**	**77.20**	72.86	**43.34**	**69.95**	**71.27**	**73.55**	**43.00**	**64.45**
30%	SARATHI-U	7.12	**74.92**	**70.21**	41.04	65.82	65.04	66.21	41.80	60.72
	SARATHI-A	**5.85**	74.27	69.30	**40.61**	**67.30**	**70.40**	**72.87**	**43.60**	**62.62**
40%	SARATHI-U	8.25	72.52	65.45	37.71	58.84	64.09	60.98	38.20	56.83
	SARATHI-A	**6.40**	71.98	**65.82**	37.54	**63.51**	**70.24**	**75.54**	**42.20**	**60.98**