

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

Sankar Behera¹, Dhruv Singh², Satyadev Ahlawat³, Yamuna Prasad¹

¹Department of Computer Science and Engineering, IIT Jammu

²Department of Mathematics, IIT Jammu

³Department of Electrical Engineering, IIT Jammu

{sankar.behera, 2023uma0210, satyadev.ahlawat, yamuna.prasad}@iitjammu.ac.in

Abstract

Structured pruning reduces the inference cost of large language models (LLMs) by permanently removing neurons and weight rows, yielding dense, hardware-agnostic speedup. Existing post-training methods couple neuron selection and weight reconstruction through a shared calibration dataset, introducing domain bias and out-of-memory failures at high sparsity. We propose SARATHI, a framework that decouples these two stages. A data-free *NMF Residual Subspace Probe* scores neurons using intrinsic weight geometry, without accessing any calibration data, and carries a formal connection to Robust PCA. *Adaptive OBS* then reconstructs the pruned weights with a localized calibration batch, operating in $\mathcal{O}(d^2)$ peak memory per layer and eliminating the global memory bottleneck of coupled methods. A *Bias Shift Compensation* step corrects systematic activation drift in post-LayerNorm architectures that would otherwise cause perplexity collapse at high sparsity. Experiments across six models from two architecture families (gated and non-gated FFN) at 15–40% sparsity show that SARATHI achieves competitive perplexity and matches or exceeds existing structured pruning baselines on zero-shot accuracy. Code: <https://github.com/snkrbhr/SARATHI>

neurons, where compression leverage is uniquely high: FFN sub-layers hold $2\times$ more parameters than attention in MHA models and $4.2\text{--}4.7\times$ more in GQA-based models; and non-linear activations (ReLU, SwiGLU) are exclusive to FFN layers, so pruning simultaneously reduces parameter count and activation compute (Mirzadeh et al., 2024). Full parameter breakdowns are in Appendix G.

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 $\mathcal{O}(d^2)$ working memory per layer; in practice, frameworks accumulate calibration activations, inverse Hessians, Cholesky factors, and OBS buffers globally, incurring substantial peak GPU memory at scale. This 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 OOM failures at high sparsity.

1 Introduction

LLMs deliver strong performance across downstream natural language 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, yielding dense smaller matrices that accelerate inference on any BLAS hardware without requiring specialised sparse kernels, unlike unstructured methods (Frantar and Alistarh, 2023; Sun et al., 2024). The proposed framework targets FFN intermediate

SARATHI. The proposed framework decouples the problem 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. 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)), which is novel in the post-training LLM structured pruning setting. The framework 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 (99.1% reduction vs. the uncorrected model).

Contributions.

- **Decoupled Probe-Reconstruct** architecture eliminating calibration bias and the $\mathcal{O}(Ld^2)$ memory bottleneck (§2.2).
- **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 six models (LLaMA-3-8B, Mistral-7B, OPT-2.7B/6.7B, LLaMA-2-7B/13B) at 15–40% sparsity (§3) shows competitive PPL and consistently higher or comparable zero-shot accuracy relative to state-of-the-art baselines across all architectures.

2 The SARATHI Framework

Figure 1 illustrates the overall decoupled architecture.

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 $\mathbf{W}_s^{(\ell)} = \mathbf{W}^{(\ell)}[\mathbf{s}, :]$. 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):

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

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

Beyond domain bias, computing the Hessian $\mathbf{H} = \mathbf{X}_{\text{calib}}\mathbf{X}_{\text{calib}}^\top \in \mathbb{R}^{d \times d}$ per layer requires $\mathcal{O}(d^2)$ working memory per layer. For OPT-6.7B with $L=32$, $d=4096$, the Hessians themselves occupy ≈ 2.1 GB in FP32; in practice, peak GPU memory is substantially higher due to inverse Hessians, Cholesky factors, OBS reconstruction buffers, and calibration activation tensors held concurrently, causing OOM failures at 40% sparsity as demonstrated empirically in §3.

2.2 Decoupled Probe-Reconstruct Architecture

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

$$\mathbf{s}^* = \arg \max_{\mathbf{s}: \|\mathbf{s}\|_0 = \lfloor (1-\rho)N \rfloor} \sum_{i=1}^N s_i \cdot \eta_i \quad (2)$$

(P2: Reconstruction)

$$\begin{aligned} \Delta\mathbf{W}^* = \arg \min_{\Delta\mathbf{W} \in \mathcal{C}(\mathbf{s}^*)} \quad & \left\| (\mathbf{W}^{(\ell)} - \Delta\mathbf{W})\mathbf{X}_{\text{local}} \right. \\ & \left. - \mathbf{W}_{\mathbf{s}^*}^{(\ell)}\mathbf{X}_{\text{local}} \right\|_F^2 \end{aligned} \quad (3)$$

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., 2021; 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, our framework 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:

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

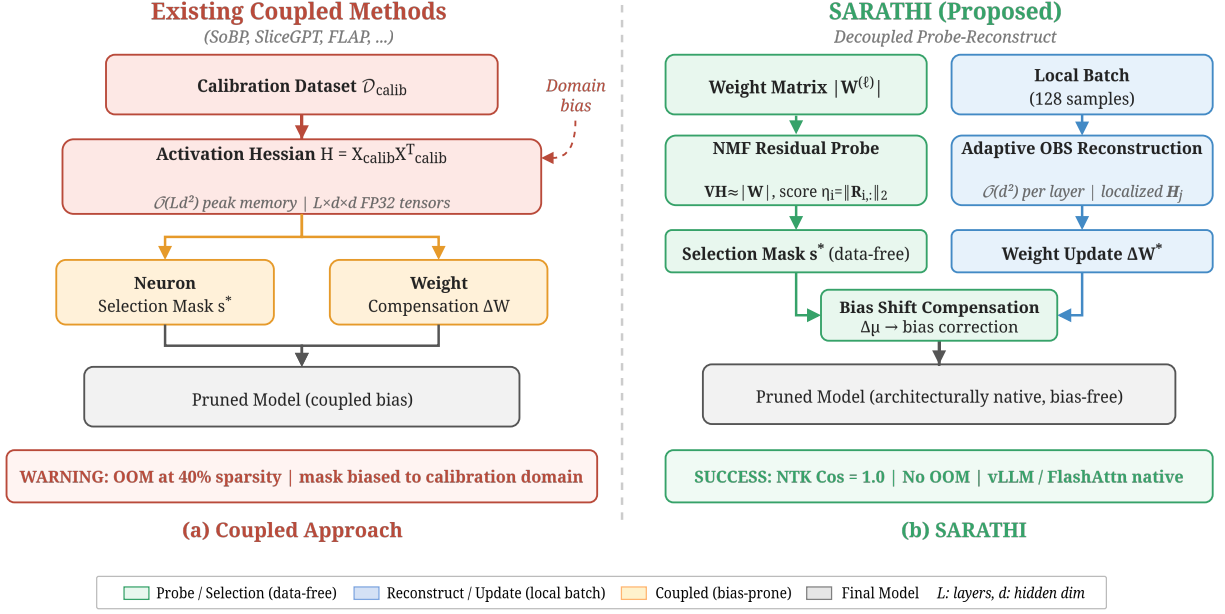


Figure 1: **The coupling flaw in existing structured pruning vs. SARATHI’s decoupled design.** (a) *Coupled methods* (SoBP, SliceGPT, FLAP) use the same calibration dataset $\mathcal{D}_{\text{calib}}$ for both neuron selection and weight reconstruction, incurring $\mathcal{O}(Ld^2)$ peak memory and 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 Neural Tangent Kernel (NTK) Cosine Similarity of 1.0000 and no OOM failures, even at 40% sparsity on OPT-6.7B.

NMF is preferred over truncated Singular Value Decomposition (SVD) because its non-negativity constraint enforces an *additive, parts-based* decomposition (Lee and Seung, 1999), aligning with the non-negative activation regime of ReLU/SwiGLU layers. Denoting the residual $\mathbf{S} = |\mathbf{W}^{(\ell)}| - \mathbf{V}\mathbf{H}$, it reliably identifies structurally irreplaceable neurons. This choice is empirically validated in Appendix F: replacing NMF with a signed SVD residual raises PPL by more than 8,500 points on OPT-6.7B despite a 98.95% overlap in top-ranked neurons, confirming that non-negativity in the factorization is the decisive factor. Updates converge to a stationary point:

$$\begin{aligned} \mathbf{V} &\leftarrow \mathbf{V} \odot \frac{|\mathbf{W}|\mathbf{H}^T}{\mathbf{V}\mathbf{H}\mathbf{H}^T + \epsilon}, \\ \mathbf{H} &\leftarrow \mathbf{H} \odot \frac{\mathbf{V}^T|\mathbf{W}|}{\mathbf{V}^T\mathbf{V}\mathbf{H} + \epsilon} \end{aligned} \quad (5)$$

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. Given residual $\mathbf{S}$ and a threshold $\tau > 0$ (derived by global MAD thresholding in §2.4), define the element-wise binary mask

$\mathcal{M}_{i,j} = \mathbb{I}(\mathbf{S}_{i,j} \geq \tau)$. The importance score of neuron i is:

$$\eta_i = \sum_j (|\mathbf{W}_{i,j}^{(\ell)}| \cdot \mathcal{M}_{i,j}) \quad (6)$$

Neurons with $\eta_i = 0$ have no weight entries escaping the NMF subspace and are pruning-safe; neurons with $\eta_i > 0$ contain structurally irreplaceable weights outside the consensus manifold.

Proposition 1 (Pruning safety bound for zero- η neurons). *Let $(\mathbf{V}^*, \mathbf{H}^*)$ be a stationary point of Eq. 4. For neuron i with $\eta_i = 0$ (i.e. all weight entries of row i are within the NMF subspace up to τ), the output reconstruction error from removing row i and applying Adaptive OBS compensation is bounded by $\mathcal{O}(\tau\sqrt{d}\|\mathbf{X}\|_F)$. Conversely, for neuron j with $\eta_j > 0$, the row $\mathbf{W}_{j,:}^{(\ell)}$ contains entries outside $\text{col}(\mathbf{V}^{*T})$, making its removal irreversible within the low-rank subspace. Full derivation in Appendix B.*

Connection to Robust PCA. The decomposition $|\mathbf{W}| = \mathbf{V}\mathbf{H} + \mathbf{S}$ 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 $\mathbf{S}$ correspond to the sparse outlier component: they

encode information not captured by the consensus weight manifold and must be retained. This 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 Adaptive OBS correction (Lemma 1).*

SwiGLU gate-value coupling. For gated FFN architectures (e.g., LLaMA, Mistral), neuron i coordinates across three matrices: gate ($\mathbf{W}^{\text{gate}}$), up ($\mathbf{W}^{\text{up}}$), and down ($\mathbf{W}^{\text{down}}$) projections (Dauphin et al., 2017). To avoid the vanishing-score problem of multiplicative coupling (where a near-zero score in one projection nullifies the neuron), we use the arithmetic mean:

$$\tilde{\eta}_i = \frac{1}{3}(\eta_i^{\text{gate}} + \eta_i^{\text{up}} + \eta_i^{\text{down}}) \quad (7)$$

This ensures balanced structural retention across coupled matrices without disproportionately penalising asymmetric feature distributions.

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. Instead, we apply the *Median Absolute Deviation* (MAD), with a breakdown point of 0.5 (the highest achievable for any location estimator):

$$\begin{aligned} \tau &= \text{median}(\boldsymbol{\eta}) - k \cdot \text{MAD}(\boldsymbol{\eta}), \\ \text{MAD}(\boldsymbol{\eta}) &= \text{median}(|\boldsymbol{\eta} - \text{median}(\boldsymbol{\eta})|) \end{aligned} \quad (8)$$

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

Unlike standard OBS frameworks (Hassibi and Stork, 1992; Frantar and Alistarh, 2023) that jointly use the inverse Hessian for pruning saliency and weight updates, SARATHI decouples structural search from compensation. By pre-determining the global mask $\mathbf{s}^*$ via the NMF probe, we bypass layer-wise saliency evaluation. This transforms OBS from a joint search-and-update problem into a fixed-constraint optimization: the Hessian strictly computes compensatory updates for the pre-determined survivors. This decoupling enables asymmetrical dimension allocation across layers while preserving local error-compensation guarantees, yielding a closed-form solution for Subproblem P2 (Eq. 3):

Lemma 1 (Column-wise OBS update). *Define the regularized empirical Hessian $\mathbf{H}_j^{\text{obs}} = \mathbf{X}_{\text{local}} \mathbf{X}_{\text{local}}^\top + \lambda \mathbf{I} \in \mathbb{R}^{d \times d}$ for column index j (distinct from the NMF factor $\mathbf{H}$). 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 $\frac{1}{2} \delta \mathbf{w}^\top \mathbf{H}_j^{\text{obs}} \delta \mathbf{w}$ via Lagrange conditions, is:*

$$\delta w_{ij} = -\frac{w_{ij}}{[\mathbf{H}_j^{\text{obs}-1}]_{jj}} \cdot [\mathbf{H}_j^{\text{obs}-1}]_{j,:} \quad (9)$$

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

This update is applied iteratively over all pruned row indices for each column $j \in \{1, \dots, d\}$ of the surviving weight matrix. The damping factor $\lambda > 0$ ensures $\mathbf{H}_j^{\text{obs}-1}$ is well-conditioned. Because $\mathbf{H}_j^{\text{obs}}$ 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:

$$\begin{aligned} \Delta \boldsymbol{\mu}^{(\ell)} &= \mathbb{E}_{\mathbf{x} \sim \mathbf{X}_{\text{local}}} \left[f(\mathbf{W}^{(\ell)} \mathbf{x}) \right] \\ &\quad - \mathbb{E}_{\mathbf{x} \sim \mathbf{X}_{\text{local}}} \left[f(\mathbf{W}_{\mathbf{s}^*}^{(\ell)} \mathbf{x}) \right] \in \mathbb{R}^d \end{aligned} \quad (10)$$

This shift propagates through the LayerNorm scale-and-shift parameters, cascading across all $L-\ell$ sub-

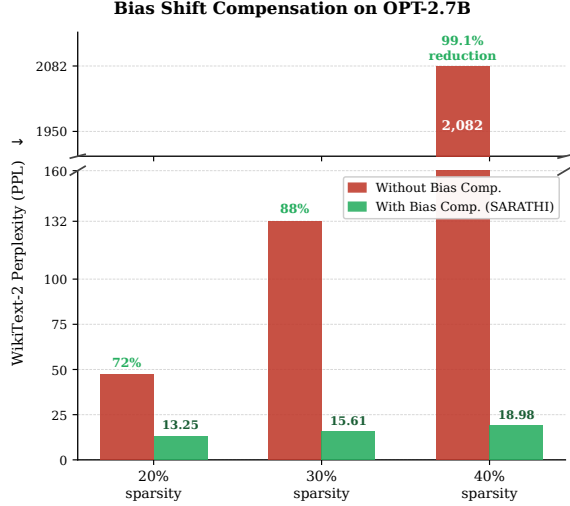


Figure 2: **Bias Shift Compensation on OPT-2.7B PPL.** Without correction (red), PPL collapses to 2,082 at 40% sparsity. SARATHI’s zero-overhead bias update (Eq. 11, green) reduces PPL by 72%, 88%, and **99.1%** at 20%, 30%, and 40%, respectively.

sequent layers and producing $\mathcal{O}(K \cdot (L - \ell))$ cumulative activation error. We correct this analytically with a zero-overhead bias update:

$$\hat{\mathbf{b}}^{(\ell)} = \mathbf{b}^{(\ell)} + \Delta\mu^{(\ell)} \quad (11)$$

computed from the same $\mathbf{X}_{\text{local}}$ used in Adaptive 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. 11.

3 Experiments

3.1 Experimental Setup

Models. Experiments span two architectural families: *gated FFN* (LLaMA-3-8B (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023), LLaMA-2-7B/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), Dynamic Slicing (Dumitru et al., 2024), and PruneNet (Sengupta et al., 2025). Two internal variants isolate the probe contribution: SARATHI-NMF (data-free NMF residual) and SARATHI-Wanda (activation-based).

Metrics. WikiText-2 PPL ($\downarrow$, seq. len. 2048); average zero-shot accuracy ($\uparrow$) on 7 commonsense

Algorithm 1 SARATHI Pruning Pipeline

Require: Model $\mathcal{M}$, target sparsity ρ , NMF rank r , min-keep α

Ensure: Pruned model $\mathcal{M}^*$

- 1: **for** each MLP layer ℓ **do**
- 2: Compute $|\mathbf{W}^{(\ell)}|$; solve NMF: $|\mathbf{W}^{(\ell)}| \approx \mathbf{V}\mathbf{H}$ (Eq. 4)
- 3: Compute residual $\mathbf{S} = |\mathbf{W}^{(\ell)}| - \mathbf{V}\mathbf{H}$; set $\mathcal{M}_{i,j} = \mathbb{I}(\mathbf{S}_{i,j} \geq \tau)$
- 4: Score neurons: $\eta_i = \sum_j |\mathbf{W}_{i,j}^{(\ell)}| \cdot \mathcal{M}_{i,j}$ (Eq. 6)
- 5: **if** SwiGLU architecture **then**
- 6: Couple gate-value scores: $\tilde{\eta}_i = \frac{1}{3}(\eta_i^{\text{gate}} + \eta_i^{\text{up}} + \eta_i^{\text{down}})$
- 7: **end if**
- 8: **end for**
- 9: Apply global MAD threshold to scores (Eq. 8); respect min-keep α
- 10: **for** each MLP layer ℓ **do**
- 11: Remove neurons where $\eta_i < \tau$; compute $\Delta\mu^\ell$
- 12: Run Adaptive OBS reconstruction (Eq. 9)
- 13: Apply bias shift compensation: $\mathbf{b}^\ell += \Delta\mu^\ell$ (Eq. 11)
- 14: Physically resize $\mathbf{W}_{\text{down}}^{(\ell)}, \mathbf{W}_{\text{up}}^{(\ell)}, \mathbf{W}_{\text{gate}}^{(\ell)}$
- 15: **end for**
- 16: **return** pruned $\mathcal{M}^*$

tasks via lm-eval-harness (Gao et al., 2022); inference throughput (tokens/sec) on NVIDIA A100 at BS=1 and BS=16.

Implementation. NMF rank $r=16$, min-keep floor $\alpha=0.5$ (sensitivity in Table 4); Adaptive OBS reconstruction uses 128 calibration samples (C4 (Raffel et al., 2020) for OPT, WikiText-2 (Merity et al., 2016) default), $\lambda=0.01$; single A100 40GB GPU, no retraining.

3.2 Main Results

Table 1 reports WikiText-2 PPL across all models and sparsity levels. Table 3 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 on LLaMA-3-8B and Mistral-7B at all tested sparsity levels (Figure 4), and succeeds at 40% on OPT-6.7B. On Mistral-7B at 30%, SARATHI-NMF achieves 7.41 PPL vs. 8.76 for Dynamic Slicing (−15%). 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. On OPT-2.7B at 30%, SoBP achieves lower PPL (13.89 vs. 15.61 for SARATHI-NMF) through calibration-dependent global reconstruction; however, SoBP fails at 40% sparsity ($\dagger$),

Zero-Shot Accuracy — LLaMA-3-8B @ 30% Sparsity
(4 tasks available for all methods)

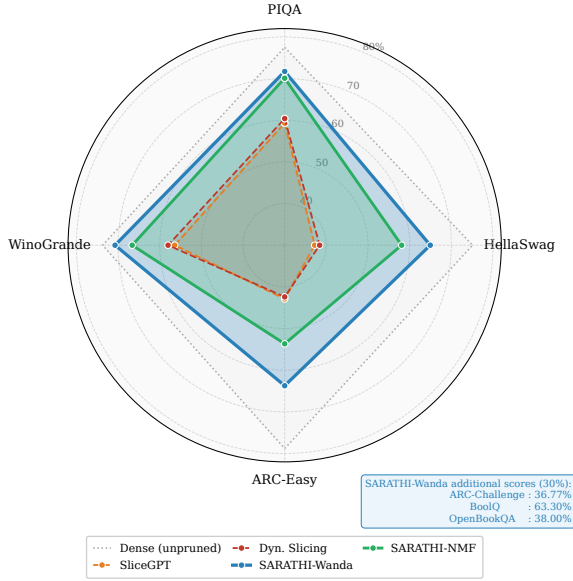


Figure 3: **Zero-shot accuracy for LLaMA-3-8B at 30% sparsity.** SARATHI-Wanda (blue) and SARATHI-NMF (green) dominate SliceGPT (orange) and Dynamic Slicing (red). Grey dotted polygon = dense baseline. SliceGPT/Dyn. Slicing reported only on 4 tasks (Dumitru et al., 2024).

whereas SARATHI-NMF succeeds at 18.98 PPL, and FLAP degrades sharply to 18.32 at 30%. On OPT-6.7B at 30% sparsity, SoBP again achieves lower PPL (12.17 vs. 13.99 for SARATHI-NMF); this gap arises because SoBP’s global reconstruction is more accurate at lower sparsity, but incurs substantial peak GPU memory due to concurrent Hessian inverses, Cholesky factors, and OBS buffers, causing OOM failure at 40%, the regime where SARATHI-NMF achieves 16.52 PPL.

Zero-shot accuracy. On OPT models at 30% sparsity, SARATHI-NMF achieves higher zero-shot accuracy than SoBP (OPT-2.7B: 52.65% vs. 51.45%; OPT-6.7B: 56.20% vs. 56.02%) without any calibration data for neuron selection, despite SoBP attaining lower PPL. 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) without calibration data for selection, and achieves better PPL at both 20% and 30% sparsity. At 40% on LLaMA-2-13B, SARATHI-NMF obtains 6.40 PPL and 60.98% zero-shot accuracy, outperforming SoBP on PPL (7.39) while using no calibration data for neuron scoring. On gated archi-

Table 1: WikiText-2 perplexity ($\downarrow$) across 15–40% sparsity. **Bold** indicates the best result. OOM = out-of-memory failure. † indicates failure due to code errors. ‡ indicates results not reported by the original authors. SliceGPT and Dynamic Slicing results for LLaMA-3-8B and Mistral-7B are from Dumitru et al. (2024). SliceGPT OPT baselines are from Wei et al. (2024). All SARATHI models are compressed without retraining.

Method	15%	20%	30%	40%
<i>LLaMA-3-8B</i> (Dense PPL: 6.14)				
SliceGPT (Ashkboos et al., 2024)	‡	‡	13.37	20.69
Dyn. Slicing (Dumitru et al., 2024)	‡	‡	12.96	19.11
SARATHI-Wanda	7.78	8.18	9.04	10.10
SARATHI-NMF	7.28	7.70	8.70	9.92
<i>Mistral-7B</i> (Dense PPL: 5.32)				
SliceGPT (Ashkboos et al., 2024)	‡	‡	8.87	13.33
Dyn. Slicing (Dumitru et al., 2024)	‡	‡	8.76	12.62
SARATHI-Wanda	6.74	7.04	7.74	8.59
SARATHI-NMF	6.45	6.71	7.41	8.38
<i>OPT-2.7B</i> (Dense PPL: 12.47)				
FLAP (An et al., 2024)	—	15.38	18.32	†
SliceGPT (Ashkboos et al., 2024)	13.28	13.89	16.31	†
SoBP (Wei et al., 2024)	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
<i>OPT-6.7B</i> (Dense PPL: 10.86)				
FLAP (An et al., 2024)	—	12.80	15.42	†
SliceGPT (Ashkboos et al., 2024)	11.26	11.60	12.80	†
SoBP (Wei et al., 2024)	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
<i>LLaMA-2-13B</i> (Dense PPL: 4.88)				
SoBP (Wei et al., 2024)	5.38	5.62	6.27	7.39
SARATHI-Wanda	—	6.36	7.12	8.25
SARATHI-NMF	—	5.42	5.85	6.40

Table 2: Zero-shot accuracy (%) on LLaMA-2-7B at 30% sparsity compared to PruneNet (Sengupta et al., 2025). **Bold** indicates the best result. SARATHI results are detailed in Appendix D.

Method	PIQA	Hella	Wino	ARC-e	ARC-c	Avg (5)
PruneNet (Sengupta et al., 2025)	71.11	58.30	61.09	53.20	33.53	55.45
SARATHI-NMF	71.33	45.62	61.48	61.24	31.57	54.25
SARATHI-Wanda	73.45	49.87	66.85	69.70	35.67	59.11

tectures (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 consistently match or exceed SliceGPT and Dynamic Slicing on zero-shot accuracy (see Figure 3 for LLaMA-3-8B).

OOM stability. At 40% sparsity on OPT-6.7B, a non-gated ReLU architecture, all coupled baselines encounter OOM failures; SARATHI-Wanda, which retains a 128 calibration batch for weight

Table 3: Zero-shot accuracy ($\uparrow$) at 30% sparsity. **Bold** indicates the best result. PPL refers to WikiText-2 perplexity. Baselines are sourced from Wei et al. (2024) and Dumitru et al. (2024). $\dagger$ indicates failure due to code errors; $\ddagger$ indicates tasks not reported by the original authors. All SARATHI models are evaluated without fine-tuning.

Method	PIQA	Hella	ARC-c	ARC-e	Wino	BoolQ	OBQA	Avg	PPL
<i>OPT-2.7B</i>									
Dense	74.81	60.58	31.14	54.42	61.01	63.20	35.40	54.37	12.47
FLAP (An et al., 2024)	72.14	54.82	29.27	48.86	57.46	61.28	33.20	51.00	18.32
SliceGPT (Ashkboos et al., 2024)	72.85	55.43	30.12	49.75	58.09	59.42	33.60	51.32	16.31
SoBP (Wei et al., 2024)	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
<i>OPT-6.7B</i>									
Dense	76.22	67.20	33.50	60.48	65.11	66.35	37.40	58.04	10.86
FLAP (An et al., 2024)	74.75	62.22	32.25	56.82	62.35	61.00	36.40	55.11	15.42
SliceGPT (Ashkboos et al., 2024)	75.63	64.20	32.94	57.24	63.22	62.78	37.80	56.26	12.80
SoBP (Wei et al., 2024)	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
<i>LLaMA-3-8B</i>									
SliceGPT (Ashkboos et al., 2024)	59.30	37.20	$\ddagger$	42.90	56.40	$\ddagger$	$\ddagger$	$\ddagger$	13.37
Dyn. Slicing (Dumitru et al., 2024)	60.40	38.40	$\ddagger$	42.40	58.00	$\ddagger$	$\ddagger$	$\ddagger$	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
<i>Mistral-7B</i>									
SliceGPT (Ashkboos et al., 2024)	62.60	38.00	$\ddagger$	51.10	59.70	$\ddagger$	$\ddagger$	$\ddagger$	8.87
Dyn. Slicing (Dumitru et al., 2024)	63.10	38.60	$\ddagger$	51.70	60.20	$\ddagger$	$\ddagger$	$\ddagger$	8.76
SARATHI-Wanda	66.16	60.92	38.74	65.53	67.40	62.20	35.40	56.62	7.74
SARATHI-NMF	66.10	56.88	32.68	57.11	65.11	60.80	34.40	53.30	7.41
<i>LLaMA-2-13B</i>									
Dense	80.52	79.37	49.23	77.48	72.30	80.58	45.20	69.24	4.88
SoBP (Wei et al., 2024)	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

reconstruction, also fails in this high-sparsity regime. Among all evaluated methods, SARATHI-NMF successfully complete compression, achieving 16.52 PPL and 54.30% zero-shot accuracy, consistent with the $\mathcal{O}(d^2)$ peak-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 C-E.

Comparison with data-free PruneNet. We compare against PruneNet (Sengupta et al., 2025), which uses a 45M-parameter RL policy trained offline to preserve spectral weight structure, making it calibration-free at inference. PruneNet reports results on LLaMA-2-7B over 5 tasks; Table 2 uses the same subset. SARATHI-Wanda outperforms PruneNet by +3.7 pp in average zero-shot accuracy using only a 128-sample calibration batch and no offline training. SARATHI-NMF, being fully data-free for selection, achieves comparable aggre-

gate accuracy (54.25% vs. 55.45%) while attaining lower perplexity (8.58 PPL; Table 8), consistent with the PPL–accuracy trade-off analysed in §4.

Efficiency and Deployment. Table 6 (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 $2.05\times$ throughput at 40% and LLaMA-3-8B achieves $1.66\times$ over the dense baseline. OPT-2.7B gains a modest $1.06\text{--}1.11\times$ because its smaller FFN is already cache-efficient. Mistral-7B shows a slight

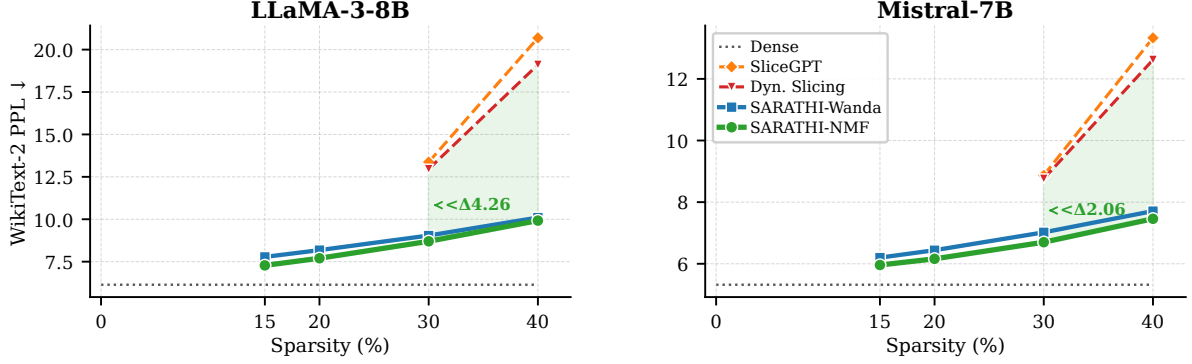


Figure 4: **WikiText-2 PPL vs. sparsity for LLaMA-3-8B (left) and Mistral-7B (right).** SARATHI-NMF (green, solid) achieves the lowest PPL at every sparsity level for these two architectures. The shaded region between SARATHI-NMF and Dynamic Slicing highlights the growing advantage at higher compression 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.

throughput regression at high batch sizes because structured removal shifts the FFN intermediate dimension away from Tensor-Core-aligned multiples of 64; at higher sparsity, BS=1 throughput can also slightly decrease as alignment overhead outweighs the reduced compute. Pruning takes 16–44 min on a single A100 with no retraining.

3.3 Ablation Studies.

NMF rank. Table 4 reports WikiText-2 PPL sensitivity to NMF rank $r \in \{7, 16, 32, 64\}$ for LLaMA-3-8B and Mistral-7B at 20–40% sparsity. PPL varies by at most 0.28 across ranks, validating $r=16$ as a robust default.

Table 4: PPL sensitivity to NMF rank r under Uniform slicing settings. The min-keep floor $\alpha = 0.5$ is constant. Best per block **bolded**.

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

Bias Shift Compensation. Table 5 quantifies the impact. Without bias correction, OPT-2.7B PPL collapses to 2,082 at 40% sparsity; with correction it drops to 18.98, a 99.1% reduction vs. the uncorrected model.

Table 5: Effect of Bias Shift Compensation on OPT-2.7B WikiText-2 PPL.

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%

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 (SARATHI-U), the adaptive strategy (SARATHI-A) achieves 5.85 PPL vs. 7.48 at 30%, and 62.62% vs. 60.15% zero-shot, with NTK Cosine Similarity 1.0000 at every sparsity. Full per-task breakdown is in Appendix E.

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): the NMF probe preserves the generalised language manifold, while the Wanda probe biases the subnet toward commonsense patterns in the calibration corpus.

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; Bias Shift Compensation and Adaptive OBS dominate performance, not the probe ranking. NTK Cosine Similarity remains 1.0000 universally, confirming Adaptive OBS fully recovers functional dynamics regardless of probe quality. The NMF vs. SVD ablation (Appendix F) further confirms that the non-negativity constraint - not factorization rank - is the decisive factor: a signed SVD residual raises PPL by over 8,500 points on OPT-6.7B despite 98.95% overlap in top-ranked neurons.

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, 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 six models at 15–40% sparsity, SARATHI achieves state-of-the-art PPL on gated architectures (LLaMA-3-8B, Mistral-7B), and consistently matches or exceeds existing methods on zero-shot accuracy across all architectures. Future work includes extending the probe to attention heads and non-gated architectures.

6 Related Work

Unstructured pruning. SparseGPT (Frantar and Alistarh, 2023) extends OBS (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. Table 2 provides a direct comparison on LLaMA-2-7B.

Matrix factorization for compression. SVD-LLM (Wang et al., 2025) and ASVD (Yuan et al., 2025) truncate singular values to obtain low-rank approximations; LoRA (Hu et al., 2021) 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. NMF has previously been explored for general DNN compression (Gurve et al., 2024) and data-free pruning at initialization (Behera et al., 2026); however, those works operate in fundamentally different settings (pre-LLM architectures, at-initialization pruning) and do not address post-training LLM structured pruning. SARATHI is distinct: it applies NMF as a *residual subspace probe* for post-training LLM pruning, with a formal Robust PCA safety bound (Proposition 1), MAD-based global budget allocation, and a decoupled Adaptive OBS reconstruction and bias-correction pipeline that are absent from prior NMF-based compression methods.

Limitations

NMF probe on non-gated architectures. For ReLU-FFN models such as OPT, NMF scores are near-uniform and offer limited discriminative signal. SARATHI still achieves strong PPL on

these models because Bias Shift Compensation and Adaptive OBS are the dominant contributors; the NMF probe’s benefit is concentrated in gated architectures. Activation-conditioned probing for non-gated models is a natural direction for future work.

Scope. PPL evaluation uses WikiText-2 only. Throughput gains at 15–20% sparsity are modest, below 10%, due to cache-line alignment effects. On Mistral-7B, structured removal shifts FFN dimensions away from Tensor-Core-aligned sizes, which introduces a throughput regression of $0.82\text{--}0.92\times$ relative to the dense baseline. NMF selection adds roughly 5–10% overhead over magnitude-only scoring but requires no gradient computation. On LLaMA-2-13B at 30%, SoBP’s calibration-dependent mask achieves higher zero-shot accuracy, 66.82% vs. 62.62% for SARATHI-NMF, though SARATHI-NMF attains lower PPL, 5.85 vs. 6.27. This trade-off reflects the broader property that data-free probing favours perplexity fidelity and memory safety over calibration-domain task alignment.

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.

Acknowledgments

The authors would like to thank Dr. Raj Kumar Choudhary, Engineering College Bikaner, for providing computational resources and access to the computing facilities under the ISEA Phase-III project, which significantly supported this research.

References

- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. [GQA: Training generalized multi-query transformer models from multi-head checkpoints](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, Singapore. Association for Computational Linguistics.
- 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 Hoefer, and James Hensman. 2024. SliceGPT: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations (ICLR)*.
- Sankar Behera, Anuj Kumar, Mahendra Kumar Gurve, Satyadev Ahlawat, and Yamuna Prasad. 2026. [You only prune once: A Zero-Shot, data-free pruning at initialization via Low-Rank residual saliency](#). *Transactions on Machine Learning Research*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandeep Sandholm, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. GPT-3.
- 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 Belkada, and Luke Zettlemoyer. 2022. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. volume 35, pages 30318–30332.
- 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*, Miami, Florida, USA. Association for Computational Linguistics.
- 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, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2022. [A framework for few-shot language model evaluation](#).
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur

- Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#).
- Mahendra Kumar Gurve, Gaurav Kumar, Sankar Behera, Yamuna Prasad, and Satyadev Ahlawat. 2024. [On exploring non-negative matrix factorization for deep neural network compression](#). In *2024 21st International SoC Design Conference (ISOCC)*, pages 115–116. IEEE.
- Babak Hassibi and David Stork. 1992. [Second order derivatives for network pruning: Optimal brain surgeon](#). In *Advances in Neural Information Processing Systems*, volume 5. Morgan-Kaufmann.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#).
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- 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.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. [Pointer sentinel mixture models](#).
- Seyed Iman Mirzadeh, Keivan Alizadeh-Vahid, Sachin Mehta, Carlo C del Mundo, Oncel Tuzel, Golnoosh Samei, Mohammad Rastegari, and Mehrdad Farajtabar. 2024. [ReLU strikes back: Exploiting activation sparsity in large language models](#). In *The Twelfth International Conference on Learning Representations*.
- B. K. Natarajan. 1995. [Sparse approximate solutions to linear systems](#). *SIAM Journal on Computing*, 24(2):227–234.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(140):1–67.
- Ayan Sengupta, Siddhant Chaudhary, and Tanmoy Chakraborty. 2025. [You only prune once: Designing calibration-free model compression with policy learning](#). In *The Thirteenth International Conference on Learning Representations*.
- Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations (ICLR)*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- 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. 2025. [Asvd: Activation-aware singular value decomposition for compressing large language models](#).
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 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 29469–29487. Association for Computational Linguistics.

A Throughput, Pruning Time, and NTK Diagnostics

Table 6 reports inference throughput (tokens/sec on A100 40GB), speedup over the dense baseline, pruning time, and NTK diagnostics across five of the evaluated 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.

Table 6: Hardware performance and structural diagnostics for SARATHI-NMF. Throughput is measured on a single A100 40GB GPU. NTK Cosine Similarity near 1.0000 indicates preserved local gradient dynamics. Mistral-7B exhibits a modest throughput regression at BS=16 due to FFN dimensions shifting 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

B 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 $\mathbf{S} = |\mathbf{W}^{(\ell)}| - \mathbf{V}^* \mathbf{H}^*$. The element-wise binary mask is $\mathcal{M}_{i,j} = \mathbb{I}(\mathbf{S}_{i,j} \geq \tau)$, and the importance score is $\eta_i = \sum_j (|\mathbf{W}_{i,j}^{(\ell)}| \cdot \mathcal{M}_{i,j})$.

Part 1: Safe removal of zero- η neurons. Suppose $\eta_i = 0$, i.e., $\mathbf{S}_{i,j} < \tau$ for all j . This implies $\|\mathbf{S}_{i,:}\|_\infty < \tau$, and therefore $\|\mathbf{S}_{i,:}\|_2 < \tau\sqrt{d}$:

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

The layer output contribution of neuron i over a token batch $\mathbf{X}$ is $\mathbf{W}_{i,:}^{(\ell)} \mathbf{X} \in \mathbb{R}^T$. Removing row i and redistributing its contribution via the Adaptive OBS weight update (Eq. 9) yields a residual reconstruction error bounded by:

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

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

Part 2: Irreplaceability of nonzero- η neurons. Conversely, suppose $\eta_j > 0$, i.e., there exists

at least one j' with $\mathbf{S}_{j,j'} \geq \tau$, meaning $|\mathbf{W}_{j,j'}^{(\ell)}|$ lies outside the NMF consensus. By definition, $(\mathbf{V}^* \mathbf{H}^*)_{j,:}$ is the best rank- r NMF approximation to $|\mathbf{W}_{j,:}^{(\ell)}|$ achievable within $\text{col}(\mathbf{V}^{*\top})$. If $\|\mathbf{S}_{j,:}\|_2 = \delta > 0$, then $|\mathbf{W}_{j,:}^{(\ell)}|$ lies at distance δ from the subspace $\text{col}(\mathbf{V}^{*\top})$: 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 $\text{col}(\mathbf{V}^{*\top})$) 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 $\mathbf{S}^*$ in place of $\mathbf{S}$.

C Extended Zero-Shot Accuracy Results

Table 7 reports zero-shot accuracy at 20% and 40% sparsity for all four evaluated models, supplementing the 30% results in the main text (Table 3). At 40% sparsity on OPT-6.7B, SARATHI-Wanda (vB)

runs out of memory; only SARATHI-NMF results are reported.

D LLaMA-2-7B Results

Table 8 reports SARATHI results on LLaMA-2-7B (Touvron et al., 2023) at 15–40% structured sparsity. On this architecture, SARATHI-Wanda (activation-based probe) outperforms SARATHI-NMF on both PPL and zero-shot accuracy. This aligns with our analysis in §4: LLaMA-2 exhibits different calibration-domain sensitivity than LLaMA-3, allowing the activation-based probe to better track its commonsense reasoning pathways.

E LLaMA-2-13B Full Results

Table 9 reports per-task zero-shot accuracy and WikiText-2 PPL for LLaMA-2-13B across 20–40% structured sparsity, comparing a uniform slicing baseline (SARATHI-U) against our adaptive global allocation strategy (SARATHI-A). The adaptive strategy substantially outperforms the uniform baseline, particularly on Winogrande (+5.14 to +6.67 pp) and BoolQ (+4.68 to +15.23 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 6), 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.

F NMF vs. SVD Probe Ablation

To isolate the effect of the factorization type in the SARATHI probe, we replace the NMF decomposition (Eq. 4) with a rank-16 truncated SVD while keeping the residual-based importance scoring (Eq. 6) identical. Critically, *Adaptive OBS reconstruction and Bias Shift Compensation are intentionally excluded* so that any performance difference reflects the probe alone rather than downstream correction steps.

Table 10 reports the results of 30% structured pruning on OPT-6.7B.

Table 10: NMF vs. SVD probe ablation on OPT-6.7B at 30% sparsity (probe-only; Adaptive OBS and Bias Shift Compensation excluded to isolate probe quality). Absolute PPL values are not representative of the full SARATHI pipeline; see Table 1. **Bold** = better per metric.

Metric	NMF Probe	SVD Probe
<i>Ranking agreement (probe-independent)</i>		
Spearman ρ (neuron ranking)	0.9987	0.9987
Top-25% neuron overlap	98.95%	98.95%
<i>Reconstruction and downstream quality</i>		
Rel. reconstruction error	0.6103	0.6078
Coefficient of variation (CV)	0.2805	0.2777
WikiText-2 PPL (probe-only)	19,695	28,268

Key finding. Despite a Spearman correlation of $\rho = 0.9987$ and a 98.95% overlap in the top-25% most important neurons, replacing NMF with an SVD-based residual raises perplexity by more than 8,500 points. This result demonstrates that *non-negativity in the factorization* - not the rank or the residual-scoring rule - is the decisive factor. The non-negative additive decomposition enforces a parts-based structure that aligns with the non-negative activation regime of SwiGLU/ReLU FFNs, capturing structurally irreplaceable neurons that a signed SVD factorization fails to identify.

Clarification on absolute PPL. The perplexity values in Table 10 are intentionally high because Adaptive OBS reconstruction and Bias Shift Compensation (§2.5) are excluded. For reference, the full SARATHI-NMF pipeline achieves 11.95, 13.99, and 16.52 PPL on OPT-6.7B at 20%, 30%, and 40% sparsity respectively (Table 1).

G Why FFN? Attention vs. FFN Parameter Budget in Modern LLMs

Structured pruning of LLMs can target different architectural components: attention heads, FFN intermediate neurons, or entire layers. This work focuses on FFN intermediate neurons, motivated by three architectural properties that make FFN layers the dominant target for structured compression. In every modern LLM evaluated here, the FFN sub-layer contains substantially more parameters per transformer layer than the attention sub-layer, and this gap has widened considerably with the adoption of GQA (Ainslie et al., 2023).

Table 11 quantifies the per-layer parameter counts for the attention projection matrices ($\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O \in \mathbb{R}^{d \times d}$ for MHA; $\mathbf{W}_Q \in \mathbb{R}^{d \times d}$, $\mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d/g}$ for GQA with g

Table 7: Extended zero-shot accuracy ($\uparrow$) 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%	71.82	70.04	44.71	71.46	71.19	80.46	39.20	64.13
	SARATHI-N	20%	70.89	65.27	37.63	64.60	68.35	62.94	39.20	58.41
	SARATHI-W	40%	61.70	51.87	30.89	54.04	62.83	62.02	31.80	50.74
	SARATHI-N	40%	62.62	49.72	29.86	52.02	62.04	58.99	31.00	49.46
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 8: 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

KV groups) and the FFN projection matrices ($\mathbf{W}_{\text{up}}, \mathbf{W}_{\text{down}} \in \mathbb{R}^{d_{\text{ff}} \times d}$, plus $\mathbf{W}_{\text{gate}}$ for SwiGLU, counted as $3d \cdot d_{\text{ff}}$ total).

Even in older MHA architectures such as OPT-66B, the FFN contains twice as many parameters as the attention sub-layer. For GQA-based models which include every state-of-the-art LLM evaluated in this work (Mistral-7B, LLaMA-3-8B) and large-scale counterparts (LLaMA-2-70B, LLaMA-3-70B), the FFN accounts for over 80% of per-layer parameters. Pruning attention in these architectures yields at most a $\sim 17\%$ per-layer reduction even if all attention parameters were removed, whereas a 30% FFN sparsity directly eliminates 30% of the dominant parameter block.

Beyond raw parameter counts, FFN layers have additional structural advantages for compression.

In LLaMA-2-7B, FFN layers collectively contribute 64% of parameters while self-attention contributes 32%; crucially, the neuron density (number of neurons directly connected to a parameter) in self-attention is only $\sim 27\%$ due to the multi-headed structure, whereas it is 100% in FFN layers (Sengupta et al., 2025). This means each removed FFN neuron eliminates a fully dense row, yielding maximum effective sparsity per removed parameter. Additionally, non-linear activations are exclusive to FFN layers (SwiGLU, ReLU), and Mirzadeh et al. (2024) highlighted that these non-linear operations comprise a significant fraction of per-layer compute; pruning FFN neurons thus simultaneously reduces both parameter count and activation overhead.

This strongly motivates the design choice to fo-

Table 9: 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	6.86	76.21	73.09	42.06	67.19	66.13	68.87	40.62	62.02
	SARATHI-A	5.42	77.20	72.86	43.34	69.95	71.27	73.55	43.00	64.45
30%	SARATHI-U	7.48	74.32	69.63	40.57	65.18	64.51	65.62	41.21	60.15
	SARATHI-A	5.85	74.27	69.30	40.61	67.30	70.40	72.87	43.60	62.62
40%	SARATHI-U	8.55	71.87	64.82	37.02	58.14	63.57	60.31	37.57	56.19
	SARATHI-A	6.40	71.98	65.82	37.54	63.51	70.24	75.54	42.20	60.98

Table 11: Per-layer parameter counts and Attn : FFN ratio for representative LLMs. Attn/layer counts $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O$. FFN/layer counts up, down (and gate for SwiGLU). GQA models use fewer KV heads (column “KV/Q heads”), sharply reducing attention parameters while leaving the FFN unchanged.

Model	Attn	d	d_{ff}	KV/Q heads	Attn/layer (M params)	FFN/layer (M params)
OPT-66B	MHA	9216	36 864	72/72	339.74	679.48
LLaMA-2-7B	MHA	4096	11 008	32/32	67.11	135.27
Mistral-7B	GQA	4096	14 336	8/32	41.94	176.16
LLaMA-3-8B	GQA	4096	14 336	8/32	41.94	176.16
LLaMA-2-70B	GQA	8192	28 672	8/64	150.99	704.64
LLaMA-3-70B	GQA	8192	28 672	8/64	150.99	704.64
<i>Attn : FFN ratio</i>						
MHA models (OPT-66B, LLaMA-2-7B)					$\approx 1 : 2.0$	
GQA models (Mistral-7B, LLaMA-3-8B)					$\approx 1 : 4.2$	
GQA models (LLaMA-2-70B, LLaMA-3-70B)					$\approx 1 : 4.7$	

cus structured pruning on the FFN intermediate dimension, where the compression leverage is largest, consistent with PruneNet (Sengupta et al., 2025), FLAP (An et al., 2024), and SoBP (Wei et al., 2024). Extending the NMF Residual Probe to attention heads (e.g., structured head or head-group pruning) is a natural direction for future work.