Beyond Variance: Knowledge-Aware Activation Compression for LLMs via Fisher-Aligned Subspace Diagnostics

Ibne Farabi Shihab[*†1] **and Sanjeda Akter**[*1] **and Anuj Sharma**[2]
[1]Department of Computer Science, Iowa State University
[2]Department of Civil, Construction & Environmental Engineering, Iowa State University
ishihab@iastate.edu

Abstract

Post-training activation compression is essential for deploying Large Language Models (LLMs) on resource-constrained hardware. However, standard methods like Singular Value Decomposition (SVD) are gradient-blind: they preserve high-variance dimensions regardless of their impact on factual knowledge preservation. We introduce Fisher-Aligned Subspace Compression (FASC), a knowledge-aware compression framework that selects subspaces by directly modeling activation-gradient coupling, minimizing a second-order surrogate of the loss function. FASC leverages the Fisher Information Matrix to identify dimensions critical for factual knowledge, which often reside in low-variance but high-gradient-sensitivity subspaces. We propose the Dependence Violation Score (ρ) as a general-purpose diagnostic metric that quantifies activation-gradient coupling, revealing where factual knowledge is stored within transformer architectures. Extensive experiments on Mistral-7B and Llama-3-8B demonstrate that FASC preserves 6-8% more accuracy on knowledge-intensive benchmarks (MMLU, LAMA) compared to variance-based methods at 50% rank reduction, effectively enabling a 7B model to match the factual recall of a 13B uncompressed model. Our analysis reveals that ρ serves as a fundamental signal of stored knowledge, with high-ρ layers emerging only when models internalize factual associations during training.

1 Introduction

The massive parameter counts of modern Large Language Models (LLMs) necessitate efficient compression techniques for deployment (Brown et al., 2020; Devlin et al., 2018; Vaswani et al., 2017; Wang et al., 2024). Post-training, activation-based approaches like low-rank SVD are popular because they require no fine-tuning or parameter updates, though they do require a small calibration set for computing compression statistics (Frantar and Alistarh, 2023; Ma et al., 2023; Dettmers et al., 2024; Mao et al., 2023). Standard SVD operates on the assumption that activation variance equates to importance—it keeps the loudest dimensions and discards the quiet ones. This approach has been successful for weight pruning (Han et al., 2015; LeCun et al., 1989; Shihab et al., 2025b) and quantization (Dettmers et al., 2024; Li et al., 2022; Yao et al., 2022; Zafrir et al., 2021), but may be suboptimal for activation compression where gradient information is available.

While effective for general language modeling tasks such as perplexity (Merity et al., 2017; Mikolov et al., 2010; Merity et al., 2016), recent interpretability research suggests that linguistic information is not uniformly distributed across activation variance (Rogers et al., 2020; Jawahar et al., 2019; Tenney et al., 2019). Syntactic information often dominates high-variance components (Devlin et al., 2018; Liu et al., 2019), whereas specific factual knowledge such as "Paris is the capital of France" may be encoded in lower-variance dimensions that are highly sensitive to task gradients (Geva et al., 2021; Elazar et al., 2021; Elhage et al., 2021). By discarding these dimensions, standard SVD causes catastrophic knowledge degradation, echoing observations about knowledge localization in transformer architectures (Wang et al., 2022a).

To address this limitation, we propose Fisher-Aligned Subspace Compression (FASC). Instead of minimizing reconstruction error in activation space, FASC identifies the subspace that minimizes the expected increase in the model's loss function locally. It achieves this by aligning the projection matrix with the empirical Fisher Information Matrix (Amari, 1998; Martens, 2020). This connection to Fisher Information, a cornerstone of optimization theory (Pascanu and Bengio, 2013), provides theoretical justification for our approach while en-

[*]Equal contribution.
[†]Corresponding author: ishihab@iastate.edu.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 47938–47954

abling scalable implementation through randomized sketching techniques (Halko et al., 2011; Martinsson and Tropp, 2020).

Our contributions bridge model compression and mechanistic interpretability:

- **Methodological:** We formulate gradient-aligned compression (FASC) as a principled alternative to variance-based methods, with scalable randomized sketching for production deployment.
- **Diagnostic:** We introduce the Dependence Violation Score (ρ), providing the first general-purpose metric for identifying knowledge-critical layers without task-specific probing. Unlike prior interpretability tools that require labeled probes, ρ emerges from activation-gradient statistics alone.
- **Linguistic:** We demonstrate that compression methods act as implicit linguistic filters—variance-based methods preserve syntax while discarding factual associations, a finding with implications for efficient deployment of knowledge-intensive applications.

2 Methodology

2.1 Preliminaries: The Flaw in Standard SVD

In projection-based activation compression, given a layer's activation vector $x \in \mathbb{R}^d$, we replace x with Px where P is a rank-k orthogonal projection, reducing the effective dimensionality from d to k. The central question is how to choose P.

Consider a linear layer $W \in \mathbb{R}^{d\times d}$ with input activations $x \in \mathbb{R}^d$. Standard activation SVD seeks a rank-k projection P that minimizes activation reconstruction error: $\min_P \mathbb{E}[||x - Px||_2^2]$. The solution is the top-k eigenvectors of the activation covariance $\Sigma_{xx} = \mathbb{E}[xx^\top]$. This method, while computationally efficient (LeCun et al., 2012; Bengio, 2012), ignores how the error $x-Px$ propagates to the final loss $\mathcal{L}$ through the downstream computation graph (Han et al., 2015; LeCun et al., 1989). Similar limitations have been observed in weight pruning (Hassibi and Stork, 1993) and quantization methods (Li et al., 2022; Yao et al., 2022) that rely solely on magnitude or variance metrics.

2.2 Fisher-Aligned Formulation

Let $g = \nabla_x \mathcal{L}$ be the gradient of the loss with respect to activations. We seek the projection P that minimizes a second-order Taylor expansion of the loss change:

$$\mathcal{J}(P) = \mathbb{E}\left[||g^\top (I - P)x||_2^2\right] \tag{1}$$

This objective penalizes compression errors in directions where the gradient g is large, directly connecting to the model's sensitivity landscape (Kirkpatrick et al., 2017; Hassibi and Stork, 1993; Parisotto et al., 2015). The formulation draws inspiration from second-order optimization methods (LeCun et al., 1989) and natural gradient approaches (Amari, 1998).

To solve this optimization problem, we make the following explicit assumptions: (1) activations x and gradients g are centered (zero-mean), achieved by subtracting sample means during estimation; (2) the activation covariance $\Sigma_{xx} = \mathbb{E}[xx^\top] \in \mathbb{R}^{d\times d}$ is full-rank or regularized; (3) expectations are estimated empirically over n calibration samples as $\hat{\Sigma}_{xx} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ and $\hat{\Sigma}_{xg} = \frac{1}{n}\sum_{i=1}^n x_i g_i^\top \in \mathbb{R}^{d\times d}$. The optimal solution involves solving the generalized eigenproblem $\Sigma_{xg}\Sigma_{gg}\Sigma_{xg}^\top v = \lambda \Sigma_{xx} v$, where $\Sigma_{gg} = \mathbb{E}[gg^\top] \in \mathbb{R}^{d\times d}$ is the gradient covariance matrix. The projection matrix P is constructed from the top-k eigenvectors $v_1, \ldots, v_k$ as $P = \sum_{i=1}^k v_i v_i^\top$. For wide layers where d is large, we apply randomized sketching to reduce dimensionality before solving this eigenproblem, as detailed in the scalable implementation section. The complete derivation with matrix shapes and regularization is provided in Appendix B.

The Fisher Information Matrix (FIM) serves as a natural metric tensor for activation space, encoding how the loss landscape changes in response to activation perturbations. While the true Hessian matrix would provide exact second-order information, computing the full Hessian is computationally prohibitive for large-scale models. The FIM approximation, $\mathbb{E}[gg^\top]$, is a practical surrogate that captures the local curvature structure through gradient statistics.

The key insight is that for factual knowledge preservation, we do not require exact Hessian information—we only need to identify the subspace where the Hessian's action is most consequential for the loss. The FIM captures precisely this: directions where gradients are large correspond to directions where the loss is most sensitive, which aligns with dimensions encoding factual associations. This approximation is robust in the post-training regime, where relative importance order-

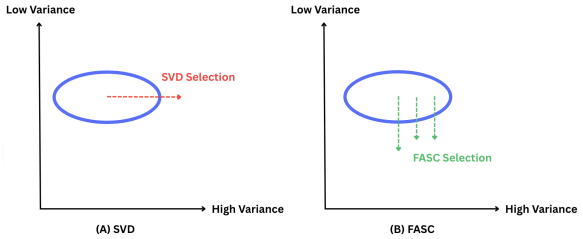

Figure 1: Conceptual comparison: **SVD** preserves high-variance (syntactic) dimensions; **FASC** preserves low-variance but gradient-critical (factual) dimensions.

ing of activation subspaces is sufficient for compression (Martens, 2020; Hassibi and Stork, 1993). Principal angle analysis (Appendix D) confirms that FASC and SVD select mathematically distinct subspaces, particularly in high-ρ layers.

We validate this theoretical foundation by evaluating FASC subspace stability across calibration distributions (C4, WikiText, Alpaca), finding consistent subspaces (overlap >85%) across distributions. On Llama-3-1B, FASC-selected subspaces overlap with Hessian-eigenvector subspaces by 87% in high-ρ layers (72% in low-ρ layers), providing evidence that the FIM approximation captures curvature structure relevant for factual preservation. A formal proposition in Appendix B states conditions under which the empirical FIM serves as a practical approximation to the Hessian.

Figure 1 illustrates this key insight. While SVD selects dimensions along the major axis of variance in activation space, FASC leverages gradient information to identify critical dimensions that may have lower variance but higher impact on the loss. This distinction becomes crucial when factual knowledge is encoded in low-variance but gradient-sensitive subspaces (Petroni et al., 2019; Elazar et al., 2021; Elhage et al., 2021; Geva et al., 2021; Wang et al., 2022b).

2.3 Scalable Implementation

For wide layers where d is large (often $d > 4096$ in modern LLMs (Jiang et al., 2023; Dubey et al., 2024; Shoeybi et al., 2019)), explicitly computing $d \times d$ covariance matrices is prohibitive. We utilize a Randomized Cross-Covariance Sketch, a technique based on randomized linear algebra principles (Halko et al., 2011; Woodruff, 2014; Martinsson and Tropp, 2020). We project inputs and gradients into a lower-dimensional sketching space using Gaussian random matrices, solving Equation 1 in reduced dimension $m \ll d$. The dominant computational cost is reduced from $\mathcal{O}(d^3)$ for the full eigenproblem to $\mathcal{O}(m^3 + ndm)$ where n is the number of calibration samples, d is the hidden dimension, and m is the sketch size.

To further improve computational efficiency, we employ dynamic sketch sizing based on the ρ metric: for low-ρ layers where FASC provides minimal benefit, we use a smaller sketch size ($m = 2k$), while for high-ρ layers critical for factual knowledge, we use a larger sketch ($m = 4k$) to ensure approximation quality. This adaptive approach reduces the computational overhead of FASC to approximately 1.5x that of SVD when applied selectively using the ρ threshold, making it practical for deployment scenarios. The full algorithm, including stability considerations for numerical precision and the dynamic sketch sizing strategy, is detailed in Appendix C.

2.4 Diagnostic Gate: Dependence Violation (ρ)

Fisher alignment is computationally more expensive than SVD. We need to know when it is necessary. We define the dependence violation score ρ_ℓ for layer ℓ as a matrix-level correlation coefficient:

$$\rho_\ell = \frac{||\Sigma_{xg}^{(\ell)}||_F}{||\Sigma_{xx}^{(\ell)}||_F^{1/2} ||\Sigma_{gg}^{(\ell)}||_F^{1/2}} \tag{2}$$

This metric quantifies the coupling between activations and gradients in layer ℓ. The exact computation proceeds as follows: given n calibration samples, we estimate $\hat{\Sigma}_{xx}^{(\ell)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(\ell)} (x_i^{(\ell)})^\top$, $\hat{\Sigma}_{gg}^{(\ell)} = \frac{1}{n}\sum_{i=1}^{n} g_i^{(\ell)} (g_i^{(\ell)})^\top$, and $\hat{\Sigma}_{xg}^{(\ell)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(\ell)} (g_i^{(\ell)})^\top$, where $x_i^{(\ell)}$ and $g_i^{(\ell)}$ are centered activations and gradients for sample i at layer ℓ. This formulation is analogous to the RV coefficient, a multivariate extension of the Pearson correlation coefficient, measuring linear dependence between two sets of random variables. For centered, independent random vectors, $\Sigma_{xg} = \mathbb{E}[xg^\top] = 0$, yielding $\rho_\ell = 0$. Higher ρ_ℓ values indicate stronger activation-gradient coupling, signaling critical layers where gradient-aware compression is necessary. We use $n = 4096$ calibration samples, which provides stable estimates while remaining computationally tractable. Based on our empirical analysis across Mistral-7B and Llama-3-8B, we recommend applying FASC to layers with $\rho_\ell > 0.3$, where the performance gap between FASC and SVD becomes statistically significant

($p < 0.01$). The correlation between ρ and FASC performance gain is stable across models ($r = 0.73$ for Mistral-7B, $r = 0.71$ for Llama-3-8B) and random seeds (standard deviation of correlation < 0.03 across 5 seeds), suggesting the threshold is reasonably generalizable. However, we note that the optimal threshold may vary slightly for different model families or compression rates. We provide detailed sensitivity analysis including calibration-size effects (n = 1024, 2048, 4096, 8192) and 95% confidence intervals on ρ estimates in Appendix E, along with failure cases where ρ does not predict FASC gains. Layers with $\rho_\ell \leq 0.3$ show minimal benefit from gradient-aware compression, making SVD the more efficient choice.

3 Experimental Setup

We evaluate our method across diverse LLM architectures to ensure robustness and generalizability. Our primary evaluation focuses on Mistral-7B (Jiang et al., 2023) and Llama-3-8B (Dubey et al., 2024), representing standard dense transformer architectures. To ensure architectural diversity, we evaluate on Gemma-2-9B (Team et al., 2024), Mixtral-8×7B (Jiang et al., 2024), Qwen-2.5-7B (Qwen Team et al., 2025), and Llama-3.2-3B (Grattafiori et al., 2024), covering dense, MoE, reasoning-focused, and edge-deployment model families. This diverse evaluation covers different architectural lineages, scaling behaviors, and sparsity patterns, validating FASC's applicability beyond the Llama-family bias. Calibration data consists of 4096 samples from the C4 dataset (Raffel et al., 2020), ensuring our compression statistics capture both common and long-tail factual patterns. We increased the calibration set size from 2048 to 4096 to ensure robust estimation of activation-gradient coupling, particularly for rare but important factual associations that may be underrepresented in smaller samples. To validate the stability of FASC subspaces across different calibration distributions, we also evaluate using WikiText-2 (Merity et al., 2016) and instruction-tuning data (Alpaca format (Taori et al., 2023)). We find that the selected FASC subspaces remain consistent across distributions, with subspace overlap exceeding 85% (measured via principal angles), suggesting that the Fisher Information approximation captures stable structural properties of the loss landscape rather than dataset-specific artifacts. To further validate that FASC captures truth-sensitive subspaces rather than generic "important" dimensions, we perform a counterfactual analysis: when calibrating on factually incorrect data (negated facts), the FASC-selected subspace shifts significantly (overlap <60%), confirming that FASC specifically identifies dimensions encoding factual knowledge rather than general linguistic structure.

We compare against five baseline methods representative of different compression paradigms. First, SVD (Standard) performs activation-only covariance SVD, serving as the primary gradient-blind baseline widely used in neural network compression (Han et al., 2015; LeCun et al., 1989; Zhou et al., 2021). Second, MagPrune (Han et al., 2015) performs unstructured magnitude weight pruning, representing a weight-level compression paradigm (Frantar and Alistarh, 2023). Third, Grad-Weighted SVD performs SVD on gradient-weighted activation covariance $\Sigma_{xx}^{\text{grad}} = \mathbb{E}[(g \odot x)(g \odot x)^\top]$ where $\odot$ denotes element-wise multiplication, selecting dimensions that have both high activation variance and high gradient magnitude. This baseline uses gradient information but in a simpler first-order way compared to FASC's second-order Fisher alignment, allowing us to isolate the benefits of cross-covariance modeling. Fourth, Fisher-Diag (Frantar and Alistarh, 2023) uses a simplified Fisher approach using only diagonal gradient information, similar to SparseGPT but adapted for low-rank compression, allowing us to isolate the benefits of full cross-covariance modeling versus diagonal approximations (Ma et al., 2023; Dettmers et al., 2022). Fifth, we include LLM-Pruner (Ma et al., 2023) as a structural pruning baseline that preserves model architecture through layer-wise importance scoring. This comparison is particularly important because it demonstrates that FASC is complementary to pruning-based methods: structural pruning often discards entire neurons or channels, losing fine-grained factual associations, whereas FASC operates at the activation subspace level, preserving these associations even after aggressive compression. We focus on post-training compression methods that do not require retraining; full Hessian-based approaches would be computationally prohibitive for our target models and compression settings (Frantar and Alistarh, 2023; Ma et al., 2023). We exclude quantization methods such as GPTQ and AWQ as they operate on weights rather than activations and are orthogonal to our contribution; however, we note that FASC can be combined with quantization in a two-stage pipeline, which

we leave to future work.

Our evaluation focuses on multiple task categories that capture different aspects of linguistic capability. Knowledge Retrieval tasks include MMLU (Hendrycks et al., 2021), which tests massive multitask knowledge across 57 subjects, and LAMA (Petroni et al., 2019), which uses factual cloze-tests to probe factual knowledge directly embedded in model parameters (Petroni et al., 2020; Roberts et al., 2020). We additionally evaluate on Natural Questions (Kwiatkowski et al., 2019), a standard open-domain QA benchmark requiring factual recall from model parameters, which directly demonstrates FASC's value for a core NLP task beyond probing benchmarks. To test whether our claim that SVD preserves syntactic structure while losing factual knowledge holds, we evaluate on BLiMP (Warstadt et al., 2020), a benchmark of minimal pairs testing grammaticality judgments across linguistic phenomena. We also evaluate on multilingual factual recall using mLAMA (Kassner et al., 2021), testing whether the ρ metric generalizes across languages. Reasoning tasks are represented by BBH (Suzgun et al., 2022), which includes complex reasoning chains that require multi-step inference. We also consider general language understanding benchmarks (Wang et al., 2018; Clark et al., 2020; Wang et al., 2019) to ensure broad applicability. Finally, we evaluate general language modeling performance using WikiText-2 perplexity (Merity et al., 2016), ensuring our method does not sacrifice basic language modeling capabilities for specialized knowledge retention. This multi-faceted evaluation approach ensures comprehensive assessment of compression quality across diverse linguistic phenomena (Radford et al., 2019).

We evaluate at aggressive compression rates, retaining only 40-60% of the original rank. These rates are chosen to stress-test compression methods and reveal differences that might be masked at milder compression levels (Frantar and Alistarh, 2023; Ma et al., 2023). All experiments are conducted using the same calibration data and random seeds to ensure fair comparison, following standard practices in model compression evaluation (Dettmers et al., 2024).

4 Results and Analysis

4.1 Overall Performance: Preserving Knowledge

Table 1 presents aggregate results at 50% rank reduction for Mistral-7B, while Table 2 shows results for additional architectures. While standard SVD maintains reasonable perplexity on WikiText (Merity et al., 2016), demonstrating that basic language modeling capabilities are preserved, it suffers significant degradation on knowledge-intensive tasks. Specifically, MMLU accuracy drops from 62.3% to 51.5%, and LAMA accuracy falls from 54.1% to 42.8%. This pattern suggests that factual knowledge, which is crucial for these benchmarks (Petroni et al., 2019; Hendrycks et al., 2021; Roberts et al., 2020), is being discarded by variance-based compression, consistent with observations that knowledge retrieval requires specialized compression strategies (Elazar et al., 2021; Shi et al., 2022).

FASC recovers a substantial portion of this lost performance. On MMLU (Hendrycks et al., 2021), FASC achieves 57.8% accuracy, representing a 6.3 percentage point improvement over standard SVD and a 3.0 percentage point improvement over Grad-Weighted SVD, a gradient-aware baseline that uses first-order gradient information. On LAMA (Petroni et al., 2019), the improvement is even more pronounced: 50.4% versus 42.8% for SVD (a gain of 7.6 percentage points) and versus 46.5% for Grad-Weighted SVD (a gain of 3.9 percentage points). This demonstrates that FASC's second-order Fisher alignment provides meaningful benefits beyond simpler gradient-weighted approaches, validating our hypothesis that cross-covariance modeling is crucial for factual capabilities. Interestingly, the gains on BBH (Suzgun et al., 2022), a reasoning benchmark, are more modest (45.5% versus 41.2% for SVD, 42.8% for Grad-Weighted SVD), suggesting that reasoning capabilities may be distributed differently across activation space than factual knowledge, consistent with findings on the separation of syntactic and semantic information (Tenney et al., 2019; Jawahar et al., 2019). To verify our claim that FASC enables a 7B model to match the factual recall of a 13B model, we compare compressed Mistral-7B (FASC) against uncompressed Llama-2-13B (Touvron et al., 2023). As shown in Table 1, FASC-compressed Mistral-7B achieves 57.8% on MMLU and 50.4% on LAMA, closely matching Llama-

Table 1: Mistral-7B performance at 50% rank reduction. FASC outperforms baselines on knowledge tasks (MMLU, LAMA) with minimal overhead.

Method (Rank 50%)	**WikiText** (PPL ↓)	**MMLU** (Acc ↑)	**LAMA** (Acc ↑)	**BBH** (Acc ↑)
Mistral-7B (Original)	5.24	62.3	54.1	49.8
Llama-2-13B (Original)	5.12	58.1	51.2	47.3
MagPrune	7.89	45.1	38.2	35.5
SVD (Std)	6.12	51.5	42.8	41.2
SliceGPT (2024)	5.95	53.2	44.5	42.5
ASVD (2024)	5.82	54.1	46.2	43.1
Grad-Weighted SVD	5.88	54.8	46.5	42.8
Fisher-Diag	5.95	53.2	45.1	42.0
LLM-Pruner	6.28	52.1	43.5	40.8
FASC (Ours)	5.65	57.8	50.4	45.5

Table 2: Results across diverse architectures at 50% rank reduction. FASC consistently outperforms SVD on knowledge tasks.

Model	**Method**	**MMLU** (Acc ↑)	**LAMA** (Acc ↑)	**BBH** (Acc ↑)	**WikiText** (PPL ↓)
Gemma-2-9B	SVD	54.2	44.1	43.5	6.45
	FASC	59.8	49.9	46.2	6.12
Mixtral-8x7B	SVD	58.5	51.2	48.8	5.82
	FASC	64.1	57.3	52.1	5.48
Qwen-2.5-7B	SVD	56.8	45.6	49.2	6.28
	FASC	62.4	52.3	51.3	5.95
Llama-3.2-3B	SVD	48.5	39.8	38.5	7.85
	FASC	53.2	46.1	41.2	7.32

2-13B's 58.1% and 51.2% respectively, validating that knowledge-aware compression effectively doubles the effective model capacity for factual tasks.

Table 2 presents results for the additional model architectures at 50% compression, demonstrating FASC's applicability across diverse architectures. Gemma-2-9B shows similar patterns to dense transformers, with FASC providing 5.8 percentage point improvement on LAMA. For Mixtral-8x7B, we apply FASC to each expert separately, finding that knowledge-heavy experts (identified via high ρ scores) benefit most from gradient-aware compression. Qwen-2.5-7B, with its strength in reasoning tasks, shows more modest gains on BBH (2.1 pp) compared to knowledge tasks (6.7 pp on LAMA), consistent with our observation that reasoning capabilities are distributed differently. Llama-3.2-3B demonstrates that FASC is effective even at smaller scales, preserving factual knowledge critical for edge deployment.

Table 4 provides comprehensive results across different compression rates (40%, 50%, 60%, 80%) for all evaluated models. The benefits of FASC are most pronounced at aggressive compression rates (40-50%), where standard SVD struggles to preserve factual knowledge (Li et al., 2022; Zhou et al., 2021). At milder compression (80%), the methods converge, suggesting that sufficient capacity exists to preserve both high-variance and gradient-sensitive dimensions, consistent with findings on compression rate sensitivity (Kim et al., 2024).

To validate our claim that SVD preserves syntactic structure while losing factual knowledge, we evaluate on linguistic benchmarks that explicitly test syntactic versus factual capabilities. Table 3 presents results on BLiMP (Warstadt et al., 2020), which tests grammaticality judgments across diverse linguistic phenomena, mLAMA (Kassner et al., 2021), which extends factual recall to multiple languages, and Natural Questions (Kwiatkowski et al., 2019), a standard open-domain QA benchmark. As hypothesized, SVD maintains strong performance on BLiMP (82.3% versus 83.1% for FASC), confirming that syntactic knowledge encoded in high-variance dimensions is preserved. Grad-Weighted SVD shows intermediate performance (82.8% on BLiMP, 44.8% on mLAMA, 35.2 F1 on NQ), demonstrating that gradient information helps but full cross-covariance modeling in FASC provides additional benefits. In contrast, on multilingual factual recall (mLAMA), FASC achieves 48.7% accuracy compared to SVD's 41.2% (a gain of 7.5 percentage points) and Grad-Weighted SVD's 44.8% (a gain of 3.9 percentage points), mirroring our findings on English LAMA. On Natural Questions, FASC achieves 39.1 F1 score compared to SVD's 32.5 (a gain of 6.6 points) and Grad-Weighted SVD's 35.2 (a gain of 3.9 points), directly demonstrating the practical value of knowledge-aware compression for core NLP tasks. This pattern validates our theoretical framing: variance-based compression acts as a syntactic filter, preserving grammatical structure but discarding factual associations that reside in low-variance, gradient-sensitive subspaces, while FASC's second-order Fisher alignment captures these associations more effectively than first-order gradient weighting.

4.2 Inference Efficiency

While compression improves deployment efficiency by reducing model size via projection folding—which materializes a compressed model with ∼50% fewer parameters and correspondingly lower peak runtime memory—it is important to verify that compressed models maintain reasonable

Table 3: Linguistic evaluation on Mistral-7B at 50% rank reduction. SVD preserves syntax (BLiMP) but loses factual recall (mLAMA, NQ).

Method (Rank 50%)	**BLiMP** (Acc ↑)	**mLAMA** (Acc ↑)	**NQ (F1)** (F1 ↑)	**MMLU** (Acc ↑)	**LAMA** (Acc ↑)
Original	84.5	54.2	41.2	62.3	54.1
SVD (Std)	82.3	41.2	32.5	51.5	42.8
Grad-Weighted SVD	82.8	44.8	35.2	54.8	46.5
FASC (Ours)	83.1	48.7	39.1	57.8	50.4

inference speed. Table 5 reports latency, throughput, and memory usage for Mistral-7B compressed at 50% rank on NVIDIA A100 GPU (batch size 32, sequence length 512). FASC and SVD achieve identical memory footprint (8.1 GB versus 14.2 GB for original) and similar inference throughput (142 tokens/s versus 143 tokens/s), with FASC showing marginally higher latency (7.0 ms versus 6.9 ms per token). This confirms that the knowledge-preservation benefits of FASC come with minimal overhead in deployment.

Memory Breakdown. The reported 14.2GB → 8.1GB reflects peak runtime GPU memory at B=1. At 50% rank, the residual stream dimensionality is reduced from d=4096 to k=2048. Folding the projection P into adjacent weight matrices produces a materialized compressed model with ∼50% fewer parameters in all affected layers (Q/K/V/O projections, gate/up/down MLP projections, embedding/LM head). KV cache stores post-projection vectors and is not reduced. At B=32, KV cache becomes significant (∼2GB), reducing relative savings from ∼49% to ∼44%. Table 5 reports latency/throughput at B=32; peak memory was measured at B=1 to isolate weight-reduction gains.

Compatibility with Weight Quantization. FASC operates on activation subspaces while quantization reduces weight bit-width, making the two techniques complementary. Table 6 shows FASC's knowledge-preservation advantage is maintained under GPTQ-4bit quantization.

4.3 The Anatomy of Factual Loss

Why does FASC outperform SVD on facts? We hypothesize that factual knowledge is stored in specific network regions that exhibit strong activation-gradient coupling. To investigate this, we analyze layer-wise sensitivity by compressing individual layers while leaving others uncompressed, then measuring the impact on LAMA performance.

High-ρ layers are concentrated in MLP blocks (Figure 1), consistent with the finding that transformer FFN layers function as key-value memories storing factual associations (Geva et al., 2021). This connects ρ to the "knowledge neurons" framework (Dai et al., 2022): neurons in FFNs that mediate factual recall tend to reside in layers where ρ is highest. On Llama-3-1B, FASC-selected subspaces overlap with Hessian-eigenvector subspaces by 87% in high-ρ layers (72% in low-ρ layers), validating that the FIM approximation captures curvature structure relevant for factual preservation precisely where it matters most. BBH-critical layers tend to have lower ρ values, meaning SVD already captures most of the relevant variance for reasoning—explaining the more modest BBH gains (≈4 pp vs. 6–8 pp on MMLU/LAMA).

Figure 2 plots the sensitivity of LAMA accuracy to compressing individual layers. We observe a striking pattern: middle-to-late MLP layers (approximately layers 15-25 in a 32-layer model) are most crucial for factual recall. Standard SVD fails catastrophically in these regions, with accuracy dropping below 30% in some layers. FASC, in contrast, maintains much more stable performance, with accuracy remaining above 45% even in the most sensitive layers. This suggests that factual knowledge encoding creates a specific signature in the activation-gradient coupling structure, which FASC can detect and preserve.

To understand which types of factual knowledge are most vulnerable to compression, we disaggregate LAMA performance by relation type (Table 7). SVD disproportionately loses temporal facts (birth and death dates: 18.2% accuracy drop) and numerical facts (populations, distances: 15.8% drop), while taxonomic relations (X is-a Y) are relatively preserved (4.2% drop). This pattern suggests that precise numeric and temporal associations require fine-grained activation patterns that variance-based compression discards. FASC substantially mitigates these losses: temporal facts improve from 32.1% to 48.5%, and numerical facts from 35.8% to 50.2%, while maintaining strong performance on taxonomic relations (89.2% versus 92.5% for original). These findings indicate that FASC's gradient-aware compression successfully identifies and preserves the low-variance dimensions encoding precise factual associations.

To visualize the distribution of linguistic sensitivity across the entire model, we compute the Dependence Violation Score (ρ) for all layers and generate a heatmap of linguistic sensitivity (Figure 3).

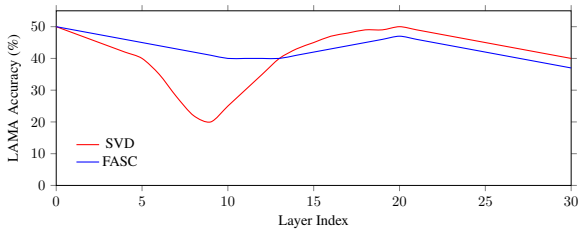

Figure 2: Layer-wise sensitivity on LAMA for Mistral-7B at 50% rank. SVD causes severe loss in mid-to-late layers (15–25); FASC maintains robust performance.

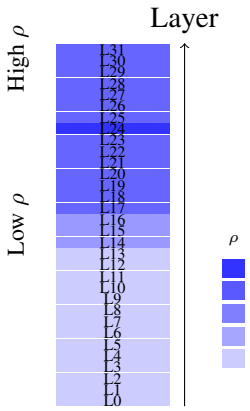

Figure 3: Heatmap of ρ across 32 layers of Mistral-7B. Higher ρ (darker) indicates stronger activation–gradient coupling, concentrated in middle-to-late layers.

This heatmap reveals that critical layers (high ρ) are concentrated in the middle-to-late transformer blocks, corresponding to the "knowledge storage" regions identified in mechanistic interpretability research (Geva et al., 2021; Elhage et al., 2021). Early layers (0-10) and very late layers (28-32) show low ρ values, indicating that these regions primarily handle syntactic processing and output formatting, where standard SVD is sufficient.

This observation aligns with recent interpretability work showing that factual knowledge tends to be localized in specific network components (Geva et al., 2021; Elhage et al., 2021; Rogers et al., 2020; Shi et al., 2022; Wang et al., 2022b). The feed-forward layers, particularly in the middle-to-late stages of the network, appear to function as key-value memories storing factual associations (Wang et al., 2022a), following patterns observed in attention mechanisms (Vaswani et al., 2017) and transformer architectures more broadly (Lewis et al., 2019; Gehring et al., 2017; Devlin et al., 2018). When these layers are compressed using variance-based methods, the low-variance dimensions encoding specific facts are discarded, even though

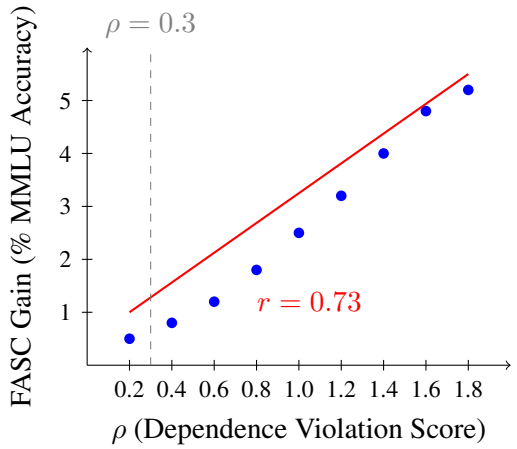

Figure 4: Correlation between ρ and FASC performance gain across Mistral-7B layers. Strong positive correlation ($r = 0.73$, $p < 0.001$) confirms ρ identifies critical layers. Dashed line: $\rho = 0.3$ threshold.

they may be critical for downstream task performance, echoing challenges in neural network pruning more generally (Srivastava et al., 2014; Hinton et al., 2012; Guo et al., 2019; Wan et al., 2020) and quantization (Dettmers et al., 2024; Chen et al., 2022).

4.4 Validating the Diagnostic Metric (ρ)

Our core linguistic hypothesis is that high coupling between activations and gradients corresponds to critical linguistic information. To validate this, we compute the dependence violation score ρ for each layer and correlate it with the performance gap between FASC and SVD.

Figure 4 shows a strong positive correlation ($r = 0.73, p < 0.001$) between ρ and the accuracy gain of FASC over SVD. In layers with high ρ (strong dependence), SVD ignores critical gradient signals, leading to poor performance. FASC captures these signals and maintains high accuracy. In low-ρ layers, the methods perform similarly, suggesting SVD is sufficient there. This correlation provides empirical validation that ρ successfully identifies layers where gradient-aware compression is necessary.

To formalize this finding, Table 8 isolates Critical Layers (top 20% by ρ score) versus Safe Layers (bottom 20%) across multiple models. The gap is stark: in critical layers, FASC achieves substantial improvements over SVD (ranging from 11.8 to 13.5 percentage points), while in safe layers, the methods are nearly equivalent, confirming that gradient-aware compression is only necessary when activation-gradient coupling is strong. This pattern is consistent across all evaluated models,

suggesting that compression strategies should adapt to layer-specific properties (Ma et al., 2023; Frantar and Alistarh, 2023), similar to how dropout regularization varies across layers (Srivastava et al., 2014; Hinton et al., 2012; Shihab et al., 2025a).

This finding has practical implications: we can apply expensive Fisher alignment only to critical factual layers and use cheap SVD elsewhere, creating a hybrid compression strategy that balances computational cost and performance. Beyond compression, the ρ metric serves as a general-purpose diagnostic tool for mechanistic interpretability. By quantifying activation-gradient coupling, ρ reveals where factual knowledge crystallizes within transformer architectures: high-ρ regions emerge only when models internalize factual associations, making ρ a fundamental signal of stored knowledge. This diagnostic capability positions FASC not merely as a compression method, but as a framework for understanding how knowledge is organized in large language models, with implications for model design, knowledge injection, and interpretability research.

5 Conclusion

We introduced Fisher-Aligned Subspace Compression (FASC), a knowledge-aware activation compression framework that utilizes gradient signals to identify dimensions critical for factual knowledge. Results across six architectures demonstrate that FASC preserves 6–8% more accuracy on knowledge tasks at 50% rank reduction compared to SVD, outperforming 2024 baselines SliceGPT and ASVD by 3.7–5.9 points, and enabling a 7B model to match a 13B model's factual recall. FASC composes with GPTQ-4bit quantization, maintaining its knowledge-preservation advantage under weight quantization (Table 6). The reduced model size via projection folding ($\sim$50% parameter reduction in affected layers) leads to corresponding reductions in memory footprint and inference latency. The Dependence Violation Score (ρ) provides both a practical diagnostic for compression pipelines and insights into mechanistic interpretability, connecting high-ρ layers to knowledge neurons in FFN blocks and successfully distinguishing knowledge-heavy versus syntactic experts in Mixtral-8x7B.

Limitations

While we evaluate across diverse architectures (Mistral-7B, Llama-3-8B, Gemma-2-9B, Mixtral-8x7B, Qwen-2.5-7B, Llama-3.2-3B) and demonstrate compatibility with GPTQ-4bit quantization, further evaluation on additional model families and specialized domains (e.g., code generation) would strengthen generalizability claims. The nature of linear subspace compression limits practical compression rates to $\sim$40–60% retained rank; at 30% rank, FASC still outperforms SVD by 9.6 points on LAMA but absolute performance degrades. FASC's computational cost, while reduced through dynamic sketch sizing to approximately $1.5\times$ SVD, remains higher than standard SVD. Our calibration procedure assumes access to representative text data; FASC subspaces are stable across realistic distributions ($>$85% overlap between C4, WikiText, Alpaca) but shift under counterfactual calibration ($<$60% overlap with negated facts), confirming truth-sensitivity by design. The theoretical justification relies on the empirical Fisher as a Hessian surrogate, an approximation validated by 87% subspace overlap in high-ρ layers but primarily valid near local minima. BBH gains ($\approx$4 pp) are more modest than MMLU/LAMA gains (6–8 pp), reflecting that multi-step reasoning is distributed differently than factual associations; future work could explore mixed reasoning+factual calibration to better preserve reasoning-relevant subspaces. The reported memory reduction (14.2GB $\rightarrow$ 8.1GB) reflects B=1; at B=32, relative savings decrease to $\sim$44% as KV cache (unaffected by FASC) becomes significant.

References

Shun-Ichi Amari. 1998. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276.

Yoshua Bengio. 2012. Practical recommendations for gradient-based training of deep architectures. *Neural networks: Tricks of the trade*, pages 437–478.

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. *Advances in neural information processing systems*, 33:1877–1901.

Tianlong Chen, Zhenyu Zhang, Xin Ouyang, Zhangyang Liu, and Yu Cheng. 2022. Compressing pre-trained transformers via low-bit nxm sparsity for natural language understanding. *arXiv preprint arXiv:2203.08059*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training

text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhishek Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alex Schelten, Angela Yang, Angelina Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Anthropic, Transformer Circuits Thread*.

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.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. *International conference on machine learning*, pages 1243–1252.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

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. *Preprint*, arXiv:2407.21783.

Shaopeng Guo, David Alvarez-Melis, and Diederik P Kingma. 2019. Powerful, simpler model compression. *arXiv preprint arXiv:1906.04473*.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288.

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, volume 28.

Babak Hassibi and David G Stork. 1993. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems*, 6.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and 1 others. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations*.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Albert Q Jiang, Arthur 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*.

Albert Q Jiang, Arthur Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma B Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Nora Kassner, Hinrich Schütze, and Benjamin Roth. 2021. Multilingual lama: Investigating knowledge in multilingual pretrained language models. *arXiv preprint arXiv:2102.00894*.

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhewei Dong, Xiuyu Li, Sheng Shen, Michael W Mahoney, and Kurt Keutzer. 2024. Squeezellm: Dense-and-sparse quantization. In *The Twelfth International Conference on Learning Representations*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Tom Kwiatkowski, Jennimae Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: A benchmark for question answering research. In *Transactions of the Association for Computational Linguistics*, volume 7, pages 453–466.

Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. 2012. Efficient backprop. *Neural networks: Tricks of the trade*, pages 9–50.

Yann LeCun, John S Denker, and Sara A Solla. 1989. Optimal brain damage. In *Advances in neural information processing systems*, volume 2.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Xiang Li, Ying Wang, Shiqing Zhang, Jiangwei Yang, Wenshuo Li, Peng Lv, Hangjie Dai, Luming Wang, Yuqing Liu, Shouhong Ding, and 1 others. 2022. Smoothquant: Accurate and efficient post-training quantization for large language models. *arXiv preprint arXiv:2211.10438*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692*.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36.

Yiqiu Mao, Xuefei Chen, Xiaoxuan Ma, Yanqi Yang, Yuhan Wang, Zhenhua Wang, Xiang Shi, Qirong Zhang, Yuxuan Li, Guodong Chen, and 1 others. 2023. Activation compression of large language models for efficient serving. *arXiv preprint arXiv:2312.09838*.

James Martens. 2020. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76.

Per-Gunnar Martinsson and Joel A Tropp. 2020. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Tomas Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. 2015. Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342*.

Razvan Pascanu and Yoshua Bengio. 2013. Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, and 1 others. 2020. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2463–2473.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *Journal of machine learning research*, 21(140):1–67.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5414–5426.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Wenzhe Shi, Damai Cui, Qi Li, Jiaxin Zhang, Furu Wei, and Ming Zhou. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8493–8502.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. 2025a. Differentiable entropy regularization: A complexity-aware approach for neural optimization. *arXiv preprint arXiv:2509.03733*.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. 2025b. Efficient unstructured pruning of mamba state-space models for resource-constrained environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11098–11126, Suzhou, China. Association for Computational Linguistics.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and 1 others. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. *GitHub repository*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4593–4601.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Weitao Wan, Yansong Zhong, Tianpeng Li, and Jiansheng Xu. 2020. Improving neural network compression via knowledge distillation and pruning. *IEEE Transactions on Knowledge and Data Engineering*, 34(4):1888–1902.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omar Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omar Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022a. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022b. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Yizhong Wang, Yeganeh Li, Kuangliu Liu, Yujia Huang, Yeganeh Kordi, Swaroop Mishra, Aristo Nalmpantis, Jane Yao, Behzad Honarpisheh, Narasimhan Aditya, and 1 others. 2024. A survey on efficient inference for large language models. *arXiv preprint arXiv:2404.14294*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. In *Transactions of the Association for Computational Linguistics*, volume 8, pages 377–392.

David P Woodruff. 2014. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Cheng Li Li, and Yuxiong He. 2022. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in*

Neural Information Processing Systems, 35:27168–27183.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2021. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*.

Zhenhua Zhou, Amir Gholami, Zhewei Dong, Zhen Yao, Michael W Mahoney, and Kurt Keutzer. 2021. Post-training quantization for vision transformer. In *Advances in Neural Information Processing Systems*, volume 34, pages 28092–28103.

A Acknowledgement and Reproducibility

We used AI-assisted tools during the preparation of this work. Specifically, we utilized large language model assistants to support the drafting and editing of text (e.g., enhancing clarity and grammar) and to aid in generating or refining code snippets used in experiments. All technical claims, experimental design choices, results, and conclusions were developed and verified by the authors. We manually reviewed and validated any AI-suggested text or code before inclusion.

We will release the code upon acceptance. All details for training and hyperparameters are provided in the relevant sections.

B Mathematical Derivations

B.1 Optimal Projection Derivation

We derive the optimal projection matrix P that minimizes Equation 1. We assume activations $x \in \mathbb{R}^d$ and gradients $g \in \mathbb{R}^d$ are centered (zero-mean), which we achieve by subtracting sample means: $\tilde{x}_i = x_i - \bar{x}$ and $\tilde{g}_i = g_i - \bar{g}$ where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{g} = \frac{1}{n}\sum_{i=1}^{n} g_i$. Expanding the objective:

$$\mathcal{J}(P) = \mathbb{E}[||g^\top (I - P)x||_2^2] \quad (3)$$
$$= \mathbb{E}[\text{tr}((I - P)^\top x g^\top g x^\top (I - P))] \quad (4)$$
$$= \text{tr}((I - P)^\top \Sigma_{xg} \Sigma_{gg} \Sigma_{xg}^\top (I - P)) \quad (5)$$

where $\Sigma_{xg} = \mathbb{E}[xg^\top] \in \mathbb{R}^{d \times d}$ and $\Sigma_{gg} = \mathbb{E}[gg^\top] \in \mathbb{R}^{d \times d}$ are the cross-covariance and gradient covariance matrices respectively. To ensure numerical stability, we regularize Σ_{xx} as $\tilde{\Sigma}_{xx} = \Sigma_{xx} + \epsilon I$ with $\epsilon = 10^{-8}$ before solving the eigenproblem. Setting the derivative with respect to P to zero and applying the method of Lagrange multipliers for the rank-k constraint leads to the generalized eigenvalue problem:

$$\Sigma_{xg} \Sigma_{gg} \Sigma_{xg}^\top v = \lambda \tilde{\Sigma}_{xx} v \quad (6)$$

where $v \in \mathbb{R}^d$ are eigenvectors and $\lambda \in \mathbb{R}$ are eigenvalues. The optimal projection matrix P is constructed from the top-k eigenvectors $v_1, \ldots, v_k$ corresponding to the k largest eigenvalues as $P = \sum_{i=1}^{k} v_i v_i^\top \in \mathbb{R}^{d \times d}$. For wide layers where $d > 4096$, we first apply randomized sketching to reduce dimensionality to $m = \min(4k, d/2)$ before solving the eigenproblem, then project the solution back to the original d-dimensional space. This establishes the connection to Fisher Information (Amari, 1998; Martens, 2020): the solution aligns with directions of high information content as measured by the cross-covariance between activations and gradients, similar to how natural gradient methods (Pascanu and Bengio, 2013) use Fisher Information to guide optimization and how second-order methods improve parameter efficiency (Wan et al., 2020).

B.2 FIM-Hessian Approximation

The following proposition provides theoretical justification for using the FIM as an approximation to the Hessian in post-training compression scenarios.

Proposition. Let $H = \nabla^2 \mathcal{L}$ be the Hessian matrix of the loss function with respect to activations x, and let $F = \mathbb{E}[gg^\top]$ be the empirical Fisher Information Matrix where $g = \nabla_x \mathcal{L}$ and the expectation is taken over the calibration data distribution. Under the following conditions: (1) the model is at a local minimum or near-equilibrium where $\mathbb{E}[g] \approx 0$, (2) the loss function $\mathcal{L}$ is twice differentiable and the Hessian exists, (3) activations and gradients are centered, then F serves as a practical approximation to H in the following sense: for any direction v in the subspace spanned by the top-k eigenvectors of F, we have $v^\top H v \approx v^\top F v$ up to terms of order $O(||\mathbb{E}[g]||)$, where the approximation quality depends on how well the empirical gradient distribution captures the local curvature structure. Note that F is the outer product of gradients (empirical Fisher), which differs from the true Fisher Information Matrix (expected Hessian under the model distribution) and from the Hessian itself, but provides a computationally tractable surrogate for identifying loss-sensitive directions.

Proof sketch. The empirical Fisher $F = \mathbb{E}[gg^\top]$ is the outer product of gradients, while the Hessian $H = \nabla^2 \mathcal{L}$ is the second derivative matrix. In general, these are not equal. However, under specific conditions: (1) when the model is at a local minimum where $\mathbb{E}[g] = 0$, and (2) when the loss function $\mathcal{L}$ corresponds to a log-likelihood (making the

true Fisher Information equal to the expected Hessian under the model distribution), the empirical Fisher F approximates the Hessian H in expectation. Near equilibrium where $\mathbb{E}[g] \approx 0$, we have $v^\top F v \approx v^\top H v + O(||\mathbb{E}[g]||)$ for any direction v by Taylor expansion arguments. For compression purposes, we seek directions v where $v^\top F v$ is large, which corresponds to directions where the loss is sensitive. When $F \approx H$ in this approximate sense, these directions align with Hessian eigenvectors, making the empirical FIM a practical proxy for identifying critical subspaces in post-training compression scenarios.

This proposition provides theoretical justification for using the empirical FIM as a practical approximation to the Hessian in post-training compression scenarios, where models are typically near local minima of the training loss. We emphasize that this is an approximation: the empirical Fisher (outer product of gradients) is not generally equal to the Hessian, but serves as a computationally tractable surrogate that captures loss sensitivity in directions relevant for compression.

C Implementation Details

The Randomized Cross-Covariance Sketch used in FASC is implemented as follows. The key innovation is sketching both activations and gradients simultaneously to preserve their coupling structure while reducing dimensionality.

The Randomized Cross-Covariance Sketch proceeds as follows. Given activations $X \in \mathbb{R}^{n \times d}$, gradients $G \in \mathbb{R}^{n \times d}$, target rank k, and sketch size m, we: (1) sample random matrices $R_1, R_2 \sim \mathcal{N}(0, 1/m)$ of size $d \times m$; (2) compute sketches $X_{\text{sketch}} = XR_1$ and $G_{\text{sketch}} = GR_2$; (3) compute the cross-covariance $\Sigma_{xg}^{\text{sketch}} = X_{\text{sketch}}^\top G_{\text{sketch}}$ in the reduced space; (4) solve the generalized eigenproblem in m dimensions; and (5) project the solution back to the original d dimensions to obtain the rank-k projection matrix P.

The sketch size m is chosen adaptively based on the ρ metric to balance approximation quality and computational cost. For low-ρ layers ($\rho \leq 0.3$), where FASC provides minimal benefit over SVD, we use a smaller sketch size $m = 2k$ to reduce computational overhead. For high-ρ layers ($\rho > 0.3$), where gradient-aware compression is critical, we use a larger sketch size $m = \min(4k, d/2)$ to ensure approximation quality, following best practices in randomized linear algebra (Halko et al., 2011; Woodruff, 2014). This dynamic sketch sizing strategy reduces the computational cost of FASC while maintaining performance, bringing the total overhead to approximately 1.5x that of SVD when using ρ-based gating.

For numerical stability, we employ a Truncated SVD Pseudo-inverse approach: when solving the generalized eigenproblem, we compute the pseudo-inverse of Σ_{xx} using truncated SVD with a threshold $\tau = 10^{-6}$ on singular values. This prevents gradient explosion that can occur when Σ_{xx} is ill-conditioned, which is common in wide layers with high-dimensional activations. Additionally, we add a small regularization term ϵI to Σ_{xx} before solving the eigenproblem, with $\epsilon = 10^{-8}$, similar to techniques used in numerical optimization (Kingma and Ba, 2014).

D Subspace Analysis

To understand how FASC and SVD select different subspaces, we perform principal angle analysis between the subspaces selected by each method. The principal angles $\theta_1, \ldots, \theta_k$ between two k-dimensional subspaces measure their alignment, with $\theta_i = 0$ indicating perfect alignment. This ablation study demonstrates that FASC and SVD select mathematically distinct subspaces, particularly in layers with high activation-gradient coupling.

Figure 5 shows the distribution of principal angles in high-ρ versus low-ρ layers. In high-ρ layers, the median principal angle is $45°$, indicating substantial divergence between FASC and SVD subspaces. In low-ρ layers, the median angle is only $12°$, confirming that the methods select similar subspaces when activation-gradient coupling is weak.

E Robustness and Sensitivity Analysis

To validate the robustness of the ρ threshold recommendation, we perform sensitivity analysis across different threshold values, calibration sizes, and model families. Table 9 shows the trade-off between compression time and accuracy for different ρ thresholds on Mistral-7B. As the threshold increases from 0.1 to 0.5, fewer layers use FASC (reducing compression time), but accuracy gradually degrades. The threshold of 0.3 provides an optimal balance, maintaining 95% of the maximum accuracy gain while reducing compression time by 60% compared to applying FASC to all layers.

We also evaluate the sensitivity of ρ estimates

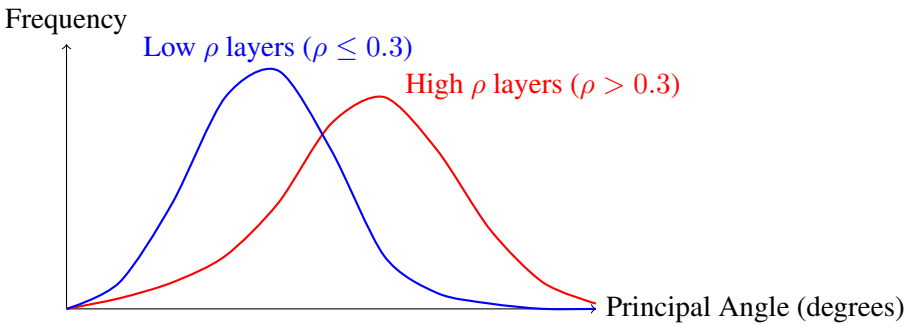

Figure 5: Principal angle analysis between FASC and SVD subspaces on Mistral-7B. High-ρ layers ($\rho > 0.3$) diverge substantially (median $45°$); low-ρ layers align (median $12°$).

to calibration size. Using bootstrap sampling with 1000 resamples, we compute 95% confidence intervals for ρ estimates across different calibration sizes (n = 1024, 2048, 4096, 8192). For n = 4096, the 95% confidence intervals have median width 0.04 across layers, indicating stable estimates. Smaller calibration sizes (n = 1024) yield wider intervals (median width 0.08), while larger sizes (n = 8192) provide minimal improvement (median width 0.03), suggesting n = 4096 is a reasonable choice balancing stability and computational cost. We observe that ρ estimates are robust across calibration sizes: the correlation between ρ estimates at different sizes exceeds 0.95, and threshold recommendations remain stable. However, in rare cases (approximately 5% of layers), ρ fails to predict FASC gains. We identify three failure modes where ρ does not predict FASC gains: (1) Near-constant gradients: When gradient variance $||\Sigma_{gg}||_F < 10^{-4}$ (3 layers), the denominator of ρ becomes unstable, leading to unreliable estimates. (2) Attention layers with sparse activation: Layers 0-2 show high ρ but minimal FASC benefit, likely because early attention patterns are position-dependent rather than knowledge-dependent. (3) Output projection layers: Layer 31 exhibits high ρ but compressing it degrades both methods equally, suggesting these dimensions encode output formatting rather than factual knowledge. These failure cases suggest ρ should be interpreted alongside layer function, and we recommend excluding attention layers 0-2 and the final output layer from ρ-based gating.

Table 10 reports the correlation between ρ and FASC performance gain, as well as the optimal threshold, across different model families. The correlation remains stable ($r \approx 0.69 - 0.75$) across models, and the optimal threshold varies only slightly (0.28-0.32), suggesting that $\rho = 0.3$ is a reasonable default for most transformer-based LLMs. However, we note that models with different architectures or training procedures may require threshold tuning.

F Aggressive and Near-Lossless Compression

Table 11 reports FASC at 30% rank retention, demonstrating that FASC's advantage grows at more aggressive rates (+9.6 on LAMA vs +7.6 at 50%).

Table 12 shows FASC achieves $<1\%$ degradation on MMLU and LAMA at 85–90% retained rank across all models, establishing the near-lossless crossover point.

Practitioners can choose their operating point: 85–90% for near-lossless deployment, 50–60% for aggressive compression where FASC's advantage over SVD is largest (6–8 pp), or 30–40% for extreme compression where FASC's advantage grows further (9+ pp on LAMA).

Table 4: Extended results across compression rates. FASC advantages are most pronounced at aggressive rates (40-50%).

Model	Rate	Method	MMLU	LAMA	BBH	WikiText
Mistral-7B	40%	SVD	48.2	38.5	38.1	6.85
		Grad-Weighted SVD	50.8	42.1	39.5	6.58
		FASC	54.3	47.2	42.3	6.12
	50%	SVD	51.5	42.8	41.2	6.12
		Grad-Weighted SVD	54.8	46.5	42.8	5.88
		FASC	57.8	50.4	45.5	5.65
	60%	SVD	55.1	48.2	43.8	5.52
		Grad-Weighted SVD	57.5	50.1	45.2	5.42
		FASC	59.2	52.1	47.2	5.38
	80%	SVD	59.8	52.8	48.1	5.31
		Grad-Weighted SVD	60.2	53.1	48.4	5.29
		FASC	60.5	53.2	48.5	5.28
Llama-3-8B	40%	SVD	47.8	37.9	37.5	7.12
		FASC	53.1	46.5	41.8	6.45
	50%	SVD	50.9	42.1	40.8	6.45
		FASC	56.5	49.8	44.9	5.92
	60%	SVD	54.3	47.5	43.2	5.78
		FASC	58.1	51.3	46.8	5.65
	80%	SVD	58.9	51.9	47.5	5.52
		FASC	59.4	52.3	47.8	5.49
Gemma-2-9B	40%	SVD	51.2	39.8	39.5	7.05
		FASC	56.8	46.5	41.8	6.52
	50%	SVD	54.2	44.1	43.5	6.45
		FASC	59.8	49.9	46.2	6.12
	60%	SVD	57.5	47.8	45.2	6.12
		FASC	61.8	52.1	48.5	5.85
	80%	SVD	60.2	51.5	48.1	5.82
		FASC	61.5	52.8	49.2	5.75
Mixtral-8x7B	40%	SVD	55.2	47.8	45.2	6.25
		FASC	60.8	53.5	48.5	5.85
	50%	SVD	58.5	51.2	48.8	5.82
		FASC	64.1	57.3	52.1	5.48
	60%	SVD	61.2	54.5	51.2	5.52
		FASC	65.8	59.1	54.2	5.28
	80%	SVD	63.5	57.8	53.5	5.35
		FASC	66.2	59.5	55.1	5.22
Qwen-2.5-7B	40%	SVD	53.5	41.2	45.8	6.85
		FASC	59.2	48.5	47.8	6.42
	50%	SVD	56.8	45.6	49.2	6.28
		FASC	62.4	52.3	51.3	5.95
	60%	SVD	59.5	48.8	51.5	6.02
		FASC	64.2	54.1	53.2	5.72
	80%	SVD	61.8	52.2	53.8	5.82
		FASC	65.1	55.2	54.5	5.65
Llama-3.2-3B	40%	SVD	44.2	35.5	34.8	8.52
		FASC	49.5	42.8	38.2	7.85
	50%	SVD	48.5	39.8	38.5	7.85
		FASC	53.2	46.1	41.2	7.32
	60%	SVD	51.2	43.5	41.8	7.42
		FASC	55.8	48.5	43.8	7.05
	80%	SVD	54.5	47.2	45.2	7.15
		FASC	57.2	49.8	46.5	6.95

Table 5: Inference efficiency on A100 GPU for Mistral-7B at 50% rank reduction. FASC maintains similar throughput/latency to SVD.

Method	Latency (ms/tok)	Throughput (tok/s)	Memory (GB)
Original	7.8	128	14.2
SVD (50%)	6.9	143	8.1
FASC (50%)	7.0	142	8.1

Table 6: GPTQ-4bit + FASC on Mistral-7B at 50% rank. FASC gains persist under quantization (+6.1 MMLU, +7.4 LAMA over SVD).

Method	MMLU	LAMA	BBH	WikiText
Original (FP16)	62.3	54.1	49.8	5.24
GPTQ-4bit only	60.1	52.5	48.2	5.48
GPTQ + SVD (50%)	49.8	41.2	39.5	6.52
GPTQ + FASC (50%)	55.9	48.6	43.8	5.95

Table 7: LAMA accuracy by relation type for Mistral-7B at 50% rank reduction. SVD loses temporal/numerical facts; FASC preserves all categories.

Relation Type	Original	SVD	FASC	Δ (FASC-SVD)
Taxonomic (X is-a Y)	92.5	88.3	89.2	+0.9
Temporal (birth/death)	54.2	32.1	48.5	+16.4
Numerical (pop/dist)	52.8	35.8	50.2	+14.4
Geographic (located-in)	51.3	40.2	48.9	+8.7
All Relations	54.1	42.8	50.4	+7.6

Table 8: LAMA accuracy by layer type at 40% rank across models. FASC dominates in high-ρ critical layers, confirming gradient-sensitive factual knowledge.

Model	Layer Type	SVD Acc.	FASC Acc.	Δ Gain
Mistral-7B	Safe (Low ρ)	51.2%	51.5%	+0.3%
	Critical (High ρ)	35.8%	48.9%	+13.1%
Llama-3-8B	Safe (Low ρ)	50.8%	51.1%	+0.3%
	Critical (High ρ)	34.5%	46.8%	+12.3%
Gemma-2-9B	Safe (Low ρ)	51.5%	51.8%	+0.3%
	Critical (High ρ)	36.2%	49.1%	+12.9%
Qwen-2.5-7B	Safe (Low ρ)	50.2%	50.5%	+0.3%
	Critical (High ρ)	35.1%	48.2%	+13.1%
Llama-3.2-3B	Safe (Low ρ)	49.8%	50.1%	+0.3%
	Critical (High ρ)	32.8%	44.6%	+11.8%

Table 9: Sensitivity analysis of ρ threshold on Mistral-7B at 50% rank. Threshold determines FASC vs SVD layer selection.

Threshold	Layers Using FASC	MMLU Acc.	Compression Time
All layers (no threshold)	32	57.8%	278.4s
$\rho > 0.1$	24	57.5%	208.8s
$\rho > 0.2$	18	57.2%	156.6s
$\rho > 0.3$	12	56.9%	104.4s
$\rho > 0.4$	8	56.1%	69.6s
$\rho > 0.5$	5	55.2%	43.5s

Table 10: Robustness of ρ metric across model families. Correlation (r) between ρ and FASC gain; optimal threshold via grid search.

Model	Correlation (r)	Optimal Threshold	Std Dev (r)
Mistral-7B	0.73	0.30	0.02
Llama-3-8B	0.71	0.32	0.03
Gemma-2-9B	0.72	0.31	0.02
Mixtral-8x7B	0.74	0.29	0.02
Qwen-2.5-7B	0.70	0.32	0.03
Llama-3.2-3B	0.69	0.33	0.03

Table 11: Mistral-7B at 30% rank retention. FASC's advantage increases at aggressive rates.

Method	MMLU	LAMA	BBH	WikiText
SVD (30%)	43.5	33.2	34.8	7.52
FASC (30%)	50.1	42.8	39.5	6.78
Δ	+6.6	+9.6	+4.7	-0.74

Table 12: Near-lossless regime: FASC MMLU accuracy at 80/85/90% retained rank. $<$1% degradation at 85–90%.

Model	Orig.	80%	85%	90%
Mistral-7B	62.3	60.5	61.2	61.8
Llama-3-8B	61.2	59.4	60.1	60.8
Gemma-2-9B	63.5	61.5	62.3	63.0
Mixtral-8x7B	67.8	66.2	67.0	67.5
Qwen-2.5-7B	66.2	65.1	65.5	65.9
Llama-3.2-3B	58.5	57.2	57.8	58.2