WHAT SURVIVES IS NOT ENOUGH: LOSS-ALIGNED DIRECTIONAL SALIENCE CALIBRATION FOR LLM COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank decomposition facilitates the deployment of large language models (LLMs) by representing dense transformations within compact low-dimensional subspaces. Existing approaches primarily optimize reconstruction objectives to preserve approximation fidelity during compression. However, preserving a direction does not guarantee that its contribution is properly calibrated to the end-to-end language-modeling objective. Towards this end, we propose a novel post-compression framework termed Loss-aligned Directional Salience Calibration (**LENS**) for LLM compression. The core of **LENS** is to recalibrate what survives compression through two complementary components: *loss-aligned direction identification* and *state-adaptive salience calibration*. In particular, we characterize the first-order loss response to directional rescaling within the retained subspace and aggregate these responses into a response matrix whose eigendirections reveal modes that locally favor amplification or suppression. Subsequently, we estimate direction-specific relative gains using a response-based quadratic surrogate and calibrate their joint magnitude through a shared stage-level scale. The resulting calibration is accepted only when it achieves reliable loss reduction on a held-out validation set and is then folded directly into the existing low-rank factors, preserving the retained subspaces and ranks without introducing additional inference-time parameters or operations. Extensive experiments across multiple LLM backbones and downstream tasks demonstrate the effectiveness of **LENS**. We further show that **LENS** can be integrated with different low-rank compressors while improving their performance, supporting its use as a general post-compression calibration framework. The code is available at https://lens1.github.io/.

1 INTRODUCTION

Despite the broad applications of large language models (LLMs) (Achiam et al., 2023; Grattafiori et al., 2024; Yang et al., 2025; Fu et al., 2026), their deployment remains constrained by the high storage and computational costs of billions of parameters (Ma et al., 2023; An et al., 2024), limiting deployment in resource-constrained settings (Park et al., 2026). To address this issue, various model compression strategies have been proposed, including pruning (Ma et al., 2023; Gao et al., 2024), quantization (Shao et al., 2024; Yao et al., 2022), low-rank decomposition (Saha et al., 2024; Huang et al., 2025), and knowledge distillation (Gu et al., 2024; Xiao et al., 2026). Among them, low-rank decomposition, typically instantiated by singular value decomposition (SVD), reparameterizes large-scale linear transformations within compact low-dimensional subspaces (Wang et al., 2025c; Xv et al., 2026), providing an efficient and direct solution for structured model compression.

Recent advances in low-rank approximation generally seek to identify a compact subspace by minimizing reconstruction error under a constrained compression budget (Saha et al., 2024; Song et al., 2024; Gong et al., 2026). Building on this objective, existing work has further refined low-rank compression through improved approximation schemes and adaptive compression-budget allocation (Wang et al., 2025c; Xv et al., 2026). Despite these advances, existing SVD-based approaches (Wang et al., 2025a; Thoma et al., 2026) primarily focus on retaining a compact subspace that minimizes the reconstruction distortion (Li et al., 2026; Xv et al., 2026). However, their performance may deteriorate when the compression budget becomes highly constrained.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

More fundamentally, preserving a set of representational directions does not necessarily ensure that their contributions are properly calibrated to the compressed model's language-modeling objective (Hsu et al., 2022; Shihab et al., 2026). While subspace preservation determines which directions are retained, it does not by itself determine how strongly they should contribute to the model's end-to-end computation. As depicted in Figure 1, this gap first raises the question: ❶ *How can we reliably identify loss-aligned directions within the retained subspace?* However, reliable direction identification addresses only which directions should be recalibrated, leaving the magnitude of their adjustments unresolved. This motivates a second question: ❷ *How can we adaptively determine the calibration strength for the identified directions?* Taken together, these observations motivate a complementary perspective on SVD-based compression: beyond retaining a compact subspace, we seek to recalibrate the contributions of individual directions while leaving the compression structure unchanged.

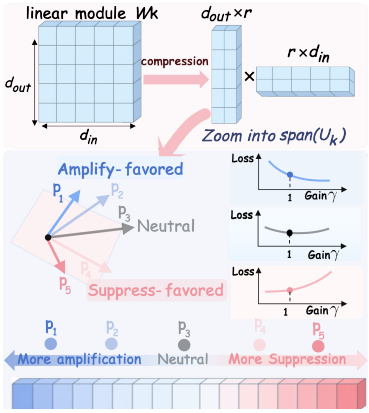

Figure 1: Motivation of LENS.

In this work, we propose a simple yet effective approach termed Loss-aligned Directional Salience Calibration (LENS) for post-compression calibration of low-rank compressed LLMs. Given a compressed model, LENS determines which retained directions should be amplified or suppressed and how strongly each direction should be adjusted. Specifically, LENS first profiles the first-order loss response along each retained direction to infer whether amplifying or suppressing that direction is locally favorable to the language-modeling objective. To improve the reliability of these directional decisions, we evaluate them on a disjoint fitting subset and retain only directions whose loss responses exhibit confidence-consistent adjustment signs. While first-order responses determine the adjustment signs, a second-order surrogate estimates response-adaptive relative gains, whose overall strength is further calibrated by stage-wise global salience scaling. Extensive experiments show that LENS consistently improves compressed LLMs across compression budgets and compression methods, supporting the robustness and broad applicability of the proposed calibration strategy. The main contributions of this work are summarized as follows:

❶ ***New Perspective.*** Beyond preserving a compact subspace for LLM compression, we reveal that retaining representational directions alone does not ensure that their contributions are properly calibrated to the language-modeling objective. This motivates a complementary perspective that considers not only which directions are retained, but also which directions warrant amplification or suppression and how strongly their contributions should be calibrated.

❷ ***Novel Framework with Local Theoretical Guarantee.*** To calibrate the retained subspace, we first identify reliable loss-aligned directions through first-order loss-response profiling, and then determine their strengths with direction-specific gains. Theoretically, we establish first-order sufficiency of the response eigenbasis for the considered retained-subspace perturbations and show that the resulting gains induce a local descent direction while preserving the compressed structure.

❸ ***Empirical Validation.*** Extensive experiments across LLM backbones and compression budgets demonstrate the effectiveness of LENS over baselines, with further results supporting its robustness to calibration data and compatibility with different low-rank compressors and quantization.

2 Related Work

Low-rank Compression for LLMs. Existing methods primarily improve reconstruction fidelity through activation-aware or task-aware decomposition (Hsu et al., 2022; Yuan et al., 2025), or adapt compression structures and budgets across model components (Wang et al., 2025b; Huang et al., 2025; Xv et al., 2026). Despite their different formulations, these methods largely focus on constructing a more faithful or better allocated low-rank representation during compression, rather than recalibrating how individual directions contribute once the retained subspace has been fixed. Complementing these efforts, LENS recalibrates directional contributions within the retained subspaces of an already compressed model without altering its compression structure or rank allocation.

Post-compression Calibration. Existing approaches refine retained low-rank factors or compensate for compression residuals (Wang et al., 2025c; Li et al., 2025; Liu et al., 2024; Jaiswal et al., 2024). More closely related to our work, SVS (Kim & Yoo, 2025) rescales retained singular values

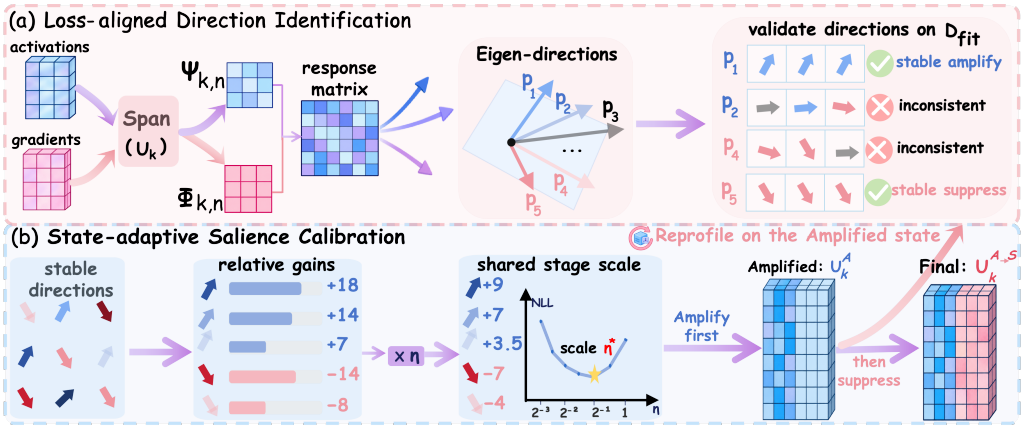

Figure 2: Overview of LENS. LENS identifies reliable loss-aligned directions and sequentially calibrates their amplification and suppression strengths within the retained low-rank subspace.

with fixed singular vectors, while SVD-Surgeon (Safari & Hutter, 2026) adjusts retained singular values using second-order loss information. Whereas these spectral methods recalibrate coefficients along predetermined singular directions, LENS identifies loss-aligned directions within retained output subspaces and then calibrates their amplification or suppression strengths.

3 The Proposed Framework: LENS

Problem Formulation. Given a calibration dataset $\mathcal{D}_{\text{cal}} = \{\mathbf{s}_n\}_{n=1}^{N}$, let $\mathbf{M}_{\text{dense}}$ be a pretrained dense language model and $\mathbf{M}_0 = \mathcal{C}(\mathbf{M}_{\text{dense}})$ its compressed counterpart. For module k, let $\mathbf{W}_k \in \mathbb{R}^{d_k^{\text{out}} \times d_k^{\text{in}}}$ denote its dense weight and $\mathbf{X}_{k,n} \in \mathbb{R}^{d_k^{\text{in}} \times L_n}$ the input activation of sequence $\mathbf{s}_n$. The rank-r_k compressed operator is $\widehat{\mathbf{W}}_k^{(0)} = \mathbf{U}_k\mathbf{V}_k$, where $\mathbf{U}_k \in \mathbb{R}^{d_k^{\text{out}} \times r_k}$ and $\mathbf{V}_k \in \mathbb{R}^{r_k \times d_k^{\text{in}}}$. Such low-rank factors are commonly obtained by minimizing an activation reconstruction objective (Wang et al., 2025c; Yuan et al., 2025; Xv et al., 2026):

$$\min_{\mathbf{U}_k,\mathbf{V}_k} \sum_{n=1}^{N} \|\mathbf{W}_k\mathbf{X}_{k,n} - \mathbf{U}_k\mathbf{V}_k\mathbf{X}_{k,n}\|_F^2 . \tag{1}$$

The compressed output is $\mathbf{H}_{k,n} = \mathbf{U}_k\mathbf{V}_k\mathbf{X}_{k,n}$. We keep the compression structure and ranks $\{r_k\}$ fixed throughout calibration. Compression determines the retained representation and its current directional contributions, but does not guarantee that these contributions are well calibrated to the end-to-end language-modeling objective. Consequently, the objective of LENS is to recalibrate these contributions without altering the compression structure. For hybrid compressors such as Duo-SVD, LENS calibrates only the low-rank branch, while explicitly preserved columns remain fixed.

3.1 Framework Overview

In this work, we introduce a post-compression calibration framework for low-rank LLM compression, with the central idea of aligning directional adjustments in the retained space with their effects on the language-modeling objective. In particular, we first perform first-order loss-response profiling to estimate the signed effect of scaling each retained direction on the language-modeling loss. Subsequently, we validate these directional preferences on a calibration set, retaining only adjustments that exhibit consistent loss responses. This process yields a signed calibration rule that distinguishes directions benefiting from amplification from those better suppressed. To determine how much each selected direction should be adjusted, we further estimate a response-adaptive relative gain using a second-order surrogate, and modulate the resulting update by reliability-gated global salience. An overview of the calibration framework is illustrated in Figure 2.

3.2 Loss-aligned Direction Identification

Low-rank decomposition determines which directions are retained, but not whether the contribution of each retained direction should be amplified or suppressed. We therefore characterize the

retained subspace through its first-order loss responses to local rescaling (Liu et al., 2025). These responses define a symmetric response matrix whose eigendirections reveal modes that locally favor amplification or suppression. We further use a separate fitting subset to retain only directions whose adjustment tendencies remain consistent beyond the discovery data.

Retained-subspace Response Profiling. For the k-th module, recall that the compressed low-rank output is $\mathbf{H}_{k,n} = \mathbf{U}_k\mathbf{V}_k\mathbf{X}_{k,n}$. For a unit direction $\mathbf{p} \in \mathrm{span}(\mathbf{U}_k)$ with $\|\mathbf{p}\|_2 = 1$, we introduce the rank-one rescaling perturbation $\mathbf{H}_{k,n}(\delta) = (\mathbf{I} + \delta\mathbf{p}\mathbf{p}^\top)\mathbf{H}_{k,n}$, where $\mathbf{I}$ denotes the identity matrix and δ controls the gain along $\mathbf{p}$. Let $\ell_n(\delta; \mathbf{p})$ be the mean negative log-likelihood (NLL) of $\mathbf{s}_n$ under a rank-one rescaling perturbation of strength δ along $\mathbf{p}$. Therefore, the directional loss response is:

$$q_{k,n}(\mathbf{p}) = \left.\frac{\partial \ell_n(\delta; \mathbf{p})}{\partial \delta}\right|_{\delta=0} = \mathbf{p}^\top \mathbf{D}_{k,n}\mathbf{H}_{k,n}^\top \mathbf{p}, \tag{2}$$

where $\mathbf{D}_{k,n} = \frac{\partial \ell_n}{\partial \mathbf{H}_{k,n}}$ denotes the loss gradient with respect to the low-rank branch output. Thus, $q_{k,n}(\mathbf{p}) < 0$ favors amplification, whereas $q_{k,n}(\mathbf{p}) > 0$ favors suppression. Assuming $\mathrm{rank}(\mathbf{U}_k) = r_k$, we compute the reduced QR decomposition $\mathbf{U}_k = \mathbf{Q}_k\mathbf{R}_k$, where $\mathbf{Q}_k \in \mathbb{R}^{d_k^{\mathrm{out}} \times r_k}$ has orthonormal columns that form a basis for $\mathrm{span}(\mathbf{U}_k)$. This parameterization reduces the search for directions from the full d_k^{out}-dimensional output space to the r_k-dimensional retained coordinate space. Any unit direction $\mathbf{p} \in \mathrm{span}(\mathbf{U}_k)$ can be uniquely parameterized as $\mathbf{p} = \mathbf{Q}_k\mathbf{v}$, where $\|\mathbf{v}\|_2 = 1$. Defining $q_{k,n}(\mathbf{v}) := q_{k,n}(\mathbf{Q}_k\mathbf{v})$, Eq. 2 gives

$$q_{k,n}(\mathbf{v}) = \mathbf{v}^\top \mathbf{\Phi}_{k,n}\mathbf{\Psi}_{k,n}^\top \mathbf{v} = \mathbf{v}^\top \frac{\mathbf{\Phi}_{k,n}\mathbf{\Psi}_{k,n}^\top + \mathbf{\Psi}_{k,n}\mathbf{\Phi}_{k,n}^\top}{2}\mathbf{v}. \tag{3}$$

The projected gradient and activation terms in Eq. 3 are defined as

$$\mathbf{\Phi}_{k,n} = \mathbf{Q}_k^\top \mathbf{D}_{k,n}, \qquad \mathbf{\Psi}_{k,n} = \mathbf{Q}_k^\top \mathbf{H}_{k,n} = \mathbf{R}_k\mathbf{V}_k\mathbf{X}_{k,n}. \tag{4}$$

The second equality in Eq. 3 holds since a quadratic form depends only on the symmetric part of its matrix. To identify directions that characterize the average first-order loss response across instances, we construct an aggregated response matrix over a discovery subset $\mathcal{D}_{\mathrm{disc}} \subset \mathcal{D}_{\mathrm{cal}}$:

$$\mathbf{S}_k = \frac{1}{|\mathcal{D}_{\mathrm{disc}}|} \sum_{n \in \mathcal{D}_{\mathrm{disc}}} \frac{\mathbf{\Phi}_{k,n}\mathbf{\Psi}_{k,n}^\top + \mathbf{\Psi}_{k,n}\mathbf{\Phi}_{k,n}^\top}{2}, \tag{5}$$

which is the symmetrized activation–gradient cross-moment in the retained coordinate space. Since $\mathbf{S}_k$ is symmetric, we compute its eigendecomposition:

$$\mathbf{S}_k = \mathbf{E}_k\mathbf{\Lambda}_k\mathbf{E}_k^\top, \quad \text{where} \quad \mathbf{E}_k = [\mathbf{v}_{k,1}, \ldots, \mathbf{v}_{k,r_k}], \quad \mathbf{\Lambda}_k = \mathrm{diag}(\lambda_{k,1}, \ldots, \lambda_{k,r_k}). \tag{6}$$

In particular, for each eigendirection $\mathbf{v}_{k,j}$, the average response over $\mathcal{D}_{\mathrm{disc}}$ is:

$$\bar{q}_{k,\mathrm{disc}}(\mathbf{v}_{k,j}) = \frac{1}{|\mathcal{D}_{\mathrm{disc}}|} \sum_{n \in \mathcal{D}_{\mathrm{disc}}} q_{k,n}(\mathbf{v}_{k,j}) = \mathbf{v}_{k,j}^\top \mathbf{S}_k \mathbf{v}_{k,j} = \lambda_{k,j}. \tag{7}$$

Each eigenvector $\mathbf{v}_{k,j}$ therefore maps back to a corresponding direction in the output space:

$$\mathbf{p}_{k,j} = \mathbf{Q}_k\mathbf{v}_{k,j}, \qquad \|\mathbf{p}_{k,j}\|_2 = 1. \tag{8}$$

Thus, the eigenvalue $\lambda_{k,j}$ directly quantifies the average first-order response of the loss to rescaling the corresponding output-space direction $\mathbf{p}_{k,j}$ over $\mathcal{D}_{\mathrm{disc}}$. A negative eigenvalue $\lambda_{k,j} < 0$ indicates that an infinitesimal increase in the gain along $\mathbf{p}_{k,j}$ decreases the loss to first order, making $\mathbf{p}_{k,j}$ a candidate for amplification. Conversely, a positive eigenvalue $\lambda_{k,j} > 0$ indicates that an infinitesimal decrease in its gain decreases the loss to first order, making $\mathbf{p}_{k,j}$ a candidate for suppression.

Reliability-guided Mode Selection. Because the candidate modes are identified from a finite discovery subset $\mathcal{D}_{\mathrm{disc}}$, their estimated response signs may be sensitive to sampling noise. To retain only modes whose amplification or suppression tendencies remain consistent on held-out data, we employ a disjoint fitting subset $\mathcal{D}_{\mathrm{fit}} \subset \mathcal{D}_{\mathrm{cal}}$, with $\mathcal{D}_{\mathrm{fit}} \cap \mathcal{D}_{\mathrm{disc}} = \varnothing$, to validate the candidate modes. For each candidate $\mathbf{v}_{k,j}$, we compute its mean first-order response over the fitting subset as:

$$\bar{q}_{k,\mathrm{fit}}(\mathbf{v}_{k,j}) = \frac{1}{|\mathcal{D}_{\mathrm{fit}}|} \sum_{n \in \mathcal{D}_{\mathrm{fit}}} q_{k,n}(\mathbf{v}_{k,j}). \tag{9}$$

We also estimate the sample standard deviation of $q_{k,n}(\mathbf{v}_{k,j})$ across $n \in \mathcal{D}_{\text{fit}}$, denoted by $\sigma_{k,j}$. Under a normal approximation to the sampling distribution of the mean directional response, we construct the following approximate $(1-\alpha)$ confidence interval:

$$\left[L_{k,j}^{1-\alpha}, U_{k,j}^{1-\alpha}\right] = \left[\bar{q}_{k,\text{fit}}(\mathbf{v}_{k,j}) - z_{1-\alpha/2}\frac{\sigma_{k,j}}{\sqrt{|\mathcal{D}_{\text{fit}}|}},\ \bar{q}_{k,\text{fit}}(\mathbf{v}_{k,j}) + z_{1-\alpha/2}\frac{\sigma_{k,j}}{\sqrt{|\mathcal{D}_{\text{fit}}|}}\right], \tag{10}$$

where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of the standard normal distribution. Furthermore, we set $\alpha = 0.05$ throughout our experiments. We retain an amplify mode if $\lambda_{k,j} < 0$ and $U_{k,j}^{1-\alpha} < 0$, and a suppress mode if $\lambda_{k,j} > 0$ and $L_{k,j}^{1-\alpha} > 0$.

3.3 State-adaptive Salience Calibration

Loss-aligned direction identification determines whether a direction favors amplification or suppression, but not its rescaling strength. Prior post-compression refinement methods have explored rescaling retained spectral components along predetermined singular directions (Kim & Yoo, 2025; Safari & Hutter, 2026). In contrast, we estimate response-adaptive relative gains over the loss-aligned directions and calibrate their joint magnitude using reliability-gated global salience scaling.

Response-adaptive Relative Gain Estimation. For each validated direction $\mathbf{p}_{k,j} = \mathbf{Q}_k\mathbf{v}_{k,j}$, let $\bar{\ell}_{\text{fit}}(\delta; \mathbf{p}_{k,j}) = \sum_{n\in\mathcal{D}_{\text{fit}}} \ell_n(\delta; \mathbf{p}_{k,j})/|\mathcal{D}_{\text{fit}}|$ denote the mean negative log-likelihood over $\mathcal{D}_{\text{fit}}$ under the rank-one rescaling perturbation with signed rescaling coefficient δ introduced in Section 3.2. Assuming sufficient second-order smoothness in δ, a Taylor expansion around $\delta = 0$ gives

$$\Delta_\delta\bar{\ell}_{\text{fit}}(\mathbf{p}_{k,j}) = \bar{\ell}_{\text{fit}}(\delta; \mathbf{p}_{k,j}) - \bar{\ell}_{\text{fit}}(0; \mathbf{p}_{k,j}) = \bar{q}_{k,\text{fit}}(\mathbf{v}_{k,j})\delta + \frac{1}{2|\mathcal{D}_{\text{fit}}|}\sum_{n\in\mathcal{D}_{\text{fit}}} \frac{\partial^2\ell_n(\delta; \mathbf{p}_{k,j})}{\partial\delta^2}\bigg|_{\delta=0}\delta^2 + o(\delta^2). \tag{11}$$

Rather than explicitly estimating the second-order directional curvature, we construct a nonnegative surrogate from the empirical second moment of the first-order directional responses. Specifically, we define the following empirical second-moment surrogate:

$$f_{k,j} = \frac{1}{|\mathcal{D}_{\text{fit}}|}\sum_{n\in\mathcal{D}_{\text{fit}}} \left[q_{k,n}(\mathbf{v}_{k,j})\right]^2. \tag{12}$$

We interpret $f_{k,j}$ as a response-energy normalizer rather than an estimator of the true directional curvature. Using this response-based surrogate in a regularized quadratic objective, we define the response-adaptive relative gain as its minimizer:

$$\widetilde{\delta}_{k,j} = \arg\min_{\delta\in\mathbb{R}}\left[\bar{q}_{k,\text{fit}}(\mathbf{v}_{k,j})\delta + \frac{1}{2}(f_{k,j}+\epsilon)\delta^2\right] = -\frac{\bar{q}_{k,\text{fit}}(\mathbf{v}_{k,j})}{f_{k,j}+\epsilon}. \tag{13}$$

Here $\epsilon > 0$ is a stabilizer that bounds the magnitude of the relative gains. The coefficients $\widetilde{\delta}_{k,j}$ specify the relative rescaling strengths across these validated directions.

Reliability-gated Global Salience Scaling. The relative gains in Eq. 13 are estimated from mode-wise local surrogates and therefore do not explicitly capture cross-direction or cross-module interactions under joint rescaling. We therefore introduce a stage-level global scale η shared across validated directions to control the joint rescaling magnitude. Let $\mathcal{J}_k$ denote the indices of the validated modes in module k for the current calibration stage. Given η, the joint rescaling transformation is

$$\mathbf{G}_k(\eta) = \mathbf{I} + \eta\sum_{j\in\mathcal{J}_k}\widetilde{\delta}_{k,j}\mathbf{p}_{k,j}\mathbf{p}_{k,j}^\top. \tag{14}$$

Accordingly, we rescale $\mathbf{U}_k$ by $\mathbf{G}_k(\eta)$ while keeping $\mathbf{V}_k$ fixed, with gain $\gamma_{k,j} = 1+\eta\widetilde{\delta}_{k,j}$. To select the stage-level scale, we use a scale-selection subset $\mathcal{D}_{\text{scale}} \subset \mathcal{D}_{\text{cal}}$ disjoint from the discovery and fitting subsets, together with a predefined geometric grid $\mathcal{E}$. Let $\mathcal{E}^+ \subseteq \mathcal{E}$ denote the feasible scales satisfying $1+\eta\widetilde{\delta}_{k,j} > 0$ for all validated directions. Let $\mathbf{M}_{\text{ref}}$ denote the reference model of the current stage and $\mathbf{M}_\eta$ the candidate model under scale η. We select:

$$\widehat{\eta} = \arg\min_{\eta\in\mathcal{E}^+}\frac{1}{|\mathcal{D}_{\text{scale}}|}\sum_{n\in\mathcal{D}_{\text{scale}}}\left[\ell_n(\mathbf{M}_\eta) - \ell_n(\mathbf{M}_{\text{ref}})\right]. \tag{15}$$

The selected scale is further evaluated on a disjoint held-out subset $\mathcal{D}_{\text{val}}$ and accepted only if it yields a negative mean NLL change on $\mathcal{D}_{\text{scale}}$ and a negative upper confidence bound on $\mathcal{D}_{\text{val}}$; otherwise, $\eta^\star = 0$. Full scale-search and validation details are provided in Appendix B.2.

Finally, the accepted rescaling is folded into the output factor, $\mathbf{U}_k^\star = \mathbf{G}_k(\eta^\star)\mathbf{U}_k$, yielding

$$\widetilde{\mathbf{W}}_k = \mathbf{U}_k^\star \mathbf{V}_k = \mathbf{G}_k(\eta^\star)\mathbf{U}_k\mathbf{V}_k. \tag{16}$$

With positive effective gains, the transformation preserves the retained span and rank, while its offline folding into $\mathbf{U}_k^\star$ introduces no additional inference-time parameters or operations.

3.4 Sequential Calibration

We integrate the above components into a sequential *amplify-then-suppress* calibration procedure. At each stage, **LENS** starts from the current reference model and forms a calibrated directional update from the reliable loss-aligned responses. The relative gains determine the direction-wise strengths, while held-out calibration data selects a shared stage-level scale for the overall update.

Starting from the initial compressed model M_0, we first calibrate the validated amplification modes to obtain an intermediate model M_A. Crucially, M_A then becomes the reference state for the second stage: the directional responses are reprofiled on the updated model before the suppression modes and their calibration strengths are determined. At the factor level, the sequential update is given by

$$\mathbf{U}_k^A = \mathbf{G}_k^A(\eta_A^\star)\mathbf{U}_k^{(0)}, \qquad \mathbf{U}_k^{A\to S} = \mathbf{G}_k^S(\eta_S^\star)\mathbf{U}_k^A, \tag{17}$$

where $\mathbf{G}_k^A$ and $\mathbf{G}_k^S$ denote the stage-specific instances of the directional rescaling in Eq. 14, with $\mathbf{G}_k^S$ constructed after reprofiling the updated model M_A. The final model $M_{A\to S}$ therefore reflects the loss geometry after the primary amplify correction, rather than applying a suppress update predetermined from the original compressed model M_0.

3.5 Theoretical Analysis

We analyze two questions underlying **LENS**: whether the full response eigenbasis is sufficient for first-order search over symmetric retained-coordinate perturbations, and whether the jointly applied relative gains define a local descent direction for the empirical NLL. We further characterize gain feasibility and explain why the second calibration stage is reprofiled at the updated model state. Full proofs are provided in Appendix A.

Theorem 3.1. *Fix a calibration stage with reference model $\mathbf{M}_{\text{ref}}$, and assume that $\bar{\ell}_{\text{disc}}$ is jointly differentiable with respect to the calibrated factors $\{\mathbf{U}_k\}_k$ at the current reference model. For each calibrated module k, let $\mathbf{A}_k = \mathbf{A}_k^\top \in \mathbb{R}^{r_k\times r_k}$ and consider*

$$\mathbf{U}_k(t) = \left(\mathbf{I} + t\mathbf{Q}_k\mathbf{A}_k\mathbf{Q}_k^\top\right)\mathbf{U}_k, \tag{18}$$

with all remaining parameters fixed. Let $\mathbf{M}(t)$ denote the resulting model. Then

$$\frac{d}{dt}\bar{\ell}_{\text{disc}}(\mathbf{M}(t))\Big|_{t=0} = \sum_k \langle \mathbf{S}_k, \mathbf{A}_k\rangle_F . \tag{19}$$

For $\mathbf{S}_k = \mathbf{E}_k\mathbf{\Lambda}_k\mathbf{E}_k^\top$, let $\Pi_{\mathbf{E}_k}(\mathbf{A}_k) = \sum_{j=1}^{r_k}(\mathbf{v}_{k,j}^\top\mathbf{A}_k\mathbf{v}_{k,j})\mathbf{v}_{k,j}\mathbf{v}_{k,j}^\top$. Then

$$\langle \mathbf{S}_k, \Pi_{\mathbf{E}_k}(\mathbf{A}_k)\rangle_F = \langle \mathbf{S}_k, \mathbf{A}_k\rangle_F , \qquad \|\Pi_{\mathbf{E}_k}(\mathbf{A}_k)\|_F \le \|\mathbf{A}_k\|_F . \tag{20}$$

Remark 3.1. *The guarantee applies to the full response eigenbasis. Reliability filtering is a subsequent restriction that retains only directions whose response signs are supported by held-out data.*

Theorem 3.2. *At the current calibration stage, let $\mathcal{I} = \{(k,j) : j \in \mathcal{J}_k\} \neq \varnothing$ denote the validated module–direction pairs. For each $i = (k,j) \in \mathcal{I}$, define*

$$g_i := \bar{q}_{k,\text{fit}}(\mathbf{v}_{k,j}), \qquad f_i := f_{k,j}, \qquad \lambda_i := \lambda_{k,j}, \qquad d_i := -\frac{g_i}{f_i+\epsilon} = \widetilde{\delta}_{k,j}.$$

Assume that the empirical NLLs on $\mathcal{D}_{\text{disc}}$ and $\mathcal{D}_{\text{fit}}$ are jointly differentiable with respect to the calibrated factors $\{\mathbf{U}_k\}_k$ at the current reference model. Under the validation conditions used for mode selection, $\lambda_i g_i > 0$ for every $i \in \mathcal{I}$. Along the joint calibration path in Eq. 14, we have

$$\frac{d}{d\eta}\bar{\ell}_{\text{fit}}(\mathbf{M}_\eta)\Big|_{\eta=0} = -\sum_{i\in\mathcal{I}}\frac{g_i^2}{f_i+\epsilon} < 0, \qquad \frac{d}{d\eta}\bar{\ell}_{\text{disc}}(\mathbf{M}_\eta)\Big|_{\eta=0} = -\sum_{i\in\mathcal{I}}\frac{\lambda_i g_i}{f_i+\epsilon} < 0. \tag{21}$$

Consequently, there exists $\eta_0 > 0$ such that every $0 < \eta < \eta_0$ strictly decreases both empirical NLLs while satisfying $1 + \eta d_i > 0$ for all $i \in \mathcal{I}$.

Table 1: Comparison of compression methods across model families. Best and second-best results are shown in **bold** and underlined, respectively. Results on Mistral-7B are in Appendix Table 3.

Model	Method	MMLU	PIQA	WinoG.	HellaS.	ARC-e	ARC-c	OBQA	Avg.↑	PPL↓
LLaMA-2 7B	Dense	45.93	79.11	68.90	76.02	74.62	46.42	44.20	62.17	5.12
	LLM-Pruner	30.49	76.55	62.27	69.34	65.57	39.76	41.60	55.08	7.68
	FLAP	31.60	74.65	65.35	65.04	59.39	34.22	38.20	52.63	6.63
	SliceGPT	30.95	68.28	64.40	58.59	62.33	36.09	37.60	51.18	6.33
	SVD-LLM	29.04	75.19	64.48	68.56	65.36	**40.70**	40.00	54.76	7.05
	BoLaCo	32.04	73.29	66.06	61.27	67.80	37.46	38.20	53.73	7.16
	SoLA	31.40	72.63	65.19	59.86	63.30	35.49	37.00	52.13	6.82
	Duo-SVD	32.09	76.55	67.56	69.97	69.32	39.08	41.60	56.59	5.95
	LENS	**34.73**	**76.71**	**67.88**	**72.68**	**71.17**	40.53	**42.80**	**58.07**	**5.61**
Qwen3-8B	Dense	76.56	79.38	72.69	78.64	80.13	57.00	42.00	69.49	6.51
	LLM-Pruner	57.42	76.88	66.77	71.59	75.80	47.78	39.40	62.23	9.89
	FLAP	62.40	72.36	68.67	63.14	68.31	44.03	41.20	60.02	8.14
	SliceGPT	42.32	72.20	69.53	64.66	59.22	40.10	37.80	55.12	7.76
	SVD-LLM	59.22	74.76	70.09	69.76	73.48	51.19	42.20	62.96	9.09
	BoLaCo	53.48	75.30	66.61	64.49	75.21	47.01	38.20	60.04	8.33
	SoLA	49.68	69.75	66.61	53.37	63.76	36.60	37.60	53.91	8.99
	Duo-SVD	63.87	**78.62**	**70.80**	71.54	81.78	56.23	**42.80**	66.52	7.54
	LENS	**64.26**	78.51	70.56	**71.77**	**81.94**	**56.57**	**42.80**	**66.63**	**7.34**
LLaMA-3.1 8B	Dense	65.35	81.28	73.95	78.91	81.14	53.41	45.00	68.44	5.84
	LLM-Pruner	35.72	**78.78**	64.01	**71.22**	70.37	45.14	39.00	57.75	9.81
	FLAP	37.17	72.63	64.56	58.68	57.07	32.94	38.40	51.64	8.23
	SliceGPT	37.23	66.43	67.88	56.69	59.72	35.41	38.80	51.74	8.33
	SVD-LLM	39.00	72.47	63.06	65.38	64.35	39.76	35.40	54.20	10.03
	BoLaCo	32.44	72.14	66.46	59.50	66.50	38.99	39.00	53.58	10.31
	SoLA	40.81	72.14	67.72	58.39	55.56	34.56	39.00	52.60	9.48
	Duo-SVD	52.49	75.41	69.61	68.13	71.04	44.20	41.20	60.30	8.68
	LENS	**54.78**	76.33	**69.85**	70.02	**73.02**	**46.42**	**42.60**	**61.86**	**7.95**

Remark 3.2. *The guarantee is local to the current model state, motivating reprofiling after amplify before constructing the subsequent suppress update.*

Corollary 3.1. *For every $i \in \mathcal{I}$,*

$$|d_i| \le \frac{1}{2\sqrt{\epsilon}}, \qquad 0 \le \eta < 2\sqrt{\epsilon} \implies 1 + \eta d_i > 0. \tag{22}$$

Hence, $\mathbf{G}_k(\eta)$ is positive definite and invertible throughout this interval. If the current low-rank operator $\widehat{\mathbf{W}}_k = \mathbf{U}_k\mathbf{V}_k$ satisfies $\operatorname{rank}(\widehat{\mathbf{W}}_k) = r_k$, then $\widehat{\mathbf{W}}_k(\eta) = \mathbf{G}_k(\eta)\widehat{\mathbf{W}}_k$ preserves the rank, column space, and row space of $\widehat{\mathbf{W}}_k$.

These guarantees ensure feasible positive gains and structural preservation, but do not imply loss decrease at an arbitrary finite scale. LENS retains scale selection and validation for finite updates.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation. Following prior work on low-rank LLM compression (Xv et al., 2026), we report perplexity (PPL) on WikiText-2 (Merity et al., 2016) and evaluate seven downstream benchmarks. We evaluate MMLU (Hendrycks et al., 2020) in a 5-shot setting and report zero-shot accuracy on PIQA (Bisk et al., 2019), WinoGrande (Sakaguchi et al., 2021), HellaSwag (Zellers et al., 2019), ARC-Easy, ARC-Challenge (Clark et al., 2018), and OpenBookQA (Mihaylov et al., 2018). We additionally report the average accuracy across the downstream benchmarks as Avg.

Baselines and Backbones. We compare LENS with representative LLM compression approaches from three categories: (i) structured pruning methods, including LLM-Pruner (Ma et al., 2023) and FLAP (An et al., 2024); (ii) low-rank decomposition methods, including SVD-LLM (Wang et al., 2025c), BoLaCo (Ji et al., 2024), SoLA (Huang et al., 2025), and Duo-SVD (Xv et al., 2026); and (iii) structured dimensionality-reduction methods, represented by SliceGPT (Ashkboos et al., 2024). We conduct experiments on four backbones: LLaMA-2-7B (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), LLaMA-3.1-8B (Grattafiori et al., 2024), and Qwen3-8B (Yang et al., 2025).

Implementation Details. Unless otherwise specified, all methods are evaluated at a 20% compression ratio. We use WikiText-2 for post-compression calibration with 128/64/64 training windows for

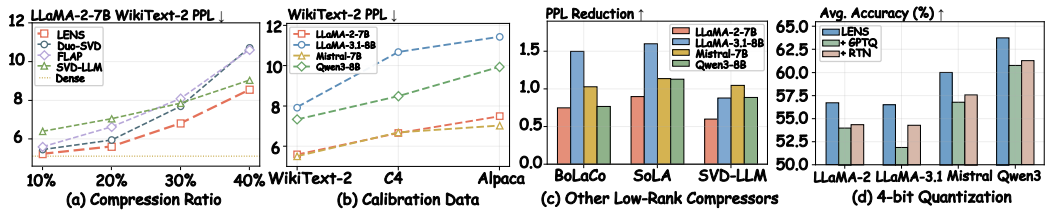

Figure 3: Generalization and robustness of LENS. (a) WikiText-2 PPL under 10–40% compression. (b) WikiText-2 PPL with different calibration datasets. (c) PPL reduction when applying LENS to other compressors. (d) Average accuracy after subsequent 4-bit GPTQ or RTN quantization.

direction discovery, response fitting, and stage-wise scale selection, respectively, and 64 validation windows for final acceptance. See Appendix B for full implementation details.

4.2 Performance Comparison

Table 1 compares LENS with seven representative LLM compression methods at a unified 20% compression ratio. As depicted in the table, we could derive three insights: ***Obs.❶ LENS consistently achieves the strongest overall performance across different model families.*** Across the three backbones, LENS obtains the lowest PPL and the highest average downstream accuracy among all compressed models, while ranking first in 16 of the 21 backbone–task combinations (including ties) and second in the remaining 5. ***Obs.❷ A compact low-rank representation alone is not sufficient to recover post-compression model quality.*** LENS consistently improves over SVD-LLM, BoLaCo, SoLA, and Duo-SVD in both PPL and average downstream accuracy, supporting the need to calibrate retained directions against the end-to-end objective. ***Obs.❸ The advantage of LENS extends beyond low-rank compression methods.*** Compared with LLM-Pruner, FLAP, and SliceGPT, LENS achieves lower PPL and higher average downstream accuracy across all three backbones, showing that its benefits are not confined to comparisons within the low-rank family.

4.3 Ablation Studies

We conduct ablation studies to examine the roles of amplify and suppress, together with the effects of their ordering and state-adaptive reprofiling. Specifically, we consider four variants of LENS: V1 removes suppress, V2 removes amplify, V3 profiles both stages on the original compressed model and applies their corrections jointly, and V4 reverses the default amplify-then-suppress order. From the results, we can find that: ***Obs. ❹ Amplify is the dominant stage, while suppress provides complementary refinement.*** V1 consistently outperforms V2, while the full LENS further improves over V1, showing that amplify contributes the larger individual effect and that suppress provides additional refinement when applied afterward. ***Obs. ❺ Both stage ordering and state-adaptive reprofiling matter for effective calibration.*** The full LENS consistently outperforms V3, supporting the benefit of reprofiling the model state between the two stages, while its advantage over V4 supports the default amplify-before-suppress order.

Table 2: Ablation study of LENS on three backbones.

Variant	LLaMA-2-7B		Qwen3-8B		LLaMA-3.1-8B	
	PPL↓	Avg.↑	PPL↓	Avg.↑	PPL↓	Avg.↑
V1	5.76	57.87	7.40	66.34	8.00	61.63
V2	5.83	56.67	7.52	66.45	8.58	60.30
V3	5.85	57.89	7.36	65.94	8.00	61.65
V4	5.67	57.68	7.36	66.53	8.03	61.73
LENS	**5.61**	**58.07**	**7.34**	**66.63**	**7.95**	**61.86**

4.4 Further Analysis

Robustness to Compression Ratio and Calibration Dataset. We vary the compression ratio from 10% to 40%, and replace the default WikiText-2 calibration dataset with C4 or Alpaca, evaluating WikiText-2 PPL, C4 PPL, and downstream tasks. Complete results are provided in Appendix C.2. We note that: ***Obs.❻ LENS remains robust across compression budgets and calibration datasets.*** As shown in Figure 3(a)(b), LENS consistently achieves the lowest WikiText-2 PPL across 10–40% compression, while WikiText-2 calibration yields the lowest WikiText-2 PPL across all backbones.

Compatibility with Other Compression Methods. We apply LENS to three additional low-rank compressors, BoLaCo, SoLA, and SVD-LLM, and further evaluate compatibility with subsequent 4-bit GPTQ (Frantar et al., 2022) and RTN quantization. We can find that: ***Obs.❼ LENS generalizes across different low-rank compressors and remains compatible with low-bit quantization.*** As

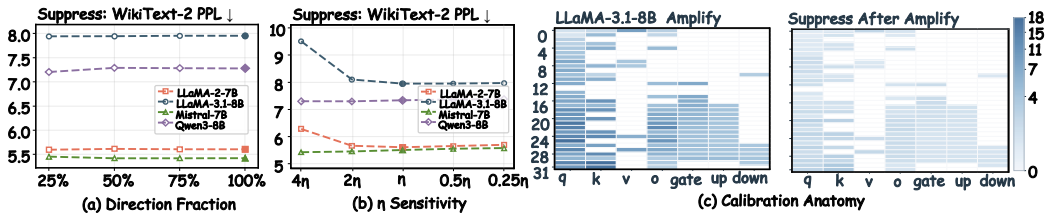

Figure 4: Further analysis of LENS. (a) Direction-fraction sensitivity. (b) Stage-wise scale sensitivity. (c) Calibration anatomy on LLaMA-3.1-8B, where darker cells indicate larger changes.

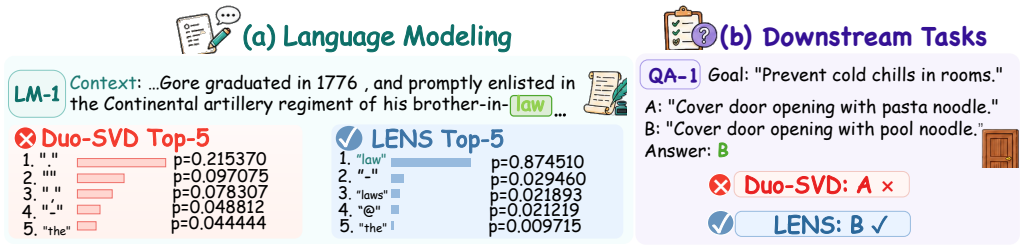

Figure 5: Qualitative examples of prediction recovery with LENS on two tasks.

shown in Figure 3(c), LENS yields consistent PPL reductions across all three low-rank compressors and four backbones. Moreover, models calibrated with LENS retain stable downstream accuracy after subsequent 4-bit GPTQ or RTN quantization, as shown in Figure 3(d).

Sensitivity to Calibration Choices. We examine the sensitivity of LENS to two key calibration choices. For direction selection, we vary the retained fraction of validated directions over $\{25\%, 50\%, 75\%, 100\%\}$; for scale sensitivity, letting η denote the selected stage-wise scale, we evaluate $\{4\eta, 2\eta, \eta, 0.5\eta, 0.25\eta\}$. Figure 4(a,b) reports representative results for the suppress stage, while complete results are provided in Appendix C.3. From the results, we note that: ***Obs.❽ LENS is stable to its key calibration choices.*** Varying the retained direction fraction results in only relatively minor PPL variations across all four backbones. Performance also remains relatively stable around the selected η, whereas overly large scales can noticeably degrade some models. These results suggest that LENS does not rely on a narrowly tuned calibration setting.

Calibration Anatomy. We examine how amplify and the subsequent suppress stage modify different layers and decoder-linear modules by measuring the relative effective-weight change of each layer–module pair. Figure 4(c) visualizes these changes on LLaMA-3.1-8B, with complete results in Appendix C.4. We can see that: ***Obs.❾ LENS induces structured and stage-specific model corrections.*** The calibration strength varies across both layers and module families rather than being uniformly distributed across the network. amplify produces consistently stronger modifications, while the subsequent suppress stage introduces smaller and more selective adjustments, consistent with the ablation results. Together, these patterns suggest that the two stages play distinct but complementary roles rather than applying uniform corrections across layers and modules.

Qualitative Examples. Figure 5 presents examples of prediction recovery after calibration. We find that: ***Obs. ❿ LENS can recover correct predictions at the model-output level.*** In language modeling, Duo-SVD misses the correct continuation "law", whereas LENS restores it as the top-1 prediction. In the QA example, LENS recovers the correct answer. These examples illustrate that the effect of directional recalibration extends beyond aggregate metrics to concrete prediction recovery.

5 Conclusion

This paper studies how retained directions in low-rank LLM compression can be recalibrated after compression to better align with the language-modeling objective. We proposed LENS, a post-compression framework that performs sequential amplify-then-suppress calibration within the retained subspace while preserving the original compression structure without additional inference-time cost. Our theoretical analysis provides first-order support for this calibration principle, while experiments show consistent benefits across different compression conditions and compressors. Overall, these results suggest that recalibrating what survives compression can complement methods that primarily focus on determining what should be retained.

AI USE STATEMENT

Generative AI tools were used only to assist with language polishing, grammatical refinement, and the retrieval and organization of publicly available information and references. All technical content, methodology, experiments, analyses, and conclusions were developed and verified by the authors.

ETHICS STATEMENT

This work focuses on post-compression calibration for pretrained large language models and does not involve human subjects, personal or sensitive data, or deployment in high-stakes settings. We do not identify additional ethical concerns beyond those generally associated with the underlying pretrained models and public datasets.

REPRODUCIBILITY STATEMENT

We provide detailed experimental settings, calibration protocols, baseline implementations, evaluation procedures, and software environments in Appendix B to facilitate reproducibility. In addition, the code and supporting materials for reproducing our experiments are available on our project page: https://lens1.github.io/.

REFERENCES

Ali Abbasi, Chayne Thrash, Haoran Qin, Shansita Sharma, Sepehr Seifi, and Soheil Kolouri. Zero sum SVD: Balancing loss sensitivity for low rank LLM compression. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=tAdsDOkRBw.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Fluctuation-based adaptive structured pruning for large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i10.28960. URL https://doi.org/10.1609/aaai.v38i10.28960.

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

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:208290939.

Hung-Yueh Chiang, Chi-Chih Chang, Yu-Chen Lu, Chien-Yu Lin, Kai-Chiang Wu, Mohamed Abdelfattah, and Diana Marculescu. Uniql: Unified quantization and low-rank compression for adaptive edge llms. In *International Conference on Learning Representations*, volume 2026, pp. 32286–32310, 2026.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018. URL https://api.semanticscholar.org/CorpusID:3922816.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Yiwei Fu, Jiaxiao Chen, Haoyu Lin, Zhonghui Gu, Qingqing Long, Hui Wan, Xiao Luo, and Minghua Deng. Out-of-distribution generalization enhances protein function annotation for low-homology sequences. *Briefings in Bioinformatics*, 27(3):bbag243, 2026.

Shangqian Gao, Chi-Heng Lin, Ting Hua, Tang Zheng, Yilin Shen, Hongxia Jin, and Yen-Chang Hsu. Disp-llm: Dimension-independent structural pruning for large language models. *Advances in Neural Information Processing Systems*, 37:72219–72244, 2024.

Wenbo Gong, Meyer Scetbon, Chao Ma, and Edward Meeds. Towards efficient optimizer design for llm via structured fisher approximation with a low-rank extension. In *International Conference on Learning Representations*, volume 2026, pp. 130611–130662, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hao Gu, Wei Li, Lujun Li, Qiyuan Zhu, Mark Lee, Shengjie Sun, Wei Xue, and Yike Guo. Delta decompression for moe-based llms compression. *arXiv preprint arXiv:2502.17298*, 2025.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300, 2020. URL https://api.semanticscholar.org/CorpusID:221516475.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uPv9Y3gmAI5.

Xinhao Huang, You-Liang Huang, and Zeyi Wen. Sola: Leveraging soft activation sparsity and low-rank decomposition for large language model compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17494–17502, 2025.

Ajay Jaiswal, Yifan Wang, Lu Yin, Shiwei Liu, Runjin Chen, Jiawei Zhao, Ananth Grama, Yuandong Tian, and Zhangyang Wang. From low rank gradient subspace stabilization to low-rank weights: Observations, theories, and applications. *arXiv preprint arXiv:2407.11239*, 2024.

Yixin Ji, Yang Xiang, Juntao Li, Qingrong Xia, Zi Ye, Xinyu Duan, Zhefeng Wang, Kehai Chen, and Min Zhang. Adaptive feature-based low-rank compression of large language models via Bayesian optimization. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4152–4168, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.240. URL https://aclanthology.org/2024.findings-emnlp.240/.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Hyeonjin Kim and Jaejun Yoo. Singular value scaling: Efficient generative model compression via pruned weights refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17859–17867, 2025.

Runchao Li, Yao Fu, Mu Sheng, Haotian Yu, Xianxuan Long, and Kenneth A Loparo. Jw-svd: Bridging the cross-modal mismatch in post-training mllm compression. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42691–42702, 2026.

Zhiteng Li, Mingyuan Xia, Jingyuan Zhang, Zheng Hui, Haotong Qin, Linghe Kong, Yulun Zhang, and Xiaokang Yang. Adasvd: Adaptive singular value decomposition for large language models. *arXiv preprint arXiv:2502.01403*, 2025.

Yuesen Liao, Zhiwei Li, Binrui Wu, Zihao Cheng, Su Zhao, Shuai Chen, and Weizhong Zhang. Compress large language models via collaboration between learning and matrix approximation. *Advances in Neural Information Processing Systems*, 38:60660–60680, 2026.

Chi-Heng Lin, Shangqian Gao, James Smith, Abhishek Patel, Shikhar Tuli, Yilin Shen, Hongxia Jin, and Yen-Chang Hsu. Modegpt: Modular decomposition for large language model compression. In *International Conference on Learning Representations*, volume 2025, pp. 101355–101390, 2025.

Kainan Liu, Yong Zhang, Ning Cheng, Zhitao Li, Shaojun Wang, and Jing Xiao. Grasp: Replace redundant layers with adaptive singular parameters for efficient model compression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26344–26359, 2025.

Shih-Yang Liu, Huck Yang, Chein-Yi Wang, Nai Chit Fung, Hongxu Yin, Charbel Sakr, Saurav Muralidharan, Kwang-Ting Cheng, Jan Kautz, Yu-Chiang Frank Wang, et al. Eora: Training-free compensation for compressed llm with eigenspace low-rank approximation. *arXiv preprint arXiv:2410.21271*, 5, 2024.

Jun Lu, Tianyi Xu, Bill Ding, David Li, and Yu Kang. Large language model compression via the nested activation-aware decomposition, 2025. URL https://arxiv.org/abs/2503.17101.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023.

Mehdi Makni, Xiang Meng, and Rahul Mazumder. 3basil: An algorithmic framework for sparse plus low-rank compression of llms. *Advances in Neural Information Processing Systems*, 38: 173917–173947, 2026.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. URL https://arxiv.org/abs/1609.07843.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018. URL https://arxiv.org/abs/1809.02789.

Mohammad Mozaffari, Amir Yazdanbakhsh, and Maryam Mehri Dehnavi. Slim: One-shot quantization and sparsity with low-rank approximation for llm weight compression. *arXiv preprint arXiv:2410.09615*, 2024.

Junyoung Park, Dalton Jones, Matthew Morse, Raghavv Goel, Mingu Lee, and Christopher Lott. Keydiff: Key similarity-based kv cache eviction for long-context llm inference in resource-constrained environments. *Advances in Neural Information Processing Systems*, 38:5983–6019, 2026.

Uliana Parkina and Maxim Rakhuba. Coala: Numerically stable and efficient framework for context-aware low-rank approximation. *Advances in Neural Information Processing Systems*, 38:71014–71041, 2026.

Mahmoud Safari and Frank Hutter. Svd-surgeon: Optimal singular-value surgery for large language model compression. *arXiv preprint arXiv:2606.23568*, 2026.

Rajarshi Saha, Naomi Sagan, Varun Srivastava, Andrea J Goldsmith, and Mert Pilanci. Compressing large language models using low rank and low precision decomposition. *Advances in Neural Information Processing Systems*, 37:88981–89018, 2024.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, August 2021. ISSN 0001-0782. doi: 10.1145/3474381. URL https://doi.org/10.1145/3474381.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Gao Peng, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 45472–45496, 2024.

Li Shen, Anke Tang, Yong Luo, Tao Sun, Han Hu, and Xiaochun Cao. Targeted low-rank refinement: Enhancing sparse language models with precision. In *Forty-second International Conference on Machine Learning*, 2025.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. Beyond variance: Knowledge-aware llm compression via fisher-aligned subspace diagnostics. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47938–47954, 2026.

Lin Song, Yukang Chen, Shuai Yang, Xiaohan Ding, Yixiao Ge, Ying-Cong Chen, and Ying Shan. Low-rank approximation for sparse attention in multi-modal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13763–13773. IEEE, 2024.

Moritz Thoma, Maximilian Groezinger, Maximilian Forstenhäusler, Emad Aghajanzadeh, Manoj Rohit Vemparala, Christos Anagnostopoulos, Pierpaolo Mori, Nael Fasfous, Alexander Frickenstein, Daniel Mueller-Gritschneder, et al. Advancing svd-based llm compression via layer-wise error model search. In *Forty-third International Conference on Machine Learning*, 2026.

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

Qinsi Wang, Jinghan Ke, Masayoshi Tomizuka, Kurt Keutzer, and Chenfeng Xu. Dobi-svd: Differentiable svd for llm compression and some new perspectives. In *International Conference on Learning Representations*, volume 2025, pp. 12561–12590, 2025a.

Xin Wang, Samiul Alam, Zhongwei Wan, Hui Shen, and Mi Zhang. Svd-llm v2: Optimizing singular value truncation for large language model compression, 2025b. URL https://arxiv.org/abs/2503.12340.

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

Yilin Xiao, Chuang Zhou, Qinggang Zhang, Bo Li, Qing Li, and Xiao Huang. Reliable reasoning path: Distilling effective guidance for llm reasoning with knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, 2026.

Lin Xv, Xian Gao, Ting Liu, and Yuzhuo fu. Beyond uniform SVD: Dual-level optimization across columns and modules for LLM compression. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 18335–18349, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.912. URL https://aclanthology.org/2026.findings-acl.912/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in neural information processing systems*, 35:27168–27183, 2022.

Zhihang Yuan, Yuzhang Shang, Yue Song, Dawei Yang, Qiang Wu, Yan Yan, and Guangyu Sun. Asvd: Activation-aware singular value decomposition for compressing large language models, 2025. URL https://arxiv.org/abs/2312.05821.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Jialin Zhao, Yingtao Zhang, and Carlo Vittorio Cannistraci. Pivoting factorization: A compact meta low-rank representation of sparsity for efficient inference in large language models. *arXiv preprint arXiv:2501.19090*, 2025.

APPENDIX CONTENTS

A Proofs for the Theoretical Analysis **15**

A.1 Proof Setup 15

A.2 Proof of Theorem 3.1 15

A.3 Proof of Theorem 3.2 16

A.4 Proof of Corollary 3.1 17

B Detailed Experimental Setup **18**

B.1 Compression Setup 18

B.2 `LENS` Calibration Protocol 19

B.3 Baseline Implementations 19

B.4 Evaluation Details 21

B.5 Reproducibility 21

C Additional Experimental Results **21**

C.1 Additional Main Comparison Results 21

C.2 Generalization and Robustness 21

C.3 Sensitivity Analysis 24

C.4 Calibration Anatomy 27

C.5 Detailed Ablation Results 27

D Additional Related Work **27**

D.1 Low-rank Compression for LLMs 27

D.2 Post-compression Refinement and Calibration 32

A Proofs for the Theoretical Analysis

A.1 Proof Setup

Throughout this appendix, we fix a calibration stage and its reference model $\mathbf{M}_{\mathrm{ref}}$. Unless stated otherwise, all response quantities and derivatives are evaluated at this reference state. We use the notation introduced in Section 3. For convenience, let $\mathbf{Z}_{k,n} := \mathbf{V}_k \mathbf{X}_{k,n}$, so that $\mathbf{H}_{k,n} = \mathbf{U}_k \mathbf{Z}_{k,n}$, and define

$$\mathrm{sym}(\mathbf{C}) := \frac{\mathbf{C} + \mathbf{C}^\top}{2}. \tag{A.1}$$

Along the perturbation paths considered below, only the calibrated parameter blocks $\{\mathbf{U}_k\}_k$ are directly perturbed; all remaining model parameters are fixed. For simultaneous perturbations of multiple modules, we differentiate the end-to-end loss with respect to each perturbed block $\mathbf{U}_k$ and sum the resulting first-order contributions. The full-network adjoint $\mathbf{D}_{k,n} = \partial \ell_n / \partial \mathbf{H}_{k,n}$ already incorporates downstream dependence through backpropagation. All results below are deterministic conditional on the realized data splits, response eigendecompositions, and validated mode sets.

A.2 Proof of Theorem 3.1

Proof. We first derive the differential of the empirical discovery loss. For each module k, since $\mathbf{H}_{k,n} = \mathbf{U}_k \mathbf{Z}_{k,n}$, we have

$$\nabla_{\mathbf{U}_k} \ell_n = \mathbf{D}_{k,n} \mathbf{Z}_{k,n}^\top. \tag{A.2}$$

Since $\mathcal{D}_{\mathrm{disc}}$ is finite,

$$\nabla_{\mathbf{U}_k} \bar{\ell}_{\mathrm{disc}} = \frac{1}{|\mathcal{D}_{\mathrm{disc}}|} \sum_{n \in \mathcal{D}_{\mathrm{disc}}} \mathbf{D}_{k,n} \mathbf{Z}_{k,n}^\top. \tag{A.3}$$

Under the perturbation in Eq. 18,

$$\dot{\mathbf{U}}_k(0) = \mathbf{Q}_k \mathbf{A}_k \mathbf{Q}_k^\top \mathbf{U}_k. \tag{A.4}$$

By the assumed joint differentiability of $\bar{\ell}_{\mathrm{disc}}$ with respect to the calibrated parameter blocks,

$$\begin{aligned}
\frac{d}{dt} \bar{\ell}_{\mathrm{disc}}(\mathbf{M}(t))\Big|_{t=0} &= \sum_k \left\langle \nabla_{\mathbf{U}_k} \bar{\ell}_{\mathrm{disc}}, \dot{\mathbf{U}}_k(0) \right\rangle_F \\
&= \sum_k \left\langle \frac{1}{|\mathcal{D}_{\mathrm{disc}}|} \sum_{n \in \mathcal{D}_{\mathrm{disc}}} \mathbf{D}_{k,n} \mathbf{Z}_{k,n}^\top, \mathbf{Q}_k \mathbf{A}_k \mathbf{Q}_k^\top \mathbf{U}_k \right\rangle_F \\
&= \sum_k \left\langle \frac{1}{|\mathcal{D}_{\mathrm{disc}}|} \sum_{n \in \mathcal{D}_{\mathrm{disc}}} \mathbf{Q}_k^\top \mathbf{D}_{k,n} \mathbf{H}_{k,n}^\top \mathbf{Q}_k, \mathbf{A}_k \right\rangle_F \\
&= \sum_k \left\langle \frac{1}{|\mathcal{D}_{\mathrm{disc}}|} \sum_{n \in \mathcal{D}_{\mathrm{disc}}} \mathrm{sym}\big(\mathbf{Q}_k^\top \mathbf{D}_{k,n} \mathbf{H}_{k,n}^\top \mathbf{Q}_k\big), \mathbf{A}_k \right\rangle_F \\
&= \sum_k \langle \mathbf{S}_k, \mathbf{A}_k \rangle_F .
\end{aligned} \tag{A.5}$$

The third equality uses $\mathbf{H}_{k,n} = \mathbf{U}_k \mathbf{Z}_{k,n}$ and cyclic invariance of the trace, while the fourth uses $\mathbf{A}_k = \mathbf{A}_k^\top$. This establishes Eq. 19.

It remains to show spectral sufficiency. Fix a module k and define

$$\mathbf{C}_k := \mathbf{E}_k^\top \mathbf{A}_k \mathbf{E}_k. \tag{A.6}$$

Using $\mathbf{S}_k = \mathbf{E}_k \mathbf{\Lambda}_k \mathbf{E}_k^\top$ and the orthogonality of $\mathbf{E}_k$,

$$\langle \mathbf{S}_k, \mathbf{A}_k \rangle_F = \mathrm{tr}\,(\mathbf{\Lambda}_k \mathbf{C}_k) = \sum_{j=1}^{r_k} \lambda_{k,j} (\mathbf{C}_k)_{jj}. \tag{A.7}$$

Thus, the first-order loss effect depends only on the diagonal coordinates of $\mathbf{A}_k$ in the response eigenbasis. By definition, $\Pi_{\mathbf{E}_k}(\mathbf{A}_k)$ retains exactly these coordinates, and therefore

$$\langle \mathbf{S}_k, \Pi_{\mathbf{E}_k}(\mathbf{A}_k) \rangle_F = \sum_{j=1}^{r_k} \lambda_{k,j} (\mathbf{C}_k)_{jj} = \langle \mathbf{S}_k, \mathbf{A}_k \rangle_F . \tag{A.8}$$

Finally, orthogonal invariance of the Frobenius norm gives $\|\mathbf{C}_k\|_F = \|\mathbf{A}_k\|_F$, whereas

$$\|\Pi_{\mathbf{E}_k}(\mathbf{A}_k)\|_F^2 = \sum_{j=1}^{r_k} (\mathbf{C}_k)_{jj}^2 \le \sum_{i=1}^{r_k} \sum_{j=1}^{r_k} (\mathbf{C}_k)_{ij}^2 = \|\mathbf{A}_k\|_F^2. \tag{A.9}$$

Taking square roots yields the norm bound in Eq. 20. The argument holds for any orthonormal eigendecomposition of $\mathbf{S}_k$, so repeated eigenvalues do not affect the result. □

A.3 Proof of Theorem 3.2

Proof. Introduce auxiliary gain coordinates $\mathbf{a} = (a_{k,j})_{(k,j)\in\mathcal{I}}$ and define

$$\mathbf{G}_k(\mathbf{a}) = \mathbf{I} + \sum_{j \in \mathcal{J}_k} a_{k,j} \mathbf{p}_{k,j} \mathbf{p}_{k,j}^\top, \qquad \mathbf{U}_k(\mathbf{a}) = \mathbf{G}_k(\mathbf{a}) \mathbf{U}_k. \tag{A.10}$$

Let $\mathbf{M}(\mathbf{a})$ denote the resulting model and define

$$F_T(\mathbf{a}) = \bar{\ell}_T(\mathbf{M}(\mathbf{a})), \qquad T \in \{\mathrm{disc}, \mathrm{fit}\}. \tag{A.11}$$

By the differentiability assumption in Theorem 3.2 and the affine parameterization above, F_{disc} and F_{fit} are differentiable at $\mathbf{a} = \mathbf{0}$.

We first identify the gradients with respect to the auxiliary gain coordinates. For $i = (k, j) \in \mathcal{I}$, write $a_i := a_{k,j}$. Differentiating Eq. A.10 with respect to a_i at $\mathbf{a} = \mathbf{0}$ gives

$$\left. \frac{\partial \mathbf{U}_k(\mathbf{a})}{\partial a_i} \right|_{\mathbf{a}=\mathbf{0}} = \mathbf{p}_{k,j} \mathbf{p}_{k,j}^\top \mathbf{U}_k. \tag{A.12}$$

Using $\nabla_{\mathbf{U}_k} \ell_n = \mathbf{D}_{k,n} \mathbf{Z}_{k,n}^\top$, the corresponding coordinate derivative of the end-to-end loss is

$$\begin{aligned} \left. \frac{\partial \ell_n(\mathbf{M}(\mathbf{a}))}{\partial a_i} \right|_{\mathbf{a}=\mathbf{0}} &= \langle \mathbf{D}_{k,n} \mathbf{Z}_{k,n}^\top, \mathbf{p}_{k,j} \mathbf{p}_{k,j}^\top \mathbf{U}_k \rangle_F \\ &= \mathbf{p}_{k,j}^\top \mathbf{D}_{k,n} \mathbf{H}_{k,n}^\top \mathbf{p}_{k,j} = q_{k,n}(\mathbf{v}_{k,j}). \end{aligned} \tag{A.13}$$

Here $\mathbf{p}_{k,j} = \mathbf{Q}_k \mathbf{v}_{k,j}$.

Averaging over the fitting and discovery subsets gives, respectively,

$$\left. \frac{\partial F_{\mathrm{fit}}}{\partial a_i} \right|_{\mathbf{a}=\mathbf{0}} = g_i, \qquad \left. \frac{\partial F_{\mathrm{disc}}}{\partial a_i} \right|_{\mathbf{a}=\mathbf{0}} = \mathbf{v}_{k,j}^\top \mathbf{S}_k \mathbf{v}_{k,j} = \lambda_i. \tag{A.14}$$

Collecting the coordinate derivatives therefore yields

$$\nabla_{\mathbf{a}} F_{\mathrm{fit}}(\mathbf{0}) = \mathbf{g}, \qquad \nabla_{\mathbf{a}} F_{\mathrm{disc}}(\mathbf{0}) = \boldsymbol{\lambda}, \tag{A.15}$$

where $\mathbf{g} = (g_i)_{i\in\mathcal{I}}$ and $\boldsymbol{\lambda} = (\lambda_i)_{i\in\mathcal{I}}$.

We next establish the sign alignment induced by reliability filtering. Write the fitting confidence interval of mode i as $[g_i - w_i, g_i + w_i]$ with $w_i \ge 0$. For a retained amplification mode, the validation condition gives $\lambda_i < 0$ and $g_i + w_i < 0$, hence $g_i < 0$. For a retained suppression mode, the validation condition gives $\lambda_i > 0$ and $g_i - w_i > 0$, hence $g_i > 0$. Thus,

$$g_i \neq 0, \qquad \lambda_i g_i > 0, \qquad \forall i \in \mathcal{I}. \tag{A.16}$$

Moreover, $f_i \ge 0$ and $\epsilon > 0$, so $f_i + \epsilon > 0$.

Let $\mathbf{d} = (d_i)_{i\in\mathcal{I}}$ with

$$d_i = -\frac{g_i}{f_i + \epsilon} = \widetilde{\delta}_{k,j}. \tag{A.17}$$

Since $d_i = \widetilde{\delta}_{k,j}$, the actual calibration path in Eq. 14 is exactly represented by $\mathbf{a} = \eta\mathbf{d}$ under the auxiliary coordinates above; in particular, $\mathbf{M}(\eta\mathbf{d}) = \mathbf{M}_\eta$.

By the chain rule,

$$\begin{aligned}
\frac{d}{d\eta} F_{\text{fit}}(\eta\mathbf{d})\Big|_{\eta=0} &= \nabla_{\mathbf{a}} F_{\text{fit}}(\mathbf{0})^\top \mathbf{d} = -\sum_{i\in\mathcal{I}} \frac{g_i^2}{f_i + \epsilon} < 0, \\
\frac{d}{d\eta} F_{\text{disc}}(\eta\mathbf{d})\Big|_{\eta=0} &= \nabla_{\mathbf{a}} F_{\text{disc}}(\mathbf{0})^\top \mathbf{d} = -\sum_{i\in\mathcal{I}} \frac{\lambda_i g_i}{f_i + \epsilon} < 0.
\end{aligned} \tag{A.18}$$

The inequalities are strict because $\mathcal{I} \neq \varnothing$ and every selected mode satisfies $g_i \neq 0$ and $\lambda_i g_i > 0$.

Finally, let

$$\phi_T(\eta) := F_T(\eta\mathbf{d}), \qquad T \in \{\text{disc}, \text{fit}\}. \tag{A.19}$$

Since ϕ_T is differentiable at $\eta = 0$ and $\phi_T'(0) < 0$, there exists $\eta_T > 0$ such that

$$0 < \eta < \eta_T \quad \Longrightarrow \quad F_T(\eta\mathbf{d}) < F_T(\mathbf{0}). \tag{A.20}$$

For positivity of the effective gains, define

$$\eta_{\text{pos}} := \min_{i\in\mathcal{I}:\, d_i<0} \left(-\frac{1}{d_i}\right), \tag{A.21}$$

with $\eta_{\text{pos}} := +\infty$ if the index set is empty. Then $\eta_{\text{pos}} > 0$ and $1 + \eta d_i > 0$ for all $i \in \mathcal{I}$ whenever $0 < \eta < \eta_{\text{pos}}$.

Therefore,

$$\eta_0 := \min\{\eta_{\text{disc}}, \eta_{\text{fit}}, \eta_{\text{pos}}\} > 0 \tag{A.22}$$

satisfies, for every $0 < \eta < \eta_0$,

$$F_{\text{disc}}(\eta\mathbf{d}) < F_{\text{disc}}(\mathbf{0}), \qquad F_{\text{fit}}(\eta\mathbf{d}) < F_{\text{fit}}(\mathbf{0}), \qquad 1 + \eta d_i > 0 \quad \forall i \in \mathcal{I}. \tag{A.23}$$

This proves the theorem. □

A.4 Proof of Corollary 3.1

Proof. For $i = (k, j) \in \mathcal{I}$, let $q_{n,i} := q_{k,n}(\mathbf{v}_{k,j})$ for $n \in \mathcal{D}_{\text{fit}}$. By the definitions of g_i and f_i,

$$g_i = \frac{1}{|\mathcal{D}_{\text{fit}}|} \sum_{n\in\mathcal{D}_{\text{fit}}} q_{n,i}, \qquad f_i = \frac{1}{|\mathcal{D}_{\text{fit}}|} \sum_{n\in\mathcal{D}_{\text{fit}}} q_{n,i}^2. \tag{A.24}$$

Hence,

$$f_i - g_i^2 = \frac{1}{|\mathcal{D}_{\text{fit}}|} \sum_{n\in\mathcal{D}_{\text{fit}}} (q_{n,i} - g_i)^2 \geq 0, \tag{A.25}$$

so $f_i \geq g_i^2$. Since $\epsilon > 0$,

$$f_i + \epsilon \geq g_i^2 + \epsilon \geq 2|g_i|\sqrt{\epsilon}, \tag{A.26}$$

where the last inequality follows from $(|g_i| - \sqrt{\epsilon})^2 \geq 0$. Therefore,

$$|d_i| = \frac{|g_i|}{f_i + \epsilon} \leq \frac{1}{2\sqrt{\epsilon}}, \tag{A.27}$$

which proves the uniform bound in Corollary 3.1.

For any $0 \leq \eta < 2\sqrt{\epsilon}$,

$$1 + \eta d_i \geq 1 - \eta|d_i| \geq 1 - \frac{\eta}{2\sqrt{\epsilon}} > 0, \qquad \forall i \in \mathcal{I}. \tag{A.28}$$

For convenience, write $d_{k,j} := d_i$ for $i = (k, j)$. Since $d_{k,j} = \widetilde{\delta}_{k,j}$, the transformation in Eq. 14 can be written as

$$\mathbf{G}_k(\eta) = \mathbf{I} + \eta \sum_{j \in \mathcal{J}_k} d_{k,j} \mathbf{p}_{k,j} \mathbf{p}_{k,j}^\top. \tag{A.29}$$

Since $\mathbf{Q}_k$ and $\mathbf{E}_k$ have orthonormal columns, the selected directions $\{\mathbf{p}_{k,j}\}_{j \in \mathcal{J}_k}$ are orthonormal. Hence, $\mathbf{G}_k(\eta)$ has eigenvalue $1 + \eta d_{k,j} > 0$ along each selected direction $\mathbf{p}_{k,j}$ and eigenvalue 1 on their orthogonal complement. Consequently,

$$\mathbf{G}_k(\eta) \succ 0, \tag{A.30}$$

and hence $\mathbf{G}_k(\eta)$ is invertible.

Let

$$\widehat{\mathbf{W}}_k := \mathbf{U}_k \mathbf{V}_k, \qquad \widehat{\mathbf{W}}_k(\eta) := \mathbf{G}_k(\eta) \widehat{\mathbf{W}}_k. \tag{A.31}$$

Assume $\operatorname{rank}(\widehat{\mathbf{W}}_k) = r_k$. Since left multiplication by an invertible matrix preserves rank and row space,

$$\operatorname{rank}\big(\widehat{\mathbf{W}}_k(\eta)\big) = \operatorname{rank}\big(\widehat{\mathbf{W}}_k\big) = r_k, \tag{A.32}$$

and

$$\operatorname{row}\big(\widehat{\mathbf{W}}_k(\eta)\big) = \operatorname{row}\big(\widehat{\mathbf{W}}_k\big). \tag{A.33}$$

It remains to verify preservation of the column space. The same rank condition implies $\operatorname{rank}(\mathbf{U}_k) = r_k$, and therefore

$$\operatorname{col}(\widehat{\mathbf{W}}_k) = \operatorname{col}(\mathbf{U}_k). \tag{A.34}$$

Moreover, every selected $\mathbf{p}_{k,j}$ lies in $\operatorname{col}(\mathbf{U}_k)$. Hence, for any $\mathbf{x} \in \operatorname{col}(\mathbf{U}_k)$,

$$\mathbf{G}_k(\eta)\mathbf{x} = \mathbf{x} + \eta \sum_{j \in \mathcal{J}_k} d_{k,j} \mathbf{p}_{k,j} (\mathbf{p}_{k,j}^\top \mathbf{x}) \in \operatorname{col}(\mathbf{U}_k). \tag{A.35}$$

Thus,

$$\mathbf{G}_k(\eta) \operatorname{col}(\mathbf{U}_k) \subseteq \operatorname{col}(\mathbf{U}_k). \tag{A.36}$$

Since $\mathbf{G}_k(\eta)$ is invertible, its restriction to $\operatorname{col}(\mathbf{U}_k)$ is injective, so both subspaces have the same dimension. Therefore,

$$\mathbf{G}_k(\eta) \operatorname{col}(\mathbf{U}_k) = \operatorname{col}(\mathbf{U}_k). \tag{A.37}$$

It follows that

$$\operatorname{col}\big(\widehat{\mathbf{W}}_k(\eta)\big) = \operatorname{col}\big(\widehat{\mathbf{W}}_k\big). \tag{A.38}$$

This completes the proof. □

B Detailed Experimental Setup

B.1 Compression Setup

Unless otherwise specified, all methods are evaluated at a 20% compression ratio. We define the compression ratio as the fraction of decoder-linear weight parameters removed relative to the dense model. The compression budget covers the decoder-linear weights, including the query, key, value, output, gate, up, and down projections in every decoder layer, while excluding bias parameters. For methods with discrete rank or structural granularity, we ensure that the realized decoder-linear parameter count does not exceed the prescribed budget.

For the default Duo-SVD initialization, we use WikiText-2 Train as the calibration corpus. We estimate activation covariance statistics using 256 randomly sampled windows of 2048 tokens each (sampling seed 3). Following the released Duo-SVD implementation, MAAS sensitivity is estimated using 32 calibration windows over candidate retained ratios $\{0.1, 0.2, \ldots, 1.0\}$. The resulting allocation is then factorized using the activation-aware column-preserving strategy (CPS). For Duo-SVD's column-preserving representation, `LENS` is applied only to the low-rank branch, while the explicitly preserved columns remain fixed throughout calibration.

B.2 LENS CALIBRATION PROTOCOL

We use 128, 64, and 64 WikiText-2 training windows for direction discovery, response fitting, and scale selection, respectively, together with 64 WikiText-2 validation windows for final acceptance. The same validation subset is reused across the two sequential stages. All windows contain 2048 tokens. The 256-window pool used for $\mathcal{D}_{\text{disc}}$ and $\mathcal{D}_{\text{fit}}$ is the same WikiText-2 training pool used to construct the default Duo-SVD initialization. Within this pool, $\mathcal{D}_{\text{disc}}$ and $\mathcal{D}_{\text{fit}}$ are disjoint. By contrast, $\mathcal{D}_{\text{scale}}$ is sampled separately with seed 17, and $\mathcal{D}_{\text{val}}$ is drawn from the WikiText-2 validation split.

Reliability Filtering. Direction discovery and reliability filtering follow the procedure described in Section 3. We use approximate two-sided 95% normal confidence intervals with $\alpha = 0.05$ and $z_{0.975} = 1.96$.

Adaptive Stabilizer. In our implementation, the stabilizer ϵ in the relative-gain estimator is determined adaptively for each calibration stage rather than fixed globally. Let $\mathcal{I}_s$ denote the set of directions retained after the discovery and fitting tests at stage $s \in \{\text{amplify}, \text{suppress}\}$. We set

$$\epsilon_s = \max\left\{10^{-12}, 10^{-6} \cdot \text{median}_{(k,j)\in\mathcal{I}_s} f_{k,j}\right\}. \tag{B.1}$$

The same ϵ_s is shared by all retained directions within the current stage. After amplify, it is recomputed from the re-profiled suppress statistics.

Stage-level Scale Search and Held-out Validation. For each calibration stage, we search over the predefined geometric grid

$$\mathcal{E} = \{2^0, 2^{-1}, \ldots, 2^{-23}\}. \tag{B.2}$$

Given the validated directions and their relative gains for the current stage, we restrict the search to

$$\mathcal{E}^+ = \left\{\eta \in \mathcal{E} : 1 + \eta\widetilde{\delta}_{k,j} > 0, \quad \forall k,\ \forall j \in \mathcal{J}_k\right\}, \tag{B.3}$$

so that all effective directional gains remain positive.

Let $\mathbf{M}_{\text{ref}}$ denote the reference model of the current stage and $\mathbf{M}_\eta$ the candidate model obtained by jointly applying Eq. 14 with scale η. For $T \in \{\text{scale}, \text{val}\}$, define the mean paired NLL change as

$$\overline{\Delta}_T(\eta) = \frac{1}{|\mathcal{D}_T|}\sum_{n\in\mathcal{D}_T}\left[\ell_n(\mathbf{M}_\eta) - \ell_n(\mathbf{M}_{\text{ref}})\right]. \tag{B.4}$$

The candidate $\widehat{\eta}$ is selected on $\mathcal{D}_{\text{scale}}$ according to Eq. 15. If multiple candidates attain the same objective, we select the smaller η.

For held-out validation, let $s_{\text{val}}(\eta)$ denote the sample standard deviation of the per-sequence paired NLL changes $\ell_n(\mathbf{M}_\eta) - \ell_n(\mathbf{M}_{\text{ref}})$ over $\mathcal{D}_{\text{val}}$. The upper endpoint of the approximate $(1-\alpha)$ confidence interval is

$$U^{1-\alpha}_{\text{val}}(\eta) = \overline{\Delta}_{\text{val}}(\eta) + z_{1-\alpha/2}\frac{s_{\text{val}}(\eta)}{\sqrt{|\mathcal{D}_{\text{val}}|}}. \tag{B.5}$$

The stage update is accepted according to

$$\eta^\star = \begin{cases}\widehat{\eta}, & \overline{\Delta}_{\text{scale}}(\widehat{\eta}) < 0 \text{ and } U^{1-\alpha}_{\text{val}}(\widehat{\eta}) < 0,\\ 0, & \text{otherwise.}\end{cases} \tag{B.6}$$

If $\mathcal{E}^+ = \varnothing$, we likewise set $\eta^\star = 0$, leaving the reference model unchanged.

The complete sequential calibration procedure is summarized in Algorithm 1.

B.3 BASELINE IMPLEMENTATIONS

We follow the released implementations and recommended configurations of all compared methods whenever available, while enforcing the same target compression ratio.

Algorithm 1 `LENS`: Loss-aligned Directional Salience Calibration for LLM Compression.

Input: Initial compressed model $\mathbf{M}_0$; calibration subsets $\mathcal{D}_{\text{disc}}$, $\mathcal{D}_{\text{fit}}$, $\mathcal{D}_{\text{scale}}$, and $\mathcal{D}_{\text{val}}$; candidate scale set $\mathcal{E}$
Output: Calibrated model $\mathbf{M}_{A\to S}$

$\mathbf{M} \leftarrow \mathbf{M}_0$
for $\tau \in \{\text{AMPLIFY}, \text{SUPPRESS}\}$ **do**
  $\mathbf{M}_{\text{ref}} \leftarrow \mathbf{M}$ `// Profile responses at the current model state`
  for each calibrated low-rank module k **do** `// Identify loss-aligned directions`
    Compute the retained-subspace response statistics using Eqs. 2–4
    Construct $\mathbf{S}_k$ on $\mathcal{D}_{\text{disc}}$ using Eq. 5
    Eigendecompose $\mathbf{S}_k$ to obtain $\{(\lambda_{k,j}, \mathbf{v}_{k,j})\}_{j=1}^{r_k}$ using Eq. 6
    Compute the discovery responses and output-space directions using Eqs. 7–8
    Compute fitted responses and confidence intervals on $\mathcal{D}_{\text{fit}}$ using Eqs. 9–10
    if $\tau = \text{AMPLIFY}$ **then**
      $\mathcal{J}_k \leftarrow \{j : \lambda_{k,j} < 0,\ U_{k,j}^{1-\alpha} < 0\}$ `// Select reliable amplification directions`
    else
      $\mathcal{J}_k \leftarrow \{j : \lambda_{k,j} > 0,\ L_{k,j}^{1-\alpha} > 0\}$ `// Select reliable suppression directions`
    end if
    Compute $f_{k,j}$ and estimate $\widetilde{\delta}_{k,j}$ for $j \in \mathcal{J}_k$ using Eqs. 12–13 `// Estimate relative gains`
  end for
  Construct $\mathbf{G}_k(\eta)$ using Eq. 14 and the feasible scale set $\mathcal{E}^+$ using Eq. B.3
  if $\mathcal{E}^+ \neq \varnothing$ **then** `// Search for a feasible stage-level scale`
    Select $\widehat{\eta} \in \mathcal{E}^+$ on $\mathcal{D}_{\text{scale}}$ using Eq. 15
    Evaluate $\mathbf{M}_{\widehat{\eta}}$ on $\mathcal{D}_{\text{val}}$
    Set $\eta^\star$ according to Eq. B.6
  else
    $\eta^\star \leftarrow 0$ `// Skip the stage if no scale is feasible`
  end if
  Fold $\mathbf{G}_k(\eta^\star)$ into $\mathbf{U}_k$ for all calibrated modules and update $\mathbf{M}$ using Eq. 16
  if $\tau = \text{AMPLIFY}$ **then**
    $\mathbf{M}_A \leftarrow \mathbf{M}$ `// Use the amplify-updated model for suppress`
  end if
end for
return $\mathbf{M}$ as $\mathbf{M}_{A\to S}$

LLM-Pruner (Ma et al., 2023). We use BookCorpus for importance estimation and Alpaca for post-pruning recovery. Following the released configuration, we adopt block-wise Type-A/Type-B structural dependency pruning with Taylor `param_first` importance and local pruning, while protecting the first four and last two decoder layers. The pruned model is recovered on Alpaca for two epochs using LoRA with rank 8, scaling factor 16, dropout 0.05, learning rate 10^{-4}, and 100 warmup steps.

FLAP (An et al., 2024). We use 256 WikiText-2 training windows of 2048 tokens each for calibration. We adopt WIFV importance estimation and AL-AM global thresholding, together with structural attention/MLP pruning and the released bias compensation strategy. No post-compression training is performed.

SliceGPT (Ashkboos et al., 2024). We calibrate SliceGPT using 256 WikiText-2 training windows of 2048 tokens each. Following the released implementation, we apply layer-wise rotations, RMSNorm fusion, uncentered Gram-PCA-based slicing, and round the resulting hidden dimensions to multiples of 8. No post-compression training is performed.

SVD-LLM (Wang et al., 2025c). We use 256 WikiText-2 training windows of 2048 tokens each for calibration and Alpaca for the released parameter-update procedure. We apply truncation-aware data whitening and low-rank decomposition to all seven decoder projections, followed by the released sequential low-rank update in the $U \to V$ order.

SoLA (Huang et al., 2025). We use 256 WikiText-2 training windows of 2048 tokens each. Following the released configuration, we preserve 15% of the prime FFN neurons and apply activation-

Table 3: Performance comparison on Mistral-7B at 20% compression. Best and second-best results among compressed models are shown in **bold** and underlined, respectively.

Method	PPL↓	Avg.↑	MMLU	PIQA	WinoG.	HellaS.	ARC-e	ARC-c	OBQA
Dense	4.91	68.20	62.64	82.15	73.95	81.06	79.59	54.01	44.00
LLM-Pruner	7.63	59.55	39.87	**79.22**	66.14	**76.45**	69.57	43.60	**42.00**
FLAP	5.97	53.37	40.38	69.42	65.67	57.52	64.39	36.60	39.60
SliceGPT	6.07	53.31	37.71	68.66	68.51	59.63	62.58	38.05	38.00
SVD-LLM	7.56	58.77	**43.60**	75.57	64.96	70.35	74.12	44.03	38.80
BoLaCo	7.01	52.73	29.28	72.96	67.48	59.02	68.22	36.35	35.80
SoLA	6.92	51.66	33.25	70.24	68.19	55.24	62.79	35.49	36.40
Duo-SVD	5.81	59.61	40.16	78.07	68.98	72.10	72.47	43.86	41.60
LENS	**5.51**	**61.59**	42.99	78.45	**70.40**	76.27	**74.20**	**47.01**	41.80

aware marginal-subspace factorization with rank granularity 32 and a 90% spectral-energy guard. No post-compression training is performed.

BoLaCo (Ji et al., 2024). We use WikiText-2 Train for calibration and follow the released output-feature PCA and grouped Bayesian allocation procedure. We use the 5×4 grouping configuration with rank granularity 8. No post-compression training is performed.

Duo-SVD (Xv et al., 2026). We follow the Duo-SVD compression configuration described in Section B.1, including MAAS-based global allocation and activation-aware CPS. No post-compression training is performed.

B.4 Evaluation Details

For WikiText-2 perplexity, we use complete non-overlapping token blocks of length 4096 with batch size 1 and model-specific special tokens enabled. For downstream evaluation, we report raw accuracy for MMLU and WinoGrande and length-normalized accuracy for PIQA, HellaSwag, ARC-Easy, ARC-Challenge, and OpenBookQA.

B.5 Reproducibility

All experiments are conducted on 8 NVIDIA GeForce RTX 4090 GPUs. Our main software environment uses PyTorch 2.6.0 with CUDA 12.4, Transformers 5.9.0, Accelerate 1.13.0, Datasets 3.6.0, and `lm-eval` 0.4.3. Evaluation is performed in FP16 with batch size 1. We disable cuDNN in the formal evaluation entry points to reduce backend-dependent numerical variation.

C Additional Experimental Results

C.1 Additional Main Comparison Results

Table 3 reports the complete Mistral-7B results corresponding to the main comparison in Table 1. **LENS** ranks among the top two compressed models on every reported metric, showing that the improvement is consistent across both language modeling and downstream evaluation.

C.2 Generalization and Robustness

Results across Compression Ratios. Figure 6(a) reports WikiText-2 PPL and downstream average accuracy under compression ratios from 10% to 40% on LLaMA-2-7B and Mistral-7B. Table 4 further provides the complete task-level results for all compared methods across the four compression ratios. Together, these results show that the effectiveness of **LENS** persists as the compression budget becomes increasingly aggressive.

Results with Different Calibration Datasets. Figure 6(b) complements the WikiText-2 PPL results reported in the main text with C4 PPL and downstream average accuracy across all four back-

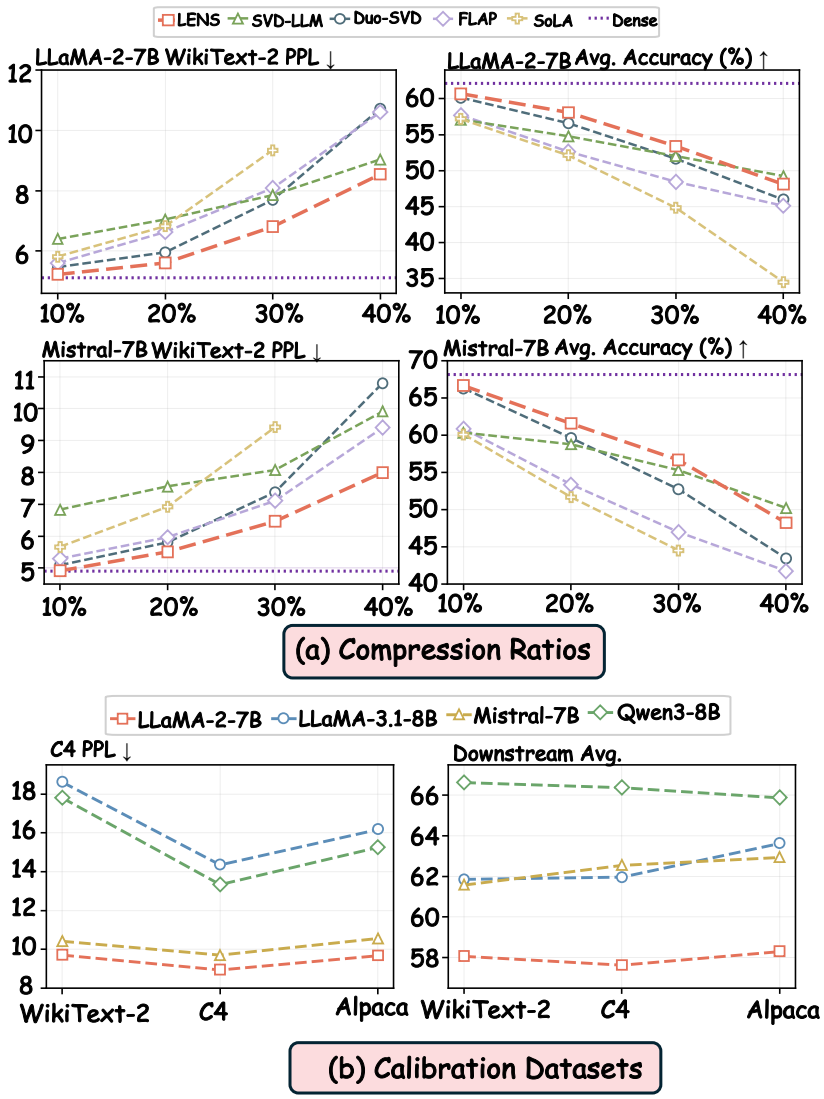

Figure 6: **Additional robustness results of LENS.** (a) WikiText-2 PPL and downstream average accuracy under compression ratios from 10% to 40%. (b) C4 PPL and downstream average accuracy when using WikiText-2, C4, or Alpaca for calibration.

bones. Table 5 further reports the complete WikiText-2 PPL, C4 PPL, average accuracy, and individual results on all seven downstream tasks for WikiText-2, C4, and Alpaca calibration. C4 calibration consistently achieves the lowest C4 PPL, while downstream performance remains comparatively stable across calibration datasets.

Compatibility with Other Low-rank Compressors. Figure 7(a) reports the complete WikiText-2 PPL and downstream average accuracy results when applying LENS to BoLaCo, SoLA, and SVD-LLM across all four backbones. Table 6 further provides the corresponding complete results on all seven downstream tasks. Across all 12 compressor–backbone combinations, applying LENS reduces WikiText-2 PPL, while average downstream accuracy improves in 11 of 12 cases. These results show that the gains of LENS are not specific to Duo-SVD and support its broad compatibility with different low-rank compressors.

Compatibility with Low-bit Quantization. Figure 7(b) reports the complete WikiText-2 PPL and downstream average accuracy results when models compressed by LENS at a 25% compression ra-

Table 4: **Complete results across different compression ratios.** "–" indicates that SoLA does not support the corresponding 40% compression setting on Mistral-7B.

		Overall		Downstream Tasks						
Ratio	**Method**	**PPL↓**	**Avg.↑**	**MMLU**	**PIQA**	**WinoG.**	**HellaS.**	**ARC-e**	**ARC-c**	**OBQA**
LLaMA-2-7B										
–	*Dense ref.*	5.12	62.17	45.93	79.11	68.90	76.02	74.62	46.42	44.20
10%	Duo-SVD	5.46	60.14	40.12	77.69	**67.80**	74.19	73.74	43.43	44.00
	FLAP	5.60	57.69	39.66	76.77	67.32	72.34	67.97	40.36	39.40
	SVD-LLM	6.40	57.08	32.98	76.71	66.14	71.11	67.68	43.52	41.40
	SoLA	5.81	57.19	36.91	77.53	67.09	68.26	69.61	39.51	41.40
	LENS	**5.23**	**60.67**	**41.43**	**78.02**	67.01	**75.44**	**74.20**	**44.37**	**44.20**
20%	Duo-SVD	5.95	56.59	32.09	76.55	67.56	69.97	69.32	39.08	41.60
	FLAP	6.63	52.63	31.60	74.65	65.35	65.04	59.39	34.22	38.20
	SVD-LLM	7.05	54.76	29.04	75.19	64.48	68.56	65.36	**40.70**	40.00
	SoLA	6.82	52.13	31.40	72.63	65.19	59.86	63.30	35.49	37.00
	LENS	**5.61**	**58.07**	**34.73**	**76.71**	**67.88**	**72.68**	**71.17**	40.53	**42.80**
30%	Duo-SVD	7.69	51.67	28.22	71.49	**65.59**	60.47	61.70	36.01	38.20
	FLAP	8.09	48.44	29.06	71.16	62.27	57.30	52.44	31.83	35.00
	SVD-LLM	7.85	52.01	29.38	72.03	62.75	63.74	59.97	**37.80**	38.40
	SoLA	9.34	44.85	24.55	67.90	58.80	47.05	52.44	29.18	34.00
	LENS	**6.81**	**53.40**	**30.89**	**72.85**	65.51	**64.68**	**63.17**	35.92	**40.80**
40%	Duo-SVD	10.72	46.00	26.45	65.78	60.77	50.48	52.57	31.74	34.20
	FLAP	10.61	45.11	**27.72**	65.94	60.06	49.13	48.48	30.03	34.40
	SVD-LLM	9.04	**49.25**	27.43	**70.24**	61.25	**58.94**	**55.81**	**33.11**	**38.00**
	SoLA	38.14	34.53	23.09	55.66	48.70	29.47	35.40	22.18	27.20
	LENS	**8.55**	48.12	27.18	67.85	**61.80**	54.63	55.26	32.51	37.60
Mistral-7B										
–	*Dense ref.*	4.91	68.20	62.64	82.15	73.95	81.06	79.59	54.01	44.00
10%	Duo-SVD	5.09	66.28	58.70	**81.61**	**73.09**	79.81	**76.94**	49.83	44.00
	FLAP	5.30	60.87	50.12	76.33	71.43	68.91	73.27	44.80	41.20
	SVD-LLM	6.83	60.35	48.38	76.93	66.46	73.43	73.36	46.08	37.80
	SoLA	5.66	60.10	48.45	76.66	70.56	68.50	73.23	44.71	38.60
	LENS	**4.91**	**66.68**	**58.72**	81.39	72.45	**80.47**	76.89	**52.22**	**44.60**
20%	Duo-SVD	5.81	59.61	40.16	78.07	68.98	72.10	72.47	43.86	41.60
	FLAP	5.97	53.37	40.38	69.42	65.67	57.52	64.39	36.60	39.60
	SVD-LLM	7.56	58.77	**43.60**	75.57	64.96	70.35	74.12	44.03	38.80
	SoLA	6.92	51.66	33.25	70.24	68.19	55.24	62.79	35.49	36.40
	LENS	**5.51**	**61.59**	42.99	**78.45**	**70.40**	**76.27**	**74.20**	**47.01**	**41.80**
30%	Duo-SVD	7.38	52.73	33.41	73.18	66.06	61.57	63.97	34.73	36.20
	FLAP	7.11	46.97	29.23	63.93	62.12	48.02	54.59	33.53	37.40
	SVD-LLM	8.07	55.30	36.55	73.88	65.51	65.35	67.05	**40.19**	**38.60**
	SoLA	9.42	44.51	26.95	64.09	61.64	44.18	51.56	28.75	34.40
	LENS	**6.46**	**56.69**	**38.09**	**75.30**	**68.35**	**68.35**	**68.14**	40.02	**38.60**
40%	Duo-SVD	10.80	43.45	24.32	65.45	59.51	44.47	50.17	27.65	32.60
	FLAP	9.41	41.74	25.89	59.47	58.09	41.79	44.87	28.07	34.00
	SVD-LLM	9.91	**50.23**	27.06	**70.46**	60.14	**60.04**	**63.17**	**37.37**	33.40
	SoLA	–	–	–	–	–	–	–	–	–
	LENS	**7.99**	48.24	**28.12**	67.36	**62.19**	55.51	56.78	33.11	**34.60**

Table 5: **Complete results with different calibration datasets.** We calibrate LENS using WikiText-2, C4, or Alpaca and report WikiText-2 perplexity (WT2 PPL), C4 perplexity (C4 PPL), average downstream accuracy (Avg.), and accuracy on seven downstream tasks. For each backbone, the best result among the three calibration datasets is shown in **bold**. The default WikiText-2 calibration setting is highlighted in blue.

	Overall			Downstream Tasks						
Calibration	**WT2 PPL↓**	**C4 PPL↓**	**Avg.↑**	**MMLU**	**PIQA**	**WinoG.**	**HellaS.**	**ARC-e**	**ARC-c**	**OBQA**
LLaMA-2-7B										
Dense ref.	5.12	7.31	62.17	45.93	79.11	68.90	76.02	74.62	46.42	44.20
WikiText-2	**5.61**	9.73	58.07	34.73	76.71	**67.88**	**72.68**	71.17	40.53	**42.80**
C4	6.67	**8.96**	57.63	34.94	**77.48**	66.69	71.47	69.82	41.04	42.00
Alpaca	7.50	9.69	**58.31**	**35.81**	77.20	66.93	71.66	**71.84**	**42.32**	42.40
LLaMA-3.1-8B										
Dense ref.	5.84	9.64	68.44	65.35	81.28	73.95	78.91	81.14	53.41	45.00
WikiText-2	**7.95**	18.64	61.86	54.78	76.33	69.85	70.02	73.02	**46.42**	42.60
C4	10.68	**14.37**	61.97	54.13	**78.45**	69.93	71.06	71.25	45.56	43.40
Alpaca	11.43	16.19	**63.64**	**58.89**	77.48	**72.45**	**71.74**	**74.96**	46.33	**43.60**
Mistral-7B										
Dense ref.	4.91	8.17	68.20	62.64	82.15	73.95	81.06	79.59	54.01	44.00
WikiText-2	**5.51**	10.42	61.59	42.99	78.45	**70.40**	76.27	74.20	47.01	41.80
C4	6.69	**9.71**	62.56	50.34	**80.09**	69.85	**76.82**	73.86	45.56	41.40
Alpaca	7.03	10.56	**62.95**	**53.62**	79.82	67.72	74.38	**75.51**	**47.61**	**42.00**
Qwen3-8B										
Dense ref.	6.51	11.73	69.49	76.56	79.38	72.69	78.64	80.13	57.00	42.00
WikiText-2	**7.34**	17.82	**66.63**	**64.26**	78.51	**70.56**	71.77	**81.94**	56.57	**42.80**
C4	8.49	**13.35**	66.38	62.68	**79.71**	70.40	**73.99**	80.72	**57.59**	39.60
Alpaca	9.94	15.27	65.89	63.51	78.84	68.59	71.75	81.19	56.74	40.60

tio are further quantized to 4 bits using GPTQ or RTN. Table 7 further provides the corresponding complete results on all seven downstream tasks. Across all four backbones, the additionally quantized models retain reasonable downstream performance, although the degradation varies with both the quantizer and backbone. RTN is generally more stable in downstream accuracy in our setting, supporting the practical compatibility of LENS with subsequent low-bit quantization.

C.3 Sensitivity Analysis

Direction-fraction Sensitivity. We provide the complete sensitivity results for the fraction of validated directions retained during calibration. For each stage $s \in \{\mathrm{A}, \mathrm{S}\}$, let $\mathcal{J}_s$ denote the set of all validated directions across modules. We rank these directions globally by

$$a_{k,j} = \left| g_{k,j} \widetilde{\delta}_{k,j} \right| = \frac{g_{k,j}^2}{f_{k,j} + \epsilon_s}, \tag{C.1}$$

and, for a retained fraction ρ, keep the top $K_\rho = \lceil \rho |\mathcal{J}_s| \rceil$ directions. Ties are resolved deterministically by the fully qualified module name and then by the original latent-direction index. Thus, the fraction is applied globally rather than independently within each module, and a module may retain no direction under a fractional setting.

For the amplify sweep, we vary the amplify fraction over $\{25\%, 50\%, 75\%, 100\%\}$ and then re-profile and execute the full suppress stage on the resulting amplify-updated model. For the suppress sweep, amplify is kept at 100%, and the fraction is applied to the validated suppress directions re-profiled after amplify. Both sweeps retain the standard scale-selection, positivity, and held-out acceptance procedures of LENS. Figure 8(a) reports WikiText-2 PPL and downstream average accuracy for both stages across all four backbones. Table 8 further provides the corresponding complete results on all seven downstream tasks. Overall, the preferred direction fraction varies across cal-

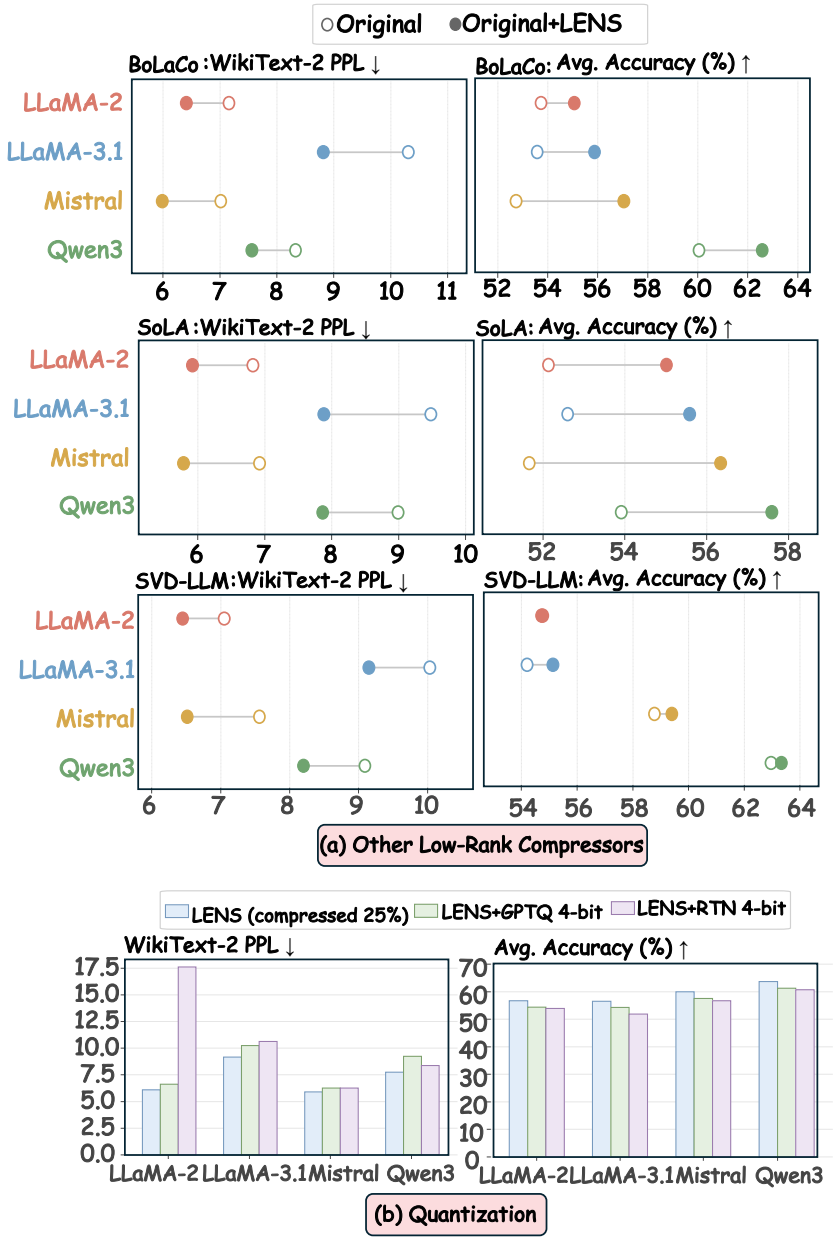

Figure 7: **Additional compatibility results of LENS.** (a) WikiText-2 PPL and downstream average accuracy when applying LENS on top of BoLaCo, SoLA, and SVD-LLM across four backbones. (b) WikiText-2 PPL and downstream average accuracy when models compressed by LENS at a 25% compression ratio are further quantized to 4 bits using GPTQ or RTN.

ibration stages and backbones, suggesting that the appropriate calibration strength is stage- and model-dependent.

Scale Sensitivity. We further report the complete results for the stage-wise scale η. Let $\eta_A^\star$ and $\eta_S^\star$ denote the amplify and suppress scales selected by the standard LENS pipeline, respectively. We evaluate multiplicative factors $c \in \{4, 2, 1, 0.5, 0.25\}$ around these selected operating points.

For the amplify sweep, each prescribed scale $c\eta_A^\star$ is applied only when it satisfies the standard positivity constraint for all retained amplify directions. The amplify acceptance gate is not re-applied for these prescribed sensitivity points. After the resulting amplify update, the model is re-profiled

Table 6: **Complete compatibility results with other low-rank compressors.** We apply LENS to BoLaCo, SoLA, and SVD-LLM across four backbones and report WikiText-2 perplexity (PPL), average downstream accuracy (Avg.), and accuracy on seven downstream tasks. For each compressor–backbone pair, the better result between the original compressed model and its LENS-calibrated counterpart is shown in **bold**.

	Overall		Downstream Tasks						
Method	**PPL↓**	**Avg.↑**	**MMLU**	**PIQA**	**WinoG.**	**HellaS.**	**ARC-e**	**ARC-c**	**OBQA**
LLaMA-2-7B									
BoLaCo	7.16	53.73	32.04	73.29	66.06	61.27	67.80	37.46	38.20
BoLaCo + LENS	**6.41**	**55.06**	**33.09**	**74.10**	**66.30**	**64.97**	**69.78**	**38.40**	**38.80**
SoLA	6.82	52.13	31.40	72.63	65.19	59.86	63.30	35.49	37.00
SoLA + LENS	**5.92**	**55.01**	**32.42**	**74.92**	**65.75**	**66.22**	**66.79**	**38.40**	**40.60**
SVD-LLM	7.05	**54.76**	29.04	**75.19**	**64.48**	68.56	**65.36**	**40.70**	40.00
SVD-LLM + LENS	**6.45**	54.72	**29.28**	74.86	63.22	**69.55**	**65.36**	40.19	**40.60**
LLaMA-3.1-8B									
BoLaCo	10.31	53.58	32.44	72.14	66.46	59.50	66.50	38.99	39.00
BoLaCo + LENS	**8.81**	**55.87**	**35.15**	**73.29**	**67.01**	**64.28**	**68.86**	**41.30**	**41.20**
SoLA	9.48	52.60	40.81	72.14	67.72	58.39	55.56	34.56	39.00
SoLA + LENS	**7.88**	**55.58**	**43.96**	**73.88**	**68.51**	**65.50**	**60.06**	**36.95**	**40.20**
SVD-LLM	10.03	54.20	**39.00**	72.47	**63.06**	65.38	64.35	39.76	35.40
SVD-LLM + LENS	**9.15**	**55.13**	38.87	**73.45**	**63.06**	**66.35**	**66.62**	**41.13**	**36.40**
Mistral-7B									
BoLaCo	7.01	52.73	29.28	72.96	67.48	59.02	68.22	36.35	35.80
BoLaCo + LENS	**5.98**	**57.04**	**34.28**	**75.03**	**67.72**	**67.75**	**72.90**	**42.83**	**38.80**
SoLA	6.92	51.66	33.25	70.24	68.19	55.24	62.79	35.49	36.40
SoLA + LENS	**5.78**	**56.34**	**38.38**	**73.45**	**69.22**	**65.21**	**68.10**	**41.21**	**38.80**
SVD-LLM	7.56	58.77	43.60	75.57	64.96	70.35	**74.12**	44.03	38.80
SVD-LLM + LENS	**6.51**	**59.40**	**44.40**	**76.50**	**65.98**	**72.36**	72.01	**45.14**	**39.40**
Qwen3-8B									
BoLaCo	8.33	60.04	53.48	75.30	66.61	64.49	75.21	47.01	38.20
BoLaCo + LENS	**7.56**	**62.57**	**57.98**	**76.33**	**68.03**	**68.92**	**77.10**	**50.43**	**39.20**
SoLA	8.99	53.91	49.68	69.75	66.61	53.37	63.76	36.60	37.60
SoLA + LENS	**7.86**	**57.59**	**53.41**	**71.55**	**68.03**	**61.45**	**68.22**	**42.06**	**38.40**
SVD-LLM	9.09	62.96	59.22	74.76	70.09	69.76	**73.48**	51.19	**42.20**
SVD-LLM + LENS	**8.20**	**63.33**	**59.63**	**74.92**	**71.51**	**71.00**	73.44	**51.62**	41.20

and the subsequent suppress stage follows the standard LENS procedure, including scale selection and held-out validation.

For the suppress sweep, we instead perform a post-hoc replay of the already folded terminal suppress update. Let $\mathbf{U}_A$ denote the frozen amplify checkpoint and $\mathbf{U}_\star$ the final A→S checkpoint obtained at the selected suppress scale $\eta_S^\star$. For a prescribed multiplier $c \in \{4, 2, 1, 0.5, 0.25\}$, we construct

$$\mathbf{U}(c) = \mathbf{U}^A + c(\mathbf{U}^\star - \mathbf{U}^A). \tag{C.2}$$

The point $\eta_S = \eta_S^\star$ therefore exactly recovers the accepted final model. For the off-center suppress points, direction profiling, scale selection, held-out acceptance, and the original direction-level positivity test are not rerun; these points serve as a post-hoc sensitivity analysis of the magnitude of the folded suppress update.

Figure 8(b) reports WikiText-2 PPL and downstream average accuracy for both stages across all four backbones, while Table 9 provides the complete task-level results. Overall, performance re-

Table 7: **Complete results with low-bit quantization.** We first compress each backbone with Duo-SVD at a 25% compression ratio and calibrate it with LENS, followed by 4-bit GPTQ or RTN quantization. We report WikiText-2 perplexity (PPL), average downstream accuracy (Avg.), and accuracy on seven downstream tasks.

	Overall		Downstream Tasks						
Setting	**PPL↓**	**Avg.↑**	**MMLU**	**PIQA**	**WinoG.**	**HellaS.**	**ARC-e**	**ARC-c**	**OBQA**
LLaMA-2-7B									
LENS (25%)	6.09	56.74	33.22	75.30	66.93	70.03	69.15	40.53	42.00
LENS + GPTQ-4bit	17.64	53.98	30.94	73.45	63.61	66.48	64.52	39.08	39.80
LENS + RTN-4bit	6.62	54.38	30.64	73.56	65.90	68.39	63.68	39.25	39.20
LLaMA-3.1-8B									
LENS (25%)	9.17	56.51	44.08	72.85	67.40	63.76	67.89	38.99	40.60
LENS + GPTQ-4bit	10.63	51.89	34.88	69.75	64.25	57.94	62.08	35.15	39.20
LENS + RTN-4bit	10.25	54.31	38.44	71.16	67.56	60.73	64.77	36.86	40.60
Mistral-7B									
LENS (25%)	5.90	60.02	41.43	77.75	69.06	73.00	72.31	44.62	42.00
LENS + GPTQ-4bit	6.27	56.78	34.33	76.55	66.93	69.90	68.39	42.75	38.60
LENS + RTN-4bit	6.27	57.59	35.25	76.82	67.32	71.96	69.91	41.89	40.00
Qwen3-8B									
LENS (25%)	7.76	63.76	60.13	76.33	67.32	68.43	80.22	53.67	40.20
LENS + GPTQ-4bit	8.38	60.77	55.01	74.43	67.72	66.81	74.12	46.93	40.40
LENS + RTN-4bit	9.25	61.29	51.20	74.65	66.93	67.10	76.64	50.94	41.60

mains relatively stable around the scale selected by LENS, while substantial rescaling can noticeably degrade some models.

C.4 Calibration Anatomy

We provide the complete calibration-anatomy results across all four backbones in Figure 9. For each backbone, we visualize the relative effective-weight change across decoder layers and linear-module families, separately for the amplify stage and the subsequent suppress stage. Across backbones, the calibration strength is clearly non-uniform across layers and module families, revealing structured and stage-specific adjustment patterns. amplify generally induces broader and stronger modifications, while the subsequent suppress stage makes smaller and more selective corrections.

C.5 Detailed Ablation Results

Table 10 provides the complete task-level results for the ablation variants evaluated in Table 2, including WikiText-2 PPL and accuracy on all seven downstream benchmarks across the four backbones. Overall, amplify contributes the larger individual PPL improvement, whereas suppress alone typically yields more modest gains. Compared with the joint A+S variant, sequential A→S achieves lower PPL on all four backbones and higher average downstream accuracy on three of four. Compared with the reversed S→A ordering, A→S yields higher average downstream accuracy on all four backbones, although S→A attains slightly lower PPL on Mistral-7B. Overall, these results support A→S as a favorable default trade-off rather than a universally dominant ordering on every metric.

D Additional Related Work

D.1 Low-rank Compression for LLMs

Low-rank decomposition has been widely explored as a structured approach to reducing the memory and computational costs of LLMs (Chiang et al., 2026). A major line of work improves decomposition fidelity by incorporating task- or data-dependent information into the low-rank approximation (Parkina & Rakhuba, 2026; Zhao et al., 2025; Liao et al., 2026; Makni et al., 2026; Gu et al.,

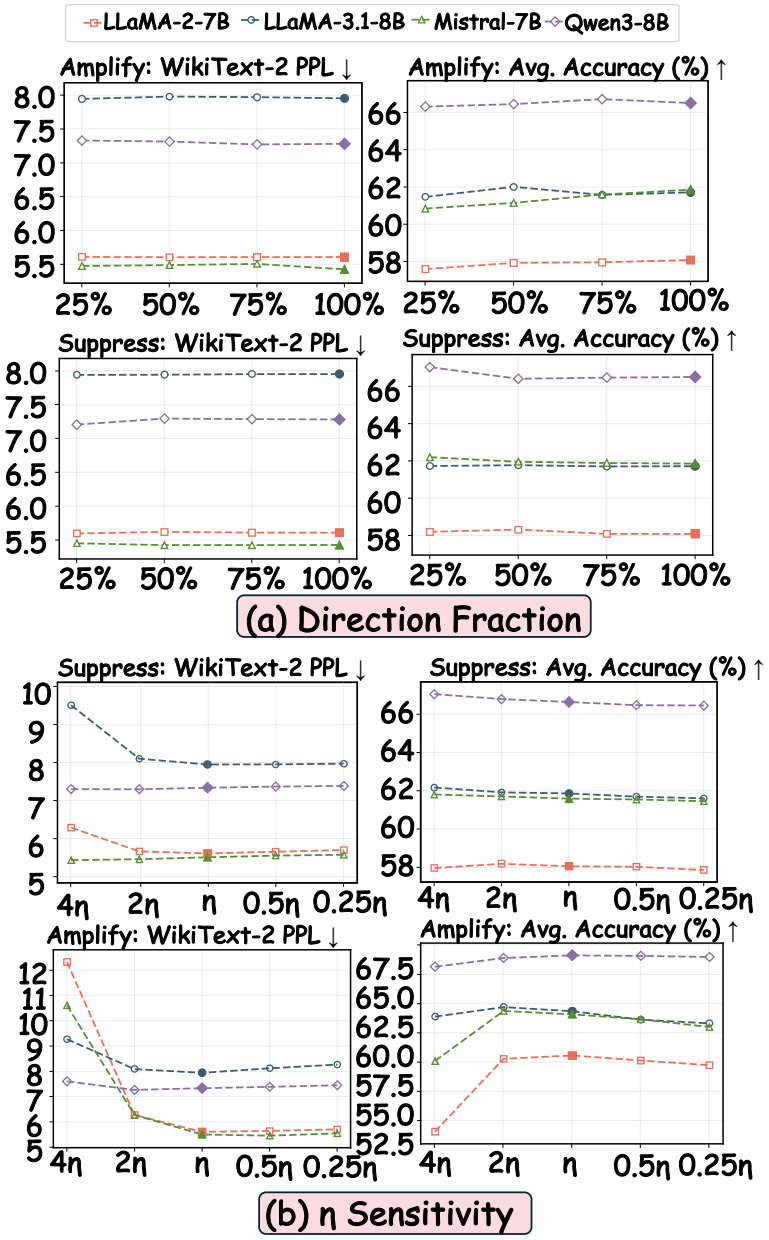

Figure 8: **Complete sensitivity analysis of LENS.** (a) WikiText-2 PPL and downstream average accuracy when retaining different fractions of validated directions. (b) WikiText-2 PPL and downstream average accuracy when varying the selected stage-wise scale over $\{4\eta, 2\eta, \eta, 0.5\eta, 0.25\eta\}$.

2025). FWSVD (Hsu et al., 2022) weights low-rank reconstruction according to Fisher information, thereby accounting for the task sensitivity of different parameters. ASVD (Yuan et al., 2025) further incorporates activation statistics into the decomposition process to better account for the input distribution encountered during inference. SVD-LLM (Wang et al., 2025c) similarly exploits activation information through truncation-aware transformations and supplements the resulting decomposition with sequential low-rank factor updates to mitigate truncation-induced reconstruction errors. More recently, COALA (Parkina & Rakhuba, 2026) develops a numerically stable and regularized formulation of context-aware low-rank approximation, avoiding explicit Gram-matrix inversion while

Table 8: **Complete results for direction-fraction sensitivity.** For the amplify sweep, we vary the retained fraction of validated amplify directions while keeping the suppress fraction at 100%. For the suppress sweep, we vary the suppress fraction while keeping the amplify fraction at 100%. We report WikiText-2 perplexity (PPL), average downstream accuracy (Avg.), and accuracy on seven downstream tasks. The default 100% setting is highlighted in blue.

		Overall		Downstream Tasks						
Stage	**Fraction**	**PPL↓**	**Avg.↑**	**MMLU**	**PIQA**	**WinoG.**	**HellaS.**	**ARC-e**	**ARC-c**	**OBQA**
LLaMA-2-7B										
amplify	25%	5.61	57.60	33.39	76.99	67.56	71.69	70.54	40.44	42.60
	50%	5.61	57.93	34.39	76.82	68.19	72.23	71.17	40.53	42.20
	75%	5.61	57.96	34.41	76.88	67.88	72.46	71.00	40.53	42.60
	100%	**5.61**	**58.07**	34.73	76.71	67.88	72.68	71.17	40.53	42.80
suppress	25%	5.60	58.20	34.70	76.93	68.11	72.98	71.42	40.61	42.60
	50%	5.62	58.31	34.73	77.15	68.43	72.97	71.30	41.21	42.40
	75%	5.61	58.09	34.67	77.04	68.03	72.81	71.30	40.19	42.60
	100%	**5.61**	**58.07**	34.73	76.71	67.88	72.68	71.17	40.53	42.80
LLaMA-3.1-8B										
amplify	25%	7.95	61.48	54.46	75.57	69.46	69.75	72.81	45.90	42.40
	50%	7.98	62.01	54.86	76.06	69.61	70.45	73.36	46.50	43.20
	75%	7.97	61.58	54.39	76.12	69.53	69.84	72.64	46.16	42.40
	100%	**7.95**	**61.86**	54.78	76.33	69.85	70.02	73.02	46.42	42.60
suppress	25%	7.94	61.73	54.71	76.28	69.38	69.93	73.02	46.59	42.20
	50%	7.94	61.77	54.79	76.22	69.38	69.95	73.19	46.67	42.20
	75%	7.95	61.71	54.76	76.06	69.53	69.90	73.06	46.42	42.20
	100%	**7.95**	**61.86**	54.78	76.33	69.85	70.02	73.02	46.42	42.60
Mistral-7B										
amplify	25%	5.48	60.85	41.92	78.51	69.53	74.99	73.32	45.48	42.20
	50%	5.49	61.15	42.46	78.56	69.46	75.74	73.70	46.16	42.00
	75%	5.51	61.60	43.33	78.67	70.09	76.06	73.78	47.10	42.20
	100%	**5.51**	**61.59**	42.99	78.45	70.40	76.27	74.20	47.01	41.80
suppress	25%	5.45	62.21	44.57	79.05	70.40	76.72	74.79	47.53	42.40
	50%	5.42	61.96	43.92	78.78	70.32	76.59	74.49	47.01	42.60
	75%	5.42	61.90	43.65	78.51	70.09	76.47	74.58	47.18	42.80
	100%	**5.51**	**61.59**	42.99	78.45	70.40	76.27	74.20	47.01	41.80
Qwen3-8B										
amplify	25%	7.33	66.31	64.11	78.18	70.40	71.29	81.73	55.89	42.60
	50%	7.32	66.45	64.33	78.40	70.40	71.43	81.90	55.89	42.80
	75%	7.27	66.72	64.47	78.45	69.85	72.45	82.41	56.40	43.00
	100%	**7.34**	**66.63**	64.26	78.51	70.56	71.77	81.94	56.57	42.80
suppress	25%	7.21	67.04	64.68	79.05	70.24	73.24	82.53	56.91	42.60
	50%	7.29	66.41	64.42	78.40	69.77	71.84	82.11	56.14	42.20
	75%	7.29	66.47	64.51	78.51	70.09	71.86	81.90	56.06	42.40
	100%	**7.34**	**66.63**	64.26	78.51	70.56	71.77	81.94	56.57	42.80

retaining activation-dependent reconstruction objectives. These methods improve the quality of the retained low-rank representation beyond conventional weight-only SVD.

Another line of work focuses on how the limited low-rank capacity should be distributed across heterogeneous model components. SVD-LLM V2 (Wang et al., 2025b) adapts singular-value truncation and compression ratios across components, while SoLA (Huang et al., 2025) exploits activation sparsity to guide low-rank compression. BoLaCo (Ji et al., 2024) formulates feature-aware rank allocation through Bayesian optimization, whereas Dobi-SVD (Wang et al., 2025a) introduces a differentiable formulation for optimizing SVD-based truncation. Beyond individual weight matrices, MoDeGPT (Lin et al., 2025) performs decomposition at the level of Transformer modules, and NSVD (Lu et al., 2025) develops a nested activation-aware decomposition strategy. Duo-SVD (Xv

Table 9: **Complete results for η sensitivity.** For each stage, η denotes the scale selected by `LENS`, and we evaluate $\{4\eta, 2\eta, \eta, 0.5\eta, 0.25\eta\}$. For the amplify sweep, after varying the amplify scale, the subsequent suppress stage is calibrated on the resulting amplify-updated model, following the sequential `LENS` pipeline. For the suppress sweep, the amplify stage is fixed and only the suppress scale is varied. We report WikiText-2 perplexity (PPL), average downstream accuracy (Avg.), and accuracy on seven downstream tasks. The default selected scale η is highlighted in blue.

		Overall		Downstream Tasks						
Stage	**Scale**	**PPL↓**	**Avg.↑**	**MMLU**	**PIQA**	**WinoG.**	**HellaS.**	**ARC-e**	**ARC-c**	**OBQA**
LLaMA-2-7B										
amplify	4η	12.33	51.57	29.30	72.09	63.38	66.06	58.63	35.75	35.80
	2η	6.27	57.79	36.29	76.33	66.77	73.25	69.70	40.78	41.40
	$\boldsymbol{\eta}$	**5.61**	**58.07**	34.73	76.71	67.88	72.68	71.17	40.53	42.80
	0.5η	5.64	57.64	33.59	76.93	68.27	71.79	70.24	40.27	42.40
	0.25η	5.71	57.24	32.97	76.61	67.48	71.17	70.20	39.85	42.40
suppress	4η	6.29	57.97	33.55	76.71	67.01	72.07	71.93	41.55	43.00
	2η	5.67	58.20	34.86	76.88	68.03	72.88	71.25	40.87	42.60
	$\boldsymbol{\eta}$	**5.61**	**58.07**	34.73	76.71	67.88	72.68	71.17	40.53	42.80
	0.5η	5.66	58.04	34.69	76.82	68.11	72.59	71.00	40.27	42.80
	0.25η	5.70	57.87	34.62	76.88	67.96	72.51	70.71	39.85	42.60
LLaMA-3.1-8B										
amplify	4η	9.27	61.40	55.28	75.79	68.90	71.98	71.38	45.48	41.00
	2η	8.10	62.21	55.93	76.12	69.14	71.19	73.06	46.84	43.20
	$\boldsymbol{\eta}$	**7.95**	**61.86**	54.78	76.33	69.85	70.02	73.02	46.42	42.60
	0.5η	8.13	61.14	53.91	75.68	69.46	69.21	72.22	45.31	42.20
	0.25η	8.27	60.82	53.47	75.46	69.53	68.76	71.72	44.80	42.00
suppress	4η	9.51	62.17	54.75	76.39	69.53	70.39	74.03	46.67	43.40
	2η	8.10	61.92	54.74	76.33	69.61	70.30	73.57	46.25	42.60
	$\boldsymbol{\eta}$	**7.95**	**61.86**	54.78	76.33	69.85	70.02	73.02	46.42	42.60
	0.5η	7.95	61.69	54.81	75.95	69.69	69.87	72.73	46.42	42.40
	0.25η	7.97	61.60	54.74	75.95	69.77	69.81	72.35	46.42	42.20
Mistral-7B										
amplify	4η	10.62	57.60	37.10	76.66	65.90	74.57	66.54	43.86	38.60
	2η	6.28	61.88	43.25	79.00	68.90	77.53	74.07	48.38	42.00
	$\boldsymbol{\eta}$	**5.51**	**61.59**	42.99	78.45	70.40	76.27	74.20	47.01	41.80
	0.5η	5.46	61.17	42.05	78.56	70.64	75.22	73.40	46.08	42.20
	0.25η	5.56	60.50	41.18	78.56	69.77	73.84	73.06	45.31	41.80
suppress	4η	5.43	61.80	43.28	78.56	70.48	76.33	74.41	46.76	42.80
	2η	5.46	61.70	43.01	78.40	70.72	76.34	74.33	46.93	42.20
	$\boldsymbol{\eta}$	**5.51**	**61.59**	42.99	78.45	70.40	76.27	74.20	47.01	41.80
	0.5η	5.55	61.55	42.84	78.51	70.32	76.31	74.16	47.10	41.60
	0.25η	5.58	61.46	42.80	78.51	69.85	76.24	74.12	47.10	41.60
Qwen3-8B										
amplify	4η	7.61	65.66	63.66	77.91	69.38	70.81	81.31	54.35	42.20
	2η	7.27	66.41	64.23	78.45	70.56	71.68	81.78	55.55	42.60
	$\boldsymbol{\eta}$	**7.34**	**66.63**	64.26	78.51	70.56	71.77	81.94	56.57	42.80
	0.5η	7.39	66.58	64.15	78.67	70.32	71.67	81.90	56.57	42.80
	0.25η	7.45	66.49	64.04	78.62	70.48	71.51	81.78	56.23	42.80
suppress	4η	7.30	67.04	64.31	78.62	70.40	73.02	82.37	57.17	43.40
	2η	7.30	66.78	64.32	78.56	70.01	72.23	82.15	57.00	43.20
	$\boldsymbol{\eta}$	**7.34**	**66.63**	64.26	78.51	70.56	71.77	81.94	56.57	42.80
	0.5η	7.37	66.46	64.31	78.51	70.24	71.49	81.94	56.14	42.60
	0.25η	7.38	66.45	64.25	78.56	70.40	71.44	81.82	56.06	42.60

et al., 2026), which serves as the default compressor in our experiments, jointly considers column-level preservation and module-level budget allocation to better exploit heterogeneous redundancy.

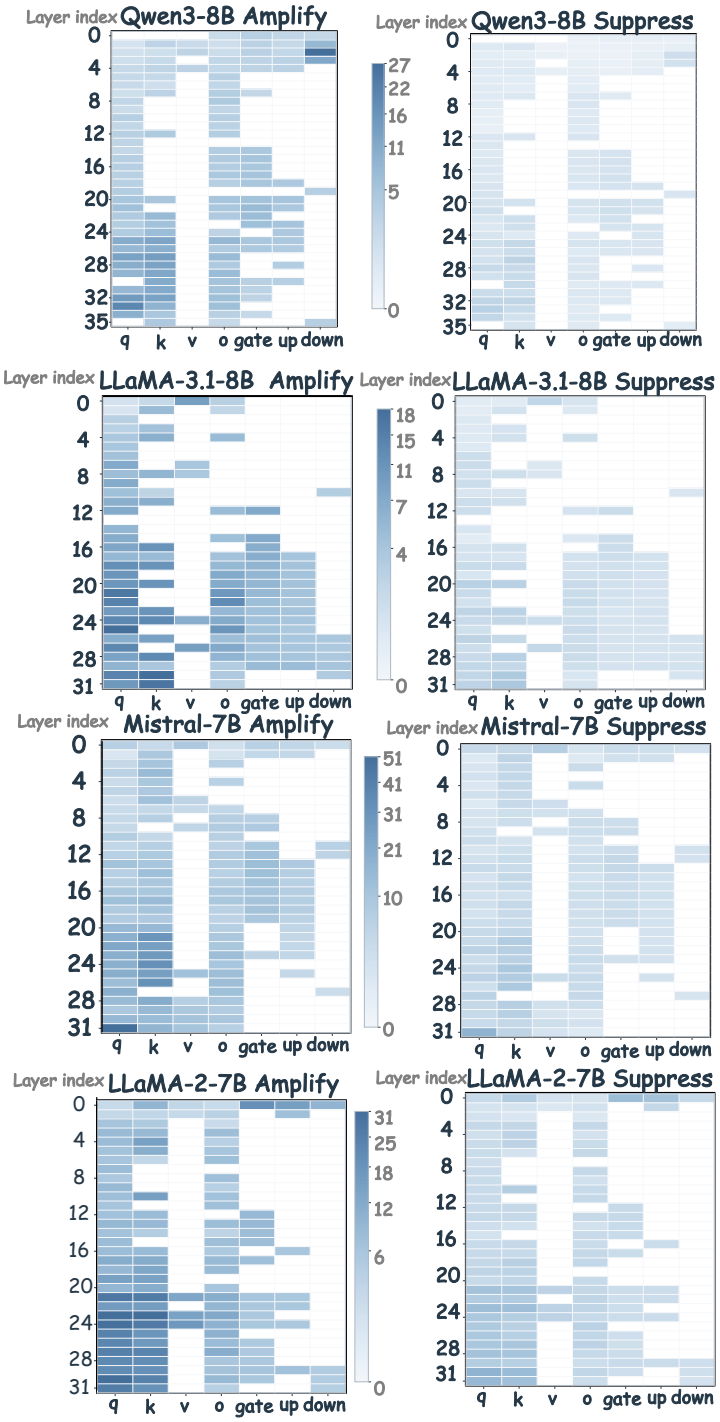

Figure 9: **Complete calibration anatomy of LENS across four backbones.** Each heatmap visualizes the relative effective-weight change across decoder layers and linear-module families induced by amplify or the subsequent suppress stage.

Recent studies have further incorporated model sensitivity into low-rank compression decisions. GRASP (Liu et al., 2025) uses adaptive singular parameters when replacing redundant layers, while Zero Sum SVD (Abbasi et al., 2026) explicitly considers loss sensitivity when determining low-rank compression. These developments increasingly move beyond purely reconstruction-based criteria

Table 10: **Detailed ablation results of LENS.** We report WikiText-2 perplexity (PPL), average downstream accuracy (Avg.), and accuracy on all seven downstream tasks for each ablation variant.

	Overall		Downstream Tasks						
Variant	**PPL↓**	**Avg.↑**	**MMLU**	**PIQA**	**WinoG.**	**HellaS.**	**ARC-e**	**ARC-c**	**OBQA**
LLaMA-2-7B									
Duo-SVD	5.95	56.59	32.09	76.55	67.56	69.97	69.32	39.08	41.60
+ amplify (A)	5.76	57.87	34.76	76.77	68.11	72.37	70.75	39.93	42.40
+ suppress (S)	5.83	56.67	31.95	76.66	67.40	70.14	69.61	39.33	41.60
+ Joint A+S	5.85	57.89	34.89	76.82	68.67	72.57	70.29	40.02	42.00
+ S→A	5.67	57.68	33.63	77.04	67.40	72.02	70.62	40.61	42.40
LENS (A→S)	**5.61**	**58.07**	34.73	76.71	67.88	72.68	71.17	40.53	42.80
Qwen3-8B									
Duo-SVD	7.54	66.52	63.87	78.62	70.80	71.54	81.78	56.23	42.80
+ amplify (A)	7.40	66.34	64.24	78.51	70.17	71.35	81.65	55.89	42.60
+ suppress (S)	7.52	66.45	63.84	78.67	70.64	71.43	81.65	56.14	42.80
+ Joint A+S	7.36	65.94	64.14	78.07	69.69	71.18	81.65	54.61	42.20
+ S→A	7.36	66.53	64.37	78.67	70.40	71.56	81.86	56.23	42.60
LENS (A→S)	**7.34**	**66.63**	64.26	78.51	70.56	71.77	81.94	56.57	42.80
Mistral-7B									
Duo-SVD	5.81	59.61	40.16	78.07	68.98	72.10	72.47	43.86	41.60
+ amplify (A)	5.60	61.42	42.66	78.67	69.93	76.25	74.07	46.93	41.40
+ suppress (S)	5.65	59.95	40.41	78.24	69.38	72.91	72.52	44.62	41.60
+ Joint A+S	5.74	**61.70**	43.51	78.94	70.09	76.54	73.61	47.18	42.00
+ S→A	**5.49**	61.31	42.29	78.73	69.93	75.45	73.91	46.25	42.60
LENS (A→S)	5.51	61.59	42.99	78.45	70.40	76.27	74.20	47.01	41.80
LLaMA-3.1-8B									
Duo-SVD	8.68	60.30	52.49	75.41	69.61	68.13	71.04	44.20	41.20
+ amplify (A)	8.00	61.63	54.81	75.90	70.09	69.77	72.10	46.33	42.40
+ suppress (S)	8.58	60.30	52.54	75.41	69.61	68.20	70.92	44.20	41.20
+ Joint A+S	8.00	61.65	54.74	76.22	69.93	69.72	72.47	46.08	42.40
+ S→A	8.03	61.73	54.58	76.17	69.61	69.98	73.06	45.90	42.80
LENS (A→S)	**7.95**	**61.86**	54.78	76.33	69.85	70.02	73.02	46.42	42.60

by incorporating information related to model behavior or task loss. Nevertheless, their primary objective remains to determine the compressed representation itself, including which structures, ranks, or spectral components should be retained. In contrast, LENS takes an already compressed model as given and addresses the question of how the contributions of directions within the retained subspace should be recalibrated with respect to the end-to-end language-modeling objective.

D.2 Post-compression Refinement and Calibration

Beyond determining the compressed representation itself, several methods further refine the model after low-rank decomposition (Shen et al., 2025). SVD-LLM (Wang et al., 2025c) complements its truncation-aware decomposition with a layer-wise parameter update to compensate for the accuracy degradation introduced by singular-value truncation. AdaSVD (Li et al., 2025) further develops this direction by adaptively compensating for truncation errors through alternating updates of the low-rank factors. These approaches therefore improve the compressed model by directly refining the retained factorization after decomposition.

A different line of work compensates for information discarded during compression rather than directly modifying the retained factors (Mozaffari et al., 2024). EoRA (Liu et al., 2024) introduces training-free low-rank residual paths to reconstruct compression-induced errors, projecting the residuals into an activation eigenspace so that more important error components receive greater represen-

tation capacity. This strategy complements the compressed operator with an additional residual correction.

More closely related to **LENS** are approaches that directly refine spectral coefficients. In broader generative-model compression, SVS (Kim & Yoo, 2025) rescales the singular values of pruned weights to reduce spectral imbalance and improve subsequent refinement. For LLM compression, SVD-Surgeon (Safari & Hutter, 2026) treats singular values as optimization variables and derives second-order loss-aware updates for the retained singular values to compensate for truncation. Despite their different objectives, these methods perform refinement along spectral directions predetermined by the underlying singular-value decomposition. In contrast, **LENS** does not restrict calibration to the original singular directions. It searches within the full retained output subspace for directions whose rescaling is aligned with the end-to-end language-modeling loss, and then determines their direction-specific relative gains together with a validated stage-level scale.