CORTEX: COMPRESSING ROUTED EXPERTS FOR MOE-BASED LLMS UNDER THE BLOCK LOSS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixture-of-Experts (MoEs) scale the sizes of LLMs, thus studying their compression becomes crucial. One distinctive feature of MoEs is that the experts of a layer are redundant in their weights and correlated in their routing: a token activates several of them at once, and the subsequent layer receives only their gated sum. Existing low-rank compression methods exploit the redundancy through parameters shared across experts, but the routing correlation rarely enters their objectives: each expert is fitted on its own inputs. In this work, we propose to fit the low-rank compression under the error of the routed MoE block, in which the errors of co-routed experts are coupled. The exact second-order metric of this loss, however, couples all experts and admits no closed-form truncation. Besides, its per-expert blocks are estimated from few tokens and overfit the calibration set. Our method therefore uses two surrogates of this metric, one keeping the co-routing interference and the other the per-expert input covariances. Under the surrogate metrics with a Stein-type shrinkage, each component of the low-rank decomposition is fitted in a closed form, and the output factors of all experts are then re-solved jointly under the exact metric. Finally, a role-split rank allocation from a first-order analysis of the block output completes the method. We demonstrate the method at 20–60% compression on Mixtral-8×7B, Qwen2-57B-A14B, and DeepSeekMoE-16B-Base, where it achieves the strongest performance in language modeling and zero-shot reasoning in most settings.

1 INTRODUCTION

Mixture-of-Experts (MoE) architectures scale the sizes of large language models (LLMs) at a fixed per-token compute: each token is routed to a few of K expert feed-forward networks, so the parameter count grows with K while the activated compute does not (Jiang et al., 2024; Dai et al., 2024; Yang et al., 2024a; Liu et al., 2024a). Despite their success, storing and deploying MoEs require high memory cost. Therefore, studying the compression of MoEs becomes crucial. In this work, we focus on a class of training-free low-rank decomposition methods (*e.g.*, Li et al., 2025), because they keep every expert and compress its matrices at a low cost.

One distinctive feature of MoEs is that the experts of a layer are related in two ways. First, their weights are redundant. Previous methods mainly study how to effectively leverage this redundancy for a better compression through different weight-sharing patterns (Li et al., 2025; Gu et al., 2025; Chen et al., 2026; XU et al., 2026). Second, the experts are also correlated in their routing: a token activates several of them at once, and the next layer receives only the gated sum. This second relation rarely enters the objective in previous methods: each expert is fitted on its own activations as if it were a dense layer, and the shared component and the truncation metric are chosen heuristically. In that gated sum, however, the errors of co-routed experts can cancel or accumulate, which a per-expert objective cannot see. A per-expert objective can therefore prefer the compression with the larger block error, and the coupling it drops grows with the number of experts a token activates.

To address this issue, we propose to fit every component of the low-rank decomposition under the error of the routed MoE block (the block loss), in which the errors of co-routed experts are coupled through the gates. Doing so directly is not straightforward. The exact second-order metric of this loss is a $Kd_{\text{in}} \times Kd_{\text{in}}$ matrix with two structures, the per-expert input covariances on its diagonal and the co-routing interference off it. This metric couples all experts and admits no closed-form

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.

truncation. Moreover, its per-expert blocks are estimated from the few tokens routed to each expert, so that matching it exactly over-fits the calibration set. A surrogate of the exact metric is therefore needed for each component of the decomposition: it must keep the structure that decides the error of that component, be estimated reliably from the few tokens per expert, and admit a closed-form fit.

We build such surrogates for the delta decomposition (Gu et al., 2025), a dense base shared by the experts plus a low-rank residual per expert. Our method, CORTEX (**CO**mpressing **R**ou**T**ed **EX**perts), uses two surrogates that each keep one of the two structures. In stage one, the shared base is fitted in closed form under the Kronecker surrogate, which keeps the co-routing interference. Each private residual is truncated under the block-diagonal surrogate, which keeps the shrunk per-expert covariances. In stage two, since the error is linear once the input factors are fixed, the output factors are re-solved jointly under the exact metric. Finally, we propose a role-split rank allocation that moves rank budget toward down projections by one ratio shared across layers, motivated by a first-order analysis of the block output. We demonstrate the method at 20–60% *whole-model* compression ratios on Mixtral-8×7B (Jiang et al., 2024), Qwen2-57B-A14B (Yang et al., 2024a), and DeepSeekMoE-16B-Base (Dai et al., 2024), evaluated on three language modeling and seven zero-shot reasoning datasets. Compared with five recent baselines, our method attains the lowest perplexity and the highest zero-shot accuracy in most settings.

1.1 Contributions

- We formulate MoE compression under the routed block loss and introduce two surrogates of its metric: the *Kronecker* surrogate keeps the co-routing interference and the *block-diagonal* surrogate keeps the per-expert covariances (Sec. 3).
- We derive closed-form fits of the delta decomposition under these surrogates, the *routed centroid* base and the shrunk per-expert truncations, and a joint re-solve of the down output factors under the exact metric (Secs. 4.1 to 4.3).
- We propose a role-split rank allocation from a first-order analysis of the block output (Sec. 4.4).

1.2 Related work

Compression of dense LLMs. Training-free compression of dense LLMs uses a small calibration set, by quantization (Frantar et al., 2022; Lin et al., 2024), pruning (Frantar & Alistarh, 2023b; Sun et al., 2024; Ma et al., 2023; Ashkboos et al., 2024), or low-rank factorization, which is the line of this work. Earlier low-rank methods focus on plain SVD with fine-tuning after compression (*e.g.*, Kim et al., 2015), which do not fit for large-scale LLMs. The remedy is to weight the factorization by data-dependent statistics such as Fisher information (Hsu et al., 2022) and activations (Denton et al., 2014; Yuan et al., 2023). Recently, SVD-LLM (Wang et al., 2025b) whitens the weights by the input covariance so that truncating the whitened singular values minimizes the function error exactly. Subsequent work refines the objective: cross-layer sharing of basis vectors (Wang et al., 2025a), anchoring each layer to the original outputs under the inputs of the compressed upstream layers (Sinha & Fleuret, 2026; Hu et al., 2026), an objective that our second stage shares within one layer but couples across co-routed experts, and tensor decompositions of stacked weights (Zagitov et al., 2026). These methods treat a matrix as a dense layer with a single input distribution. However, they do not consider the redundancy and correlations across MoE blocks.

Compression of MoE-based LLMs. The compression of the MoE block has been studied at three granularities. Expert-level methods remove or merge whole experts via expert pruning (Lu et al., 2024; Liu et al., 2024b; Lasby et al., 2026; Tseng et al., 2026) and merging (Chen et al., 2025; Li et al., 2024b). Among these, the concurrent HOPE (Tseng et al., 2026) also starts from the squared block-output error, but reduces it to pairwise gated output norms and only decides which whole experts to delete. Weight-level methods keep the experts and reduce their storage through techniques like quantization (Kim et al., 2023; Frantar & Alistarh, 2023a; Li et al., 2024a) and unstructured pruning with router-informed scores (Xie et al., 2024). This work lies in the third granularity, that keeps every expert and compresses the expert matrices by designing different low-rank factorizations (Yang et al., 2024b; Li et al., 2025; Gu et al., 2025; Chen et al., 2026; Mi et al., 2026; XU et al., 2026). These methods differ in what they share but fit it under routing-blind objectives, the weight- or activation-space error of each expert on its own tokens. We instead keep the low-rank structure and fit every component under the routed block loss, which couples the co-routed experts.

2 BACKGROUND

Mixture-of-experts. In MoE-based LLMs, the output of an attention layer, $\boldsymbol{x} \in \mathbb{R}^{d_{\text{in}}}$, is fed into an MoE feed-forward network (FFN) layer,

$$\text{MoE-FFN}(\boldsymbol{x}) = \sum_{k=1}^{K} g_k(\boldsymbol{x})\, e_k(\boldsymbol{x}), \tag{1}$$

where K is the number of experts, $g_k(\cdot)$ is the router weight of the k-th expert, and $e_k(\cdot)$ is the k-th expert function. The router is a linear layer followed by a softmax and a top-k selection, so that $\boldsymbol{g}(\boldsymbol{x}) = [g_1(\boldsymbol{x}), \ldots, g_K(\boldsymbol{x})]^\top$ is sparse. Each expert is a gated MLP,

$$e_k(\boldsymbol{x}) = \boldsymbol{W}_k^{\text{down}}\big(\text{SiLU}(\boldsymbol{W}_k^{\text{gate}}\boldsymbol{x}) \odot (\boldsymbol{W}_k^{\text{up}}\boldsymbol{x})\big), \tag{2}$$

where $\boldsymbol{W}_k^{\text{gate}}, \boldsymbol{W}_k^{\text{up}} \in \mathbb{R}^{d_{\text{mid}} \times d_{\text{in}}}$ and $\boldsymbol{W}_k^{\text{down}} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{mid}}}$, $\odot$ denotes the Hadamard product. The expert matrices dominate the parameter count of MoE-based LLMs and are the memory bottleneck; thus, following prior work (Li et al., 2025; Gu et al., 2025), we only compress MoE FFN layers.

Training-free low-rank compression of LLMs. One effective training-free technique for a linear layer $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ is the whitened truncation of SVD-LLM (Wang et al., 2025b). Given a set of calibration activations $\boldsymbol{X} \in \mathbb{R}^{d_{\text{in}} \times N}$, it finds the rank-$r$ approximation that minimizes the *activation-aware* error $\|\boldsymbol{W}\boldsymbol{X} - \hat{\boldsymbol{W}}\boldsymbol{X}\|_F^2$,

$$\hat{\boldsymbol{W}}_* = \text{SVD}_r(\boldsymbol{W}\boldsymbol{\Sigma}^{1/2}) \cdot \boldsymbol{\Sigma}^{-1/2}, \qquad \boldsymbol{\Sigma} = \boldsymbol{X}\boldsymbol{X}^\top. \tag{3}$$

where $\text{SVD}_r(\cdot)$ is the rank-r SVD. After whitening, the reconstruction error equals the function error, so truncating the smallest singular values is exact. Applied to each expert in isolation, however, Eq. (3) degrades quickly on MoEs (Li et al., 2025). Existing methods exploit the redundancy across experts through a *structure* shared within the layer (Li et al., 2025; Chen et al., 2026; XU et al., 2026), but still fit every factor under Eq. (3) one expert at a time, so the routing correlation stays out of the *objective*. We keep the structure and change the objective. The structure is the *delta decomposition* of D^2-MoE (Gu et al., 2025), a dense base $\boldsymbol{B}$ shared by the K experts plus a low-rank *residual* per expert (D^2-MoE's *delta*),

$$\boldsymbol{W}_k = \boldsymbol{B} + \boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^\top, \qquad k = 1, \ldots, K. \tag{4}$$

D^2-MoE obtains $\boldsymbol{B}$ by Fisher merging (Matena & Raffel, 2022) and truncates each residual under Eq. (3) on its own expert's tokens. We fit every component of Eq. (4) under the routed block loss instead, which couples the errors of co-routed experts through the gates and lifts the identity of Eq. (3) from one matrix to the routed block of K experts in Secs. 3 and 4.

3 THE BLOCK LOSS AND ITS TWO SURROGATES

3.1 THE BLOCK LOSS

Let $\mathcal{D}_c$ denotes the calibration distribution of input activations. For a linear role (up/gate/down in Eq. (2)), the error of the MoE block output on the calibration distribution is the *routed block loss*,

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_c}\Big\| \sum_{k=1}^{K} g_k(\boldsymbol{x})\, \Delta\boldsymbol{W}_k\, \boldsymbol{x}\Big\|_2^2 = \sum_{i,j=1}^{K} \text{Tr}\left(\Delta\boldsymbol{W}_i\, \boldsymbol{M}_{i,j}\, \Delta\boldsymbol{W}_j^\top\right), \tag{5}$$

where $\boldsymbol{M}_{i,j} := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_c}\left[g_i(\boldsymbol{x})\, g_j(\boldsymbol{x})\, \boldsymbol{x}\boldsymbol{x}^\top\right] \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}$ and $\Delta\boldsymbol{W}_k := \boldsymbol{W}_k - \hat{\boldsymbol{W}}_k$.

Stacking the errors side by side, $\Delta\boldsymbol{W} := [\Delta\boldsymbol{W}_1, \ldots, \Delta\boldsymbol{W}_K] \in \mathbb{R}^{d_{\text{out}} \times K d_{\text{in}}}$, gives $\mathcal{L} = \text{Tr}(\Delta\boldsymbol{W}\, \boldsymbol{M}\, \Delta\boldsymbol{W}^\top)$ with the exact metric $\boldsymbol{M} := [\boldsymbol{M}_{i,j}]_{i,j=1}^{K} \in \mathbb{R}^{K d_{\text{in}} \times K d_{\text{in}}}$, which couples two structures. Its diagonal blocks $\boldsymbol{M}_{k,k} = \mathbb{E}[g_k^2(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\top] \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}$ are the *routed* per-expert input covariances, computed on the tokens the expert receives and in proportion to its gate; the routing-blind objective of previous methods (*e.g.*, Gu et al., 2025) keeps only these blocks, with $g_k \equiv 1$. Its off-diagonal blocks carry the *co-routing interference*: on a token routed to both i and j the block emits the gated *sum* of the two errors, which can cancel or amplify (Sec. B.1). $\mathcal{L} = \|\Delta\boldsymbol{W}\, \boldsymbol{M}^{1/2}\|_F^2$ is a whitened norm as in Eq. (3), but $\boldsymbol{M}^{1/2}$ mixes the expert and input indices, so the truncation of the K experts under it is not a whitened SVD and has no closed form. Moreover, it is hard to estimate due to the curse-of-dimensionality (Sec. D.1).

3.2 TWO SURROGATES

We use two surrogates of $\boldsymbol{M}$, each of which keeps one of its two structures exactly and simplifies the other. Let $\boldsymbol{\Gamma} := \mathbb{E}[\boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^\top] \in \mathbb{R}^{K\times K}$ be the gate Gram matrix and $\bar{\boldsymbol{\Sigma}} := \sum_k \boldsymbol{M}_{k,k}/\operatorname{Tr}(\boldsymbol{\Gamma})$ the pooled routed input covariance, the gate-weighted mean of the per-expert covariance $\boldsymbol{M}_{k,k}/\boldsymbol{\Gamma}_{kk}$.

The Kronecker surrogate. The coarsest surrogate that keeps the *co-routing structure* is

$$\boldsymbol{M}_{i,j} \;\approx\; \boldsymbol{\Gamma}_{ij}\,\bar{\boldsymbol{\Sigma}}, \qquad \textit{i.e.,} \qquad \boldsymbol{M} \approx \boldsymbol{\Gamma} \otimes \bar{\boldsymbol{\Sigma}}. \tag{6}$$

It retains the full second moment of the gates, including all co-routing correlations, and pools only the per-expert input covariances $\boldsymbol{M}_{k,k}/\boldsymbol{\Gamma}_{kk}$ across experts.

Proposition 1 (Whitened Frobenius norm)**.** *Under* covariance homogeneity, *i.e.,* $\boldsymbol{M}_{i,j} = \boldsymbol{\Gamma}_{ij}\,\bar{\boldsymbol{\Sigma}}$ *for all* $i, j \in [K]$ *with a single SPD matrix* $\bar{\boldsymbol{\Sigma}}$*, for every error tensor* $\Delta\mathbf{T} \in \mathbb{R}^{K\times d_{out}\times d_{in}}$ *with slices* $\Delta\boldsymbol{W}_k$*, we have*[1]

$$\mathcal{L}(\Delta\mathbf{T}) \;=\; \left\|\Delta\mathbf{T} \times_1 \boldsymbol{\Gamma}^{1/2} \times_3 \bar{\boldsymbol{\Sigma}}^{1/2}\right\|_F^2 \;=:\; \hat{\mathcal{L}}(\Delta\mathbf{T}). \tag{7}$$

Proposition 1 is the routed analogue of the SVD-LLM identity Eq. (3): the *input* mode is whitened by $\bar{\boldsymbol{\Sigma}}$ and the *expert* mode by $\boldsymbol{\Gamma}$, whose off-diagonal is exactly the interference term. Setting $\boldsymbol{\Gamma} = \boldsymbol{I}$ recovers the routing-blind objective of XU et al. (2026), which therefore coincides with the routed loss only under covariance homogeneity with uncorrelated, equal-mass gates.

The block-diagonal surrogate. The second surrogate,

$$\hat{\boldsymbol{M}} = \operatorname{blockdiag}(\boldsymbol{M}_{1,1}, \ldots, \boldsymbol{M}_{K,K}), \tag{8}$$

is exact on the diagonal and zero off it. It does not factorize across the expert and input modes, so it admits no mode-wise whitening as in Eq. (7), but it is separable across experts: a problem whose degrees of freedom are private to one expert splits into K independent, exactly solvable pieces.

Considering a structure with shared and private components, *e.g.* Eq. (4), Sec. 4 fits each under the surrogate that keeps the structure of $\boldsymbol{M}$ it reads: the shared base under the Kronecker surrogate (Sec. 4.1), each private residual under the block-diagonal surrogate, shrunk toward the pooled covariance (Sec. 4.2), and the exact metric where the remaining unknowns are linear (Sec. 4.3).

4 COMPRESSION UNDER THE ROUTED METRIC

We now derive the method, **CORTEX** (**CO**mpressing **R**ou**T**ed **EX**perts). Its structure is the delta decomposition in Eq. (4), but every component is obtained from the block loss of Sec. 3; Alg. 1 in Sec. C summarizes the pipeline. Compression acts independently on each *item* $t = (\ell, p)$, a layer ℓ and a role p (gate/up/down); we may omit the item index when it is clear from context.

Profiling. We collect the following statistics through a single forward pass over the calibration dataset $\mathcal{D}_c$: the gate Gram $\boldsymbol{\Gamma}^{(\ell)}$ for all layers and the routed per-expert covariances $\boldsymbol{M}^{(t)}_{k,k} = \mathbb{E}[g_k(\boldsymbol{x})^2\,\boldsymbol{x}^{(t)}_k(\boldsymbol{x}^{(t)}_k)^\top]$ for all items. The SiLU-gain grams of Sec. 4.4 and the refit moments of Sec. 4.3 are accumulated in two further forward passes (Sec. D.2).

4.1 THE SHARED BASE UNDER THE KRONECKER SURROGATE

Stack the experts as $\boldsymbol{W} := [\boldsymbol{W}_1, \ldots, \boldsymbol{W}_K]$ and write $\mathbf{1}^\top \otimes \boldsymbol{B} = [\boldsymbol{B}, \ldots, \boldsymbol{B}]$ for the shared base in Eq. (4). The base is fitted first, by minimizing the block loss Eq. (5) under the Kronecker surrogate Eq. (6). What decides its surrogate is the error a shared component makes. Denoting the deviation of base $\boldsymbol{B}$ from the optimal one as $\boldsymbol{D} := \boldsymbol{B} - \boldsymbol{B}^\star$, the block loss of the shift is $\operatorname{Tr}\left(\boldsymbol{D}\left(\sum_{i,j}\boldsymbol{M}_{i,j}\right)\boldsymbol{D}^\top\right)$. The off-diagonal part is the co-routing interference; a surrogate that drops it, as the block-diagonal one Eq. (8) does, fits the base under $\sum_k \boldsymbol{M}_{k,k}$ instead and returns a different base. Thus a surrogate for the base must keep the co-routing structure such as the Kronecker surrogate.

[1] Mode-d product: for $\mathbf{T} \in \mathbb{R}^{I_1\times\cdots\times I_D}$ and $\boldsymbol{A} \in \mathbb{R}^{J\times I_d}$, $(\mathbf{T} \times_d \boldsymbol{A})_{i_1,\ldots,j,\ldots,i_D} = \sum_{i_d} \mathbf{T}_{i_1,\ldots,i_D}\,\boldsymbol{A}_{j,i_d}$.

Proposition 2 (Routed centroid). *Let* $\|\boldsymbol{E}\|^2_{\boldsymbol{\Gamma},\bar{\boldsymbol{\Sigma}}} := \hat{\mathcal{L}}(\boldsymbol{E}) = \sum_{i,j} \boldsymbol{\Gamma}_{ij} \operatorname{Tr}(\boldsymbol{E}_i \bar{\boldsymbol{\Sigma}} \boldsymbol{E}_j^\top)$ *be the surrogate loss Eq. (7), with* $\bar{\boldsymbol{\Sigma}} \succ 0$ *and* $\mathbf{1}^\top \boldsymbol{\Gamma} \mathbf{1} > 0$. (i) *The unique minimizer of* $\|\boldsymbol{W} - \mathbf{1}^\top \otimes \boldsymbol{B}\|^2_{\boldsymbol{\Gamma},\bar{\boldsymbol{\Sigma}}}$ *is*

$$\boldsymbol{B}^\star \;=\; \sum_{k=1}^{K} c_k \, \boldsymbol{W}_k, \qquad \boldsymbol{c} \;=\; \frac{\boldsymbol{\Gamma}\mathbf{1}}{\mathbf{1}^\top \boldsymbol{\Gamma} \mathbf{1}}. \tag{9}$$

(ii) *The residuals* $\boldsymbol{R}_k := \boldsymbol{W}_k - \boldsymbol{B}^\star$*, stacked as* $\boldsymbol{R}$*, are* routing-centered, $\sum_k (\boldsymbol{\Gamma}\mathbf{1})_k \boldsymbol{R}_k = \mathbf{0}$*, and for every* $\boldsymbol{B}$ *and every routing-centered residual approximation* $\hat{\boldsymbol{R}}$ *the loss splits as* $\|\boldsymbol{W} - \mathbf{1}^\top \otimes \boldsymbol{B} - \hat{\boldsymbol{R}}\|^2_{\boldsymbol{\Gamma},\bar{\boldsymbol{\Sigma}}} = \|\boldsymbol{R} - \hat{\boldsymbol{R}}\|^2_{\boldsymbol{\Gamma},\bar{\boldsymbol{\Sigma}}} + (\mathbf{1}^\top \boldsymbol{\Gamma} \mathbf{1}) \, \|(\boldsymbol{B} - \boldsymbol{B}^\star) \, \bar{\boldsymbol{\Sigma}}^{1/2}\|^2_F$.

The base is thus the orthogonal projection of the expert stack onto the common-expert subspace in the routed geometry, and the residual lives in its $\boldsymbol{\Gamma}$-orthogonal complement: base and residuals do not compete for the same error directions. The second term of (ii) is the block loss $\operatorname{Tr}\left(\boldsymbol{D} \left(\sum_{i,j} \boldsymbol{M}_{i,j}\right) \boldsymbol{D}^\top\right)$ of the shift $\boldsymbol{D}$ above, evaluated under the surrogate. With renormalized gates (*e.g.*, Mixtral-8x7B), $(\boldsymbol{\Gamma}\mathbf{1})_k = \mathbb{E}[g_k]$ and $\mathbf{1}^\top \boldsymbol{\Gamma} \mathbf{1} = 1$, so $c_k = \mathbb{E}[g_k(\boldsymbol{x})]$ and the base is the *gate-weighted* mean expert; without renormalization (*e.g.*, Qwen2-57B-A14B), $c_k \propto \mathbb{E}[g_k \sum_i g_i]$ also depends on the co-routing correlations, and the routing-blind endpoint $\boldsymbol{\Gamma} = \boldsymbol{I}$ gives the plain average. Unlike the Fisher-merged base of D^2-MoE, which requires gradients over the calibration set and is tied to no block-level objective, Eq. (9) is the optimal shared matrix under the Kronecker surrogate of the block loss and needs no backward pass.

4.2 The private residuals under the shrunk diagonal blocks

The residuals $\boldsymbol{R}_k = \boldsymbol{W}_k - \boldsymbol{B}^\star$ are compressed by independent per-expert low-rank factors,

$$\hat{\boldsymbol{W}}_k \;=\; \boldsymbol{B}^\star \;+\; \boldsymbol{P}_k \boldsymbol{Q}_k^\top, \qquad \boldsymbol{P}_k \in \mathbb{R}^{d_{\text{out}} \times L_k}, \;\; \boldsymbol{Q}_k \in \mathbb{R}^{d_{\text{in}} \times L_k}. \tag{10}$$

Similar to Sec. 4.1, what decides the surrogate for these factors is the error they make. Denoting $\boldsymbol{E}_k := \boldsymbol{R}_k - \boldsymbol{P}_k \boldsymbol{Q}_k^\top$, the block loss contains it through its own block $\operatorname{Tr}(\boldsymbol{E}_k \boldsymbol{M}_{k,k} \boldsymbol{E}_k^\top)$ and cross terms $\operatorname{Tr}(\boldsymbol{E}_k \boldsymbol{M}_{k,j} \boldsymbol{E}_j^\top)$. The own block decides the input directions the truncation keeps, whereas the cross terms only couple the K truncations. The block-diagonal surrogate keeps the own block exactly and drops the coupling. The Kronecker surrogate pools the own blocks. On all three models the pooling error exceeds the dropped coupling by an order of magnitude (Prop. 4, Sec. B.2). One reason for this gap may lie in Prop. 2: the common component, on which the cross terms add coherently, has already been removed by the base, and the residual is $\boldsymbol{\Gamma}$-orthogonal to it. Thus the surrogate here must keep each expert's own covariance, such as the block-diagonal surrogate Eq. (8).

In the profiling stage, $\boldsymbol{M}_{k,k} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}$ is estimated on average a k_{top}/K fraction of the calibration data. To avoid the curse-of-dimensionality, one remedy is the Stein-type shrinkage of covariance estimates (Ledoit & Wolf, 2004), classically toward the pooled covariance (Friedman, 1989). In our case, the target is at hand: $\boldsymbol{\Gamma}_{kk} \bar{\boldsymbol{\Sigma}}$, the diagonal block of the Kronecker surrogate, is estimated from every token. For expert k we therefore use

$$\boldsymbol{\Sigma}_k(\mu) \;=\; (1 - \mu) \, \boldsymbol{M}_{k,k} \;+\; \mu \, \boldsymbol{\Gamma}_{kk} \, \bar{\boldsymbol{\Sigma}}, \qquad \mu \in [0, 1], \tag{11}$$

the convex combination of the diagonal blocks of the two surrogates. With $\boldsymbol{S}_k = \operatorname{chol}(\boldsymbol{\Sigma}_k(\mu) + \delta \boldsymbol{I})$, the minimizer of the loss under $\operatorname{blockdiag}(\boldsymbol{\Sigma}_k(\mu))$ over Eq. (10), folded back to native coordinates,

$$\boldsymbol{R}_k \boldsymbol{S}_k \;\approx\; \tilde{\boldsymbol{P}}_k \tilde{\boldsymbol{Q}}_k^\top \quad (\text{rank } L_k), \qquad \boldsymbol{P}_k = \tilde{\boldsymbol{P}}_k, \qquad \boldsymbol{Q}_k = \boldsymbol{S}_k^{-\top} \tilde{\boldsymbol{Q}}_k. \tag{12}$$

This completes stage one: for gate/up these are the final factors, whereas for down only $\boldsymbol{Q}_k$ is final and stage two (Sec. 4.3) re-solves $\boldsymbol{P}_k$ under the exact metric. Table 4 confirms the estimation reading: mixing in the pooled covariance gives the better compression although its error on the calibration statistics is *higher*, by a margin that narrows as $\mathcal{D}_c$ grows.

4.3 The down output factors under the exact metric

For the down projection, whose output is the gated sum $\sum_k g_k(\boldsymbol{x}) \, \boldsymbol{W}_k \boldsymbol{h}_k(\boldsymbol{x})$ over the active experts, stage two restores the cross terms $\operatorname{Tr}(\boldsymbol{E}_k \boldsymbol{M}_{k,j} \boldsymbol{E}_j^\top)$ that stage one (Sec. 4.2) drops. With the input

factors $\boldsymbol{Q}_k$ fixed by Eq. (12), the error is linear in the output factors $\boldsymbol{P}_k$, and its minimizer under any PSD metric is a least-squares problem.

Let $\hat{\boldsymbol{h}}_k(\boldsymbol{x})$ be the intermediate activation of expert k under the compressed gate/up, $\boldsymbol{z}_k(\boldsymbol{x}) := g_k(\boldsymbol{x})\,\boldsymbol{Q}_k^\top \hat{\boldsymbol{h}}_k(\boldsymbol{x}) \in \mathbb{R}^{L_k}$, stacked into $\boldsymbol{z} \in \mathbb{R}^{\sum_k L_k}$, and $\boldsymbol{y}(\boldsymbol{x}) := \sum_k g_k(\boldsymbol{x})\big[\boldsymbol{W}_k \boldsymbol{h}_k(\boldsymbol{x}) - \boldsymbol{B}^\star \hat{\boldsymbol{h}}_k(\boldsymbol{x})\big]$. The error of the down projection, $\mathbb{E}\|\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{P}\,\boldsymbol{z}(\boldsymbol{x})\|^2$ with $\boldsymbol{P} := [\boldsymbol{P}_1, \ldots, \boldsymbol{P}_K]$, is quadratic in $\boldsymbol{P}$ with the Gram $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top]$. The Gram is the exact metric of the down projection, including co-routing blocks, projected onto the retained subspaces of size $\sum_k L_k$ rather than $K d_{\text{mid}}$. As in Sec. 4.2, this Gram is a calibration estimate and the solution is shrunk accordingly, here toward the stage-one output factors $\tilde{\boldsymbol{P}}_k$ of Eq. (12), stacked as $\tilde{\boldsymbol{P}} := [\tilde{\boldsymbol{P}}_1, \ldots, \tilde{\boldsymbol{P}}_K]$:

$$\boldsymbol{P}(\lambda) \;=\; \big(\boldsymbol{C} + \lambda\,\tilde{\boldsymbol{P}}\boldsymbol{D}\big)\big(\boldsymbol{G} + \lambda\,\boldsymbol{D}\big)^{-1}, \qquad \boldsymbol{G} := \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top],\;\; \boldsymbol{C} := \mathbb{E}[\boldsymbol{y}\boldsymbol{z}^\top],\;\; \boldsymbol{D} := \operatorname{diag}(\boldsymbol{G}), \tag{13}$$

with λ selected per layer on held-out calibration sequences; $\lambda = \infty$ returns $\tilde{\boldsymbol{P}}$, so the held-out error never increases (derivation in Sec. C.1). For gate/up the experts are summed only after the nonlinearity and the down projection, so no such linear stage exists.

How much the refit can recover is bounded by the weight of the cross terms, at most $k_{\text{top}} - 1$ times that of the diagonal blocks (Prop. 4 (ii)). On Mixtral-8×7B the two gates of a token are renormalized to sum to one, so the off-diagonal mass of $\boldsymbol{\Gamma}$ cannot exceed its diagonal mass; Qwen2-57B-A14B and DeepSeekMoE-16B activate eight and six experts without renormalization, and their off-diagonal mass is several times the diagonal (Table 5). We therefore expect the refit to bring little on Mixtral and to improve Qwen2 and DeepSeekMoE, which is consistent with the results in Secs. 5.2 and 5.3.

4.4 Role-split rank allocation

The ranks $\{L_k^{(t)}\}$ can vary across layers, experts, and the three roles of a layer. D^2-MoE keeps all three uniform, and so do we except along the role axis. In the MoE-FFN block Eq. (2), the errors of gate and up are attenuated by the nonlinearity and the down projection, whereas a down error arrives as it is, so a uniform ratio under-allocates the down projection. Specifically, expanding Eq. (2) to first order in the weight errors, with $\boldsymbol{a} = \boldsymbol{W}_k^{(\text{gate})}\boldsymbol{x}$, $\boldsymbol{b} = \boldsymbol{W}_k^{(\text{up})}\boldsymbol{x}$, and $\boldsymbol{h} = \text{SiLU}(\boldsymbol{a}) \odot \boldsymbol{b}$,

$$\Delta e_k(\boldsymbol{x}) \;\approx\; \Delta\boldsymbol{W}_k^{(\text{down})}\,\boldsymbol{h} \;+\; \boldsymbol{W}_k^{(\text{down})}\operatorname{diag}(\boldsymbol{s})\,\Delta\boldsymbol{W}_k^{(\text{gate})}\,\boldsymbol{x} \;+\; \boldsymbol{W}_k^{(\text{down})}\operatorname{diag}(\boldsymbol{t})\,\Delta\boldsymbol{W}_k^{(\text{up})}\,\boldsymbol{x}, \tag{14}$$

with the per-unit gains $\boldsymbol{s} = \text{SiLU}'(\boldsymbol{a}) \odot \boldsymbol{b}$ and $\boldsymbol{t} = \text{SiLU}(\boldsymbol{a})$. The gate and up errors thus reach the block output only after being scaled by these gains, small and sparse under SiLU, and by $\boldsymbol{W}_k^{(\text{down})}$; on Mixtral the measured scaling is 10^{-3} to 0.5, mostly from the gains (Sec. C.2). Minimizing the block loss of the first-order error Eq. (14) over the split of a fixed rank budget among the three roles gives the ratio between the down and gate/up ranks, denoted as s_κ (Sec. C.2).

Since the first-order model is only a proxy for the loss, we interpolate between the uniform ratio and s_κ with a single hyper-parameter β. The ratio between the down rank and the gate/up rank is $s_{\text{down}} = 1 + \beta\,(s_\kappa - 1)$, so that $\beta = 0$ is the uniform ratio and $\beta = 1$ the first-order prediction.

5 Experiments

5.1 Setup

Models. We evaluate the compression results on three MoE-based LLMs: (1) Mixtral-8×7B (Jiang et al., 2024, $K = 8$, top-2 experts activated); (2) Qwen2-57B-A14B (Yang et al., 2024a, $K = 64$ routed experts, top-8 experts activated, plus 8 shared experts); and (3) DeepSeekMoE-16B-Base (Dai et al., 2024, $K = 64$, top-6 experts activated, plus 2 shared experts). We test different compression methods under the *whole-model* compression ratios.

Calibration and evaluation. Following previous work (Li et al., 2025; Gu et al., 2025), we use 256 sequences of 2 048 tokens from WikiText-2 (Merity et al., 2017) as the calibration dataset. For evaluation, we report perplexity on three language datasets: WikiText-2, PTB (Marcus et al., 1993), and C4 (Raffel et al., 2020), and zero-shot accuracy on 7 reasoning datasets: OpenBookQA (Mihaylov et al., 2018), ARC-e/c (Clark et al., 2018), WinoGrande (Sakaguchi et al., 2021), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), and MathQA (Amini et al., 2019).

Table 1: PPL↓ and ACC↑ results at $40/60\%$ compression rates. Best results are in **bold** and second best results are underlined. Entries are printed to one or two decimal but ranked at full precision. †Results are quoted from the respective paper.

Ratio	Method	Wiki.	PTB	C4	Openb.	ARC_e	WinoG.	HellaS.	ARC_c	PIQA	MathQA	Avg.
Mixtral-8×7B												
0%	Original	3.84	14.73	7.19	35.4	84.2	76.5	64.9	57.1	82.4	42.4	63.3
40%	MoE-SVD†	6.74	27.73	12.41	27	72	67	43	38	71	32	50
	D^2-MoE	5.45	22.68	9.72	30.6	76.6	73.4	56.2	45.3	77.4	33.7	56.2
	MoBE	39.28	156.56	36.26	28.2	65.1	66.9	43.4	37.0	72.1	27.3	48.6
	RFID-MoE	41.09	176.15	39.15	27.8	64.2	66.1	42.9	35.5	71.9	27.4	47.9
	TD-MoE	24.86	100.02	30.58	17.2	55.6	59.4	32.7	24.1	67.0	24.5	40.1
	Ours	**5.41**	21.07	**9.55**	**32.4**	77.9	73.8	**57.2**	**46.6**	**78.5**	**34.5**	**57.3**
	Ours + refit	5.42	**20.88**	9.60	32.0	**78.0**	**73.9**	56.8	45.6	78.1	33.9	56.9
60%	MoE-SVD†	13.52	130.26	39.54	19	45	55	33	23	62	25	37
	D^2-MoE	7.14	27.27	12.45	26.6	70.0	70.2	49.4	38.1	73.2	30.3	51.1
	TD-MoE	1382.39	1296.35	828.65	14.4	35.6	51.1	27.6	21.2	56.4	21.5	32.5
	Ours	**6.87**	**24.21**	**12.15**	**28.0**	**70.8**	**71.6**	**51.5**	**38.9**	**74.0**	**31.3**	**52.3**
	Ours + refit	7.01	25.53	12.55	27.6	**70.8**	71.0	50.4	38.8	73.6	30.9	51.9
Qwen2-57B-A14B												
0%	Original	5.92	9.34	8.87	32.8	74.8	74.3	62.9	46.3	80.8	39.0	58.7
40%	D^2-MoE	7.04	11.15	12.90	29.6	77.4	76.7	54.9	47.0	77.6	39.7	57.6
	MoBE	8.40	13.38	13.14	30.8	78.2	75.1	54.9	48.0	79.1	37.0	57.6
	RFID-MoE	7.47	11.64	11.95	31.2	**79.2**	76.4	56.8	**48.9**	**79.8**	**39.9**	**58.9**
	TD-MoE	8.04	12.31	13.05	29.6	75.3	70.4	53.9	45.5	77.5	35.9	55.4
	Ours	6.84	10.75	12.05	31.2	77.5	**77.3**	56.6	45.9	79.0	38.9	58.1
	Ours + refit	**6.80**	**10.64**	**11.76**	**32.4**	77.6	77.0	**57.0**	47.0	78.9	39.1	58.4
60%	D^2-MoE	10.43	19.42	21.58	26.0	66.0	**74.0**	46.0	39.3	72.3	30.3	50.6
	TD-MoE	12.72	19.88	21.11	**28.6**	**72.5**	67.6	44.4	**42.2**	**74.9**	29.9	51.4
	Ours	10.18	18.97	20.76	26.8	69.1	73.5	**47.3**	39.3	72.0	31.6	51.4
	Ours + refit	**9.24**	**15.72**	**19.11**	26.0	68.9	73.8	46.8	40.4	72.7	**31.8**	**51.5**
DeepSeekMoE-16B-Base												
0%	Original	6.51	9.72	9.65	32.0	76.0	70.5	58.1	44.5	78.6	31.3	55.8
40%	D^2-MoE	9.06	14.44	16.95	27.0	67.7	65.2	43.8	36.9	70.6	27.5	48.4
	MoBE	12.61	17.06	18.07	25.4	67.3	65.5	44.0	34.7	71.9	26.5	47.9
	RFID-MoE	10.12	14.11	15.18	**27.4**	68.6	**69.1**	46.7	36.0	**73.8**	27.9	49.9
	TD-MoE	16.84	22.39	24.69	22.8	57.3	64.6	42.4	32.1	69.6	24.4	44.7
	Ours	8.56	13.16	14.92	25.6	70.6	68.1	46.9	**37.6**	73.2	28.5	50.1
	Ours + refit	**8.18**	**12.76**	**14.45**	25.2	**71.0**	68.2	**47.3**	37.2	72.7	**29.2**	**50.1**
60%	D^2-MoE	14.32	28.35	33.89	20.4	55.6	60.3	34.1	26.3	63.5	24.5	40.7
	TD-MoE	45.17	57.16	64.07	14.2	40.5	53.5	30.5	18.3	61.9	22.6	34.5
	Ours	13.47	25.40	29.43	21.0	59.2	61.3	35.5	28.5	**65.5**	**25.1**	42.3
	Ours + refit	**11.18**	**22.06**	**26.37**	**21.8**	**60.4**	**62.9**	**35.8**	**29.4**	65.3	24.8	**42.9**

Baselines. We choose 5 recent low-rank compression methods for MoE-based LLMs as baselines: (1) MoE-SVD (Li et al., 2025): a method extends SVD-LLM on MoEs by introducing V-matrix sharing; (2) D^2-MoE (Gu et al., 2025): a direct comparison of CORTEX by sharing a similar structure; (3) MoBE (Chen et al., 2026): a method introduces basis sharing across experts and non-linear activations inside the low-rank factorization; (4) RFID-MoE (Mi et al., 2026): a method shares the same structure with MoBE, while additionally introduces a rank allocation strategy and residual approximation; (5) TD-MoE (XU et al., 2026): a tensor decomposition approach that minimizes the routing-blind objective and high-order SVD. Except for MoE-SVD, we re-ran other methods following their instructions to match the evaluation pipeline and parameter budgets.

More experimental details are shown in Sec. D.

5.2 Main Results

Table 1 reports the perplexity (PPL) on language modeling tasks and the accuracy (ACC) on zero-shot reasoning tasks. We test compression ratios 40/60% on all three models; the Mixtral-8×7B block at 20% is in Table 9 (Sec. E.1). For MoBE and RFID-MoE, since they only compress the up and gate projections, at compression ratio 60%, they are either unusable or unreachable.

Table 3: Component ladder on DeepSeek-MoE-16B at 40% compression ratio.

Configuration	Wiki.	PTB	C4	Acc.
D^2-MoE (reference)	9.06	14.44	16.95	48.4
Ours (μ=0, β=0)	8.54	13.74	15.78	49.0
+ shrinkage (μ=0.5)	8.49	13.09	14.71	49.9
+ role split (β=0.25)	8.56	13.16	14.92	50.1
+ refit	**8.18**	**12.76**	**14.45**	**50.1**

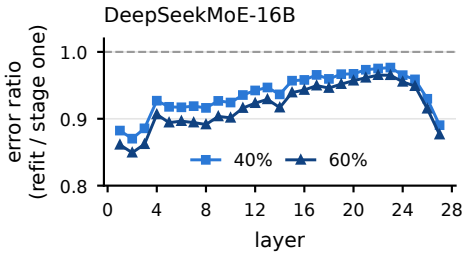

Figure 1: The coupled refit at the block level.

Table 4: Calibration set sizes and shrinkage scales on DeepSeekMoE-16B at 40% compression ratio. Green indicates improvement over the μ=0 row and red indicates degradation.

$\lvert\mathcal{D}_c\rvert$	μ	Wiki.↓	PTB↓	C4↓	Acc.↑
128	0.0	8.83	14.11	16.40	48.8
	0.5	8.63 (−2.2%)	13.17 (−6.7%)	14.92 (−9.1%)	49.9 (+1.1)
	1.0	9.31 (+5.5%)	13.58 (−3.7%)	15.12 (−7.8%)	49.7 (+0.9)
256	0.0	8.54	13.74	15.78	49.0
	0.5	8.49 (−0.6%)	13.09 (−4.7%)	14.71 (−6.8%)	49.9 (+0.9)
	1.0	9.25 (+8.4%)	13.54 (−1.4%)	15.02 (−4.8%)	49.9 (+0.8)
512	0.0	8.39	13.55	15.40	49.5
	0.5	8.41 (+0.2%)	12.97 (−4.3%)	14.55 (−5.5%)	50.3 (+0.7)
	1.0	9.22 (+9.8%)	13.50 (−0.4%)	14.95 (−2.9%)	49.9 (+0.4)
1024	0.0	8.33	13.53	15.23	49.9
	0.5	8.38 (+0.5%)	12.97 (−4.1%)	14.48 (−4.9%)	50.3 (+0.4)
	1.0	9.21 (+10.5%)	13.47 (−0.4%)	14.91 (−2.1%)	49.8 (−0.1)

Language modeling. Our methods achieve the best PPL on all three domains in every block of Table 1. For Mixtral, the refit procedure does not seem to help the approximation. However, for Qwen2 and DeepSeekMoE, the advantage of CORTEX on PPL is more profound as expected. Moreover, for these two models, the refit procedure can further decrease the PPL, by 0.6–2.4% at 40% and 8–17% at 60% on Qwen2 and by 3–4.5% and 10–17% on DeepSeekMoE, which extends the lead over D^2-MoE to 3.4–19.1% and 9.7–22.2%. The split between the models follows the co-routing mass of $\mathbf{\Gamma}$ (Sec. 4.3); Sec. 5.3 traces it to the block error layer by layer.

Zero-shot reasoning. Our methods also achieve the best average ACC in five of the six blocks, the exception being RFID-MoE at Qwen2 40% (and at Mixtral 20% in Table 9), which keeps its down projections intact, an advantage at low ratios that leaves it unusable at 60%. Compared to D^2-MoE, our methods improve the average ACC by 0.5–2.2 points in every block while achieving lower PPL at the same time. As for PPL, the refit helps more on Qwen2 and DeepSeekMoE, but by less: it reduces the routed block error of the down projection on the calibration distribution, which PPL tracks directly and zero-shot accuracy only loosely.

Compression cost. Table 2 shows the compressing time on Qwen2 at 40% compression. The whole pipeline of CORTEX costs about 128 GPU minutes, 1.1× less than D^2-MoE, 3.2× less than MoBE, and 6.8× less than RFID-MoE. MoBE/RFID-MoE run 10^4 Adam steps per weight matrix and are thus much slower. Our method also needs no backward pass at all, while for D^2-MoE, it needs 4 NVIDIA GB200 GPUs to run in the Fisher profiling stage for 31 minutes. TD-MoE is the fastest one, because of the computational advantage of tensor HOSVD. More importantly, it only factorizes 14 layers that are chosen beforehand; other methods instead factorize nearly all 28 layers.

Table 2: Compressing time (in minutes) on Qwen2-57B-A14B at 40% compression ratio.

Method	Profiling	Fit	Total
D^2-MoE	80.5	56.8	137.3
MoBE	–	414.8	414.8
RFID-MoE	15.9	853.0	868.9
TD-MoE	18.3	30.0	48.3
Ours	11.6	97.7	109.3
Ours + refit	11.6	116.1	127.7

5.3 Ablation Studies

In this subsection, we ablate the components of Sec. 4, the calibration scale, and the hyperparameters μ (shrinkage) and β (rank allocation). More results are shown in Sec. F.

Ablation on different components. Table 3 adds the components one at a time on DeepSeekMoE-16B. Replacing D^2-MoE's Fisher-merged base and unweighted per-expert metric by the routed centroid and the exact per-expert block improves every column (first row). The shrinkage then helps most on the two held-out domains and on accuracy, the generalization gain anticipated in Sec. 4.2;

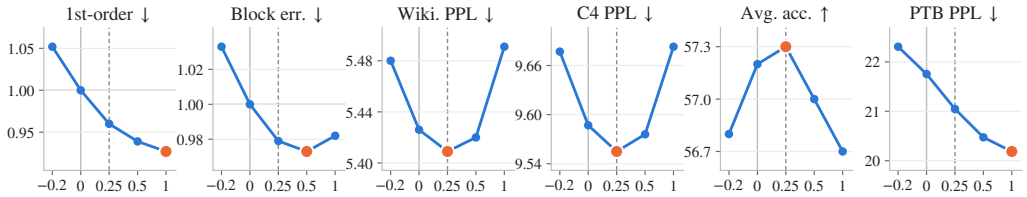

Figure 2: Role-split ablation on Mixtral-8×7B at 40% compression ratio, stage one, against β (Sec. 4.4). "Block err." is the block-output error of the compressed blocks on the original model's inputs and "1st-order" the prediction $\sum \kappa\sigma^2$ of Eq. (17), both relative to the uniform ratio (β=0).

the role split is neutral here and is isolated in Fig. 2; the refit procedure lowers all three PPL at unchanged accuracy. Table 10 repeats the ladder on the other two models.

Shrinkage of the residual metric and calibration scale. Table 4 varies the shrinkage μ of Eq. (11) and the number of calibration sequences $|\mathcal{D}_c|$. At the default 256 sequences, $\mu = 0.5$ beats the exact block ($\mu = 0$) on every column (on Qwen2 as well, Table 10; on Mixtral on all but PTB, Table 11). More data narrows the gap, as shrinkage theory predicts: the optimal intensity falls with the sample size, so the fixed μ may over-shrink. On the calibration domain (Wiki.), the advantage indeed turns into a small deficit from 512 sequences on, while on the two held-out domains it persists at 1 024. The pooled covariance alone ($\mu = 1$) is the mirror image. Since it is already well estimated on the smallest set, it barely moves with $|\mathcal{D}_c|$ and stays behind $\mu = 0.5$ on all three datasets; what it lacks is the expert-specific directions provided by the exact diagonal blocks. The two ends of Eq. (11) thus fail for opposite reasons: the exact block ($\mu = 0$) is estimated from the few tokens routed to one expert and over-fits them (large variance), the pooled covariance ($\mu = 1$) is well estimated but not expert-specific (large bias), and the shrinkage in between balances the two.

The coupled refit at the block level. Figure 1 shows the effect of the refit procedure on the held-out error of the down output, relative to the stage-one factors. More results on other models and compression ratios are in Fig. 4. The refit lowers this error in every layer. The magnitude follows the weight of the co-routing term: on Mixtral (two active experts) the reduction is small and the refit is neutral or slightly harmful (Tables 1 and 9); on Qwen2 and DeepSeekMoE (six to eight active experts), it is several times larger and improves all three PPL at unchanged or slightly higher accuracy. These results show that the refit lowers the block error everywhere and the loss wherever the co-routing term is heavy, as Sec. 4.3 predicts.

The role split. Figure 2 varies only β of Sec. 4.4 on Mixtral at 40%, *i.e.*, how much rank each layer moves from the gate/up projections to the down projection. Moving rank away from down is worse in every panel and moving it toward down is better, so the uniform ratio does under-allocate the down projection, whose error reaches the residual stream unattenuated. The magnitude, however, is over-predicted: the first-order objective keeps falling up to $\beta = 1$, the block error bottoms out at $\beta = 0.5$, and the loss turns at $\beta = 0.25$, which is also what the held-out criterion of Sec. D.2 picks.

6 Discussion and Limitations

CORTEX fits every component of the delta decomposition under the routed block loss, through two surrogates of its metric and an exact re-solve of the down factors. In experiments, it attains the best perplexity and, in most settings, the best accuracy on three MoEs. Its limitations lie in what the surrogates leave out. First, the block loss is stated per projection, so the exact co-routing blocks enter only the down factors; for gate/up projections, the coupling passes through the nonlinearity and has no closed form. Second, all statistics are estimates on the calibration dataset $\mathcal{D}_c$. So the fits of Sec. 4 minimize block errors on the calibration set, not the end-to-end loss. Besides, hyper-parameters such as μ and β are fixed once on one model, although the optimal shrinkage falls with the per-expert sample size (Table 4), which motivates a per-expert intensity (Ledoit & Wolf, 2004). Finally, the stage-two gain is bounded by the co-routing mass of $\mathbf{\Gamma}$, so under renormalized top-2 routing (*e.g.*, Mixtral-8×7B) there is little to recover.

AI USE STATEMENT

In this work, we used generative AI tools for *assist in the writing of proofs, design or provide feedback on research methodology or experiments, implement methods*. We have not used generative AI tools for *generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims,propose or refine hypotheses, interpret results*, and *assist with translation, clean and reformat dataset, support qualitative and thematic data analysis* are not applicable to this work. Additionally, we used generative AI tools for *create or edit software code, brainstorming, edit a research paper to improve readability, propose a title or keywords for a research paper*. We have reviewed all AI-assisted work. We checked LLM-generated research ideas with related papers by searching keywords and reading them. We checked code generated by LLMs for correctness. We checked all contents that are polished by LLMs for accuracy and consistency. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on training-free compression of pre-trained and open-source LLMs. The authors are not aware of any potential ethical issues that are specific to this work.

REPRODUCIBILITY STATEMENT

The proofs and derivations are provided in Secs. A and C. The whole pipeline is provided in Alg. 1. The experimental details including device configurations, hyperparameters, dataset descriptions, and evaluation protocols are provided in Sec. D. We believe these materials are sufficient to ensure the reproducibility of our results. Besides, we will release the code upon publication.

REFERENCES

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2357–2367, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1245. URL https://aclanthology.org/N19-1245/.

Saleh Ashkboos, Maximilian Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. Slicegpt: Compress large language models by deleting rows and columns. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 11682–11701, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/316648eb8b4ffb6010f531b07848c300-Paper-Conference.pdf.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

I-Chun Chen, Hsu-Shen Liu, Wei-Fang Sun, Chen-Hao Chao, Yen-Chang Hsu, and Chun-Yi Lee. Retraining-free merging of sparse MoE via hierarchical clustering. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8594–8620. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/chen25aq.html.

Xiaodong Chen, Mingming Ha, Zhenzhong Lan, Jing Zhang, and Jianguo Li. Mobe: Mixture-of-basis-experts for compressing moe-based llms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 40010–40028, 2026. URL

https://proceedings.iclr.cc/paper_files/paper/2026/file/431d53d513461ff155d5bc8faa9a440c-Paper-Conference.pdf.

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 preprint arXiv:1803.05457*, 2018.

Damai Dai, Chengqi Deng, Chenggang Zhao, R.X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y.K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1280–1297, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.70. URL https://aclanthology.org/2024.acl-long.70/.

Emily Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/1adaeb993eba95859121a43ea61bd858-Paper.pdf.

Elias Frantar and Dan Alistarh. Qmoe: Practical sub-1-bit compression of trillion-parameter models. *arXiv preprint arXiv:2310.16795*, 2023a.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10323–10337. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/frantar23a.html.

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.

Jerome H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989. doi: 10.1080/01621459.1989.10478752. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478752.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, September 2021. URL https://doi.org/10.5281/zenodo.5371629.

Hao Gu, Wei Li, Lujun Li, Zhu Qiyuan, Mark G. Lee, Shengjie Sun, Wei Xue, and Yike Guo. Delta decompression for MoE-based LLMs compression. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20497–20514. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/gu25c.html.

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.

Xing Hu, Dawei Yang, Yuan Cheng, Zhixuan Chen, and Zukang Xu. Saes-svd: Self-adaptive suppression of accumulated and local errors for svd-based llm compression. *arXiv preprint arXiv:2602.03051*, 2026.

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

Yong-Deok Kim, Eunhyeok Park, Sungjoo Yoo, Taelim Choi, Lu Yang, and Dongjun Shin. Compression of deep convolutional neural networks for fast and low power mobile applications. *arXiv preprint arXiv:1511.06530*, 2015.

Young Jin Kim, Raffy Fahim, and Hany Hassan Awadalla. Mixture of quantized experts (moqe): Complementary effect of low-bit quantization and robustness. *arXiv preprint arXiv:2310.02410*, 2023.

Mike Lasby, Ivan Lazarevich, Nish Sinnadurai, Sean Lie, Yani Ioannou, and Vithursan Thangarasa. Reap the experts: Why pruning prevails for one-shot moe compression. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 146883–146912, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ed93b2b5722acc2341d421b8916404a1-Paper-Conference.pdf.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. ISSN 0047-259X. doi: https://doi.org/10.1016/S0047-259X(03)00096-4. URL https://www.sciencedirect.com/science/article/pii/S0047259X03000964.

Pingzhi Li, Xiaolong Jin, Zhen Tan, Yu Cheng, and Tianlong Chen. Quantmoe-bench: Examining post-training quantization for mixture-of-experts. *arXiv preprint arXiv:2406.08155*, 2024a.

Pingzhi Li, Zhenyu Zhang, Prateek Yadav, Yi-Lin Sung, Yu Cheng, Mohit Bansal, and Tianlong Chen. Merge, then compress: Demystify efficient smoe with hints from its routing policy. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 14234–14256, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/3d09a88c3372cdb79401fde592ca4db8-Paper-Conference.pdf.

Wei Li, Lujun Li, Hao Gu, You-Liang Huang, Mark G. Lee, Shengjie Sun, Wei Xue, and Yike Guo. MoE-SVD: Structured mixture-of-experts LLMs compression via singular value decomposition. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35209–35230. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25az.html.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In P. Gibbons, G. Pekhimenko, and C. De Sa (eds.), *Proceedings of Machine Learning and Systems*, volume 6, pp. 87–100, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/file/42a452cbafa9dd64e9ba4aa95cc1ef21-Paper-Conference.pdf.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Enshu Liu, Junyi Zhu, Zinan Lin, Xuefei Ning, Matthew B Blaschko, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. Efficient expert pruning for sparse mixture-of-experts language models: Enhancing performance and reducing inference costs. *arXiv preprint arXiv:2407.00945*, 2024b.

Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pp. 6159–6172, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.334. URL https://aclanthology.org/2024.acl-long.334/.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 21702–21720. Curran Associates, Inc., 2023. doi: 10.52202/075280-0950. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/44956951349095f74492a5471128a7e0-Paper-Conference.pdf.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

Michael S Matena and Colin Raffel. Merging models with fisher-weighted averaging. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17703–17716. Curran Associates, Inc., 2022. doi: 10.52202/068431-1287. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/70c26937fbf3d4600b69a129031b66ec-Paper-Conference.pdf.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Zhendong Mi, Yixiao Chen, Pu Zhao, Xiaodong Yu, Hao Wang, Yanzhi Wang, and Shaoyi Huang. Effective moe-based LLM compression by exploiting heterogeneous inter-group experts routing frequency and information density. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=czym69s5GU.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

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.

Atul Kumar Sinha and François Fleuret. AA-SVD: Anchored and adaptive SVD for large model compression, 2026. URL https://openreview.net/forum?id=fIpDd5UlFP.

Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. A simple and effective pruning approach for large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 4942–4964, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/14c856c7a41297804de4c4890e846b25-Paper-Conference.pdf.

Alex M Tseng, Prannay Kaul, Luca Zancato, Wei Xia, and Stefano Soatto. Higher-order pruning of experts in mixture-of-experts language models. *arXiv preprint arXiv:2609.18916*, 2026.

Jingcun Wang, Yu-Guang Chen, Ing-Chao Lin, Bing Li, and Grace Li Zhang. Basis sharing: Cross-layer parameter sharing for large language model compression. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 13386–13404, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/238c98450b1d9e8055f94d22f303bb57-Paper-Conference.pdf.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. Svd-llm: Truncation-aware singular value decomposition for large language model compression. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 19299–19319, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3104e1ab39875cf54fe1eb4473e7c5a1-Paper-Conference.pdf.

Yanyue Xie, Zhi Zhang, Ding Zhou, Cong Xie, Ziang Song, Xin Liu, Yanzhi Wang, Xue Lin, and An Xu. Moe-pruner: Pruning mixture-of-experts large language model using the hints from its router. *arXiv preprint arXiv:2410.12013*, 2024.

Yuebin XU, YANHONG WANG, Xuemei Peng, Hui Zang, Chen Minghao, Pengfei Xia, and Zeyi Wen. TD-moe: Tensor decomposition for moe models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=D9cnZNZfxX.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024a. URL https://arxiv.org/abs/2407.10671.

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Yuanlin Duan, Wenqi Jia, Miao Yin, Yu Cheng, and Bo Yuan. MoE-i^2: Compressing mixture of experts models through inter-expert pruning and intra-expert low-rank decomposition. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10456–10466, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.612. URL https://aclanthology.org/2024.findings-emnlp.612/.

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. *arXiv preprint arXiv:2312.05821*, 2023.

Artur Zagitov, Alexander Miasnikov, Maxim Krutikov, Vladimir Aletov, Gleb Molodtsov, Nail Bashirov, Artem Tsedenov, and Aleksandr Beznosikov. Rethinking the role of tensor decompositions in post-training llm compression. *arXiv preprint arXiv:2606.03465*, 2026.

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/.

APPENDIX CONTENTS

A Proofs **15**

B The block loss: co-routing interference and covariance heterogeneity **15**

B.1 Co-routing interference 15

B.2 Measured covariance heterogeneity 17

C Method details **18**

C.1 The coupled refit: projected Gram and ridge solution 18

C.2 Rank allocation: rule, live-unit cap, and first-order analysis 18

D Experimental details **20**

D.1 Models, calibration, and evaluation 20

D.2 Details of our method 22

E Additional results **23**

E.1 Mixtral-8×7B at 20% 23

E.2 Per-layer held-out errors of the refit 23

F Additional ablations **24**

A PROOFS

This section proves the propositions stated in Secs. 3 and 4.

Proof of Prop. 1. Under covariance homogeneity $\boldsymbol{M} = \boldsymbol{\Gamma} \otimes \bar{\boldsymbol{\Sigma}}$, so $\mathcal{L} = \operatorname{Tr}(\Delta \boldsymbol{W}\, \boldsymbol{M}\, \Delta \boldsymbol{W}^\top) = \|\Delta \boldsymbol{W}\, \boldsymbol{M}^{1/2}\|_F^2$. Since $\boldsymbol{\Gamma} \succeq 0$ and $\bar{\boldsymbol{\Sigma}} \succ 0$, $(\boldsymbol{\Gamma} \otimes \bar{\boldsymbol{\Sigma}})^{1/2} = \boldsymbol{\Gamma}^{1/2} \otimes \bar{\boldsymbol{\Sigma}}^{1/2}$. The k-th block of $\Delta \boldsymbol{W}(\boldsymbol{\Gamma}^{1/2} \otimes \bar{\boldsymbol{\Sigma}}^{1/2})$ is $\sum_i (\boldsymbol{\Gamma}^{1/2})_{ki}\, \Delta \boldsymbol{W}_i \bar{\boldsymbol{\Sigma}}^{1/2}$, the k-th slice of $\Delta \mathbf{T} \times_1 \boldsymbol{\Gamma}^{1/2} \times_3 \bar{\boldsymbol{\Sigma}}^{1/2}$, which gives Eq. (7). Expanding the blocks gives $\hat{\mathcal{L}} = \sum_{i,j} \boldsymbol{\Gamma}_{ij} \operatorname{Tr}(\Delta \boldsymbol{W}_i \bar{\boldsymbol{\Sigma}} \Delta \boldsymbol{W}_j^\top)$. □

Proof of Prop. 2. (i) The objective is a strictly convex quadratic in $\boldsymbol{B}$ because $\bar{\boldsymbol{\Sigma}} \succ 0$. Its gradient is $-2 \sum_{i,j} \boldsymbol{\Gamma}_{ij} (\boldsymbol{W}_j - \boldsymbol{B}) \bar{\boldsymbol{\Sigma}}$, and setting it to zero gives $(\mathbf{1}^\top \boldsymbol{\Gamma} \mathbf{1}) \boldsymbol{B} = \sum_j (\boldsymbol{\Gamma} \mathbf{1})_j \boldsymbol{W}_j$. (ii) The centering is the stationarity condition of (i). For the decomposition, write $\boldsymbol{W} - \mathbf{1}^\top \otimes \boldsymbol{B} - \hat{\boldsymbol{R}} = (\boldsymbol{R} - \hat{\boldsymbol{R}}) + \mathbf{1}^\top \otimes (\boldsymbol{B}^\star - \boldsymbol{B})$. The cross term $2 \operatorname{Tr}\big[\big(\sum_i (\boldsymbol{\Gamma} \mathbf{1})_i (\boldsymbol{R}_i - \hat{\boldsymbol{R}}_i)\big) \bar{\boldsymbol{\Sigma}} (\boldsymbol{B}^\star - \boldsymbol{B})^\top\big]$ vanishes because $\boldsymbol{R}$ and $\hat{\boldsymbol{R}}$ are both routing-centered, and $\|\mathbf{1}^\top \otimes \boldsymbol{D}\|_{\boldsymbol{\Gamma}, \bar{\boldsymbol{\Sigma}}}^2 = (\mathbf{1}^\top \boldsymbol{\Gamma} \mathbf{1}) \operatorname{Tr}(\boldsymbol{D} \bar{\boldsymbol{\Sigma}} \boldsymbol{D}^\top)$. □

B THE BLOCK LOSS: CO-ROUTING INTERFERENCE AND COVARIANCE HETEROGENEITY

This section supplements Sec. 3. Section B.1 states what the block loss measures and shows on a two-expert example that a routing-blind objective can select the opposite direction; Sec. B.2 states and measures the mismatch of the two surrogates.

B.1 CO-ROUTING INTERFERENCE

What the block loss measures. The block loss Eq. (5) is the squared error of the block output, $\sum_k g_k(\boldsymbol{x})\, \Delta \boldsymbol{W}_k\, \boldsymbol{x}$ on each token, and this gated sum is all that the next layer receives. A routing-

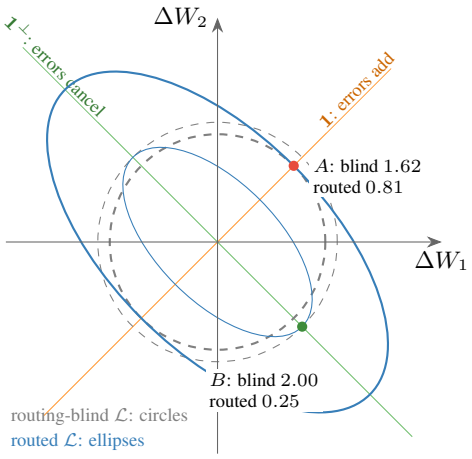

Figure 3: The two scalar experts of Example 3 in the plane of their weight errors. The routing-blind loss has circular level sets; the routed loss $\Delta \boldsymbol{W}\, \boldsymbol{\Gamma}\, \Delta \boldsymbol{W}^\top$ has elliptical ones, long along $\mathbf{1}^\perp$, where the errors of co-routed experts cancel, and short along $\mathbf{1}$, where they add. The blind loss is lower for A; the routed loss is lower for B, by a factor of three. A shared-base error lies on $\mathbf{1}$ (weight $\mathbf{1}^\top \boldsymbol{\Gamma} \mathbf{1} = 1$); the routing-centered residual of Prop. 2 (ii) lies on $\mathbf{1}^\perp$ (weight $\boldsymbol{\Gamma}_{11} + \boldsymbol{\Gamma}_{22} - 2\boldsymbol{\Gamma}_{12} = \frac{1}{4}$). The expert-mode whitening $\boldsymbol{\Gamma}^{1/2}$ of Prop. 1 maps the ellipses to circles.

blind objective measures the error of each expert separately, as if the expert acted alone with gate one; under top-k routing with $k \geq 2$ the block never computes that function. The errors of the experts routed to the same token are added before anything downstream sees them, so they can cancel or accumulate, and a compression with the smaller error on every expert can have the larger error at the block output. This coupling is what the signed off-diagonal blocks of the exact metric carry, and it can make the two objectives select opposite directions, as the following example shows.

Example 3. Let $K = 2$ with scalar experts ($d_{\text{in}} = d_{\text{out}} = 1$, $\mathbb{E}[x^2] = 1$), so that a compression is a pair of weight errors $\Delta \boldsymbol{W} = (\Delta W_1, \Delta W_2)$. Let the router alternate between the gates $(\frac{3}{4}, \frac{1}{4})$ and $(\frac{1}{4}, \frac{3}{4})$, so that both experts are active on every token and $\boldsymbol{\Gamma} = \frac{1}{16}\left(\begin{smallmatrix} 5 & 3 \\ 3 & 5 \end{smallmatrix}\right)$, whose off-diagonal mass is 0.6 of the diagonal, as on Mixtral-8×7B (Table 5). The routing-blind objective is $\Delta W_1^2 + \Delta W_2^2$. The block loss Eq. (5) is $\mathbb{E}\left[(g_1 \Delta W_1 + g_2 \Delta W_2)^2 x^2\right] = \Delta \boldsymbol{W}\, \boldsymbol{\Gamma}\, \Delta \boldsymbol{W}^\top$, which the Kronecker surrogate reproduces exactly, since covariance homogeneity holds in one dimension. Compare A: $\Delta \boldsymbol{W} = (0.9, 0.9)$ with B: $\Delta \boldsymbol{W} = (1, -1)$. The routing-blind objective ranks A first ($1.62 < 2$), and so does any reweighting by $\operatorname{diag}(\boldsymbol{\Gamma})$, which scales both by the same factor. The block loss is 0.81 for A and 0.25 for B: the block output deviates from the original by $0.9\,x$ on every token under A and by $\pm 0.5\,x$ under B. The router adds the two errors on every token, so aligned errors add and opposing errors cancel, and the ranking flips (Fig. 3): A has the smaller error on each expert and the larger one at the block output. The discriminating quantity is $\boldsymbol{\Gamma}_{12}$. With fixed gates $(\frac{1}{2}, \frac{1}{2})$, $\boldsymbol{\Gamma} = \frac{1}{4}\mathbf{1}\mathbf{1}^\top$ is singular, the ellipses degenerate into lines, and the block loss of B is 0.

The axes of the ellipses in Fig. 3 are the eigenvectors of $\boldsymbol{\Gamma}$. Along $\mathbf{1}$ the errors of co-routed experts add: an error (c, c) has block loss $c^2\, \mathbf{1}^\top \boldsymbol{\Gamma} \mathbf{1}$. This is the direction of an error shared by all experts, which is why the base is fitted under a surrogate that keeps the co-routing structure (Sec. 4.1). Along $\mathbf{1}^\perp$ they cancel: an error $(c, -c)$ has block loss $c^2(\boldsymbol{\Gamma}_{11} + \boldsymbol{\Gamma}_{22} - 2\boldsymbol{\Gamma}_{12})$, four times smaller here, and the routing-centered residuals of Prop. 2(ii), $\sum_k (\boldsymbol{\Gamma}\mathbf{1})_k \boldsymbol{R}_k = \mathbf{0}$, lie on this direction in the symmetric example. The block-diagonal surrogate has circular level sets and does not distinguish the two directions; the Kronecker surrogate has the elliptical ones, and its expert-mode whitening $\boldsymbol{\Gamma}^{1/2}$ (Prop. 1) maps them to circles.

Under top-1 routing $g_i g_j \equiv 0$ for $i \neq j$ and the interference vanishes. With top-2 (Mixtral) every token couples one pair of experts and the off-diagonal mass of $\boldsymbol{\Gamma}$ is 0.5–0.7 of the diagonal; with top-8 of 64 (Qwen2) it is 3–7 times the diagonal. For the base, Eq. (9) uses this structure directly: the off-diagonal of $\boldsymbol{\Gamma}$ enters the weights when the gates are not renormalized. For the private residuals, the co-routing coupling enters only through the refit of Sec. 4.3.

B.2 MEASURED COVARIANCE HETEROGENEITY

For an SPD surrogate $\hat{\boldsymbol{M}}$ with loss $\hat{\mathcal{L}}$, let $\varepsilon(\hat{\boldsymbol{M}}) := \|\hat{\boldsymbol{M}}^{-1/2}\,\boldsymbol{M}\,\hat{\boldsymbol{M}}^{-1/2} - \boldsymbol{I}\|_2$, the smallest constant with $(1-\varepsilon)\,\hat{\mathcal{L}} \le \mathcal{L} \le (1+\varepsilon)\,\hat{\mathcal{L}}$ for every $\Delta\boldsymbol{W}$. For both surrogates, ε is bounded by quantities that need no off-diagonal block of $\boldsymbol{M}$.

Proposition 4 (Mismatch of the two surrogates)**.** *Let* $\varepsilon_{\otimes} := \varepsilon(\boldsymbol{\Gamma} \otimes \bar{\boldsymbol{\Sigma}})$ *and* $\varepsilon_{\text{bd}} := \varepsilon(\text{blockdiag}(\boldsymbol{M}_{k,k}))$*, let* $\varepsilon_k := \left\|\bar{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{M}_{k,k}/\boldsymbol{\Gamma}_{kk})\,\bar{\boldsymbol{\Sigma}}^{-1/2} - \boldsymbol{I}\right\|_2$ *be the relative deviation of the per-expert covariance* $\boldsymbol{M}_{k,k}/\boldsymbol{\Gamma}_{kk}$ *from the pooled* $\bar{\boldsymbol{\Sigma}}$ *and* $\varepsilon_{\text{cov}} := \max_k \varepsilon_k$ *the covariance heterogeneity, and let* $k_{\text{top}} \ge 2$ *be the number of activated experts.* (i) $\varepsilon_{\otimes} \ge \varepsilon_{\text{cov}}$ *unconditionally, and* $\varepsilon_{\otimes} = 0$ *iff covariance homogeneity holds.* (ii) $\varepsilon_{\text{bd}} \le k_{\text{top}} - 1$ *unconditionally, and the bound is tight. Under covariance homogeneity,* $\varepsilon_{\text{bd}} = \|\text{diag}(\boldsymbol{\Gamma})^{-1/2}\,\boldsymbol{\Gamma}\,\text{diag}(\boldsymbol{\Gamma})^{-1/2} - \boldsymbol{I}\|_2$*, the distance of the gate correlation matrix from the identity.*

Proof. Write $\mathcal{L} = \sum_o \boldsymbol{v}_o^\top \boldsymbol{M}\boldsymbol{v}_o$ and $\hat{\mathcal{L}} = \sum_o \boldsymbol{v}_o^\top \hat{\boldsymbol{M}}\boldsymbol{v}_o$ over the rows o of the error matrices. The two-sided bound is $\|\boldsymbol{R} - \boldsymbol{I}\|_2 \le \varepsilon$ for $\boldsymbol{R} := \hat{\boldsymbol{M}}^{-1/2}\boldsymbol{M}\hat{\boldsymbol{M}}^{-1/2}$, and the operator norm is the smallest such ε. (i) For $\boldsymbol{v}$ whose only nonzero block is $\boldsymbol{v}_k$, $\boldsymbol{v}^\top\boldsymbol{M}\boldsymbol{v} = \boldsymbol{v}_k^\top\boldsymbol{M}_{k,k}\boldsymbol{v}_k$ and $\boldsymbol{v}^\top(\boldsymbol{\Gamma}\otimes\bar{\boldsymbol{\Sigma}})\boldsymbol{v} = \boldsymbol{\Gamma}_{kk}\,\boldsymbol{v}_k^\top\bar{\boldsymbol{\Sigma}}\boldsymbol{v}_k$, so on these $\boldsymbol{v}$ the two-sided bound holds iff $\varepsilon \ge \varepsilon_k$. Since $\varepsilon_{\otimes}$ gives the bound on all $\boldsymbol{v}$, $\varepsilon_{\otimes} \ge \varepsilon_{\text{cov}}$; and $\varepsilon_{\otimes} = 0$ iff $\boldsymbol{R} = \boldsymbol{I}$ iff $\boldsymbol{M} = \boldsymbol{\Gamma}\otimes\bar{\boldsymbol{\Sigma}}$. (ii) For the block-diagonal surrogate, $\boldsymbol{R}$ is PSD with identity diagonal blocks, so $\lambda_{\min}(\boldsymbol{R}) \ge 0$. For $\boldsymbol{v} = [\boldsymbol{v}_1;\dots;\boldsymbol{v}_K]$, $\boldsymbol{v}^\top\boldsymbol{M}\boldsymbol{v} = \mathbb{E}\big[(\sum_i g_i(\boldsymbol{x})\,\boldsymbol{v}_i^\top\boldsymbol{x})^2\big]$. At most k_{top} gates are nonzero on any token, so by Cauchy–Schwarz $\big(\sum_i g_i(\boldsymbol{x})\,\boldsymbol{v}_i^\top\boldsymbol{x}\big)^2 \le k_{\text{top}}\sum_i g_i(\boldsymbol{x})^2(\boldsymbol{v}_i^\top\boldsymbol{x})^2$; taking expectations, $\boldsymbol{v}^\top\boldsymbol{M}\boldsymbol{v} \le k_{\text{top}}\,\boldsymbol{v}^\top\text{blockdiag}(\boldsymbol{M}_{k,k})\,\boldsymbol{v}$, *i.e.*, $\lambda_{\max}(\boldsymbol{R}) \le k_{\text{top}}$. Hence $\varepsilon_{\text{bd}} = \max(\lambda_{\max}(\boldsymbol{R}) - 1,\, 1 - \lambda_{\min}(\boldsymbol{R})) \le k_{\text{top}} - 1$. Cauchy–Schwarz is tight when the active terms coincide on every token, which the fixed-gate limit $(\frac{1}{2},\frac{1}{2})$ of Example 3 realizes for $k_{\text{top}} = 2$. Under covariance homogeneity, $\boldsymbol{R} = \text{diag}(\boldsymbol{\Gamma})^{-1/2}\boldsymbol{\Gamma}\,\text{diag}(\boldsymbol{\Gamma})^{-1/2} \otimes \boldsymbol{I}$. □

Table 5 reports the two computable sides of Prop. 4 on three layers of each model (gate/up inputs, $d = 4096$, 3584, and 2048; routed statistics from the 256 WikiText-2 calibration sequences; $\varepsilon_k = \max(|\lambda_{\max} - 1|, |\lambda_{\min} - 1|)$ over the generalized eigenvalues λ of $\boldsymbol{M}_{k,k}/\boldsymbol{\Gamma}_{kk}$ with respect to $\bar{\boldsymbol{\Sigma}}$, computed in fp64 with a relative ridge of 10^{-6}).

- **Covariance homogeneity holds only approximately.** The lower bound ε_{cov} on the Kronecker mismatch is 7–12 on Mixtral, 24–77 on Qwen2, and 67–113 on DeepSeekMoE, and the median of ε_k over experts is 5–29, so the deviation of the per-expert covariances $\boldsymbol{M}_{k,k}/\boldsymbol{\Gamma}_{kk}$ from the pooled $\bar{\boldsymbol{\Sigma}}$ is a property of most experts rather than of a few cold ones. Thus, the Kronecker surrogate is used where the approximation does not enter the solution or improves it. The base of Eq. (9) has weights that depend on $\boldsymbol{\Gamma}$ alone, so $\bar{\boldsymbol{\Sigma}}$ does not affect it. In the shrinkage target $\boldsymbol{\Gamma}_{kk}\bar{\boldsymbol{\Sigma}}$ of Eq. (11), the pooled covariance is the well-estimated quantity, and mixing it in improves the compression (Table 4).
- **The dropped coupling is bounded, and the bound is below the pooling error on every layer.** The block-diagonal mismatch is at most $k_{\text{top}} - 1 = 1$, 7, and 5 (Prop. 4(ii)); ε_{cov} exceeds it by 7–12× on Mixtral, 3.4–11× on Qwen2, and 13–23× on DeepSeekMoE, on Mixtral and DeepSeek-MoE for every expert individually. The ordering $\varepsilon_{\otimes} \ge \varepsilon_{\text{cov}} > k_{\text{top}} - 1 \ge \varepsilon_{\text{bd}}$ therefore holds without any off-diagonal block of $\boldsymbol{M}$. Under homogeneous covariances the coupling alone would give $\varepsilon_{\text{bd}} = 0.5$–$0.7$ on Mixtral and 3.3–7.0 on the other two models (column ε_{bd} (homog.); Qwen2 layer 2 is at the bound). This is the selection rule of Sec. 4.2: stage one truncates each private residual under its own covariance, *i.e.*, under the block-diagonal surrogate, and drops the coupling, the smaller of the two errors. The Kronecker surrogate enters the residuals only through the shrinkage target of Eq. (11).
- **The co-routing mass is set by whether the gates are renormalized.** The off-/diagonal mass ratio of $\boldsymbol{\Gamma}$ is 0.5–0.7 on Mixtral, whose two gates per token sum to one, so that the ratio cannot exceed 1, against 3–7 on Qwen2 and 2–3 on DeepSeekMoE, whose top-8 and top-6 gates are not renormalized. This ratio is the weight of the cross terms that stage two (Sec. 4.3) restores for the down projection. The refit is therefore expected to change little on Mixtral and to improve Qwen2 and DeepSeekMoE, which is what Secs. 5.2 and 5.3 find.

Table 5: Mismatch of the two surrogates on the gate/up inputs: the lower bound $\varepsilon_{\text{cov}} \le \varepsilon_{\otimes}$ on the Kronecker mismatch and the median of the per-expert terms ε_k, the upper bound $k_{\text{top}} - 1 \ge \varepsilon_{\text{bd}}$ on the block-diagonal mismatch, ε_{bd} under homogeneous covariances, and the off-/diagonal mass ratio of $\mathbf{\Gamma}$.

Model	Layer	$\varepsilon_{\text{cov}} = \max_k \varepsilon_k$	$\text{median}_k\, \varepsilon_k$	$k_{\text{top}} - 1$	ε_{bd} (homog.)	$\mathbf{\Gamma}$ off/diag
Mixtral-8×7B (top-2)	2	9.58	5.60	1	0.67	0.66
	15	12.25	5.13	1	0.68	0.68
	27	7.17	5.35	1	0.54	0.52
Qwen2-57B-A14B (top-8)	2	24.0	5.2	7	6.99	6.98
	15	71.5	7.6	7	6.62	5.44
	27	77.1	25.8	7	3.72	3.15
DeepSeekMoE-16B (top-6)	2	67.2	28.6	5	3.43	3.13
	15	79.9	21.9	5	3.49	3.09
	27	113.2	28.7	5	3.33	2.36

C METHOD DETAILS

Algorithm 1 lists the full pipeline of CORTEX. This section gives the derivations behind Secs. 4.3 and 4.4.

C.1 THE COUPLED REFIT: PROJECTED GRAM AND RIDGE SOLUTION

Projected Gram. The Gram $\boldsymbol{G}$ of Eq. (13) has the blocks $\boldsymbol{G}_{(i,j)} = \boldsymbol{Q}_i^\top \mathbb{E}\big[g_i g_j\, \hat{\boldsymbol{h}}_i \hat{\boldsymbol{h}}_j^\top\big]\, \boldsymbol{Q}_j$: the exact co-routing blocks of the down projection, with the expert-specific inputs that the surrogates of Sec. 3.2 do not represent, projected onto the retained subspaces. It is positive semidefinite by construction, has size $\sum_e L_e$ rather than $K\, d_{\text{mid}}$ (51 968 against 163 840 on Qwen2-57B-A14B at 40%), and is accumulated from the token stream without forming any cross block.

Ridge solution. Equation (13) is the minimizer of the down-projection error with a penalty toward the stage-one factors,

$$J(\boldsymbol{P}) \;=\; \mathbb{E}\,\big\|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{z}\big\|^2 \;+\; \lambda\, \mathrm{Tr}\,\big((\boldsymbol{P} - \tilde{\boldsymbol{P}})\, \boldsymbol{D}\, (\boldsymbol{P} - \tilde{\boldsymbol{P}})^\top\big), \qquad \boldsymbol{D} = \mathrm{diag}(\boldsymbol{G}). \tag{15}$$

Expanding the first term as $\mathrm{Tr}(\boldsymbol{P}\boldsymbol{G}\boldsymbol{P}^\top) - 2\,\mathrm{Tr}(\boldsymbol{P}\boldsymbol{C}^\top) + \mathbb{E}\|\boldsymbol{y}\|^2$ with $\boldsymbol{G} = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top]$ and $\boldsymbol{C} = \mathbb{E}[\boldsymbol{y}\boldsymbol{z}^\top]$, J is a convex quadratic in $\boldsymbol{P}$ with gradient $2\boldsymbol{P}\boldsymbol{G} - 2\boldsymbol{C} + 2\lambda(\boldsymbol{P} - \tilde{\boldsymbol{P}})\boldsymbol{D}$. Setting it to zero gives $\boldsymbol{P}(\boldsymbol{G} + \lambda\boldsymbol{D}) = \boldsymbol{C} + \lambda\tilde{\boldsymbol{P}}\boldsymbol{D}$, which is Eq. (13); the minimizer is unique for every $\lambda > 0$ since $\boldsymbol{G} \succeq 0$ and $\boldsymbol{D} \succ 0$. The ends of the grid are the unregularized least squares $\boldsymbol{C}\boldsymbol{G}^{-1}$ at $\lambda = 0$ and the stage-one factors $\tilde{\boldsymbol{P}}$ as $\lambda \to \infty$. Scaling the penalty by $\mathrm{diag}(\boldsymbol{G})$ rather than by $\boldsymbol{I}$ makes λ a relative strength: each column of $\boldsymbol{P}$ is pulled toward $\tilde{\boldsymbol{P}}$ in proportion to the energy of its coordinate of $\boldsymbol{z}$, so one grid serves every layer and model. In the implementation the diagonal is floored at 10^{-6} of its mean, so that an expert with no calibration token leaves the system nonsingular, and $\lambda = 0$ is solved with the same minimal ridge. The ridge strength λ is chosen per layer from the grid of Sec. D.2 on 32 of the 256 calibration sequences, which are held out from the moments.

C.2 RANK ALLOCATION: RULE, LIVE-UNIT CAP, AND FIRST-ORDER ANALYSIS

This section gives the first-order analysis that defines the weight κ and the ratio s_κ, the measured κ profile, the rank rule, and the live-unit cap. The held-out selection of β is in Sec. D.2; the effect of the cap is isolated in Sec. F.

The first-order analysis. The three roles of a layer share one output, the block output added to the residual stream in Eq. (1), and the path from that output to the loss is the same for gate, up, and down. The whitened error of stage one Eq. (12) is measured in each role's own output space, and the roles reach the block output differently, as Eq. (14) shows: a gate or up error passes through the per-unit gains $\boldsymbol{s} = \mathrm{SiLU}'(\boldsymbol{a}) \odot \boldsymbol{b}$ and $\boldsymbol{t} = \mathrm{SiLU}(\boldsymbol{a})$, which are small and sparse under SiLU, and through $\boldsymbol{W}_k^{(\text{down})}$, which maps d_{mid} to d_{model} (14 336 → 4 096 on Mixtral); a down error reaches

Algorithm 1 CORTEX: **CO**mpressing **R**ou**T**ed **EX**perts

Require: Experts $\{\boldsymbol{W}_k^{(t)}\}_{k=1}^K$ of every compressed item $t = (\ell, p)$; calibration set $\mathcal{D}_c$; removal ratio ρ; shrinkage μ; role-split fraction β; refit grid Λ

Ensure: The dense base $\boldsymbol{B}^{(t),\star}$ and the factors $\{(\boldsymbol{P}_k^{(t)}, \boldsymbol{Q}_k^{(t)})\}_{k=1}^K$

▷ Profiling (Sec. 4)

Accumulate $\boldsymbol{\Gamma}^{(\ell)}$, $\boldsymbol{M}_{k,k}^{(t)}$, and $\boldsymbol{\Omega}_k$ in one forward pass over $\mathcal{D}_c$ ▷ Eqs. (16) and (20)

$\bar{\boldsymbol{\Sigma}}^{(t)} \leftarrow \sum_k \boldsymbol{M}_{k,k}^{(t)} / \operatorname{Tr}(\boldsymbol{\Gamma}^{(\ell)})$

▷ Stage one: base and whitened SVD (Secs. 4.1 and 4.2)

for all items t **do**

  $\boldsymbol{c} \leftarrow \boldsymbol{\Gamma}\mathbf{1} / \mathbf{1}^\top \boldsymbol{\Gamma}\mathbf{1}$

  $\boldsymbol{B}^\star \leftarrow \sum_k c_k \boldsymbol{W}_k$

  for all experts k **do**

    $\boldsymbol{R}_k \leftarrow \boldsymbol{W}_k - \boldsymbol{B}^\star$

    Drop the dead-unit rows/columns of $\boldsymbol{R}_k$, if any ▷ Live-unit cap in Sec. C.2

    $\boldsymbol{\Sigma}_k \leftarrow (1-\mu)\,\boldsymbol{M}_{k,k} + \mu\,\boldsymbol{\Gamma}_{kk}\,\bar{\boldsymbol{\Sigma}}$

    $\boldsymbol{S}_k \leftarrow \operatorname{chol}(\boldsymbol{\Sigma}_k + \delta \boldsymbol{I})$

    $\boldsymbol{U}_k, \boldsymbol{\sigma}_k, \boldsymbol{V}_k \leftarrow \operatorname{SVD}(\boldsymbol{R}_k \boldsymbol{S}_k)$

  end for

end for

▷ Rank allocation (Sec. 4.4)

$s_\kappa \leftarrow$ first-order role split from the SVDs and $\boldsymbol{\Omega}_k$ ▷ Sec. C.2

$s_{\text{gate}} = s_{\text{up}} \leftarrow 1,\ s_{\text{down}} \leftarrow 1 + \beta(s_\kappa - 1)$

$f \leftarrow$ largest keep fraction whose ranks $L_k^{(t)}$ fit the budget B set by ρ ▷ Eqs. (18) and (19)

▷ Stage one: truncation (Sec. 4.2)

for all items t and experts k **do**

  $\tilde{\boldsymbol{P}}_k, \tilde{\boldsymbol{Q}}_k \leftarrow$ first L_k columns of $\boldsymbol{U}_k \operatorname{diag}(\boldsymbol{\sigma}_k)^{1/2}$, $\boldsymbol{V}_k \operatorname{diag}(\boldsymbol{\sigma}_k)^{1/2}$

  $\boldsymbol{P}_k \leftarrow \tilde{\boldsymbol{P}}_k,\ \boldsymbol{Q}_k \leftarrow \boldsymbol{S}_k^{-\top} \tilde{\boldsymbol{Q}}_k$

end for

▷ (Optional) Stage two: refit (Sec. 4.3)

Run the original model over $\mathcal{D}_c$, with 32 sequences held out

for all down items $t = (\ell, \text{down})$ **do**

  $\boldsymbol{z}_k \leftarrow g_k \boldsymbol{Q}_k^\top \hat{\boldsymbol{h}}_k$, with $\hat{\boldsymbol{h}}_k$ from the stage-one gate/up

  $\boldsymbol{y} \leftarrow \sum_k g_k (\boldsymbol{W}_k \boldsymbol{h}_k - \boldsymbol{B}^\star \hat{\boldsymbol{h}}_k)$

  Accumulate $\boldsymbol{G}, \boldsymbol{C}$ on the remaining sequences ▷ Sec. C.1

  $\tilde{\boldsymbol{P}} \leftarrow [\tilde{\boldsymbol{P}}_1, \ldots, \tilde{\boldsymbol{P}}_K],\ \boldsymbol{D} \leftarrow \operatorname{diag}(\boldsymbol{G})$

  $\boldsymbol{P}(\lambda) \leftarrow (\boldsymbol{C} + \lambda \tilde{\boldsymbol{P}} \boldsymbol{D})(\boldsymbol{G} + \lambda \boldsymbol{D})^{-1}$ for $\lambda \in \Lambda,\ \boldsymbol{P}(\infty) \leftarrow \tilde{\boldsymbol{P}}$

  $\lambda^\star \leftarrow$ the $\lambda \in \Lambda \cup \{\infty\}$ minimizing the held-out $\mathbb{E}\|\boldsymbol{y} - \boldsymbol{P}(\lambda)\boldsymbol{z}\|^2$

  $[\boldsymbol{P}_1, \ldots, \boldsymbol{P}_K] \leftarrow \boldsymbol{P}(\lambda^\star)$

end for

the output unchanged. Separating the token dependence of the gains from that of $\boldsymbol{x}$ (the same separability the surrogates of Sec. 3.2 use) turns the first-order block-output error of a gate residual into $\operatorname{Tr}\left(\boldsymbol{\Omega}_k \Delta \boldsymbol{W}_k^{(\text{gate})} \boldsymbol{M}_{k,k} (\Delta \boldsymbol{W}_k^{(\text{gate})})^\top\right)$ with the output-side metric

$$\boldsymbol{\Omega}_k \;=\; \left((\boldsymbol{W}_k^{(\text{down})})^\top \boldsymbol{W}_k^{(\text{down})}\right) \odot \mathbb{E}\left[g_k^2\, \boldsymbol{s}\boldsymbol{s}^\top\right] / \mathbb{E}[g_k^2], \tag{16}$$

$\odot$ is the Hadamard product and $\boldsymbol{t}$ replaces $\boldsymbol{s}$ for up projections. The grams $\mathbb{E}[g_k^2 \boldsymbol{s}\boldsymbol{s}^\top]$ and $\mathbb{E}[g_k^2 \boldsymbol{t}\boldsymbol{t}^\top]$ are accumulated per gate/up expert by a side pass of the profiler (Sec. D.2) and enter only the solve for s_κ, not the truncation: truncating under the two-sided metric $\boldsymbol{\Omega}_k \otimes \boldsymbol{\Sigma}_k$ would tie the retained subspaces to d_{mid}^2 calibration-estimated entries, with the same over-fitting risk as the $\mu = 0$ end of Eq. (11). The first-order weight of a whitened direction $\boldsymbol{u}_{k,l}$ of $\boldsymbol{R}_k \boldsymbol{S}_k = \sum_l \sigma_{k,l}\, \boldsymbol{u}_{k,l} \boldsymbol{v}_{k,l}^\top$ is

$$\kappa_{k,l} \;=\; \boldsymbol{u}_{k,l}^\top \boldsymbol{\Omega}_k \boldsymbol{u}_{k,l} \;\; (\text{gate, up}), \qquad \kappa_{k,l} \;=\; 1 \;\; (\text{down}), \tag{17}$$

so that dropping direction l adds $\sigma_{k,l}^2$ to the role's own error and, to first order, $\kappa_{k,l}\,\sigma_{k,l}^2$ to the block-output error; the two are on the same scale because $\boldsymbol{M}_{k,k}$ and $\boldsymbol{\Omega}_k$ are normalized by the same

routed mass. Given κ, the first-order optimum under a layer's rank total is a common threshold on $\kappa\sigma^2$ across the three roles; restricted to the layer-shared family of Sec. 4.4 it is the ratio s_κ, computed from the stored spectra ($s_\kappa = 2.05$ on Mixtral at 40%; Table 8 lists every model and ratio).

The measured κ profile. On Mixtral at 40%, the σ^2-weighted mean of κ of Eq. (17) over the directions the uniform ratio drops is 1.3×10^{-3} (gate) and 1.6×10^{-3} (up) at layer 0 and rises monotonically with depth to 0.48 and 0.31 at layer 31; of this, $\boldsymbol{W}_k^{(\text{down})}$ contributes a layer-independent factor of about 0.5 (the mean squared norm of its columns) and the gains the rest. The raw whitened marginal σ^2 at the uniform rank runs the other way ($59\times$ larger for gate than for down at layer 0, equal at layer 31), so after the weighting by κ the marginal block-output error at the uniform rank is 2–13$\times$ larger for down than for gate or up in every layer, and the uniform ratio under-allocates the down projection.

Rank rule. Every residual keeps a fraction of its own dense parameter count, $s_p\, f$: a global keep fraction f shared by all layers, times the per-role multiplier s_p ($s_{\text{gate}} = s_{\text{up}} = 1$ and s_{down} of Sec. 4.4). Its rank is

$$L_k^{(\ell,p)} \;=\; \text{clip}\left(\left\lfloor \frac{s_p\, f \cdot d_{\text{out}} d_{\text{in}}}{d_{\text{out}} + d_{\text{in}}} \right\rfloor,\; 1,\; \min(d_{\text{out}}, d_{\text{in}}, n_{\ell,k})\right), \tag{18}$$

where $n_{\ell,k} \le d_{\text{mid}}$ is the number of live intermediate units of expert k in layer ℓ. The keep fraction f is not a second ratio next to ρ but is solved from it: the cost of the bases, $d_{\text{out}}d_{\text{in}}$ per item, is subtracted from the budget B of Eq. (19), a residual whose factors would cost at least $d_{\text{out}}d_{\text{in}}$ is stored dense, and for given $(s_p)_p$ the largest feasible f is found by bisection.

Live-unit cap. In Qwen2-57B-A14B, part of the rows of the gate and up projections and the corresponding columns of the down projection are exactly zero, so the intermediate unit is dead; the number of live units $n_{\ell,k}$ of an expert ranges from 10 to 2 560 within a layer. This motivates the clip at $n_{\ell,k}$ in Eq. (18), which uses the checkpoint, not the calibration set. During decomposition the dead rows (gate/up) and columns (down) of the residual $\boldsymbol{R}_k = \boldsymbol{W}_k - \boldsymbol{B}^\star$ are dropped before the whitening and truncation of Eq. (12), and a $K \times d_{\text{mid}}$ boolean mask per layer is stored with the payload and applied to the intermediate activation before the down projection at inference. The masked residual has rank at most $n_{\ell,k}$ for all three roles, so a capped residual is exact and any rank above the cap would be unused. Mixtral-8$\times$7B and DeepSeekMoE-16B have no dead units, so the rule is inactive on them.

D Experimental details

This section gives the settings of the evaluation, of our method, and of the baselines.

D.1 Models, calibration, and evaluation

Hardware and software. Every run uses one NVIDIA GB200 (192 GB of HBM) with 32 cores of the node's Grace CPU and 400 GB of host memory; the only exception is the Fisher pass of D^2-MoE, which needs two (Mixtral) or four (Qwen2) GPUs for memory. The stack is PyTorch 2.13 with CUDA 13, `transformers` 4.57.6, `datasets` 4.4.1, and `lm-evaluation-harness` v0.4.3 (Gao et al., 2021). Each model is kept in the precision of its released checkpoint, fp16 for Mixtral-8$\times$7B, bf16 for Qwen2-57B-A14B and DeepSeekMoE-16B; every row of Table 1, including the "Original" row, is compressed and evaluated in that precision.

Models. Table 6 lists the three architectures. They differ in the two properties that enter the routed statistics: expert granularity ($K = 8$ wide experts in Mixtral, $K = 64$ narrow ones in Qwen2 and DeepSeekMoE) and gate normalization (Mixtral renormalizes the softmax over the two selected experts; Qwen2 and DeepSeekMoE take the top-k of a softmax over all K experts without renormalization). Both 64-expert models have shared experts, which no method compresses. The first layer of DeepSeekMoE is a dense MLP, so it has 27 MoE layers. On Qwen2 the code of D^2-MoE excludes layers 12–15, where its truncation-aware SVD fails; we compress the same 24 layers, so that the two methods are compared at the same budget on the same items. The Qwen2 checkpoint stores exact zeros for about a quarter of its routed-expert parameters, as dead intermediate units

Table 6: Architectures of the three models. K/top-k: routed experts per layer and experts activated per token; "norm.": whether the top-k gates are renormalized; "shared experts": their number as reported by the authors and the total intermediate width of the stored shared MLP (Qwen2 stores its 8 shared experts of width 2 560 as one MLP of width 20 480 behind a sigmoid gate; DeepSeekMoE its 2 of width 1 408 as one of width 2 816); "compr.": the MoE layers on which the methods of a block are compared (MoBE and RFID-MoE compress the gate/up projections of all 28 Qwen2 layers, and TD-MoE uses its own layer subsets); N_{sel}: the routed-expert parameters of those layers, from which the whole-model removal ρN_{total} is taken.

Model	K/top-k	norm.	shared experts	d_{model}	d_{mid}	MoE layers	compr.	N_{total} (N_{sel})
Mixtral-8×7B-v0.1	8/2	yes	none	4 096	14 336	32	32	46.70B (45.10B)
Qwen2-57B-A14B	64/8	no	8 (width 20 480, gated)	3 584	2 560	28	24	57.41B (42.28B)
DeepSeekMoE-16B-Base	64/6	no	2 (width 2 816, ungated)	2 048	1 408	27	27	16.38B (14.95B)

concentrated in layers 11–16 (Sec. C.2); the live-unit counts $n_{\ell,k}$ of Eq. (18) are read from the safetensors shards. Mixtral and DeepSeekMoE have no dead units.

Compression ratio. Every method leaves attention, embeddings, the router, the shared experts, and (on DeepSeekMoE) the dense first layer unchanged, so all removed parameters come from the routed-expert pool (N_{sel} in Table 6). The target is the whole-model removal ratio ρ: with N_{total} the stored parameter count of the released checkpoint and $N_{\text{sel}} = \sum_t K d_{\text{out}} d_{\text{in}}$ the count of the compressed items, the compressed representation must fit

$$B = N_{\text{sel}} - \lfloor \rho\, N_{\text{total}} \rceil. \tag{19}$$

The keep fraction f of Eq. (18) is below $1-\rho$ because ρ counts the whole model while only the routed pool is compressed, the bases are stored dense, and $s_{\text{down}} > 1$: on Mixtral-8×7B at 40%, $1-\rho = 0.60$ of the model is 0.59 of the pool, 0.46 after the dense bases, and $f = 0.42$ once the down projection keeps $s_{\text{down}}\, f$ of its own count.

Calibration data. $\mathcal{D}_c$ is 256 windows of 2 048 tokens from the training split of WikiText-2 (raw): the articles are joined with blank lines, a starting character offset is drawn uniformly at random for each window under a fixed seed, and the first 2 048 tokens from that offset are kept, with the tokenizer's beginning-of-sequence token prepended where the tokenizer has one. The same 256 windows are used by every stage of CORTEX on a model: the routed-statistics pass, the SiLU-gain grams of the role split, and the refit, whose moments are accumulated on 224 windows while the other 32 select λ. The held-out single-layer criterion that selects β (Sec. D.2) uses 32 further training windows disjoint from $\mathcal{D}_c$. Table 7 gives the number of calibration samples per dimension for the three candidate metrics of Sec. 3, the pooled covariance $\bar{\boldsymbol{\Sigma}}$, a per-expert block $\boldsymbol{M}_{k,k}$, and the full metric $\boldsymbol{M}$, per model and role. The ratios between the rows are structural: the samples per dimension fall by a factor K/k from the pooled covariance to a per-expert block and by a factor K to the full metric; in absolute terms from 37–205, to 9–32 (median over experts and layers), to 2.3–16 on Mixtral-8×7B and Qwen2-57B-A14B. At the low end the sample estimate is not a usable estimate of the population metric: on Qwen2 the full metric has 2.3 samples per dimension, 2% of the (layer, expert) pairs receive fewer calibration tokens than their block has dimensions, and five receive none. A structured surrogate is a shrinkage toward a target estimated K/k to K times better; Sec. 5.3 and Sec. F measure its effect end-to-end.

Perplexity. We follow the SVD-LLM/TD-MoE protocol. The test split of WikiText-2 (raw), the test split of PTB, and the first 2 000 documents of the C4 validation set (the SparseGPT/SVD-LLM dump of its first shard) are each joined with blank lines, tokenized once, and cut into consecutive non-overlapping windows of 2 048 tokens; the remainder is dropped and no beginning-of-sequence token is inserted at window boundaries. Perplexity is the token-weighted $\exp(\sum \text{NLL}/\#\text{tokens})$ over all windows of a corpus. Every row of Table 1 except the quoted MoE-SVD rows is measured with this evaluator.

Zero-shot accuracy. The seven tasks are run with `lm-evaluation-harness` v0.4.3 in the zero-shot setting (`num_fewshot` = 0, no sample limit, batch size 16). We report the harness's `acc`

Table 7: Calibration samples per dimension for the three candidate metrics ($|\mathcal{D}_c| = 524\,288$ tokens: 256 WikiText-2 sequences of length 2 048): $|\mathcal{D}_c|/d$ for the pooled covariance, n_k/d for a per-expert block (median over experts and profiled layers, 5th percentile in brackets), and $|\mathcal{D}_c|/(Kd)$ for the full metric. Pooled-to-per-expert is K/k and pooled-to-full is K, with k the number of active experts per token; the g_k^2 weighting lowers the effective per-expert count further, to between $0.30\,n_k$ and n_k on Mixtral.

Metric	Mixtral-8×7B (K=8, top-2)		Qwen2-57B-A14B (K=64, top-8)	
	gate/up (d=4096)	down (d=14336)	gate/up (d=3584)	down (d=2560)
Pooled $\bar{\boldsymbol{\Sigma}}$	128	37	146	205
Per-expert $\boldsymbol{M}_{k,k}$	32 [27]	9.1 [7.8]	18 [5.6]	26 [7.8]
Full $\boldsymbol{M}$	16	4.6	2.3	3.2

metric, not `acc_norm`, for every task and method; "Avg." is the unweighted mean of the seven task accuracies.

D.2 Details of our method

Pipeline. Algorithm 1 lists the pipeline. *(i)* The routed-statistics pass of Sec. 4: one forward pass over $\mathcal{D}_c$ in micro-batches of 8 sequences with bf16 activations and fp64 accumulators, storing per MoE layer ℓ the gate Gram $\boldsymbol{\Gamma}^{(\ell)}$ (gates taken after the model's own top-k selection and normalization) and per item $t = (\ell, p)$ the routed covariances

$$\boldsymbol{M}_{k,k}^{(t)} \;=\; \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_c}\big[g_k(\boldsymbol{x})^2\, \boldsymbol{x}_k^{(t)} (\boldsymbol{x}_k^{(t)})^\top\big], \tag{20}$$

where $\boldsymbol{x}_k^{(t)}$ is the attention output $\boldsymbol{x}$ for gate/up and the intermediate activation $\boldsymbol{h}_k$ of expert k for down, and $\bar{\boldsymbol{\Sigma}}^{(t)} = \sum_k \boldsymbol{M}_{k,k}^{(t)} / \operatorname{Tr}(\boldsymbol{\Gamma}^{(\ell)})$; a side pass of the same profiler accumulates the SiLU-gain grams of Eq. (16). *(ii)* Rank allocation: the whitened spectra of every residual under $\boldsymbol{\Sigma}_k(\mu)$ are computed once and stored, the uniform-ratio manifest is found by bisection on f under Eq. (19), s_κ is solved from the stored spectra and the grams of *(i)*, and the manifest is re-cut at β (on Qwen2 under the live-unit cap). *(iii)* Decomposition: the routed centroid Eq. (9) and the stage-one truncation Eq. (12) for every item. *(iv)* The stage-two refit of the down output factors (Sec. 4.3). *(v)* Evaluation. Stages *(i)* and *(ii)* do not depend on the ratio except for the final re-cut, so one profiling run and one set of spectra serve every ratio and every ablation arm of a model. There is no stochastic step after the calibration windows are drawn.

Numerics. All second-order statistics are accumulated and factorized in fp64. $\boldsymbol{\Sigma}_k(\mu)$ of Eq. (11) is symmetrized and its eigenvalues are floored at 10^{-3} before the Cholesky factorization. $\boldsymbol{\Gamma}^{1/2}$ and $\boldsymbol{\Gamma}^{-1/2}$ are the symmetric roots from an fp64 eigendecomposition with eigenvalues clamped at 10^{-6} of the largest, so an expert with no calibration tokens causes no failure. The spectra of the whitened residuals $\boldsymbol{R}_k\boldsymbol{S}_k$ are eigenvalues of fp64 Gram matrices.

Hyperparameters. CORTEX has three hyper-parameters: the shrinkage μ in Eq. (11), the role-split fraction β in Sec. 4.4, and the ridge strength λ in Eq. (13). Unless stated otherwise, all results use $\mu = 0.5$ and $\beta = 0.25$. The held-out criterion below selects $s_{\text{down}} = 1.25$, *i.e.* $\beta \approx 0.25$. The refit uses the grid $\Lambda = \{0, 0.01, 0.1, 1, 10\}$ with the ridge scaled by $\operatorname{diag}(\boldsymbol{G})$ (Eq. (13)) and 32 held-out windows.

Selection of β. The fraction β is selected on held-out text from the calibration domain, not on the evaluation sets. For a candidate s_{down}, the compressed block of one layer is put into the original model, with the original model everywhere else, and the criterion is the KL divergence of the logits from the original model's on 32 held-out WikiText-2 training windows (Sec. D.1), summed over the layers; it needs one forward pass per layer and candidate. On Mixtral at 40% the criterion selects $s_{\text{down}} = 1.25$ (summed KL relative to the uniform ratio 0.996, against 1.006 at 1.5), and the same choice results on held-out C4 and PTB text. The selected point corresponds to $\beta = 0.24$; we use $\beta = 0.25$ (ranks 1 349/1 349/1 704 against a uniform 1 467), fixed once on Mixtral at 40% and used unchanged for the other ratios and models, where only s_κ is re-solved (Table 8).

Table 8: Per-(model, ratio) quantities of the CORTEX rows of Table 1. s_κ is the first-order optimum of the role split (Eq. (17)) and $s_{\text{down}} = 1 + 0.25\,(s_\kappa - 1)$ the multiplier used; "UR rank" is the uniform rank at $\beta = 0$; ranks are per expert, gate and up sharing one value (on Qwen2, of the experts on which the live-unit cap does not bind; a capped expert keeps its live-unit count, down to 10, and "cap-bound" counts the (layer, expert) pairs on which the cap binds).

	Mixtral-8×7B			**Qwen2-57B-A14B**		**DeepSeekMoE-16B**	
	20%	40%	60%	40%	60%	40%	60%
s_κ	1.67	2.05	2.62	1.85	1.82	2.03	2.49
s_{down}	1.168	1.263	1.405	1.213	1.205	1.258	1.373
UR rank ($\beta = 0$)	2 127	1 467	808	659	253	455	272
ranks (gate(up) / down)	2015/2353	1349/1704	712/1000	670/812	252/303	419/528	242/333
cap-bound experts	–	–	–	169	131	–	–
compression	0.2000	0.4002	0.6000	0.4001	0.6001	0.4004	0.6006

Refit. Stage two runs one forward pass of the original model over the 256 calibration windows in micro-batches of 8 (bf16), 224 for the moments and 32 held out. Inside every compressed MoE block, the compressed gate/up of that layer are evaluated next to the original experts on the same routed inputs, and the layer accumulates the projected Gram $\boldsymbol{G}$ and the cross-moment $\boldsymbol{C}$ of Eq. (13) from the intermediate activations $\hat{\boldsymbol{h}}_k$ of the compressed gate/up and the routed down output of the original experts; the gate/up error of the layer is thus included, while no error from upstream layers enters. The per-layer solve Eq. (13) is closed-form, and the λ of each layer is the grid point with the smallest relative error $\mathbb{E}\|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{z}\|^2/\mathbb{E}\|\boldsymbol{y}\|^2$ on the 32 held-out windows, with $\lambda = \infty$ (*i.e.* $\tilde{\boldsymbol{P}}$) admissible. Only the down output factors change; the input factors, the ranks, the gate/up payloads, and the base are those of stage one, so the "+ refit" rows have the parameter count and the kernel of the "Ours" rows.

E Additional results

E.1 Mixtral-8×7B at 20%

Table 9 gives results of Mixtral-8×7B at 20% compression ratio. The methods are closer at this ratio than at 40% and 60%: CORTEX has the lowest PPL on WikiText-2 and C4 and D^2-MoE is 2% lower on PTB; RFID-MoE and MoBE, whose down projections are intact, have the highest average accuracy (61.2 and 61.0 against 59.9); the refit changes no perplexity here, since on Mixtral the co-routing blocks contribute little to the block loss (Sec. 4.3).

Table 9: PPL↓ and ACC↑ results on Mixtral-8×7B at 20% compression rate. Best results are in **bold** and second best results are underlined. Entries are printed to one or two decimal but ranked at full precision. †Results are quoted from the respective paper.

Ratio	Method	Wiki.	PTB	C4	Openb.	ARC_e	WinoG.	HellaS.	ARC_c	PIQA	MathQA	Avg.
	Mixtral-8×7B											
0%	Original	3.84	14.73	7.19	35.4	84.2	76.5	64.9	57.1	82.4	42.4	63.3
20%	MoE-SVD†	4.86	19.42	8.98	33	79	74	56	49	78	37	58
	D^2-MoE	4.78	**18.66**	8.50	32.4	79.6	73.7	59.9	49.2	80.3	37.3	58.9
	MoBE	5.99	23.59	9.20	34.4	**81.4**	75.5	**61.4**	53.0	**81.5**	**39.5**	61.0
	RFID-MoE	5.68	23.08	8.97	**36.6**	81.0	**76.1**	61.2	**53.4**	81.3	38.7	**61.2**
	TD-MoE	7.05	33.96	11.47	27.4	74.5	73.6	53.8	42.8	77.4	34.0	54.8
	Ours	**4.75**	19.04	**8.41**	33.0	81.1	74.9	60.7	50.8	80.6	37.6	59.8
	Ours + refit	**4.75**	19.04	**8.41**	32.8	81.0	75.6	60.5	51.1	80.7	37.4	59.9

E.2 Per-layer held-out errors of the refit

Figure 4 reports, for every compressed layer of the three models at every ratio with a refit run, the relative error of the down output, $\mathbb{E}\|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{z}\|^2/\mathbb{E}\|\boldsymbol{y}\|^2$ in the notation of Sec. 4.3, on the 32 held-

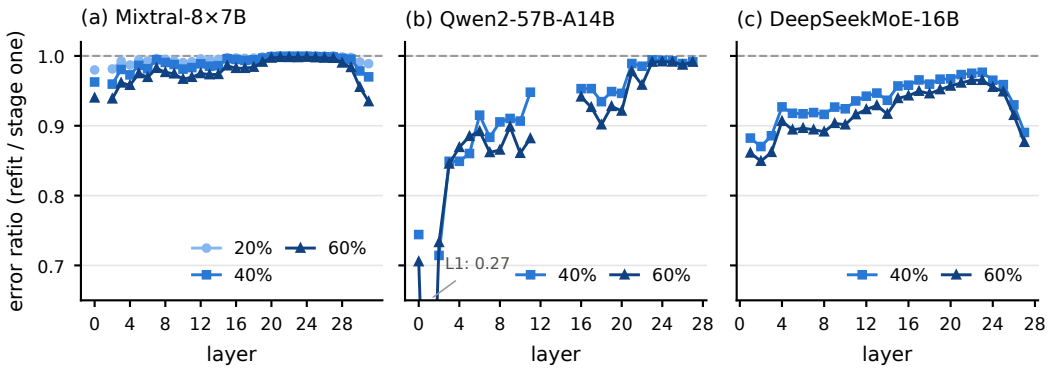

Figure 4: The coupled refit at the block level on the three models: held-out relative error of the down output, $\mathbb{E}\|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{z}\|^2/\mathbb{E}\|\boldsymbol{y}\|^2$ in the notation of Sec. 4.3, on the 32 held-out calibration sequences, after the refit (ridge at the selected λ) relative to before it (stage-one factors, $\lambda{=}\infty$), per compressed layer, at every ratio with a refit run. Layers whose stage-one residuals are exact (error 0) are omitted; the layer-1 point of Qwen2 at 60% (0.27) is below the axis.

out calibration sequences. The reduction is largest in the first layers and decreases with depth: the late-layer error is not shared across co-routed experts, so the coupling does not remove it.

Across models, the median per-layer reduction of this held-out error at the selected λ is $0.4/0.6/1.8\%$ on Mixtral-8×7B at $20/40/60\%$, where the renormalized top-2 gates keep the off-diagonal mass of the exact metric below the diagonal, against $5.4/10.0\%$ on Qwen2-57B-A14B (unnormalized top-8, off-diagonal mass 5–7× the diagonal) and $6.4/8.3\%$ on DeepSeekMoE-16B (unnormalized top-6) at $40/60\%$ (Fig. 4). On Mixtral there is little for the exact blocks to correct, and the small calibration-side reduction does not transfer to the test sets; on the other two models the reduction grows with the ratio, because the stage-one factors are further from the exact solution.

F Additional ablations

Component ladder on all models. Table 10 repeats the ladder of Table 3 on Mixtral-8×7B and Qwen2-57B-A14B. On these two models the first CORTEX row matches D^2-MoE rather than beating it: within 0.2% on Mixtral WikiText-2 and C4 (-6% on PTB, $+0.5$ accuracy) and 1% behind on Qwen2 (-0.1 accuracy). The gains come from the later rows. On Qwen2 they are the shrinkage ($-1.4/-1.4/-4.1\%$ on the three domains), the live-unit cap ($-2.5/-4.6/-4.6\%$), and the refit ($-0.5/-1.0/-2.4\%$, $+0.4$ accuracy); the role split is neutral ($+1.6\%$ and $+1.1\%$ on PTB and C4 at stage one, a tie after the refit). On Mixtral the steps are within 1% on WikiText-2 and C4 and within 5% on PTB: the shrinkage improves WikiText-2, C4, and the accuracy and loses on PTB, the role split improves all three domains and the accuracy, and the refit is neutral. The three models order the components differently for the reasons of Sec. 4: the estimation regime is harder on Qwen2 and DeepSeekMoE ($K/k_{\text{top}} = 8$ and 10.7 against 4 on Mixtral), and these are the two models on which the shrinkage and the refit produce the gains; Qwen2 additionally has dead units, and the co-routing mass its refit uses is 5–7× larger than Mixtral's.

Shrinkage on Mixtral. Table 11 varies only μ of Eq. (11) at 256 calibration sequences on Mixtral; the uniform-ratio ranks do not depend on the spectra, so a change of metric leaves the allocation, the base, and the parameter count unchanged. $\mu = 0.5$ is better than the exact block $\mu = 0$ on WikiText-2, C4, and the accuracy and behind on PTB. The pooled covariance alone ($\mu = 1$) is worse than $\mu = 0.5$ on every domain, with accuracy within noise; on DeepSeekMoE it is 8–10% behind on WikiText-2 and 1–4% on PTB and C4 at every calibration scale.

The role-split rank allocation. Table 12 lists the ranks, the block-output error, the first-order prediction, and the end-to-end results plotted in Fig. 2; the held-out selection of β is in Sec. D.2.

Table 10: Component ladder at 40% on all three models; the DeepSeekMoE block is that of Table 3. Each row adds one component of Sec. 4 to the row above it at the same structure and parameter count: the routed-centroid base with the exact per-expert block (μ=0) at uniform-ratio ranks (β=0), the shrinkage toward the pooled routed covariance (μ=0.5), the role split (β=0.25), the live-unit cap (Qwen2 only), and the coupled refit of the down factors.

Model	Configuration	Wiki.	PTB	C4	Avg. acc.
Mixtral-8×7B	D^2-MoE (reference)	5.45	22.68	9.72	56.2
	Ours (μ=0, β=0)	5.46	21.34	9.72	56.7
	+ shrinkage (μ=0.5)	5.43	22.13	9.59	57.2
	+ role split (β=0.25)	**5.41**	21.07	**9.55**	**57.3**
	+ refit	5.42	**20.88**	9.60	56.9
Qwen2-57B-A14B	D^2-MoE (reference)	7.04	11.15	12.90	57.6
	Ours (μ=0, β=0)	7.11	11.25	13.02	57.5
	+ shrinkage (μ=0.5)	7.01	11.09	12.49	57.7
	+ role split (β=0.25)	7.01	11.27	12.63	57.9
	+ live-unit cap	6.84	10.75	12.05	58.1
	+ refit	**6.80**	**10.64**	**11.76**	**58.4**
DeepSeekMoE-16B	D^2-MoE (reference)	9.06	14.44	16.95	48.4
	Ours (μ=0, β=0)	8.54	13.74	15.78	49.0
	+ shrinkage (μ=0.5)	8.49	13.09	14.71	49.9
	+ role split (β=0.25)	8.56	13.16	14.92	50.1
	+ refit	**8.18**	**12.76**	**14.45**	**50.1**

Table 11: Shrinkage of the private-residual metric at 40% whole-model removal on Mixtral-8×7B (DeepSeekMoE-16B is in Table 4). Only the shrinkage μ of Eq. (11) changes; uniform-ratio ranks (β=0) and parameter count are fixed, so the rows differ only in the metric under which each residual is truncated. Best CORTEX row per column in bold.

Model	Private-residual metric	Wiki.	PTB	C4	Avg. acc.
Mixtral-8×7B	D^2-MoE (reference)	5.45	22.68	9.72	56.2
	μ=0 (exact per-expert block)	5.46	**21.34**	9.72	56.7
	μ=0.5 (default)	**5.43**	22.13	**9.59**	**57.2**
	μ=1 (pooled covariance)	5.55	23.13	9.68	57.1

Table 12: Role-split ablation on Mixtral-8×7B at 40% compression ratio. "Block err." is the block-output error of the compressed MoE blocks on the original model's inputs, relative to the uniform ratio; "1st-order" is the prediction $\sum \kappa\sigma^2$ of Eq. (17) on the same scale.

β	Ranks gate/up/down	Block err.	1st-order	Wiki.	PTB	C4	Avg. acc.
−0.2	1578/1578/1247	1.033	1.052	5.480	22.96	9.677	56.8
0 (uniform)	1467/1467/1467	1.000	1.000	5.426	22.13	9.587	57.2
0.25	1349/1349/1704	0.979	0.960	**5.409**	21.07	**9.555**	**57.3**
0.5	1249/1249/1905	**0.973**	0.939	5.420	20.21	9.576	57.0
1	1087/1087/2229	0.982	**0.927**	5.491	**19.78**	9.683	56.7