BLoM: Bayesian Inference over Merge Operators for LoRA Model Merging

Anonymous authors
Paper under double-blind review

Abstract

Merging low-rank adaptation (LoRA) adapters efficiently combines task-specific capabilities, but common practice selects a discrete merge configuration specifying coefficients, calibration strengths, and ranks, using validation accuracy. Such selection can overfit when task updates interfere. We quantify this merge-selection gap and observe losses up to 5.6 accuracy points on a high-interference task pair. BLoM replaces selection among discrete configurations with posterior inference over the continuous merge operator that each configuration induces. A lossless core-space projection makes this operator compact and enables a full-covariance Gaussian approximation within a representative layer of models with up to 8 billion parameters. The posterior supports a regularized mean predictor, Bayesian model averaging (BMA), confidence-based abstention, and PAC-Bayes analysis of Gibbs risk, the expected loss of a posterior-sampled operator. Across five standard settings spanning three language models and two vision transformers, BLoM-Mean improves the NLI suite average by 6.9 points over the strongest of twelve baselines, including task arithmetic, TIES, DARE-TIES, RegMean, aligned-subspace, and LoRA-specific methods. In cross-domain merging, it gains 13.9 points over the strongest of three evaluated baselines, task arithmetic, TIES, and TSV-M. On the pilot pair, BLoM-Mean improves validation-selected TSV-M by 4.8 points and exceeds its finite-grid test oracle. Nominal balanced-design risk curves remain below one across nine verified ranks. These results recast LoRA merging as inference over plausible merge operators.

1 Introduction

Low-Rank Adaptation (LoRA) (Hu et al., 2022) is a standard way to specialize pretrained models, and merging several adapters offers a practical route to combine their capabilities without retraining on the original task data (Ilharco et al., 2022). Recent methods have progressed from weight-space arithmetic (Wortsman et al., 2022a; Yadav et al., 2023) to aligned singular subspaces (Stoica et al., 2025; Gargiulo et al., 2025) and direct treatments of LoRA geometry (Panariello et al., 2026; Tang & Yang, 2026; Jeong et al., 2026).

These methods differ in representation and merge rule, but their configurations are commonly chosen using held-out accuracy. Coefficients, calibration strengths, and ranks are selected from candidate sets or fixed by hand. This choice is usually treated as an implementation detail. When the validation maximum is partly determined by finite-sample noise, however, the selected merge can incur a systematic test-time loss. The risk should be largest when tasks interfere and the candidate space is difficult to distinguish reliably.

We show that this selection effect is measurable. On a high-interference task pair, the validation-selected configuration loses up to 5.6 accuracy points relative to the test oracle over the same candidate pool. The gap depends on rank and merge rule, while validation resampling on a separate six-task suite characterizes its sensitivity to sample budget. It is small on a near-orthogonal suite, where task operators have negligible subspace overlap and the merge rules nearly coincide. These results expose a selection gap under finite held-out data and motivate inference over the operator itself.

We introduce BLoM, short for Bayesian LoRA Merging, which replaces accuracy-based configuration selection with inference over the merge operator. A lossless core-space projection (Panariello

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.

et al., 2026) compresses each layer to an operator with at most $(Kr)^2$ entries. BLoM fits this compact object on held-out task data within one fixed projection layer and constructs a Gaussian approximation retaining the full empirical-Fisher covariance defined in Section 4.2. The resulting distribution supports an accuracy-oriented mean predictor, Bayesian model averaging (BMA) over plausible operators, selective prediction that abstains on low-confidence inputs, and PAC-Bayes analysis of the Gibbs predictor that samples one operator per prediction.

The experiments support this inference view across five standard suite-model settings and an additional cross-domain suite. BLoM-Mean improves the six-task natural-language inference (NLI) suite average by 6.9 points over the strongest of twelve baselines and improves an eight-task cross-domain suite by 13.9 points over the strongest of three compared methods. On the high-interference pilot pair, BLoM-Mean improves validation-selected TSV-M by 4.8 points and exceeds its finite-grid test oracle. The posterior also provides competitive selective prediction on NLI, and the corresponding PAC-Bayes rank curve remains below the unit-loss ceiling across nine verified ranks under the nominal balanced design. On the vision suites, BLoM attains accuracy comparable to several aligned-subspace methods, providing a complementary test of the same operator formulation. The benefit therefore concentrates in the interference-dominated regime that motivates the method.

Our contributions are threefold.

- We quantify the gap between validation-selected merges and finite-grid test oracles, and examine its dependence on task setting, adapter rank, and validation budget.
- We propose BLoM, a full-covariance posterior approximation over losslessly projected merge operators in a representative layer that supports mean prediction, model averaging, selective confidence, and PAC-Bayes analysis in one framework.
- We evaluate the framework across language and vision models, showing strong accuracy in interference-dominated merging, improvements over validation-selected merges, useful selective prediction, and informative PAC-Bayes rank analysis.

2 Related Work

Deterministic model merging. Weight-space combination of fine-tuned models has roots in federated averaging (McMahan et al., 2017) and weight interpolation (Wortsman et al., 2022b), and was popularized by model soups and task arithmetic (Wortsman et al., 2022a; Ilharco et al., 2022). Parameter-space methods trim, drop, or reweight task-vector entries (Yadav et al., 2023; Yu et al., 2024; Jin et al., 2023; Yang et al., 2024c). Later approaches align singular subspaces, as in KnOTS, Task Singular Vector (TSV) merging, and Iso-C (Stoica et al., 2025; Gargiulo et al., 2025; Marczak et al., 2025), or work directly with LoRA geometry by projecting adapters into shared bases and calibrating their factors or spectra (Panariello et al., 2026; He et al., 2026; Li et al., 2026b; Jeong et al., 2026; Li et al., 2026c; Tang & Yang, 2026). We use TSV-M for the evaluated TSV baseline and TSV-Core for the corresponding deterministic operator constructed in the lossless core space used as BLoM's language-model initialization and prior center. The methods in our comparison ultimately produce a single merged operator, commonly with method-specific settings chosen from held-out accuracy or fixed by hand. BLoM instead represents plausible merge operators with a distribution.

Posteriors and probabilistic merging. BLoM builds on subspace inference in Bayesian deep learning (MacKay, 1992; Daxberger et al., 2021; Izmailov et al., 2019; Maddox et al., 2019). Its subspace is defined by the merging problem and preserves the LoRA updates exactly. The inferred object is a merge operator across trained adapters rather than the training posterior of one model or adapter (Yang et al., 2024a; Wang et al., 2024). Other probabilistic approaches use uncertainty to construct merge weights or guide an outer search while returning a point estimate (Li et al., 2026a; Daheim et al., 2024; Bui et al., 2026). Complementary work predicts merge success (Rahamim et al., 2026; Zhou et al., 2026), composes or routes adapters (Huang et al., 2024a; Wu et al., 2024), benchmarks merge accuracy (Yang et al., 2024b; He et al., 2025), or adds marginal conformal coverage (Angelopoulos & Bates, 2023). BLoM maintains a distribution over the merge operator and uses it for prediction, selective confidence, and Gibbs-risk analysis. Appendix C provides the broader comparison and method taxonomy.

3 PRELIMINARIES AND MOTIVATION

3.1 MERGE OPERATORS IN A SHARED CORE SPACE

Consider a pretrained weight matrix $\boldsymbol{W}_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. LoRA fine-tuning on task $k \in \{1, \ldots, K\}$ produces an update $\Delta \boldsymbol{W}_k = \boldsymbol{B}_k \boldsymbol{A}_k$, where $\boldsymbol{B}_k \in \mathbb{R}^{d_{\text{out}} \times r}$, $\boldsymbol{A}_k \in \mathbb{R}^{r \times d_{\text{in}}}$, and $r \ll \min(d_{\text{out}}, d_{\text{in}})$ (Hu et al., 2022). Merging combines these independently trained updates into one model that serves all K tasks.

For $Kr \leq \min(d_{\text{out}}, d_{\text{in}})$, let $\boldsymbol{U}^{\text{ref}} \in \mathbb{R}^{d_{\text{out}} \times Kr}$ and $\boldsymbol{V}^{\text{ref}} \in \mathbb{R}^{d_{\text{in}} \times Kr}$ be orthonormal bases spanning the concatenated LoRA factors. Every adapter has the exact core-space representation (Panariello et al., 2026)

$$\Delta \boldsymbol{W}_k = \boldsymbol{U}^{\text{ref}} \, \boldsymbol{M}_k \, \boldsymbol{V}^{\text{ref}^\top}, \qquad \boldsymbol{M}_k \in \mathbb{R}^{Kr \times Kr}. \tag{1}$$

The projection has zero reconstruction error. A weighted merge can therefore be written as

$$\boldsymbol{M}_{\boldsymbol{w}} = \sum_{k=1}^{K} w_k \boldsymbol{M}_k, \qquad \Delta \boldsymbol{W}_{\boldsymbol{w}} = \boldsymbol{U}^{\text{ref}} \, \boldsymbol{M}_{\boldsymbol{w}} \, \boldsymbol{V}^{\text{ref}^\top}. \tag{2}$$

This representation separates the high-dimensional model weights from the compact merge operator. Deterministic methods choose one operator through their merge rule and method-specific settings. BLoM instead infers a distribution over operators in this shared core space.

3.2 MERGE-SELECTION OVERFITTING

Let $A_{\text{val}}(\boldsymbol{w})$ and $A_{\text{test}}(\boldsymbol{w})$ be the validation and test accuracy of a merged model, averaged uniformly over tasks (task-macro accuracy). Grid search over a candidate set $\mathcal{G}$ returns $\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w} \in \mathcal{G}} A_{\text{val}}(\boldsymbol{w})$. For retrospective diagnosis, define the test oracle over the same finite set as $\boldsymbol{w}^\star = \arg\max_{\boldsymbol{w} \in \mathcal{G}} A_{\text{test}}(\boldsymbol{w})$. The merge-selection gap is

$$G_{\text{sel}}(\mathcal{G}) = A_{\text{test}}(\boldsymbol{w}^\star) - A_{\text{test}}(\hat{\boldsymbol{w}}) \geq 0. \tag{3}$$

A larger candidate set creates more opportunities for finite-sample fluctuations to determine the validation winner. Task interference can amplify the resulting test loss because nearby configurations differ more sharply in their behavior. We use the subspace alignment ratio (SAR) of Panariello et al. (2026) as an interference diagnostic, where larger values indicate greater overlap between task-update subspaces. We call a setting interference-dominated when this overlap is accompanied by sharp accuracy variation across merge configurations.

3.3 PILOT MEASUREMENT

We measure G_{sel} on LoRA adapters for Financial PhraseBank (FPB) and FiQA Sentiment Analysis (FiQA-SA) using a Qwen2.5-1.5B base model. The pair shares a three-class label space, yet small changes in the merge operator produce large accuracy differences, making it an interference-dominated setting. At each rank, we evaluate 15 task-arithmetic, 15 TSV-M, and 25 TIES configurations, selecting within each method with up to 200 validation examples per task and comparing against its finite-grid test oracle.

Table 1: Selection gap on the high-interference pair. Entries are task-macro test-accuracy losses relative to each method's grid oracle, using up to 200 validation examples per task.

Rank r	Task Arithmetic	TSV-M	TIES
2	0.0273	0.0164	0.0144
4	0.0109	0.0159	0.0071
8	0.0150	0.0408	**0.0557**

Table 1 shows that the single-split gap depends jointly on merge rule and adapter rank, reaching 5.6 points for TIES at rank 8 rather than increasing monotonically for every rule. On the separate six-task suite, across $B = 200$ validation-bootstrap replicates for three adapter seeds, the $n = 200$ mean gaps range from 0.0005 to 0.0143, the maximum across replicates within each cell ranges from 0.004 to 0.043, and the selected configuration differs from the test oracle in 10% to 99% of replicates. Reducing the validation budget leaves the mean gap near 0.006 to 0.007 but increases the worst observed tail from 0.043 at $n = 200$ to 0.089

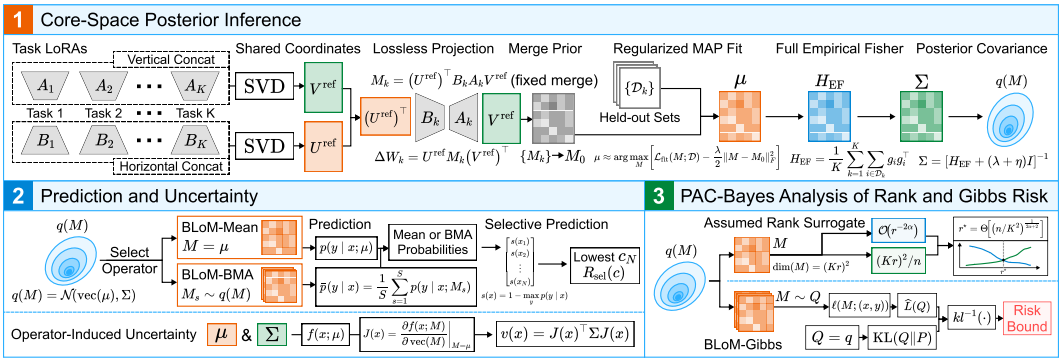

Figure 1: BLoM at one representative projection layer. (1) LoRA factors are losslessly projected into shared core coordinates, where a deterministic merge defines the prior center and regularized fitting with full empirical-Fisher curvature yields the Gaussian operator posterior $q(\boldsymbol{M})$. (2) The posterior mean and samples define BLoM-Mean and BLoM-BMA, selective prediction, and operator-induced uncertainty. (3) The same posterior supports the complexity-optimal rank analysis and the PAC-Bayes bound on BLoM-Gibbs risk.

at $n = 50$. In the original full-validation evaluation, the near-orthogonal six-task suite has within-method gaps below 0.007 and a union-grid oracle gap of 0.0091 over 165 configurations spanning three ranks. The original validation and test splits can also differ in distribution, so the observed gaps diagnose selection instability rather than isolate sampling noise alone. These measurements motivate replacing configuration selection with inference over the compact operator rather than defining another configuration search rule. Appendix D gives the resampling protocol, subgrid analysis, and complete auxiliary results.

4 Inference over Merge Operators

4.1 Overview

BLoM replaces discrete selection among merge configurations with continuous inference over a merge operator in the shared core space of one representative projection layer. A conventional configuration fixes method-specific hyperparameters and therefore produces one operator, whereas BLoM refines a deterministic initialization and infers a distribution over the operator entries. The method has three parts. First, regularized operator fitting and a full empirical-Fisher approximation construct the posterior in compact core space. Second, its mean and posterior samples define the Mean and BMA prediction modes together with operator-induced uncertainty. Third, a PAC-Bayes analysis relates posterior complexity to rank and gives a finite-sample guarantee for the corresponding Gibbs predictor. Figure 1 gives this three-stage overview, and the following subsections develop the stages in the same order.

4.2 Core-Space Posterior Inference

Let $\mathcal{D} = \cup_k \mathcal{D}_k$ be the operator-fitting held-out set, disjoint from adapter training and any calibration or test split, with n_k labeled examples for task k. We request up to $n_k = 200$ examples for the primary accuracy and bound analyses and use up to 100 per task for posterior fitting in the three-split uncertainty protocol. For an operator $\boldsymbol{M}$, define the task-scaled fitting score

$$\mathcal{L}_{\text{fit}}(\boldsymbol{M}; \mathcal{D}) = -\frac{1}{K} \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} \ell_i(\boldsymbol{M}). \tag{4}$$

Here ℓ_i is classification cross-entropy or mean completion-token cross-entropy for language-model scoring. The score sums examples with weight $1/K$, giving equal task contributions at equal sample budgets and defining a generalized posterior.

We use an isotropic Gaussian prior centered at a deterministic core-space initialization,

$$\text{vec}(\mathbf{M} - \mathbf{M}_0) \sim \mathcal{N}(\mathbf{0}, \lambda^{-1} \boldsymbol{I}), \qquad \boldsymbol{M}_0 = \text{Merge}(\boldsymbol{M}_1, \ldots, \boldsymbol{M}_K). \tag{5}$$

Algorithm 1 Posterior over merge operators in core space

Require: adapters $\{(\boldsymbol{B}_k, \boldsymbol{A}_k)\}_{k=1}^{K}$, held-out data $\mathcal{D} = \cup_k \mathcal{D}_k$, prior precision λ, samples S
Ensure: posterior $q(\mathbf{M}) = \mathcal{N}(\mathrm{vec}(\boldsymbol{\mu}), \boldsymbol{\Sigma})$
1: $\boldsymbol{U}^{\mathrm{ref}}, \boldsymbol{V}^{\mathrm{ref}} \leftarrow$ orthonormal bases of the stacked factors and cores $\{\boldsymbol{M}_k\}$ ▷ Eq. 1
2: $\boldsymbol{M}_0 \leftarrow \mathrm{Merge}(\boldsymbol{M}_1, \ldots, \boldsymbol{M}_K)$; $\boldsymbol{\mu} \leftarrow \boldsymbol{M}_0$ ▷ Initialization in Eq. 5
3: $\boldsymbol{\mu} \approx \arg\max_{\boldsymbol{M}} \mathcal{L}_{\mathrm{fit}}(\boldsymbol{M}; \mathcal{D}) - \frac{\lambda}{2}\|\boldsymbol{M} - \boldsymbol{M}_0\|_F^2$ ▷ Adam for 30 steps, Eq. 7
4: $\boldsymbol{H}_{\mathrm{EF}} \leftarrow 1/K$-weighted per-sample gradient outer products at $\boldsymbol{\mu}$ ▷ Eq. 8
5: $\boldsymbol{\Sigma} \leftarrow (\boldsymbol{H}_{\mathrm{EF}} + (\lambda + 10^{-3})\boldsymbol{I})^{-1}$ by eigendecomposition ▷ Gram path for large operators
6: Draw $\{\boldsymbol{M}_s\}_{s=1}^{S}$ by Eq. 10; **return** $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\{\boldsymbol{M}_s\}$, and $D_{\mathrm{KL}}(q\|p)$

The center uses TSV-Core for language models and uniform core averaging for vision, with all other projections fixed at that merge. The generalized posterior target is

$$\pi(\boldsymbol{M} \mid \mathcal{D}) \propto p_0(\boldsymbol{M}) \exp\{\mathcal{L}_{\mathrm{fit}}(\boldsymbol{M}; \mathcal{D})\}. \tag{6}$$

We use suite-specific prior precisions, with the reported values and available selection records given in Appendix H. The pilot sweep characterizes how precision controls posterior dispersion, while local Gaussian evidence provides a separate sensitivity analysis in Appendix F. The numerical PAC-Bayes analysis retains a conservative finite-grid union penalty.

Full Gaussian approximation. We initialize the operator at $\boldsymbol{M}_0$ and approximately optimize its maximum a posteriori (MAP) objective of Eq. 6,

$$\boldsymbol{\mu} \approx \arg\max_{\boldsymbol{M}} \; \mathcal{L}_{\mathrm{fit}}(\boldsymbol{M}; \mathcal{D}) - \frac{\lambda}{2}\|\boldsymbol{M} - \boldsymbol{M}_0\|_F^2. \tag{7}$$

Adam performs 30 refinement steps using the layer indices and optimizer settings in Appendix H. The mean predictor reports this regularized refinement separately from Bayesian model averaging, so its point-accuracy gains are not attributed to averaging.

At $\boldsymbol{\mu}$, let $\boldsymbol{g}_i = \nabla_{\mathrm{vec}\,\boldsymbol{M}} \ell_i(\boldsymbol{M})|_{\boldsymbol{\mu}}$. BLoM uses the correspondingly scaled empirical-Fisher curvature

$$\boldsymbol{H}_{\mathrm{EF}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} \boldsymbol{g}_i \boldsymbol{g}_i^{\top}. \tag{8}$$

This is a positive semidefinite curvature approximation, not the exact sample Hessian or exact generalized Gauss-Newton matrix. We retain its full, non-factorized structure in core space and form

$$\boldsymbol{\Sigma} = (\boldsymbol{H}_{\mathrm{EF}} + (\lambda + \eta)\boldsymbol{I})^{-1}, \qquad q(\mathbf{M}) = \mathcal{N}(\mathrm{vec}(\boldsymbol{\mu}), \boldsymbol{\Sigma}). \tag{9}$$

Here $\eta = 10^{-3}$ is numerical damping, and the inverse is computed by eigendecomposition. When the operator dimension exceeds twice the held-out sample count, an equivalent decomposition of the weighted per-sample gradient Gram matrix avoids materializing a dense square curvature matrix. Unlike diagonal and Kronecker-factored alternatives, the reported posterior preserves all empirical-Fisher couplings within the selected core layer. Algorithm 1 gives the implementation and Appendix G evaluates the approximation.

Scalability. The operator has at most $(Kr)^2$ parameters, with smaller rectangular cores when a projection width limits the shared bases. For $K = 6$ and rank 16, the posterior dimension is 9216, and the complete pipeline runs on a 7B model with one 24GB GPU. The posterior is computed once per suite and reused across predictions, while grid search repeats merge construction and evaluation for each candidate. Runtime measurements and the separately identified memory estimates appear in Table 18.

4.3 Prediction and Uncertainty

BLoM-Mean evaluates the posterior center, $p(y \mid \boldsymbol{x}, \boldsymbol{\mu})$, and serves as the primary accuracy mode. BLoM-BMA averages predictive distributions over posterior samples,

$$\bar{p}(y \mid \boldsymbol{x}) = \frac{1}{S} \sum_{s=1}^{S} p(y \mid \boldsymbol{x}, \boldsymbol{M}_s), \qquad \boldsymbol{M}_s = \boldsymbol{\mu} + \mathrm{unvec}(L\boldsymbol{z}_s), \quad \boldsymbol{z}_s \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{10}$$

where $LL^\top = \mathbf{\Sigma}$ and $S = 30$ unless stated otherwise. Averaging integrates predictive distributions over plausible operators under q, while Mean requires one forward pass per query.

The selective-prediction experiments rank inputs by one minus the maximum predictive probability of the corresponding Mean or BMA predictor. At coverage c, selective risk is the error rate on the most confident c fraction of inputs under this ranking. The posterior also admits a first-order variance for a scalar network score f,

$$v(\boldsymbol{x}) = \boldsymbol{J}(\boldsymbol{x})^\top \mathbf{\Sigma} \boldsymbol{J}(\boldsymbol{x}), \qquad \boldsymbol{J}(\boldsymbol{x}) = \left. \frac{\partial f(\boldsymbol{x}; \boldsymbol{M})}{\partial \operatorname{vec} \boldsymbol{M}} \right|_{\boldsymbol{M}=\boldsymbol{\mu}}. \tag{11}$$

When f is fixed without using an unknown test label, this auxiliary quantity isolates sensitivity to the merge operator and can define the selective rule of Theorem 2. Our covariance diagnostics instead instantiate f with the labeled loss, while the main selective experiments use predictive confidence, so neither empirical result is presented as an application of that theorem.

4.4 PAC-Bayes Analysis of Rank and Gibbs Risk

Complexity-optimal rank. For an approximation-estimation surrogate $Ar^{-2\alpha} + B(Kr)^2/n$, with decay exponent $\alpha > 0$, balancing the terms gives $r^\star = \Theta((n/K^2)^{1/(2\alpha+2)})$. The assumption and proof appear in Assumption 1 and Proposition 1. This relation minimizes the assumed surrogate, not the unknown population risk or the deployed Mean error.

Gibbs-risk bound. BLoM-Gibbs draws one operator $\boldsymbol{M} \sim Q$ and predicts with the sampled model, where the reported analysis takes $Q = q$ from Eq. 9. Its population risk is

$$L_{\mathrm{Gibbs}}(Q) = \mathbb{E}_{\boldsymbol{M}\sim Q}[L(\boldsymbol{M})]. \tag{12}$$

For a data-independent candidate prior P, Theorem 1 bounds this risk through the empirical Gibbs risk and the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(Q\|P)$ under the balanced independent-sampling protocol stated there. The closed-form Gaussian KL depends on the $(Kr)^2$ operator dimension rather than the full weight count. The numerical analysis uses the conservative finite-grid adjustment of Eq. 17 with $m = 300$. For the equal-size setting of Theorem 1, the value in Table 15 is the direct upper inversion of this binary-KL inequality. Its empirical Gibbs term is estimated from 200 posterior draws using a one-sided Student-t normal-approximation upper limit. The theorem gives a finite-sample event under its equal-size assumption, while the reported numerical curve uses the nominal balanced design and additionally inherits this explicit Monte Carlo approximation. Full confidence allocation and diagnostics are given in Appendix A and Appendix F.

The population-risk bound applies only to the randomized BLoM-Gibbs predictor. BLoM-Mean and BLoM-BMA are deployed prediction modes and are evaluated empirically. This separation lets the same operator distribution support practical prediction and a precisely defined theoretical guarantee without identifying the two objects.

5 Experiments

5.1 Experimental Settings

Task suites. We evaluate five standard suite-model settings and one additional cross-domain setting. The finance suite contains six FLARE tasks (Xie et al., 2023) on Qwen2.5-1.5B and Qwen2.5-7B. The NLI suite uses six public Llama-3-8B checkpoints and the evaluation protocol of Stoica et al. (2025). The vision suite covers eight image-classification tasks on CLIP ViT-B/32 and ViT-L/14 under the same protocol. A cross-domain suite merges the six NLI tasks with Financial PhraseBank (FPB) and Twitter Financial News Sentiment (TFSent) on Llama-3-8B. Main comparisons use rank-16 adapters. The PAC-Bayes rank analysis evaluates a preregistered twelve-rank grid on Finance-1.5B.

Comparisons. The twelve merging baselines cover parameter space with task arithmetic, TIES, DARE-TIES, and RegMean (Ilharco et al., 2022; Yadav et al., 2023; Yu et al., 2024; Jin et al., 2023), aligned subspaces with KnOTS, TSV-M, and Iso-C (Stoica et al., 2025; Gargiulo et al., 2025;

Table 2: Suite-average raw accuracy across four representative standard settings and the cross-domain setting. All twelve baselines are shown in the standard-setting columns, while the cross-domain column contains the three evaluated baselines. Bold marks the best merging method in each column, and Appendix E reports ViT-L/14 and the complete per-task results.

Method	Finance-1.5B	Finance-7B	NLI	Vision-B/32	Cross-domain
Reference models					
Pretrained base	0.083	0.000	0.434	0.469	–
Individual LoRA	0.626	0.671	0.817	0.851	–
Joint multi-task LoRA	0.633	0.683	0.837	0.661	–
Parameter-space baselines					
Task Arithmetic	0.559	0.371	0.746	0.509	0.638
TIES	0.547	0.531	0.740	0.510	0.734
DARE-TIES	0.519	0.528	0.698	0.218	–
RegMean	0.603	**0.673**	0.824	**0.674**	–
Aligned-subspace baselines					
KnOTS	0.549	0.532	0.746	0.513	–
TSV-M	0.565	0.616	0.733	0.540	0.672
Iso-C	0.575	0.635	0.666	0.537	–
LoRA-geometry baselines					
CtM	0.563	0.370	0.745	0.518	–
SVC	0.559	0.373	0.739	0.511	–
TARA	0.550	0.556	0.748	0.495	–
TSPA	0.558	0.371	0.745	0.510	–
Pico	0.564	0.614	0.725	0.528	–
TSV-Core initialization	0.478	0.544	0.758	–	–
BLoM-Mean	**0.606** $\pm$ 0.016	0.594	**0.893**	0.535	**0.873**

Marczak et al., 2025), and LoRA geometry with CtM, SVC, TARA, TSPA, and Pico (He et al., 2026; Li et al., 2026b; Jeong et al., 2026; Li et al., 2026c; Tang & Yang, 2026). Pretrained, individual-task, and joint multi-task models serve as contextual references rather than merging baselines. TSV-Core initializes the language operators, while vision uses uniform core averaging. Full implementation details and method-specific selection rules appear in Appendix H.4.

Metrics and protocol. We report raw accuracy averaged across tasks. Uncertainty metrics are expected calibration error (ECE), negative log likelihood (NLL), area under the risk-coverage curve (AURC), and selective accuracy at 90% coverage. Lower AURC indicates better error ranking as progressively less confident examples are rejected, while selective accuracy reports accuracy at the stated retained fraction. The Gibbs-risk analysis reports the upper-value margin and its rank correlation with the BLoM-Gibbs test-risk estimate across ranks that pass a preregistered reproducibility check. This check requires source checkpoints to reproduce reference accuracy within a fixed tolerance, and Appendix F preserves the excluded-rank records. Finance-1.5B BLoM results average three adapter seeds, while the other suite-model settings use one adapter set. Deterministic baselines use their validation-selected settings. BLoM uses held-out labels for gradient-based operator fitting, whereas most baselines use them for configuration selection. BMA uses $S = 30$ posterior samples. All experiments run on one 24GB GPU. Appendix H gives splits, budgets, hyperparameters, and per-seed results.

5.2 Main Results

Interference-dominated suites. Table 2 compares BLoM-Mean with all twelve baselines on four representative standard settings and with the three evaluated baselines in the cross-domain setting. On NLI, it reaches 0.893 and improves RegMean, the strongest of twelve baselines, by 6.9 points. The largest task-level gains over RegMean are 0.300 on QNLI and 0.143 on RTE. Table 3 resolves the cross-domain result by task, where BLoM-Mean reaches 0.873 and improves TIES, the strongest compared method, by 13.9 points. Every cross-domain task remains above 0.82. The paired task-

Table 3: Cross-domain scaling to K=8 with six NLI tasks and two finance-sentiment tasks, fpb and tfsent, on Llama-3-8B at rank 16 and λ=10. BLoM-Mean keeps every task above 0.82 and improves the suite average by 0.139 over the best of three compared baselines. Bold marks the best entry per column.

Method	SNLI	MNLI	SICK	QNLI	RTE	SciTail	FPB	TFSent	Avg
Task Arithmetic	0.787	0.473	0.658	0.287	0.584	0.757	0.760	0.800	0.638
TIES	0.863	0.783	0.885	0.193	0.584	**0.933**	0.820	0.813	0.734
TSV-M	0.780	0.817	0.559	0.223	0.520	0.823	0.803	**0.847**	0.672
BLoM-Mean	**0.890**	**0.907**	0.875	**0.880**	**0.844**	0.927	**0.823**	0.840	**0.873**
BLoM-BMA	0.817	0.903	**0.888**	0.813	**0.844**	0.923	0.780	0.810	0.847

level effects for both suites have test-set bootstrap intervals disjoint from zero under the fixed public checkpoints. The remaining per-task panels are reported in Appendix E.

Regime dependence. On Finance-1.5B, BLoM-Mean is comparable to the strongest baseline in suite accuracy at 0.606 ± 0.016 versus 0.603, and it leads four of six task columns. On Finance-7B, RegMean leads the six-task table, while BLoM-Mean improves its TSV-Core initialization by 5.0 points and remains positive on every task. On vision, BLoM is comparable to several aligned-subspace methods, reaching 0.535 on ViT-B/32 and 0.689 on ViT-L/14. The largest accuracy gains therefore occur in the interference-dominated suites.

Beyond discrete selection. On the high-interference pilot pair, BLoM-Mean reaches 0.798, improving validation-selected TSV-M by 4.8 points and exceeding its finite-grid test oracle, as shown in Figure 2. The comparison tests regularized operator refinement beyond the discrete candidate grid. Across the full suites where TSV-Core is directly evaluated, BLoM-Mean improves its initialization by 12.8 points on Finance-1.5B, 5.0 points on Finance-7B, and 13.5 points on NLI. The Finance-1.5B value uses completion-token exact match and is concentrated in the Headlines column, whereas the matched five-task label-probability comparison in Table 17 yields much smaller differences among supervised variants. The NLI gain therefore provides the clearer evidence for refinement beyond a fixed deterministic merge. The rank and validation-size stress tests are analyzed in Section 5.3.

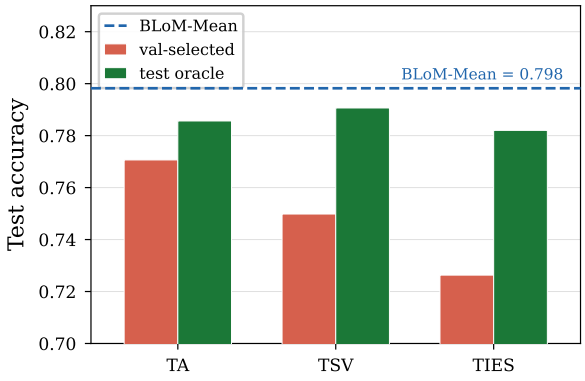

Figure 2: Continuous fitting versus discrete selection on the high-interference pair at rank 8. BLoM-Mean improves validation-selected TSV-M by 4.8 points and exceeds its finite-grid oracle.

Selective prediction. Table 14 compares BLoM with fitted calibrators under a three-split protocol. The available held-out pool is divided into an operator-fitting set for posterior inference and a disjoint calibration set for the fitted calibrators, and all metrics are measured on the untouched test split. On NLI, BLoM-Mean obtains the best risk ranking with AURC 0.028, alongside selective accuracy 0.923 at 90% coverage without a separate calibration fit. Vector scaling, which fits per-class logit scales and offsets, reaches 0.928 selective accuracy, while Figure 7 in Appendix E shows that BLoM-Mean has the lowest selective risk at 50% and 70% coverage. BLoM-BMA remains competitive but does not improve on the Mean predictor in this shared protocol.

Directly fitted calibrators lead on the finance calibration metrics, where vector scaling reaches ECE 0.089 and AURC 0.204 compared with 0.128 and 0.255 for BLoM-Mean. This setting favors methods optimized for calibration, while the NLI result shows that the BLoM predictor can provide a strong risk ordering without that extra fit. Leave-one-task-out detection remains near chance on

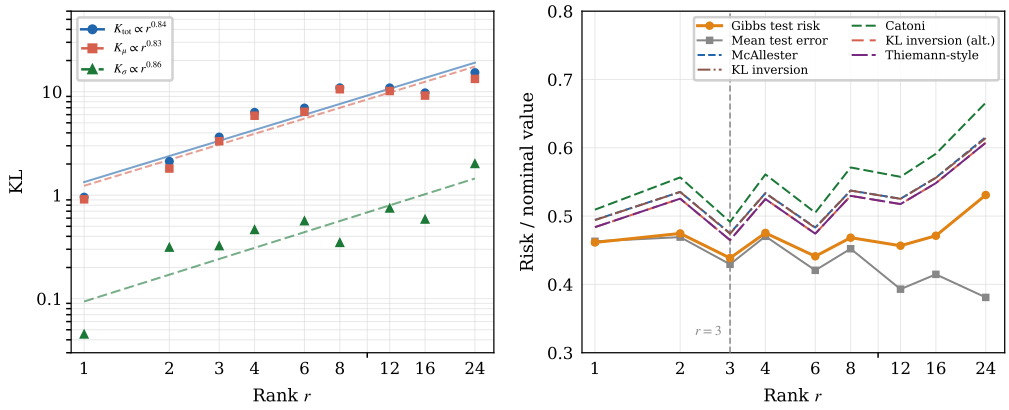

Figure 3: Posterior KL decomposition and PAC-Bayes sensitivity on the Finance-1.5B rank grid (λ=100, nominal n=1200, m=300 union penalty). **Left.** The KL to the prior splits into mean and covariance terms, with sublinear fitted growth across the verified ranks. **Right.** The direct inversion of Eq. 17, labelled KL inversion, is the nominal balanced-design curve reported in Table 15, while the other stored analytical curves are shown only as sensitivity references. The nominal curve is minimized at r=3 (grey dashed), with descriptive continuous interpolation 2.99.

both suites. We therefore interpret these scores as selective confidence within the merged task family rather than as out-of-distribution detection.

5.3 Analysis

Posterior construction. BLoM-Mean improves TSV-Core in all nine rank-precision cells of the pilot sweep. At rank 8, increasing λ from 10 to 1000 reduces mean posterior coordinate standard deviation from 0.314 to 0.031, showing how precision controls dispersion. A separate layer-factorized posterior over 28 projections improves six-task Mean accuracy from 0.4781 to 0.5451. Thus, the one-layer choice controls computation rather than limiting the operator formulation. Refinement is stable from one to one hundred steps, while a full-weight diagonal approximation increases AURC from 0.257 to 0.340. The matched supervised controls in Table 17 separate scalar learning, operator fitting, and averaging on five finance tasks. Together, these diagnostics connect the compact operator formulation to its practical accuracy, uncertainty, and computational behavior. The precision sweep, multi-layer runtime, and covariance-quality comparisons appear in Appendix D and Appendix G.

Rank and PAC-Bayes behavior. Under the nominal balanced-sample design, the Monte Carlo-adjusted binary-KL inversions remain below the unit-loss ceiling across nine verified ranks and exceed the BLoM-Gibbs test-risk estimates by 3.3 to 8.5 points. They track Gibbs test risk with Spearman rank correlation $\rho = 0.833$ and $p = 5.3 \times 10^{-3}$. Their correlation with the deterministic BLoM-Mean error is not significant, which agrees with the bounded object being BLoM-Gibbs. The mean and covariance contributions to KL grow sublinearly with fitted exponents 0.83 and 0.86, and all nine task-stratified bootstrap upper limits remain below the reported curve. The discrete minimum of this curve occurs at rank 3 and matches the measured Gibbs-test minimum. Mean error is lowest at rank 24, so the Gibbs-curve minimum is not a deployment rule for Mean. Appendix F reports the KL decomposition, Monte Carlo diagnostics, halted-rank records, and ablations.

6 Conclusion

BLoM reframes LoRA merging as inference over a compact merge operator rather than selection among candidate configurations. A lossless core-space representation makes a full-covariance Gaussian approximation practical at large-model scale. The resulting distribution supports an accuracy-oriented mean predictor, predictive model averaging, selective confidence, and PAC-Bayes analysis of Gibbs risk within one framework. More broadly, the results show that the statistical treatment of merge configurations complements the geometry used to represent them.

AI USE STATEMENT

Generative AI tools supported writing, literature retrieval, research development, and mathematical reasoning. This assistance included section drafting and polishing, literature discovery, research ideation and execution, code and figure-script development, and work on mathematical claims and proofs. The tools were not used to generate datasets or experimental measurements, translate research materials, clean or relabel datasets, or conduct qualitative or thematic data analysis. The authors reviewed all AI-assisted material against the manuscript, code, and experiment records, made all final decisions, and take responsibility for the final text, claims, and artifacts.

ETHICS STATEMENT

This work studies adapter merging with publicly available models and benchmark datasets and introduces no new data collection or human subject study. The finance datasets serve only as research benchmarks, and the results should not be interpreted as financial advice. Adapter merging can combine biases and privacy risks inherited from source models and adapters. The PAC-Bayes theorem applies only to the Gibbs predictor under its stated assumptions, while the realized unequal task strata make the reported numerical curve a nominal balanced-design analysis. The Monte Carlo adjustment uses the normal approximation described in Appendix A, and the selective prediction results do not establish deployment safety or reliable out of distribution detection. Any released artifacts should preserve the licenses and attribution requirements of the source models, adapters, and datasets.

REPRODUCIBILITY STATEMENT

Sections 3 and 4, together with Algorithm 1, specify the method and its assumptions. Appendix A provides complete proofs. Appendix D, Appendix E, Appendix F, Appendix G, and Appendix H document the preregistered pilot, complete results, posterior diagnostics, ablations, data splits, hyperparameters, hardware, software, and baseline selection rules. An anonymized supplementary archive will contain the code, adapter configurations, evaluation scripts, random seeds, reproducibility-check records, and Monte Carlo diagnostics needed to reproduce the reported experiments.

REFERENCES

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2209.04836.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(236):1–41, 2016.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning*, 16(4): 494–591, 2023.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *International Conference on Machine Learning*, 2015.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Long Minh Bui, Tuan Anh Le Van, Tung Phi Duc, Phi Le Nguyen, Jana Doppa, and Trong Nghia Hoang. Model merging as probabilistic inference in fine-tuning parameter space. In *Conference on Uncertainty in Artificial Intelligence*, 2026.

Olivier Catoni. *PAC-Bayesian Supervised Classification*. IMS Lecture Notes-Monograph Series, 2007.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Nico Daheim, Thomas Möllenhoff, Edoardo Maria Ponti, Iryna Gurevych, and Mohammad Emtiyaz Khan. Model merging by uncertainty-based gradient matching. In *International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2310.12808.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux: Effortless Bayesian deep learning. In *Advances in Neural Information Processing Systems*, 2021.

Pala Tej Deep, Rishabh Bhardwaj, and Soujanya Poria. DELLA-merging: Reducing interference in model merging through magnitude-based sampling, 2024. arXiv preprint arXiv:2406.11617.

Monroe D. Donsker and S. R. Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975. doi: 10.1002/CPA.3160290405.

Felix Draxler, K. Veschgini, M. Salmhofer, and F. Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning*, 2018.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367.

Rahim Entezari, Samuel K. Ainsworth, Eugene Belilovsky, and Alexander A. Alemi. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. doi: 10.1109/CVPR52734.2025.01742.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems*, 2018.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Yifei He, Siqi Zeng, Yuzheng Hu, Rui Yang, Tong Zhang, and Han Zhao. MergeBench: A benchmark for merging domain-specialized LLMs. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.48550/arXiv.2505.10833.

Zhengbao He, Ruiqi Ding, Zhehao Huang, Ruikai Yang, Tao Li, and Xiaolin Huang. Compress then merge: From multiple LoRAs into one low-rank adapter. In *International Conference on Machine Learning*, 2026.

Wassily Hoeffding. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, 27(3):713–721, 1956. doi: 10.1214/aoms/1177728178.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. LoRAHub: Efficient cross-task generalization via dynamic LoRA composition. In *Conference on Language Modeling*, 2024a. doi: 10.48550/arXiv.2307.13269.

Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. EMR-merging: Tuning-free high-performance model merging. In *Advances in Neural Information Processing Systems*, pp. 122741–122769, 2024b. doi: 10.48550/arXiv.2405.17461.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2022. doi: 10.48550/arXiv.2212.04089.

Pavel Izmailov, Wesley J. Maddox, Polina Kirichenko, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Subspace inference for Bayesian deep learning. In *Conference on Uncertainty in Artificial Intelligence*, 2019.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Wooseong Jeong, Wonyoung Lee, and Kuk-Jin Yoon. Preference-aligned LoRA merging: Preserving subspace coverage and addressing directional anisotropy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. doi: 10.48550/arXiv.2603.26299.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2212.09849.

Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. REPAIR: Renormalizing permuted activations for interpolation registration. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2211.08403.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, T. S. Filho, Hao Song, and Peter A. Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems*, 2019.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, 2019. doi: 10.1088/1742-5468/abc62b.

Sanwoo Lee, Jiahao Liu, Qifan Wang, Jingang Wang, Xunliang Cai, and Yunfang Wu. Dynamic fisher-weighted model merging via Bayesian optimization. In *North American Chapter of the Association for Computational Linguistics*, pp. 4923–4935, 2025. doi: 10.48550/arXiv.2504.18992.

Kaiyang Li, Shaobo Han, Qing Su, and Shihao Ji. Bayesian model merging, 2026a. arXiv preprint arXiv:2605.12843.

Yayuan Li, Ze Peng, Jian Zhang, Jintao Guo, Yue Duan, and Yinghuan Shi. When shared knowledge hurts: Spectral over-accumulation in model merging. In *International Conference on Machine Learning*, 2026b. doi: 10.48550/arXiv.2602.05536.

Zijian Li, Xiachong Feng, Weitao Ma, Yichong Huang, Xiaocheng Feng, and Bing Qin. Two-stage parameter alignment for multi-LoRA merging in large language models. In *Findings of the Association for Computational Linguistics*, pp. 30079–30093, 2026c. doi: 10.18653/v1/2026.findings-acl.

Zehua Liu, Han Wu, Yuxuan Yao, Ruifeng She, Xiongwei Han, Tao Zhong, and Mingxuan Yuan. LoRE-Merging: Exploring low-rank estimation for large language model merging. In *Findings of the Association for Computational Linguistics*, pp. 21919–21926, 2025. doi: 10.48550/arXiv.2502.10749.

Christos Louizos and Max Welling. Structured and efficient variational deep learning with matrix Gaussian posteriors. In *International Conference on Machine Learning*, 2016.

Chenyang Lu, Jiaru Li, Jinman Zhao, Xinran Chen, Yining Wang, Renyi Cai, Yuchen Li, and Chao He. On the representation geometry of LoRA model merging. In *Findings of the Association for Computational Linguistics*, pp. 5289–5304, 2026. doi: 10.18653/v1/2026.findings-acl.

David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3), 1992. doi: 10.1162/neco.1992.4.3.415.

Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. A simple baseline for Bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, 2019.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *International Conference on Machine Learning*, 2025. doi: 10.48550/arXiv.2502.04959.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *International Conference on Machine Learning*, 2015.

Michael S. Matena and Colin A. Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, 2022. doi: 10.52202/068431-1287.

Andreas Maurer. A note on the PAC-Bayesian theorem, 2004. arXiv preprint arXiv:cs/0411099.

David A. McAllester. Some PAC-Bayesian theorems. In *Conference on Computational Learning Theory*, pp. 230–234, 1998. doi: 10.1023/A:1007618624809.

David A. McAllester. Simplified PAC-Bayesian margin bounds. *Conference on Computational Learning Theory*, pp. 203–215, 2003. doi: 10.1007/978-3-540-45167-9_16.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Aaron Mishkin, Frederik Kunstner, Didrik Nielsen, Mark Schmidt, and Mohammad Emtiyaz Khan. SLANG: Fast structured covariance approximations for Bayesian deep learning with natural gradient. In *Advances in Neural Information Processing Systems*, 2018.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996. doi: 10.1007/978-1-4612-0745-0.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems*, pp. 66727–66754, 2023. doi: 10.48550/arXiv.2305.12827.

Aniello Panariello, Daniel Marczak, Simone Magistri, Angelo Porrello, Bartłomiej Twardowski, Andrew D. Bagdanov, Simone Calderara, and Joost van de Weijer. Accurate and efficient low-rank model merging in core space. In *Advances in Neural Information Processing Systems*, 2026. doi: 10.48550/arXiv.2509.17786.

Roi Paul. Spectral geometry of LoRA adapters encodes training objective and predicts harmful compliance, 2026. arXiv preprint arXiv:2604.08844.

Angelo Porrello, Pietro Buzzega, Felix Dangel, Thomas Sommariva, Riccardo Salami, Lorenzo Bonicelli, and Simone Calderara. Dataless weight disentanglement in task arithmetic via Kronecker-factored approximate curvature. In *International Conference on Learning Representations*, 2026. doi: 10.48550/arXiv.2602.17385.

Akshara Prabhakar, Yuanzhi Li, Karthik Narasimhan, Sham Kakade, Eran Malach, and Samy Jelassi. LoRA soups: Merging LoRAs for practical skill composition tasks. In *International Conference on Computational Linguistics*, 2025. doi: 10.48550/arXiv.2410.13025.

Fuli Qiao and Mehrdad Mahdavi. Merge before forget: A single LoRA continual learning via continual merging, 2025. arXiv preprint arXiv:2512.23017.

Adir Rahamim, Asaf Yehudai, Boaz Carmeli, Leshem Choshen, Yosi Mass, and Yonatan Belinkov. Will it merge? on the causes of model mergeability. In *Findings of the Association for Computational Linguistics*, pp. 26551–26570, 2026. doi: 10.18653/v1/2026.findings-acl.

Siobhan Reid, Zhixiang Chi, Li Gu, Omid Reza Heidari, Ziqiang Wang, and Yang Wang. DA-MergeLoRA: Hypernetwork-based LoRA merging for few-shot test-time domain adaptation. In *European Conference on Computer Vision*, 2026.

Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured laplace approximations for overcoming catastrophic forgetting. In *Advances in Neural Information Processing Systems*, 2018.

Keumseo Ryum and Joonhyuk Kang. CT-Merging: Consensus directions and task-level scaling for LoRA adapter merging, 2026. arXiv preprint arXiv:2607.20561.

John Shawe-Taylor and Robert C. Williamson. A PAC analysis of a Bayesian estimator. In *Conference on Computational Learning Theory*, pp. 2–9, 1997. doi: 10.1145/267460.267466.

Donald Shenaj, Ondrej Bohdal, Taha Ceritli, Mete Ozay, Pietro Zanuttigh, and Umberto Michieli. K-Merge: Online continual merging of adapters for on-device large language models. In *Annual Meeting of the Association for Computational Linguistics*, 2026. doi: 10.48550/arXiv.2510.13537.

Mingyang Song and Mao Zheng. Model merging in the era of large language models: Methods, applications, and future directions, 2026. arXiv preprint arXiv:2603.09938.

George Stoica, Daniel Bolya, Jakob Bjorner, Pratik Ramesh, Taylor Hearn, and Judy Hoffman. ZipIt! merging models from different tasks without training. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2305.03053.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with SVD to tie the Knots. In *International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2410.19735.

Daichi Sugita, Huu-Long Pham, Makoto P. Kato, Hiroaki Ohshima, Sumio Fujita, and Yoshiyuki Shoji. Which LoRA should be merged next? retrieving an additional LoRA from a target image. In *ACM International Conference on Multimedia Retrieval*, pp. 99–107, 2026. doi: 10.1145/3805622.3810865.

Yixuan Tang and Yi Yang. Crowded in b-space: Calibrating shared directions for LoRA merging, 2026. arXiv preprint arXiv:2604.16826.

Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. doi: 10.1145/1968.1972.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. BLoB: Bayesian low-rank adaptation by backpropagation for large language models. In *Advances in Neural Information Processing Systems*, pp. 67758–67794, 2024. doi: 10.48550/arXiv.2406.11675.

Zhengxuan Wei, Yi Dong, Zonghui Li, Xianhui Lin, Xing Liu, Hong Gu, Shaofeng Zhang, Wenbin Li, and Qi Fan. SSR-Merge: Subspace signal routing for training-free LoRA merging in diffusion models. In *International Conference on Machine Learning*, 2026.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, 2022a. doi: 10.48550/arXiv.2203.05482.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022b. doi: 10.1109/CVPR52688.2022.00780.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of LoRA experts. In *International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2404.13628.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. In *Advances in Neural Information Processing Systems*, pp. 33469–33484, 2023. doi: 10.48550/arXiv.2306.05443.

Bo Xu, Haotian Wu, Hehai Lin, Weiquan Huang, Beier Zhu, Yao Shu, and Chengwei Qin. ACE-merging: Data-free model merging with adaptive covariance estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. doi: 10.48550/arXiv.2603.02945.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, pp. 7093–7115, 2023. doi: 10.52202/075280-0310.

Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. In *International Conference on Learning Representations*, 2024a. doi: 10.48550/arXiv.2308.13111.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in LLMs, MLLMs, and beyond: Methods, theories, applications and opportunities, 2024b. arXiv preprint arXiv:2408.07666.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. AdaMerging: Adaptive model merging for multi-task learning. In *International Conference on Learning Representations*, 2024c. doi: 10.48550/arXiv.2310.02575.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super Mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning*, 2024. doi: 10.48550/arXiv.2311.03099.

Ziyu Zhao, Tao Shen, Didi Zhu, Zexi Li, Jing Su, Xuwu Wang, Kun Kuang, and Fei Wu. Merging LoRAs like playing LEGO: Pushing the modularity of LoRA to extremes through rank-wise clustering. In *International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2409.16167.

Shenghe Zheng, Hongzhi Wang, Chenyu Huang, Xiaohui Wang, Tao Chen, Jiayuan Fan, Shuyue Hu, and Peng Ye. Decouple and orthogonalize: A data-free framework for LoRA merging, 2025. arXiv preprint arXiv:2505.15875.

Luca Zhou, Bo Zhao, Rose Yu, and Emanuele Rodolà. Demystifying mergeability: Interpretable properties to predict model merging success, 2026. arXiv preprint arXiv:2601.22285.

APPENDIX

CONTENTS

A Statements and Proofs

A.1 Statement and Proof of Theorem 1

A.2 Statement and Proof of Lemma 1

A.3 Statement and Proof of Proposition 1

A.4 Statement and Proof of Proposition 2

A.5 Statement and Proof of the Selective-Risk Bound

A.6 Remark on Assumption 1

B Notation

C Extended Related Work

D Pilot Study Details

D.1 Adapters and Grids

D.2 Selection Gaps on the Near-Orthogonal Suite

D.3 Selection-Gap Resampling on Six Tasks

D.4 Operator-Refinement Gains on Six Tasks

D.5 Multi-Layer Posterior

D.6 Repair on the High-Interference Pair

D.7 Complexity Scaling

D.8 Prior-Precision Scaling

E Complete Main-Result Tables

F Additional Analysis

G Posterior Approximation Quality

G.1 Gaussian Approximation and Langevin Diagnostics

G.2 Covariance Structure

H Implementation and Reproducibility Details

H.1 Posterior Computation

H.2 Hardware and Budgets

H.3 Adapter Training

H.4 Baselines and External Code

H.5 Selected Hyperparameters

A Statements and Proofs

A.1 Statement and Proof of Theorem 1

Throughout, $\ell(\boldsymbol{M}, (\boldsymbol{x}, y)) \in \{0, 1\}$ is the zero-one loss of the merged model on one sample, and $\hat{L}(\boldsymbol{M})$ and $L(\boldsymbol{M})$ are its empirical and population risks. For a distribution Q over operators, $\hat{L}(Q) = \mathbb{E}_{\mathbf{M}\sim Q}[\hat{L}(\boldsymbol{M})]$ and $L(Q) = \mathbb{E}_{\mathbf{M}\sim Q}[L(\boldsymbol{M})]$. For a balanced stratified suite, $n = \sum_k n_k$ and $L(\boldsymbol{M})$ is the average of the task-specific population risks with weights n_k/n, which reduces to the uniform task-macro risk when all n_k are equal.

Theorem 1 (PAC-Bayes bound (McAllester, 1998; Maurer, 2004))**.** *Suppose the observations are independent within and across tasks and all task strata have the same fixed size. With probability at least $1 - \delta$ over $\mathcal{D}$, simultaneously for all Q over $\mathbf{M}$,*

$$\mathrm{kl}\big(\hat{L}(Q) \,\big\|\, L(Q)\big) \;\le\; \frac{D_{\mathrm{KL}}(Q\|P) + \ln \frac{2\sqrt{n}}{\delta}}{n}, \tag{13}$$

where P is any prior independent of $\mathcal{D}$ and $\mathrm{kl}(p\|q)$ is the binary KL.

The argument combines the standard PAC-Bayes change of measure (Valiant, 1984; Shawe-Taylor & Williamson, 1997; McAllester, 1998; 2003; Catoni, 2007) with a stratified extension of the concentration step in Maurer (2004). For a fixed operator $\mathbf{M}$, write the losses as independent Bernoulli variables $X_1, \ldots, X_n$ with possibly task-dependent means and average mean $L(\boldsymbol{M})$. The convex-order result of Hoeffding (1956) states that their sum is dominated by a binomial variable with the same average mean for every convex function. Because $s \mapsto \exp\{n\,\mathrm{kl}(s/n\|L(\boldsymbol{M}))\}$ is convex, the binomial exponential-moment bound of Maurer (2004) therefore gives, for $n \ge 8$,

$$\mathbb{E}_{\mathcal{D}} \exp\Big(n\,\mathrm{kl}\big(\hat{L}(\boldsymbol{M}) \,\big\|\, L(\boldsymbol{M})\big)\Big) \;\le\; 2\sqrt{n}. \tag{14}$$

Taking expectation under the prior P and applying Markov's inequality gives, with probability at least $1 - \delta$ over $\mathcal{D}$,

$$\mathbb{E}_{\mathbf{M}\sim P} \exp\Big(n\,\mathrm{kl}\big(\hat{L}(\boldsymbol{M}) \,\big\|\, L(\boldsymbol{M})\big)\Big) \;\le\; \frac{2\sqrt{n}}{\delta}. \tag{15}$$

The Donsker-Varadhan change-of-measure inequality (Donsker & Varadhan, 1975) gives $\mathbb{E}_Q[f] \le D_{\mathrm{KL}}(Q\|P) + \ln \mathbb{E}_P[e^f]$ for any admissible Q and measurable f. Applying it after joint convexity of binary KL yields

$$n\,\mathrm{kl}\big(\hat{L}(Q) \,\big\|\, L(Q)\big) \;\le\; n\,\mathbb{E}_{\mathbf{M}\sim Q}\,\mathrm{kl}\big(\hat{L}(\boldsymbol{M}) \,\big\|\, L(\boldsymbol{M})\big) \;\le\; D_{\mathrm{KL}}(Q\|P) + \ln \mathbb{E}_P e^{n\mathrm{kl}(\hat{L}\|L)}, \tag{16}$$

where the first step uses joint convexity of kl and the second is the Donsker-Varadhan inequality applied to the function $n\,\mathrm{kl}(\hat{L}(\boldsymbol{M})\|L(\boldsymbol{M}))$. Combining Eq. 15 and Eq. 16 yields Eq. 13 simultaneously for all Q with probability $1 - \delta$. □

The bound concerns Gibbs risk $L(Q) = \mathbb{E}_{\mathbf{M}\sim Q}[L(\mathbf{M})]$, which is the expected loss of a randomized operator draw. Mean evaluates at $\boldsymbol{\mu}$, whereas BMA averages $S{=}30$ predictive distributions before the argmax. These two prediction modes differ from the Gibbs classifier and are evaluated empirically in Section 5.

The theorem requires P to be independent of $\mathcal{D}$, while the Finance rank analysis inspects $\lambda \in \{10, 100, 1000\}$ using data-dependent diagnostics and reports $\lambda{=}100$. The prior center $\boldsymbol{M}_0$ depends on the adapters alone, so only precision is data dependent. Applying Eq. 13 at confidence $1-\delta/m$ to every candidate prior and taking a union bound gives, with probability at least $1 - \delta$, simultaneously for all Q and grid priors,

$$\mathrm{kl}\big(\hat{L}(Q) \,\big\|\, L(Q)\big) \;\le\; \frac{D_{\mathrm{KL}}(Q\|P) + \ln \frac{2\sqrt{n}\,m}{\delta}}{n}. \tag{17}$$

Under the equal-size premise of Theorem 1, the direct upper inversion of Eq. 17 bounds the Gibbs risk of Eq. 12. The factor $m{=}300$ conservatively upper-bounds five precision slots, twelve preregistered ranks, and the five analytical variants retained in the sensitivity figures. The Finance rank analysis uses only the three recorded precision values above, so this multiplicity factor overcounts its actual precision grid. Only the direct binary-KL inversion determines the reported bound, so

retaining the larger multiplicity penalty is conservative. Table 15 uses $\delta_{\mathrm{PAC}}{=}0.04$ and replaces $\hat{L}(Q)$ with a one-sided familywise Student-t upper limit from $S{=}200$ posterior draws. This Monte Carlo step has nominal coverage under the stated normal approximation and is not a distribution-free finite-sample inequality. Allocating $\delta_{\mathrm{MC}}{=}0.01$ across twelve ranks therefore gives a nominal overall level of 0.95 for the reported numerical values, while Eq. 17 itself retains its finite-sample PAC-Bayes guarantee for the exact empirical Gibbs risk under the theorem assumptions. The stored numerical analysis uses the preregistered nominal design of 200 examples for each of six tasks, hence $n{=}1200$. The realized Finance-1.5B operator-fitting strata contain 200, 200, 188, 200, 100, and 100 unique examples. Because these sizes are unequal, Table 15 is a nominal balanced-design sensitivity analysis rather than a finite-sample instantiation of Theorem 1 for the realized sample.

A.2 Statement and Proof of Lemma 1

Lemma 1 (Closed-form complexity)**.** *For the prior in Eq. 5 and Gaussian approximation in Eq. 9,*

$$D_{\mathrm{KL}}(q\|p) \;=\; \frac{1}{2}\Big[\,\lambda\|\boldsymbol{\mu}-\boldsymbol{M}_0\|_F^2+\lambda\,\mathrm{Tr}(\boldsymbol{\Sigma})-(Kr)^2\big(1+\ln\lambda\big)-\ln\det\boldsymbol{\Sigma}\,\Big]. \tag{18}$$

Both distributions are Gaussian over $\mathrm{vec}(\mathbf{M})\in\mathbb{R}^k$ with $k=(Kr)^2$. The prior has mean $\mathrm{vec}(\boldsymbol{M}_0)$ and covariance $\boldsymbol{\Sigma}_p=\lambda^{-1}\boldsymbol{I}_k$, and the posterior has mean $\mathrm{vec}(\boldsymbol{\mu})$ and covariance $\boldsymbol{\Sigma}_q=\boldsymbol{\Sigma}$. The Gaussian KL formula gives

$$D_{\mathrm{KL}}(q\|p) \;=\; \frac{1}{2}\Big[\,\mathrm{Tr}\left(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q\right)+\mathrm{vec}(\boldsymbol{\mu}-\boldsymbol{M}_0)^\top\boldsymbol{\Sigma}_p^{-1}\,\mathrm{vec}(\boldsymbol{\mu}-\boldsymbol{M}_0)-k+\ln\frac{\det\boldsymbol{\Sigma}_p}{\det\boldsymbol{\Sigma}_q}\Big]. \tag{19}$$

Now substitute the three ingredients. First, $\boldsymbol{\Sigma}_p^{-1}=\lambda\boldsymbol{I}_k$, so $\mathrm{Tr}(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q)=\lambda\,\mathrm{Tr}(\boldsymbol{\Sigma})$. Second, $\mathrm{vec}(\boldsymbol{\mu}-\boldsymbol{M}_0)^\top\boldsymbol{\Sigma}_p^{-1}\,\mathrm{vec}(\boldsymbol{\mu}-\boldsymbol{M}_0)=\lambda\|\boldsymbol{\mu}-\boldsymbol{M}_0\|_F^2$. Third, $\det\boldsymbol{\Sigma}_p=\lambda^{-k}$, so $\ln\det\boldsymbol{\Sigma}_p=-k\ln\lambda$. Collecting terms and using $k=(Kr)^2$ gives Eq. 18. □

A.3 Statement and Proof of Proposition 1

The optimal rank trades approximation error against estimation error. Under the spectral-decay condition detailed in Appendix A.6, the Eckart-Young theorem (Eckart & Young, 1936) gives an $O(r^{-2\alpha})$ matrix-reconstruction rate. Assumption 1 separately assumes that this behavior transfers to task risk, while the $(Kr)^2/n$ estimation term follows the dimension dependence of Lemma 1.

Assumption 1 (Approximation-estimation decomposition)**.** *There exist $\alpha, A, B>0$ and $L_\infty\geq 0$ such that the best rank-r operator inferred from n samples satisfies*

$$L(r,n) \;\leq\; L_\infty + A\,r^{-2\alpha}+B\,\frac{(Kr)^2}{n}+o\Big(r^{-2\alpha}+\frac{(Kr)^2}{n}\Big). \tag{20}$$

Proposition 1 (Complexity-optimal rank law)**.** *Under Assumption 1,*

$$r^\star \;=\; \Theta\Big(\big(n/K^2\big)^{\frac{1}{2\alpha+2}}\Big). \tag{21}$$

Minimize the r-dependent part of Eq. 20, $f(r)=Ar^{-2\alpha}+BK^2r^2/n$, over $r>0$. The derivative is

$$f'(r) \;=\; -2\alpha A\,r^{-2\alpha-1}+\frac{2BK^2}{n}\,r. \tag{22}$$

Setting $f'(r)=0$ gives $r^{2\alpha+2}=\alpha An/(BK^2)$, hence

$$r^\star \;=\; \Big(\frac{\alpha A}{B}\Big)^{\frac{1}{2\alpha+2}}\Big(\frac{n}{K^2}\Big)^{\frac{1}{2\alpha+2}} \;=\; \Theta\Big(\big(n/K^2\big)^{\frac{1}{2\alpha+2}}\Big). \tag{23}$$

Since $f''(r)=2\alpha(2\alpha+1)Ar^{-2\alpha-2}+2BK^2/n>0$, the stationary point is the unique minimum. The constants A and B affect only the prefactor, which is why the law is stated up to Θ. □

A.4 Statement and Proof of Proposition 2

Proposition 2 (PAC-Bayes oracle inequality (Alquier et al., 2016)). *With probability at least* $1-\delta$,

$$L(\hat{q}) \;\le\; \inf_{Q} \Big\{ L(Q) + 2\, s_\delta(Q) \Big\}, \tag{24}$$

where

$$\hat{q} \;=\; \arg\min_{Q} \; \hat{L}(Q) + s_\delta(Q), \qquad s_\delta(Q) \;=\; \sqrt{\frac{D_{\mathrm{KL}}(Q\|P) + \ln \frac{2\sqrt{n}}{\delta}}{2n}}. \tag{25}$$

Start from Theorem 1 and relax the binary divergence by Pinsker's inequality in the form $\mathrm{kl}(p\|q) \ge 2(p-q)^2$. With probability at least $1-\delta$, simultaneously for all Q,

$$\big|\hat{L}(Q) - L(Q)\big| \;\le\; s_\delta(Q), \qquad s_\delta(Q) = \sqrt{\frac{D_{\mathrm{KL}}(Q\|P) + \ln \frac{2\sqrt{n}}{\delta}}{2n}}. \tag{26}$$

The bound holds in both directions because it controls the absolute value. Apply the upper direction to $\hat{q}$ and the lower direction to an arbitrary Q, then use the optimality of $\hat{q}$ for the objective $\hat{L}(Q) + s_\delta(Q)$ from Eq. 25:

$$L(\hat{q}) \;\le\; \hat{L}(\hat{q}) + s_\delta(\hat{q}) \;\le\; \hat{L}(Q) + s_\delta(Q) \;\le\; L(Q) + 2\, s_\delta(Q). \tag{27}$$

Taking the infimum over Q on the right side gives Eq. 24. This is the PAC-Bayes oracle inequality in the form used by Alquier et al. (2016) for variational approximations of Gibbs posteriors. □

Scope. The optimizer $\hat{q}$ is the distribution selected by minimizing the penalized objective within the considered family. Proposition 2 provides general context for bound-based distribution ranking. It is not an optimization guarantee for the Gaussian approximation of Section 4.2, which does not minimize Eq. 25.

A.5 Statement and Proof of the Selective-Risk Bound

For any label-free scalar score f fixed before calibration, the variance $v(\boldsymbol{x})$ of Eq. 11 induces a selective rule that abstains when $v(\boldsymbol{x}) > \tau$. Because the posterior and thus v use the operator-fitting set $\mathcal{D}$, the bound is stated on an independent calibration sample $\mathcal{D}_{\mathrm{cal}}$ of size n. Relative to this sample, v and the associated predictor are fixed. Let $\ell(\boldsymbol{x}, y)$ be that predictor's zero-one loss. Let $\phi(\tau) = \Pr(v(\boldsymbol{x}) \le \tau)$ denote coverage and $R(\tau) = \mathbb{E}[\ell \mid v(\boldsymbol{x}) \le \tau]$ denote selective risk, with empirical counterparts $\hat{\phi}(\tau)$ and $\hat{R}(\tau)$ on $\mathcal{D}_{\mathrm{cal}}$.

Theorem 2 (Selective-risk bound (Chow, 1970; Geifman & El-Yaniv, 2017)). *Fix a threshold* τ *independently of* $\mathcal{D}_{\mathrm{cal}}$*, and let* $\epsilon_n = \sqrt{\ln(2/\delta)/(2n)}$*. With probability at least* $1-\delta$ *over the draw of* $\mathcal{D}_{\mathrm{cal}}$*, whenever* $\hat{\phi}(\tau) > \epsilon_n$*,*

$$R(\tau) \;\le\; \frac{\hat{R}(\tau)\,\hat{\phi}(\tau) + \epsilon_n}{\hat{\phi}(\tau) - \epsilon_n}. \tag{28}$$

The bound transfers one fixed point of a calibration-sample risk-coverage curve to the population within a band that shrinks as $\sqrt{1/n}$. Selecting among multiple thresholds requires an additional simultaneous or union correction. It applies when the score and predictor are fixed relative to an independent calibration sample. The main experiments in Section 5 instead compare selective prediction using one minus maximum predictive probability.

Proof. Fix τ and write $Z_i = \ell(\boldsymbol{x}_i, y_i)\,\mathbf{1}\{v(\boldsymbol{x}_i) \le \tau\}$ and $C_i = \mathbf{1}\{v(\boldsymbol{x}_i) \le \tau\}$. Both are independent bounded variables because the predictor and v are fixed relative to $\mathcal{D}_{\mathrm{cal}}$. We have $Z_i \in [0,1]$, $C_i \in \{0,1\}$, $\mathbb{E}[Z_i] = R(\tau)\phi(\tau)$, and $\mathbb{E}[C_i] = \phi(\tau)$. Hoeffding's inequality (Boucheron et al., 2013) gives, with probability at least $1-\delta/2$ for each event,

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \;\ge\; R(\tau)\phi(\tau) - \epsilon_n \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} C_i \;\le\; \phi(\tau) + \epsilon_n, \tag{29}$$

with $\epsilon_n = \sqrt{\ln(2/\delta)/(2n)}$. Note $\frac{1}{n}\sum Z_i = \hat{R}(\tau)\hat{\phi}(\tau)$ and $\frac{1}{n}\sum C_i = \hat{\phi}(\tau)$ by definition of the empirical selective risk and coverage. A union bound over the two events, followed by solving

$$R(\tau)\left(\hat{\phi}(\tau) - \epsilon_n\right) \;\le\; R(\tau)\,\phi(\tau) \;\le\; \hat{R}(\tau)\,\hat{\phi}(\tau) + \epsilon_n \tag{30}$$

for $R(\tau)$ under the condition $\hat{\phi}(\tau) > \epsilon_n$, yields Eq. 28. The condition is met whenever the coverage is not vanishingly small relative to $1/\sqrt{n}$. □

A.6 Remark on Assumption 1

The assumption combines two standard ingredients. The approximation term $Ar^{-2\alpha}$ is the empirical form of power-law spectral decay. If the singular values of the concatenated factor matrices decay as $\sigma_j \asymp j^{-\alpha-1/2}$, then the squared reconstruction tail of the best rank-r representation decays as $r^{-2\alpha}$ by the Eckart-Young theorem (Eckart & Young, 1936). The estimation term $B(Kr)^2/n$ matches the parametric rate for estimating an object with $(Kr)^2$ free entries, and it is the same scaling that enters the bound through Lemma 1. The pilot measurements in Appendix D support the spectral-decay ingredient, with fitted exponents $\alpha \in [0.05, 0.07]$. The quadratic dimension term remains a structural surrogate, while the measured posterior KL grows sublinearly with rank and is not pointwise monotone.

B Notation

Table 4: Symbols used in the main text.

Symbol	Meaning
$\boldsymbol{W}_0$, $\Delta\boldsymbol{W}_k$	pretrained weight matrix, LoRA update of task k
$\boldsymbol{B}_k$, $\boldsymbol{A}_k$, r	LoRA factors of task k and their rank
K	number of merged tasks
$\boldsymbol{U}^{\mathrm{ref}}$, $\boldsymbol{V}^{\mathrm{ref}}$	shared orthonormal bases spanning the concatenated factors
$\boldsymbol{M}_k$	core matrix of task k in the shared basis, size $Kr \times Kr$ under the width condition in Section 3.1
$\boldsymbol{M}_{\boldsymbol{w}}$, $\boldsymbol{w}$	merged operator and merge coefficients
$\mathcal{G}$	candidate pool of merge configurations
A_{val}, A_{test}	validation and test accuracy averaged over tasks
$\hat{\boldsymbol{w}}$, $\boldsymbol{w}^\star$	validation-selected configuration, test oracle
G_{sel}	selection gap of Eq. 3
$\mathcal{D}$, $\mathcal{D}_k$, n_k, n	operator-fitting set, per-task subset, per-task size, total size
λ, $\boldsymbol{M}_0$, m	prior precision, deterministic prior center of Eq. 5, selection-event count of Eq. 17
$q(\mathbf{M})$, Q, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$	Gaussian approximation, the operator distribution with $Q = q$ in reported bounds, its mean and covariance
$\boldsymbol{H}_{\mathrm{EF}}$	$1/K$-weighted empirical-Fisher curvature of Eq. 8
η, S, $\boldsymbol{M}_s$	covariance damping, BMA sample count and posterior samples
$\boldsymbol{J}(\boldsymbol{x})$, $v(\boldsymbol{x})$	score Jacobian at $\boldsymbol{\mu}$ and auxiliary linearized score variance
$L(\boldsymbol{M})$, $\hat{L}(\boldsymbol{M})$	expected and empirical zero-one loss
$L(Q)$, $\hat{L}(Q)$	expected and empirical loss under a distribution over operators
kl, D_{KL}	binary and general Kullback-Leibler divergence
$\hat{q}$, $s_\delta(Q)$	bound-optimal posterior and the bound residual
$\mathcal{D}_{\mathrm{cal}}$, n_{cal}	calibration sample and its size of Theorem 2, disjoint from $\mathcal{D}$
$\phi(\tau)$, $R(\tau)$	coverage and selective risk at threshold τ
α, $r^\star$	spectral decay exponent and surrogate-optimal rank of Proposition 1
SAR, σ_q^2	subspace alignment ratio, averaged auxiliary loss variance

C Extended Related Work

This appendix expands the compressed summary of Section 2.

Table 5: Related methods by generation, merge space, and properties. *TF* and *DF* denote training-free and data-free merging. *Output* distinguishes point and distribution-valued results. *UQ/Bound* records uncertainty quantification or a generalization bound. The listed prior methods return point estimates, while BLoM represents an operator distribution. TF/DF describe merge construction, excluding validation-based selection; AdaMerging is related work rather than an evaluated baseline.

Method	Merge space	TF	DF	Output	UQ/Bound
Generation 1, parameter space					
Task arithmetic (Ilharco et al., 2022)	weights	✓	✓	point	×
TIES (Yadav et al., 2023)	weights	✓	✓	point	×
DARE-TIES (Yu et al., 2024)	weights	✓	✓	point	×
RegMean (Jin et al., 2023)	weights	✓	×	point	×
AdaMerging (Yang et al., 2024c)	coefficients	×	×	point	×
Generation 2, aligned singular subspaces					
KnOTS (Stoica et al., 2025)	SVD-aligned	✓	✓	point	×
TSV-M (Gargiulo et al., 2025)	SVD-aligned	✓	✓	point	×
Iso-C (Marczak et al., 2025)	spectral	✓	✓	point	×
Generation 3, LoRA-internal geometry					
Core Space (Panariello et al., 2026)	core $(Kr)^2$	✓	✓	point	×
CtM (He et al., 2026)	shared subspace	✓	✓	point	×
SVC (Li et al., 2026b)	singular values	✓	✓	point	×
TARA (Jeong et al., 2026)	directions	×	×	point	×
TSPA (Li et al., 2026c)	two-stage	✓	✓	point	×
Pico (Tang & Yang, 2026)	output factors	✓	✓	point	×
BLoM	core posterior	×	×	*posterior*	✓

Deterministic model merging. Linear mode connectivity motivates weight-space combination for models fine-tuned from a shared initialization (Garipov et al., 2018; Draxler et al., 2018; Frankle et al., 2020; Entezari et al., 2022; Ainsworth et al., 2023; Jordan et al., 2023; Foret et al., 2021). Model averaging has roots in federated learning (McMahan et al., 2017) and was extended through weight interpolation, model soups, and task arithmetic (Wortsman et al., 2022b;a; Ilharco et al., 2022), with a tangent-space account in Ortiz-Jimenez et al. (2023). Parameter-space methods include TIES, DARE, DELLA, Fisher merging, RegMean, ZipIt, EMR, and AdaMerging (Yadav et al., 2023; Yu et al., 2024; Deep et al., 2024; Matena & Raffel, 2022; Jin et al., 2023; Stoica et al., 2023; Huang et al., 2024b; Yang et al., 2024c). Subspace methods include KnOTS, TSV, and Iso-C (Stoica et al., 2025; Gargiulo et al., 2025; Marczak et al., 2025). LoRA-specific methods include Core Space, CtM, Pico, SVC, TARA, and TSPA (Panariello et al., 2026; He et al., 2026; Tang & Yang, 2026; Li et al., 2026b; Jeong et al., 2026; Li et al., 2026c). The benchmarked methods return one merged operator, often using held-out selection or fixed settings. BLoM studies the statistical cost of point selection and replaces it with operator inference. Additional directions cover clustering, skill composition, low-rank task-vector estimation, magnitude-direction decoupling, consensus scaling, continual and diffusion-model merging, test-time adaptation, adapter retrieval, and LoRA geometry (Zhao et al., 2025; Prabhakar et al., 2025; Liu et al., 2025; Zheng et al., 2025; Ryum & Kang, 2026; Shenaj et al., 2026; Qiao & Mahdavi, 2025; Wei et al., 2026; Reid et al., 2026; Sugita et al., 2026; Lu et al., 2026; Paul, 2026).

Posteriors for neural network weights. Bayesian neural networks (Neal, 1996) include Laplace approximations (MacKay, 1992; Ritter et al., 2018; Daxberger et al., 2021), variational posteriors (Blundell et al., 2015; Louizos & Welling, 2016; Mishkin et al., 2018), and subspace methods (Izmailov et al., 2019; Maddox et al., 2019). BLoM follows the subspace lineage with a space defined by adapter merging rather than an SGD trajectory. It infers an operator across trained adapters rather than the training posterior of one model. Related work places posteriors over individual adapters,

Table 6: Winner's curse and oracle gap on the near-orthogonal six-task suite, from 300 resampled random subgrids at each pool size (n validation samples per task). The winner's curse grows with the pool size and with n, while the oracle gap stays below 0.01 even at the full 165-configuration pool.

Pool size	Winner's curse			Oracle gap
	n=50	n=100	n=200	(n=200)
5	−0.006	+0.027	+0.030	0.001
10	−0.007	+0.031	+0.033	0.002
20	−0.005	+0.035	+0.037	0.004
40	+0.005	+0.039	+0.041	0.006
80	+0.012	+0.044	+0.045	0.008
165	+0.015	+0.049	+0.050	0.009

where uncertainty reflects adapter training (Yang et al., 2024a; Wang et al., 2024). BLoM instead models uncertainty in merging and supports a Gibbs-risk certificate.

Probabilistic readings of merging. Bayesian Model Merging uses Bayesian regression in activation space and Bayesian optimization for module-level settings, then returns a point estimate (Li et al., 2026a). Daheim et al. (2024) use Fisher-weighted posterior matching to obtain one deterministic solution. Product-of-experts and Bayesian-optimization approaches also use probabilistic structure to improve point estimates (Bui et al., 2026; Lee et al., 2025). Kronecker-factored curvature supports deterministic reweighting and disentanglement (Porrello et al., 2026; Xu et al., 2026). BLoM instead maintains a distribution over the merge operator and uses distinct functionals of that distribution for Mean, BMA, and Gibbs prediction. A complementary line predicts mergeability from model properties (Rahamim et al., 2026; Zhou et al., 2026). Our target is the configuration-selection procedure used in common merging pipelines.

Composition, evaluation and bounds. Composition methods keep adapters separate and route them at inference time (Huang et al., 2024a; Wu et al., 2024), while our setting produces one merged model. Existing surveys and benchmarks emphasize accuracy evaluation (Yang et al., 2024b; Song & Zheng, 2026; He et al., 2025). Conformal methods provide marginal coverage for a fixed predictor (Angelopoulos & Bates, 2023). Our PAC-Bayes analysis instead concerns the Gibbs risk of a distribution over merge operators. Table 5 summarizes the comparison.

D Pilot Study Details

This appendix documents the pilot stage behind Section 3.3. All runs use Qwen2.5-1.5B, finance-task LoRA adapters in FLARE format with completion-only cross-entropy, and the pipeline of Section 4.2 on one 24GB GPU.

D.1 Adapters and Grids

We trained 19 adapters covering FPB, FinQA, and Headlines at ranks $\{2, 4, 8, 16, 32\}$, FiQA-SA at ranks $\{8, 16\}$, and FOMC and German at rank 16. The six-task suite combines FPB, FinQA, FiQA-SA, Headlines, FOMC, and German. The high-interference pair combines FPB and FiQA-SA, which share a three-class sentiment label space but come from different corpora. At each rank, the selection grid contains 15 task-arithmetic, 15 TSV-M, and 25 TIES configurations, totaling 55, with requested validation sizes $n \in \{50, 100, 200\}$ per task. Posterior experiments sweep ranks $\{2, 4, 8\}$ against prior precisions $\lambda \in \{10, 100, 1000\}$.

D.2 Selection Gaps on the Near-Orthogonal Suite

Resampling 300 random subgrids of growing size shows a winner's curse between 0.015 and 0.05 that grows with the pool size (Table 6 and Figure 4). These subgrid measurements complement the

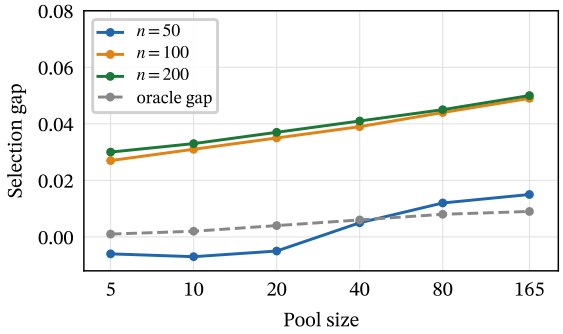

Figure 4: Winner's curse and oracle gap against the candidate pool size on the near-orthogonal six-task suite, from 300 resampled random subgrids at each pool size. The winner's curse grows with the pool size and with the validation size n (the less regular n=50 trace reflects resampling variance around a near-zero true curse), while the oracle gap stays below 0.01 even at the full 165-configuration pool.

cross-regime summary in Section 3.3 by isolating candidate-pool size within the near-orthogonal suite.

D.3 Selection-Gap Resampling on Six Tasks

The high-interference pair in Table 1 is complemented by a separate bootstrap analysis of the six-task pilot suite, using cached predictions from its configuration grids. For three adapter seeds and ranks $\{2, 4, 8\}$, we draw B=200 validation bootstrap samples per cell at $n \in \{50, 100, 200\}$ per task. We recompute the validation maximizer against cached test predictions and record both its gap and identity. Because validation and test follow the original split, this analysis measures their combined finite-sample and distribution-shift effect.

D.4 Operator-Refinement Gains on Six Tasks

Table 7 gives task-macro BLoM-Mean gains over TSV-Core on the six-task suite using one representative layer. The early pilot stores BMA accuracy pooled over examples, whereas Mean and TSV-Core use task-macro accuracy. We therefore compare Mean with TSV-Core and report the stored BMA scores separately in Table 8. The table provides the measurements underlying the posterior-construction summary in Section 5.3.

Table 7: BLoM-Mean gain over TSV-Core on the six-task suite with a single-layer posterior. Entries are task-macro accuracy differences. Gains are positive in all nine configurations.

Rank r	$\lambda = 10$	$\lambda = 100$	$\lambda = 1000$
2	+0.0438	+0.0381	+0.0139
4	+0.0859	+0.0552	+0.0229
8	**+0.1176**	**+0.0810**	**+0.0329**

D.5 Multi-Layer Posterior

A layer-factorized posterior over 28 `k_proj` layers at rank 4 and $\lambda = 100$ takes 649 minutes on one GPU without adapter retraining. Each layer retains full within-layer covariance, and posterior draws are independent across layers. Its pooled BLoM-BMA accuracy is 0.5226 and negative log likelihood is 0.8842. These pooled quantities are not task-macro comparisons against the Mean result. Mean also improves the initialization in all nine single-layer configurations by 0.014 to 0.118. These diagnostics separate Mean refinement from BMA averaging beyond the accuracy summary in Section 5.3.

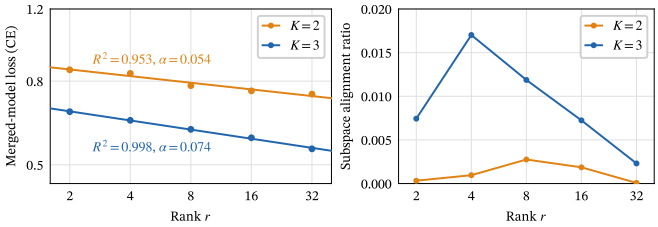

Figure 5: **Left.** Merged-model loss against rank for a two-task pair and a three-task triple of Qwen2.5-1.5B finance adapters (log-log axes), with power-law fits. Most of the loss reduction is attained at low rank. **Right.** The subspace alignment ratio stays below 0.017 at every rank, so the task operators are near-orthogonal in core space.

D.6 Repair on the High-Interference Pair

The repair result in Figure 2 uses the rank-8 setting with the largest observed selection gaps. The displayed operating point uses $\lambda = 100$. The validation-selected TSV-M model scores 0.7498, its finite-grid test oracle scores 0.7906, and BLoM-Mean scores 0.7982 under the same task-macro metric. The stored BMA value 0.7888 pools 300 FPB and 235 FiQA-SA examples and is not used to calculate a task-macro repair percentage.

D.7 Complexity Scaling

The dedicated finance rank grid in Table 15 spans posterior KL values from 0.96 at rank 1 to 15.30 at rank 24. Its fitted growth is sublinear and not pointwise monotone. The rank law concerns the surrogate in Assumption 1, rather than asserting that the measured KL must grow proportionally to dimension.

D.8 Prior-Precision Scaling

Table 8: Posterior coordinate standard deviation against prior precision at rank 8. Accuracy columns use the six-task suite, with task-macro Mean gain over TSV-Core and pooled BMA accuracy reported separately.

	Posterior std			Six-task results	
λ	Six-task	Hi-pair	Vision	Mean gain	BMA pooled acc.
10	0.314	0.313	0.309	+0.1176	0.4662
100	0.099	0.099	0.099	+0.0810	**0.5303**
1000	0.031	0.032	0.031	+0.0329	0.4711

Table 8 characterizes precision sensitivity in Section 4.2. Mean posterior coordinate standard deviation follows prior precision similarly on the six-task suite, the high-interference pair, and the vision suite. The accuracy columns report six-task Mean gains and pooled BMA accuracy separately, so their numerical values should not be subtracted.

E Complete Main-Result Tables

This appendix reports complete per-task results. Tables separate reference models, merging baselines, and BLoM, with language-model TSV-Core initializations shown where directly evaluated. All primary entries are raw accuracies. Finance-1.5B BLoM results average three adapter seeds, while the remaining suites use public or seed-0 adapters. BLoM-BMA uses $S = 30$ samples. Each finance table separates the six-task exact-match comparison from a five-task label-probability panel comparing Mean and BMA. Raw accuracy is the main protocol because normalized scores become unstable when an individual-task denominator is near zero.

Table 9: Per-task results on six FLARE finance tasks at Qwen2.5-1.5B, rank 16, and λ=100. BLoM rows are three-seed means, while baselines are single deterministic evaluations. TSV-Core is the deterministic initialization. The upper panel uses six-task teacher-forced completion-token exact match. The separate lower panel compares Mean and BMA under five-task label-probability scoring, excluding FinQA. Merging methods tie at 0.660 on German. Bold marks the best merging method in the six-task panel.

Method	FPB	FinQA	FiQA-SA	Head.	FOMC	German	Avg (6)
Reference models							
Pretrained	0.000	0.000	0.000	0.497	0.000	0.000	0.083
Individual LoRA	0.820	0.047	0.826	0.903	0.500	0.660	0.626
Joint multi-task	0.830	0.047	0.821	0.870	0.570	0.660	0.633
Parameter space (Gen 1)							
Task Arithmetic	0.693	0.017	0.791	0.733	0.460	0.660	0.559
TIES	0.737	0.013	0.774	0.640	0.460	0.660	0.547
DARE-TIES	0.737	0.023	0.796	0.390	0.510	0.660	0.519
RegMean	0.773	0.050	0.804	**0.850**	0.480	0.660	0.603
Aligned subspace (Gen 2)							
KnOTS	0.707	0.013	0.783	0.683	0.450	0.660	0.549
TSV-M	0.713	0.017	0.796	0.747	0.460	0.660	0.565
Iso-C	0.723	0.017	0.800	0.780	0.470	0.660	0.575
LoRA geometry (Gen 3)							
CtM	0.707	0.017	0.787	0.747	0.460	0.660	0.563
SVC	0.683	0.017	0.787	0.747	0.460	0.660	0.559
TARA	0.710	0.017	0.788	0.667	0.460	0.660	0.550
TSPA	0.693	0.017	0.787	0.730	0.460	0.660	0.558
Pico	0.723	0.020	0.800	0.723	0.460	0.660	0.564
TSV-Core (init)	0.747	0.057	0.855	0.030	0.520	0.660	0.478
Ours							
BLoM-Mean	**0.823**	**0.058**	**0.830**	0.676	**0.587**	0.660	**0.606**
Five-task label-probability scoring (separate protocol)							
Method	FPB	FinQA	FiQA-SA	Head.	FOMC	German	Avg (5)
BLoM-Mean	0.788	–	0.752	0.743	0.463	0.660	0.681
BLoM-BMA	0.786	–	0.718	0.667	0.463	0.660	0.659

The two BLoM predictors serve different decisions. BLoM-Mean is the primary accuracy mode. BLoM-BMA averages predictive distributions over plausible operators. Under matched finance label-probability scoring, Mean/BMA accuracies are $0.681/0.659$ at 1.5B and $0.723/0.717$ at 7B. On NLI and vision, Mean and BMA differ by at most 1.8 points. These comparisons support reporting the modes separately rather than ranking both by point accuracy alone.

The uncertainty evaluation uses the predictive confidence of each mode. On NLI, BLoM-Mean has the best AURC on four of six tasks and the lowest selective risk at 50% and 70% coverage. Fitted calibrators lead most finance calibration columns. Finance-7B follows the same pattern, with vector scaling at ECE 0.097 and AURC 0.135, compared with AURC 0.238 for both BLoM modes. On ViT-B/32, Mean leads risk ranking with AURC 0.409, while on ViT-L/14 BMA leads with 0.207. A one-parameter calibration fit further improves ECE on both vision suites. Conformal sets attain coverage near the nominal 90% target across finance, NLI, and vision.

The cross-domain result in Table 3 reinforces the interference-dependent accuracy pattern. BLoM keeps every task above 0.82 and improves the suite average by 13.9 points over the best of the three compared baselines. The paired bootstrap intervals are $[+5.0, +8.7]$ points against RegMean on NLI and $[+11.8, +15.9]$ on the cross-domain suite. Leave-one-task-out detection remains near chance at AUROC 0.516 on finance and 0.512 on NLI. The confidence results are therefore interpreted within the merged task family rather than as an out-of-distribution detector.

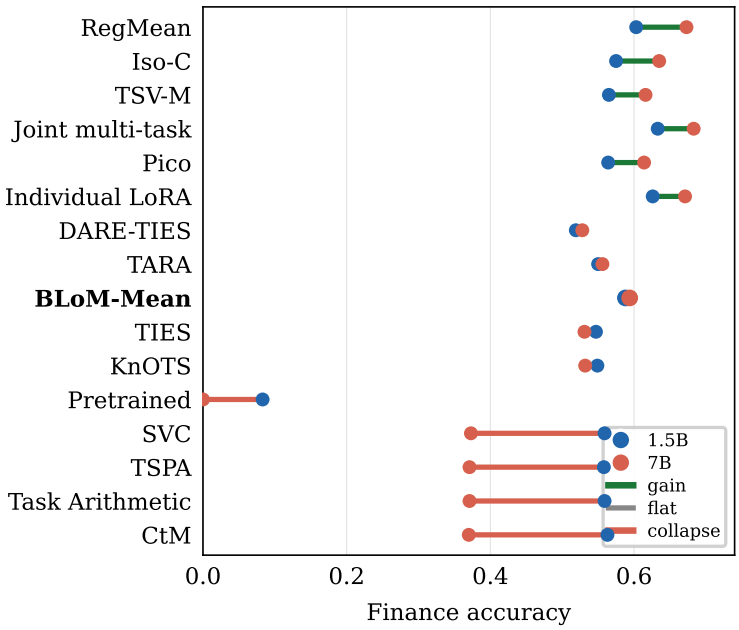

Figure 6: Finance accuracy from 1.5B to 7B, sorted by the change across scales. All rows use six-task teacher-forced completion-token exact-match accuracy. The plotted BLoM values use the seed-0 adapter pool, while Table 9 reports the three-seed Finance-1.5B results.

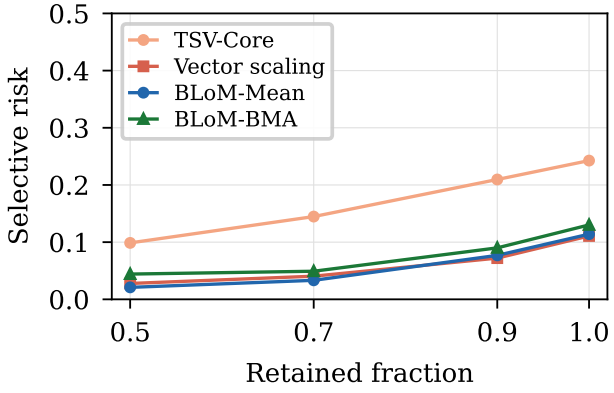

Figure 7: Selective risk on NLI under the three-split protocol. The four representative curves use the same test results as Table 14. BLoM-Mean is lowest at 50% and 70% coverage and remains close to vector scaling elsewhere.

Table 10: Per-task finance results for Qwen2.5-7B at rank 16 and λ=100 using one seed. The base model scores zero under the reported exact-match protocol. BLoM-Mean remains positive on every task, including FiQA-SA and FOMC where several weight-space methods score zero. The upper panel uses six-task teacher-forced completion-token exact match. The separate lower panel compares Mean and BMA under five-task label-probability scoring, excluding FinQA. Bold marks the best merging method in the six-task panel.

Method	FPB	FinQA	FiQA-SA	Head.	FOMC	German	Avg (6)
Reference models							
Pretrained	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Individual LoRA	0.837	0.140	0.838	0.927	0.580	0.705	0.671
Joint multi-task	0.843	0.150	0.855	0.957	0.640	0.655	0.683
Parameter space (Gen 1)							
Task Arithmetic	0.770	0.050	0.000	0.817	0.000	0.590	0.371
TIES	0.800	0.067	0.791	0.490	0.390	0.645	0.531
DARE-TIES	0.803	0.060	0.817	0.290	0.570	0.630	0.528
RegMean	**0.827**	**0.147**	**0.847**	**0.920**	0.630	0.670	**0.673**
Aligned subspace (Gen 2)							
KnOTS	0.790	0.057	0.838	0.270	0.590	0.645	0.532
TSV-M	0.803	0.090	0.800	0.817	0.510	0.675	0.616
Iso-C	0.770	0.057	0.834	0.797	0.650	**0.705**	0.635
LoRA geometry (Gen 3)							
CtM	0.773	0.047	0.000	0.810	0.000	0.590	0.370
SVC	0.773	0.047	0.000	0.813	0.000	0.605	0.373
TARA	0.790	0.063	0.791	0.720	0.300	0.670	0.556
TSPA	0.777	0.043	0.000	0.813	0.000	0.590	0.371
Pico	0.787	0.060	0.804	0.817	0.550	0.665	0.614
TSV-Core (init)	0.800	0.097	0.830	0.167	0.690	0.680	0.544
Ours							
BLoM-Mean	0.810	0.107	0.826	0.493	**0.660**	0.670	0.594
Five-task label-probability scoring (separate protocol)							
Method	FPB	FinQA	FiQA-SA	Head.	FOMC	German	Avg (5)
BLoM-Mean	0.823	–	0.715	0.897	0.520	0.660	0.723
BLoM-BMA	0.817	–	0.689	0.900	0.520	0.660	0.717

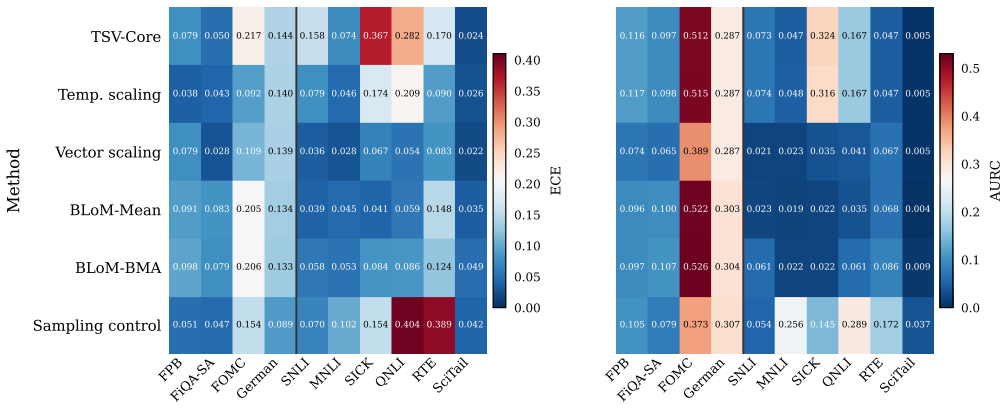

Figure 8: Per-task uncertainty quality under the three-split protocol on the finance label-probability subset and NLI. The panels show ECE and AURC, where lower is better. The sampling control is a three-seed deep ensemble on finance and MC-Dropout on NLI. Vector scaling leads five of six NLI ECE columns. Fitted calibrators lead the finance AURC columns except FOMC, while BLoM gives the best AURC on four of six NLI tasks.

Table 11: Per-task NLI results for Llama-3-8B with the Stoica et al. (2025) checkpoints, six rank-16 community single-seed adapters, and λ=10. Results use the corrected TSV-M merge implementation. Bold marks the best merging method per column.

Method	SNLI	MNLI	SICK	QNLI	RTE	SciTail	Avg
Reference models							
Pretrained	0.417	0.330	0.553	0.163	0.558	0.580	0.434
Individual LoRA	0.933	0.907	0.929	0.660	0.494	0.980	0.817
Joint multi-task	0.917	0.883	0.915	0.513	0.818	0.973	0.837
Parameter space (Gen 1)							
Task Arithmetic	0.853	0.843	0.841	0.347	0.649	0.943	0.746
TIES	0.870	0.860	0.776	0.257	0.727	0.950	0.740
DARE-TIES	0.820	0.860	0.664	0.170	0.779	0.897	0.698
RegMean	**0.913**	0.880	**0.915**	0.570	0.688	**0.977**	0.824
Aligned subspace (Gen 2)							
KnOTS	0.850	0.840	0.834	0.307	0.701	0.947	0.746
TSV-M	0.863	0.850	0.831	0.203	0.701	0.950	0.733
Iso-C	0.813	0.797	0.776	0.070	0.597	0.940	0.666
LoRA geometry (Gen 3)							
CtM	0.853	0.840	0.844	0.343	0.649	0.940	0.745
SVC	0.857	0.833	0.827	0.323	0.662	0.933	0.739
TARA	0.847	0.840	0.841	0.313	0.701	0.947	0.748
TSPA	0.853	0.840	0.840	0.343	0.649	0.943	0.745
Pico	0.853	0.807	0.871	0.217	0.649	0.950	0.725
TSV-Core (init)	0.777	0.837	0.563	0.593	0.818	0.960	0.758
Ours							
BLoM-Mean	0.893	**0.910**	0.902	**0.870**	**0.831**	0.950	**0.893**
BLoM-BMA	0.853	0.907	0.905	0.830	0.870	0.950	0.886

Table 12: Per-task results on the vision suite (K=8 CLIP tasks, rank 16). Top panel ViT-B/32 (λ=1000), bottom panel ViT-L/14 (λ=100, the archived evidence-selected value). Both BLoM vision runs initialize from uniform core averages. Bold marks the best merging method per column within each panel.

Method	Cars	DTD	EuroSAT	GTSRB	MNIST	RESISC45	SUN397	SVHN	Avg
ViT-B/32, reference models									
Pretrained	0.590	0.427	0.443	0.283	0.447	0.590	0.647	0.323	0.469
Individual LoRA	0.737	0.603	0.997	0.923	0.990	0.893	0.693	0.973	0.851
Joint multi-task	0.573	0.497	0.827	0.583	0.773	0.663	0.693	0.680	0.661
ViT-B/32, merging methods									
Task Arithmetic	0.607	0.423	0.483	0.343	0.507	0.620	0.643	0.443	0.509
TIES	0.617	0.397	0.463	0.263	0.573	0.600	0.637	0.530	0.510
DARE-TIES	0.317	0.170	0.153	0.053	0.187	0.193	0.407	0.267	0.218
RegMean	**0.653**	**0.440**	**0.773**	**0.550**	**0.877**	**0.697**	0.627	**0.777**	**0.674**
KnOTS	0.610	0.423	0.480	0.363	0.507	0.627	0.637	0.460	0.513
TSV-M	0.600	0.437	0.513	0.357	0.583	0.663	0.647	0.520	0.540
Iso-C	0.613	0.437	0.440	0.437	0.550	0.683	**0.657**	0.477	0.537
CtM	0.617	0.423	0.483	0.370	0.513	0.630	0.637	0.467	0.518
SVC	0.607	0.423	0.483	0.350	0.503	0.620	0.640	0.460	0.511
TARA	0.623	0.400	0.383	0.327	0.493	0.610	0.627	0.493	0.495
TSPA	0.607	0.423	0.480	0.350	0.503	0.620	0.640	0.460	0.510
Pico	0.627	0.413	0.497	0.387	0.533	0.637	0.647	0.487	0.528
BLoM-Mean	0.607	0.413	0.600	0.343	0.557	0.643	0.637	0.480	0.535
BLoM-BMA	0.620	0.413	0.507	0.317	0.577	0.593	0.633	0.480	0.518
ViT-L/14, reference models									
Pretrained	0.763	0.563	0.597	0.527	0.757	0.743	0.723	0.583	0.657
Individual LoRA	1.000	0.697	0.997	0.983	0.997	0.967	0.830	0.980	0.931
Joint multi-task	0.780	0.690	0.943	0.873	0.980	0.907	0.757	0.950	0.860
ViT-L/14, merging methods									
Task Arithmetic	0.763	0.563	0.617	0.590	0.803	0.800	0.717	0.607	0.683
TIES	0.777	0.557	0.530	0.613	0.683	0.823	0.707	0.667	0.670
DARE-TIES	0.380	0.330	0.160	0.137	0.193	0.460	0.590	0.157	0.301
RegMean	**0.817**	**0.593**	**0.710**	**0.780**	**0.907**	**0.847**	0.717	**0.800**	**0.771**
KnOTS	0.763	0.570	0.580	0.617	0.817	0.807	**0.720**	0.630	0.688
TSV-M	0.787	0.560	0.587	0.710	0.860	0.827	0.717	0.690	0.717
Iso-C	0.783	0.577	0.620	0.717	0.877	0.827	0.713	0.647	0.720
CtM	0.770	0.577	0.613	0.620	0.810	0.810	**0.720**	0.617	0.692
SVC	0.763	0.563	0.613	0.590	0.803	0.800	0.717	0.607	0.682
TARA	0.770	0.563	0.543	0.637	0.723	0.827	0.700	0.593	0.670
TSPA	0.767	0.570	0.590	0.610	0.810	0.807	0.720	0.637	0.689
Pico	0.810	0.563	0.607	0.687	0.850	0.833	0.710	0.633	0.712
BLoM-Mean	0.770	0.570	0.600	0.613	0.840	0.813	0.707	0.600	0.689
BLoM-BMA	0.763	0.563	0.597	0.623	0.840	0.813	0.707	0.570	0.685

Table 13: Leave-one-task-out out-of-distribution (OOD) detection with BLoM-BMA and $S{=}30$ on Finance-1.5B and NLI-8B. Each row treats one task as OOD relative to a model merged from the other tasks. AUROC denotes area under the receiver operating characteristic curve, and H denotes predictive entropy for in-distribution (ID) and OOD examples. Suite-average AUROC is 0.516 on finance and 0.512 on NLI. German reaches 0.890, while FOMC reaches 0.062 under closely overlapping sentiment labels. This diagnostic delimits the scope of the within-suite selective prediction results in Table 14.

Held-out task	AUROC (entropy)	AUROC ($1 - \max_p$)	H (ID)	H (OOD)
FPB	0.691	0.654	0.904	1.020
FiQA-SA	0.420	0.391	0.830	0.750
FOMC	0.062	0.065	0.616	0.136
German	0.890	0.711	0.650	1.038
Mean (finance)	0.516	0.455	0.750	0.736
NLI suite				
SNLI	0.502	0.490	0.436	0.436
MNLI	0.419	0.419	0.538	0.470
SICK	0.511	0.507	0.686	0.717
QNLI	0.611	0.599	0.654	0.703
RTE	0.477	0.490	0.729	0.695
SciTail	0.551	0.539	0.742	0.751
Mean (NLI)	0.512	0.507	0.631	0.629

Table 14: Uncertainty quality on the untouched test split, macro-averaged over tasks, under a three-split protocol (the fit half drives the posterior, the calibration half fits all calibrators). The predictor-controlled block fits calibrators on the same BLoM MAP logits, and the end-to-end block scores full pipelines. SA@90 denotes selective accuracy at 90% coverage, and calibration baselines are specified in Appendix H.4. Deep ensemble averages the three Finance seeds and is unavailable on single-seed NLI. †Finance uses the label-probability protocol on its four tasks with at most eight label words.

	Finance-1.5B†				NLI			
Method	ECE ↓	NLL ↓	AURC ↓	SA@90 ↑	ECE ↓	NLL ↓	AURC ↓	SA@90 ↑
Predictor-controlled (calibrators on the BLoM MAP logits)								
+ temperature	0.084	0.733	0.258	0.660	0.047	0.272	0.028	0.924
+ vector scaling	0.100	0.671	0.206	0.703	**0.045**	**0.269**	0.029	**0.928**
End-to-end (full pipelines)								
TSV-Core (softmax)	0.122	0.780	0.253	0.673	0.179	0.793	0.111	0.790
+ temperature	**0.078**	0.737	0.254	0.670	0.104	0.470	0.110	0.789
+ vector scaling	0.089	**0.671**	**0.204**	**0.713**	0.048	0.279	0.032	**0.928**
+ MC-Dropout	–	–	–	–	0.193	0.881	0.159	0.714
deep ensemble	0.085	0.739	0.216	0.701	–	–	–	–
BLoM-Mean	0.128	0.779	0.255	0.663	0.061	0.302	**0.028**	0.923
BLoM-BMA	0.129	0.780	0.258	0.664	0.076	0.342	0.043	0.910

F Additional Analysis

This appendix collects PAC-Bayes rank checks, posterior diagnostics, ablations, and efficiency measurements supporting Section 5.

PAC-Bayes rank behavior. Figure 3 visualizes Table 15 over the finance rank grid. Mean and covariance contributions to KL grow sublinearly, with fitted exponents 0.83 and 0.86 over the nine verified ranks. The reported balanced-design curve has its discrete minimum at $r{=}3$ and matches the measured Gibbs-test minimum. The existing continuous interpolation gives 2.99 and is reported only as a descriptive fit. BLoM-Mean error is not monotone and reaches its lowest value at rank 24, distinct from the Gibbs-curve minimum at rank 3.

The right panel of Figure 3 retains five stored analytical curves for sensitivity, while the direct upper inversion of Eq. 17 determines the reported curve. Using nominal $n{=}1200$, $\delta_{\mathrm{PAC}}{=}0.04$, and the conservative $m{=}300$ union correction, this curve exceeds the corresponding Gibbs test-risk estimate by 3.3 to 8.5 points across the nine verified ranks. The empirical term uses a one-sided familywise Student-t Monte Carlo upper limit from $S{=}200$ draws with $\delta_{\mathrm{MC}}{=}0.01$ across twelve preregistered ranks. Per-draw macro errors are near symmetric, with skewness from -0.51 to 0.80 and excess kurtosis at most 4.2, which supports but does not remove the normal-approximation assumption. All nine task-stratified bootstrap upper limits remain below the reported values. Ranks 32, 48, and 64 are excluded under the final reproducibility gate. Exploratory Gibbs runs at ranks 32 and 48 were retained in the archive but not admitted to the reported analysis, and the rank-64 run was interrupted. Their raw measurements are archived, and no bound is reported for them.

Interference diagnostics. Across twelve task pairs from two suites, the Mean gain over its deterministic initialization rises with subspace alignment under Pearson correlation ($r{=}0.71$, $p{=}0.010$). The trend is concentrated in high-alignment NLI pairs and is weaker under rank correlation ($\rho{=}0.34$, $p{=}0.28$). The BMA trend is marginal ($r{=}0.55$, $p{=}0.065$), so we interpret alignment as a diagnostic rather than a universal predictor of gains.

On finance pairs, Mean ECE tends to decline with alignment ($\rho{=}-0.77$, $p{=}0.072$), although the deterministic initialization improves faster. The TSV-Core-minus-Mean ECE difference therefore becomes more favorable to TSV-Core with alignment ($r{=}-0.95$, $p{=}0.004$). BMA improves ECE on the two most aligned pairs, from 0.138 to 0.125 on FPB+FOMC and from 0.137 to 0.120 on FiQA-SA+FOMC. The remaining changes lie within ± 0.011. The three German-containing pairs have alignment below 0.004 and ECE above the single-task average of 0.128, identifying German as the low-overlap anchor in this suite.

The off-diagonal covariance captures task coupling but correlates only moderately with pairwise interference ($\rho{=}0.39$, $p{=}0.15$ over fifteen pairs). Posterior scale is governed by prior precision and remains stable over the tested Gram-set sizes, as do calibration metrics. BMA's risk-ranking gain is not significantly associated with alignment ($r{=}0.40$, $p{=}0.20$ over twelve pairs).

Auxiliary variance diagnostic. A neural tangent kernel (NTK) isotropy assumption (Jacot et al., 2018; Lee et al., 2019) would be required for a conditional theorem connecting interference to auxiliary variance, but the pilot does not support this condition. We therefore examine the empirical relation

$$\sigma_q^2 \;\approx\; c \cdot \mathrm{SAR}^{\gamma}, \tag{31}$$

where SAR is the subspace alignment ratio of Panariello et al. (2026) and σ_q^2 is the auxiliary linearized loss variance averaged over inputs. The twelve-pair fit gives Pearson $r{=}0.40$ and $p{=}0.20$. We therefore use Eq. 31 only to characterize the near-orthogonal regime.

Refinement and gain decomposition. Figure 10 visualizes the refinement ablation of Table 16. Mean accuracy is stable from one to one hundred steps, and BMA changes by 0.011. The production setting of thirty steps lies in this stable region.

Table 17 compares supervised tiers on the same five finance classification tasks, held-out budget, test split, and metric. All learned tiers remain within 0.7 points of TSV-Core. Scalar coefficient learning adds 0.3 points, full-operator maximum likelihood changes accuracy by -0.4 points, and BLoM-

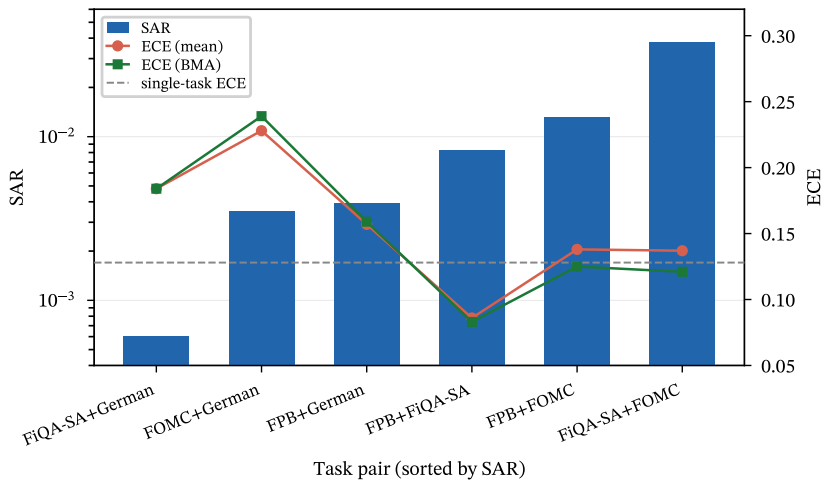

Figure 9: Subspace alignment and calibration on the six finance task pairs (rank 16, λ=100), sorted by the subspace alignment ratio (bars, log axis). The ECE advantage of the deterministic initialization over the un-recalibrated mean mode grows with alignment (Pearson $r = -0.95$, p=0.004 on the TSV-minus-mean ECE difference), and the BMA mode improves ECE on the two highest-alignment pairs. The three German-containing pairs sit below 0.004 alignment and above the average ECE of 0.128 (grey dashed), marking German as the low-overlap interference anchor of the suite.

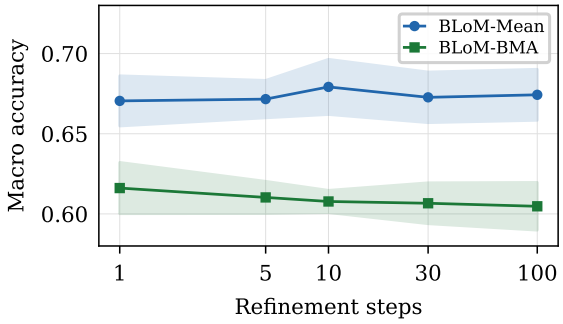

Figure 10: Macro accuracy on the three-task finance subset (FiQA-SA, FOMC, German) against the number of MAP refinement steps, mean $\pm$ std over three adapter seeds. The mean mode is flat between 0.670 and 0.679 across the sweep and the BMA mode drifts down by only 0.011, so the production default of 30 steps sits in the flat tail and refinement is not the source of the BLoM gains (Table 16).

Mean matches the scalar control. The posterior additionally supports the selection and calibration results of Table 14 and the Gibbs-risk analysis of Table 15.

Precision sensitivity illustrates the effect of posterior dispersion. At the fixed λ=100, BMA trails Mean by 0.9 points in this matched diagnostic. The evidence-selected λ=10 produces 0.516 ± 0.040 BMA accuracy with a minimum of 0.413, although its KL is close to that at λ=100 (9.7 versus 10.4). A coarser finance grid places the evidence maximum at λ=1000 and reduces accuracy by 7.6 points relative to the fixed default. On ViT-B/32, evidence selects λ=100, giving BMA accuracy 0.307 compared with Mean accuracy 0.464. The main ViT-B/32 run instead uses λ=1000. These diagnostics show how precision and grid choice affect posterior dispersion.

Table 15: Monte Carlo-adjusted PAC-Bayes sensitivity values on the Finance-1.5B rank grid (λ=100, nominal n=1200, S=200 posterior draws). The reported value is the direct upper inversion of Eq. 17 at δ_{PAC}=0.04 with the conservative union penalty m=300. Its empirical term is the one-sided Student-t normal-approximation upper limit $\widehat{L}^{U}_{\mathrm{val}}$ at δ_{MC}=0.01 across twelve preregistered ranks. Nine ranks passed the preregistered reproducibility check, and margin denotes the nominal value minus the Gibbs test-risk estimate. The realized task strata are unequal, so these values are a nominal sensitivity calculation rather than an instantiation of Theorem 1 on the realized sample. Excluded ranks contribute no risk values, and BLoM-Mean test error is an empirical reference outside the bound. The dimension column is nominal; thin bases become rectangular when Kr exceeds a projection width.

r	gate	nom. $(Kr)^2$	KL	$\widehat{L}^{U}_{\mathrm{val}}$	$\widehat{L}^{\mathrm{Gibbs}}_{\mathrm{test}}$	nominal	margin	Mean test (ref.)
1	pass	36	0.96	0.4177	0.4615	0.4942	+0.0327	0.4631
2	pass	144	2.12	0.4556	0.4746	0.5353	+0.0607	0.4693
3	pass	324	3.64	0.3911	0.4386	0.4742	+0.0356	0.4294
4	pass	576	6.32	0.4435	0.4754	0.5335	+0.0582	0.4705
6	pass	1296	6.95	0.3917	0.4412	0.4828	+0.0416	0.4207
8	pass	2304	10.90	0.4373	0.4684	0.5373	+0.0689	0.4523
12	pass	5184	10.89	0.4253	0.4566	0.5253	+0.0686	0.3931
16	pass	9216	9.75	0.4586	0.4713	0.5561	+0.0848	0.4147
24	pass	20736	15.30	0.5068	0.5309	0.6142	+0.0833	0.3810
32	halted	36864	–	–	–	–	–	–
48	halted	82944	–	–	–	–	–	–
64	halted	147456	–	–	–	–	–	–

Table 16: Component ablations at rank 16 for Finance-1.5B seed 0 and NLI in Mean mode. The full Finance result is 0.588 for this seed and 0.606 over three seeds. The tested Kronecker surrogate is not prior-matched to the full covariance (Appendix G) and has trace relative error 0.99. The full-space diagonal gives 0.639 Mean accuracy but 0.524 BMA accuracy and AURC 0.340 versus 0.257. The TSV-Core center preserves task agreement and yields lower KL than a zero center. Refinement uses a three-task subset over three seeds. Sample count reports NLI BMA accuracy with Monte Carlo standard deviation 0.003 at S=30. Task-count cells begin with fpb+finqa, and the precision row compares fixed λ=100 with configured λ=10 on NLI.

Variant	Finance	NLI	KL
Full configuration	0.588	0.893	9.75
covariance: Kronecker surrogate	not retained	–	×168
covariance: full-space diagonal	0.639	–	9.02
reference basis: TIES	0.571	–	2.29
reference basis: mean	0.557	–	4.35
prior center: zero	0.587	–	×1.19
refinement steps 1 / 10 / 100	0.671 / 0.679 / 0.673	–	–
posterior samples S=10 / 100	–	0.875 / 0.882	–
task count K=2 / 4	0.438 / 0.597	–	–
prior precision λ=100 fixed	0.591	0.872	–
stacked +Pico / +SVC head	0.601 / 0.589	–	–

Table 17: Matched-protocol gain decomposition on Finance-1.5B at rank 16. MLE denotes maximum likelihood estimation. The learned tiers remain within 0.7 points of TSV-Core. The two posterior precisions expose the distinction between fixed precision and evidence selection, analyzed in Appendix F.

Tier	FPB	FiQA-SA	Head.	FOMC	German	Macro
TSV-Core initialization (no learning)	0.743	0.749	0.597	0.440	0.660	0.638
+ scalar coefficients, frozen basis	0.747	0.749	0.607	0.440	0.660	0.640
+ full-operator MLE, no prior	0.780	0.702	0.647	0.380	0.660	0.634
BLoM-Mean (λ=100)	0.753	0.719	0.630	0.440	0.660	0.640
BLoM-BMA (λ=100, S=30)	0.780	0.698	0.600	0.420	0.660	0.632
BLoM-BMA, evidence $\lambda^\star$=10	0.690	0.604	0.353	0.250	0.660	0.512

Table 18: Wall-time in seconds per suite on one 24GB GPU. The BLoM row contains archived full-run timings, including setup, posterior fitting, and Mean/BMA evaluation, not fitting time alone. Vision timings precede the final head/precision updates and are profiling references rather than timings of the final accuracy rows. Separate precision sweeps are excluded. Afterwards one query costs one forward pass for the mean mode and S=30 passes for BMA. Baseline rows are one merge-plus-evaluation pass at their selected configuration and exclude their validation search. The joint row excludes training. Extrapolating the measured memory components gives an estimated 19.5GB peak on the 7B model (Appendix H).

Method	Finance-1.5B	NLI	Vision-B/32	Vision-L/14	Finance-7B
BLoM (full pipeline)	2027	1352	436	3346	5733
Task Arithmetic	261	192	9	143	657
TIES	404	845	20	260	1322
DARE-TIES	115	382	6	66	348
RegMean	99	160	4	58	308
KnOTS	435	2284	27	219	1888
TSV-M	415	1421	30	212	1363
Iso-C	378	1134	27	201	1203
CtM	358	1009	25	194	1121
SVC	128	441	6	64	361
TARA	104	84	5	58	273
TSPA	153	1962	39	143	1237
Pico	98	118	4	57	272
Joint multi-task (evaluation)	24	13	1	13	64

G Posterior Approximation Quality

The core-space projection is exact, while inference introduces two approximations. BLoM replaces local curvature with the empirical Fisher of Eq. 8 and uses a Gaussian approximation around a numerically refined center of the generalized posterior target. Reported BLoM results retain the full, non-factorized empirical-Fisher covariance. This section reports small-scale Langevin diagnostics and evaluates two cheaper covariance structures.

G.1 Gaussian Approximation and Langevin Diagnostics

The pilot uses preconditioned Metropolis-adjusted Langevin sampling (MALA) at dimension 64 and preconditioned unadjusted Langevin sampling (ULA) at dimension 1024. Within each run, the Gaussian approximation and sampler share data and prior. This earlier diagnostic uses task-mean loss normalization and a zero-centered prior, rather than the main sum-scaled target of Eq. 6. The sliced Wasserstein distances normalized by mean marginal standard deviation are 0.117 and 0.121, below the pilot threshold of 0.2. The mean marginal Kolmogorov-Smirnov p-values are 0.068 and 0.142. The composite pilot checks remain borderline, and ULA additionally has discretization error. These measurements are small-scale diagnostics rather than validation of the main posterior at larger dimensions.

G.2 Covariance Structure

We compare the full core-space covariance with an implemented Kronecker surrogate and a full-weight diagonal approximation. The surrogate is inspired by Kronecker-factored approximate curvature (K-FAC) (Martens & Grosse, 2015), but its gradient-moment factors and prior treatment are not a prior-matched implementation of standard K-FAC. In particular, adding $\lambda+\eta$ to each precision factor yields a zero-data covariance proportional to $(\lambda + \eta)^{-2}$ rather than $(\lambda + \eta)^{-1}$. Across the two small-scale runs, its linearized loss variances correlate with the full empirical-Fisher values at 0.63 and 0.71 Pearson. At dimension 1024, its KL to the prior is about 34 times the full-covariance KL. At rank 4, this surrogate has covariance-trace relative error 99%, KL inflation $168\times$, and loss-variance correlation 0.52. These values diagnose the tested surrogate, not the general adequacy of K-FAC.

The full-weight diagonal approximation reaches point accuracy 0.639, but its BMA accuracy is 0.524 and its AURC is 0.340 compared with 0.257 for the full core-space construction. The diagonal result favors the full core-space construction in this setting, without isolating covariance structure from parameterization. Table 16 reports the complete ablation.

H Implementation and Reproducibility Details

H.1 Posterior Computation

The MAP refinement of Eq. 7 uses Adam (Kingma & Ba, 2015) for 30 steps, with learning rate 10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. Each step accumulates gradients over the operator-fitting minibatches before updating the operator. The fixed projection is `model.layers.13.self_attn.v_proj` for both Qwen models, `layers.15.self_attn.v_proj` for NLI and cross-domain Llama, and `encoder.layers.5.self_attn.v_proj` for both CLIP models. These are configured layer indices, not a depth search. The main finance and NLI evaluation runner uses posterior minibatches of 4 and sequence length 768, with posterior minibatches reduced to 1 for Finance-7B. The vision posterior runner uses minibatches of 32. Language-model projections outside the fitted layer retain TSV-Core merges, while vision retains uniform core averages. The 28-layer construction is the factorized extension described in Appendix D.5. Per-sample gradients for Eq. 8 are computed batch by batch, limiting activation memory; storing the gradient matrix still scales with the held-out set size. BMA uses $S = 30$ posterior samples unless stated. The main selective-prediction results rank examples by one minus maximum predictive probability. The labeled loss-variance instantiation of Eq. 11 is used only for covariance diagnostics and is computed through quadratic forms with the empirical-Fisher covariance.

Main runs use prior precisions 100 for both finance scales, 10 for NLI and cross-domain, 1000 for ViT-B/32, and 100 for ViT-L/14. The archived ViT-L/14 evidence search compares $\{1, 10, 100, 1000, 10000\}$ and selects 100. Separate finance and ViT-B/32 evidence sweeps are analyzed in Appendix F. The dispersion sweep characterizes sensitivity to these precision settings. On Finance-1.5B at rank 16, one grid point requires 594 seconds for refinement, 87 seconds for curvature accumulation, and 20 seconds for spectral decomposition and sampling. Core-space projection takes 28 seconds in this profiling run. At inference, Mean costs one forward pass and BMA with $S = 30$ costs $28.5\times$ more per sample in the measured operator-switch implementation. Both modes can be amortized by batching.

When the operator dimension exceeds twice the held-out sample count, we compute the spectrum from the $n \times n$ Gram matrix of weighted per-sample gradients. This is algebraically equivalent to eigendecomposition of $\boldsymbol{H}_{\mathrm{EF}} + (\lambda + \eta)\boldsymbol{I}$ and uses $O((Kr)^2 n)$ memory instead of $O((Kr)^4)$.

H.2 Hardware and Budgets

Pilot runs use one 24GB GPU. The full 9216-dimensional Gaussian pipeline peaks at 18.6GB on the 1.5B base model. Its thirty-step refinement takes 594 seconds at the main sequence budget. The separate 28-layer posterior extension completes in 649 minutes without adapter retraining. Extrapolation to the 7B base model gives about 19.5GB peak memory; this is an estimate, not a measured peak. Per-sample gradient computation costs 0.106 seconds on the 1.5B model. Accumulating 1000 gradients takes about 1.8 minutes, while an auxiliary Hessian-vector product benchmark takes 43.8 seconds on 300 samples.

H.3 Adapter Training

Finance adapters are trained in FLARE format (Xie et al., 2023) with completion-only cross-entropy, 1200 training samples per task where available, and validation splits of at least 200 unique samples per task. Tasks with fewer unique samples use all of them. Training uses AdamW with zero weight decay, learning rate 10^{-4}, a cosine schedule with 5% warmup, three epochs, batch size 8 and a maximum sequence length of 768. The LoRA alpha equals the rank (unit scaling), the dropout is 0, and the adapters attach to the query, key, value and output projections. The 7B adapters use the identical recipe with the base model switched to Qwen2.5-7B-Instruct. The NLI and vision per-task adapters are the public checkpoints released by Stoica et al. (2025) and carry no training cost on our side. The cross-domain experiment combines the six NLI adapters with FPB and TFSent adapters trained on the same Llama-3-8B base using the finance recipe and training batch size 4. TFSent is the three-class Twitter Financial News Sentiment dataset, loaded from the archived `zeroshot/twitter-financial-news-sentiment` CSV splits, with negative, positive, and neutral labels.

Four data roles are separated throughout the pipeline. The adapter-training data (1200 samples per task where available) determine the adapters, and through them the reference bases and the prior center $\boldsymbol{M}_0$. For the primary accuracy and PAC-Bayes analyses, the operator-fitting set $\mathcal{D}$ requests up to 200 examples per task and drives MAP refinement and curvature estimation. Separate precision-selection diagnostics use held-out evidence as detailed above. In FPB, FinQA, FiQA-SA, Headlines, FOMC, and German order, the realized operator-fitting sizes are 200, 200, 188, 200, 100, and 100. For the uncertainty protocol of Table 14, each held-out pool is divided into disjoint operator-fitting and calibration sets with up to 100 examples each, for posterior fitting and fitted calibrators respectively. Finance tasks with smaller pools use $94/94$ examples for FiQA-SA and $50/50$ for FOMC and German. The auxiliary theorem of Theorem 2 concerns a label-free score fixed before examining $\mathcal{D}_{\mathrm{cal}}$ and is not used to certify the main confidence-based curves. The final uncertainty metrics are reported on the separate test split, which is not used for operator or calibrator fitting. Finance accuracy tables use completion-token exact match under teacher forcing, while their separate five-task panels use label-probability predictions and exclude FinQA. The joint multi-task references are trained on the pooled training data of each suite with the same recipe, with batch size 4 at 7B (4139.6 seconds of training). The vision joint reference uses two epochs and batch size 64.

H.4 Baselines and External Code

Deterministic baselines run through the official Core Space repository (Panariello et al., 2026), which provides the projection and the task arithmetic, TIES, TSV-M, and Iso-C merge rules. DARE-TIES (Yu et al., 2024) and RegMean (Jin et al., 2023) use the local implementations in `baselines_gen.py`, with fixed-seed drop-and-rescale and activation-Gram merging respectively. KnOTS checkpoints and evaluation protocols come from Stoica et al. (2025). SVC, TARA, CtM, and TSPA use local adaptations of their authors' public implementations (Li et al., 2026b; Jeong et al., 2026; He et al., 2026; Li et al., 2026c). Pico is reimplemented from its published description because no public implementation is available (Tang & Yang, 2026). The reference models are the pretrained model, individual task adapters, and a joint multi-task adapter trained on pooled task data. Uncertainty comparisons include temperature scaling (Guo et al., 2017), vector scaling (Kull et al., 2019), MC-Dropout (Gal & Ghahramani, 2016), and a full-weight diagonal Gaussian approximation as a covariance control.

H.5 Selected Hyperparameters

Table 19: Selected hyperparameters per suite. Baseline coefficients $\lambda^\star$ use held-out validation accuracy. Parentheses give the TIES trimming rate, TARA searches only λ, and KnOTS is hyperparameter-free. BLoM entries are the precisions used in the reported runs. ViT-L/14 uses the archived evidence-selected value; separate evidence and dispersion diagnostics are described in Appendix H and Appendix F.

Method	Fin-1.5B	NLI	B/32	L/14	Fin-7B
BLoM (prior precision)	100	10	1000	100	100
Task Arithmetic	1.5	1.5	0.7	0.5	1.5
TIES	0.75 (0.1)	0.75 (0.1)	0.5 (0.5)	0.5 (0.5)	1.25 (0.3)
DARE-TIES	0.5	0.5	0.5	0.5	0.75
RegMean	1.0	1.25	0.75	0.5	1.0
KnOTS	hyperparameter-free				
TSV-M	0.5	0.6	0.3	0.3	0.8
Iso-C	0.7	0.9	0.2	0.3	1.0
CtM	1.5	1.5	0.8	0.6	1.5
SVC	1.5	1.5	0.75	0.5	1.5
TARA	1.0	1.0	0.5	0.5	1.5
TSPA	1.5	1.5	0.75	0.75	1.5
Pico	1.0	0.75	0.5	0.5	1.25