A Rate-Distortion Function for Model Merging

Sankalp Pathak
pathaksankalp04@gmail.com

Sanjay Garg
gargsv@gmail.com

May 26, 2026

Abstract

Practitioners merge many LoRA fine-tunes into one deployable model, but no principled limit says how well this can be done at a fixed storage budget. This paper casts merging as multi-source lossy source coding under worst-task distortion and establishes what is, to the authors' knowledge, the first rate-distortion theorem for it. For T tasks with rank-r updates, a per-task error radius B, and a storage budget of R bits, the worst-task distortion under a Stiefel-random worst-case distribution has a closed-form floor $B^2(1 - d_{\text{eff}}/(Tr))$, set by an effective dimension $d_{\text{eff}} := \mathbb{E}[\text{rank}(\sum_t P_{V_t})]$ (the expected rank of the summed projectors P_{V_t} onto the task subspaces), plus a compression term scaling as $\Theta(2^{-2R/d_{\text{eff}}})$. An explicit Gaussian-QR rotation followed by uniform scalar quantization matches the lower bound up to a factor $C = Tc^2/3$ (with c a clipping constant) for a fixed task count T (tightening the T-dependence remains open).

Across 16 real LoRA adapters (4 base models $\times$ 4 tasks), the task subspaces are linearly *independent* ($d_{\text{eff}} = Tr$ in every layer), so the bound's floor is zero. Current merging methods nonetheless leave 0.10–0.22 nats/token of worst-task error. This is a loose lower envelope on the gap to optimal, since real task vectors are far more benign than the worst case. The operative question is therefore algorithmic, not informational. Among standard methods only TIES beats naive averaging, and only via the bottleneck task, whereas subspace alignment (KnOTS) is statistically indistinguishable from naive averaging (as the theory predicts at zero overlap), and a universal "less-is-more" regime appears, in which 2-bit task-vector quantization beats higher-bit quantization on all four architectures. All empirical claims carry 95% bootstrap confidence intervals (CIs).

1 Introduction

Model merging is everywhere, but nobody knows the limit. Low-Rank Adaptation (Hu et al., 2022) has made it routine to maintain dozens of task-specific fine-tunes per base model, and it has made *merging* those fine-tunes into a single deployable model the dominant practical question. A now-standard

workflow: fine-tune T LoRA adapters on T downstream tasks, combine them post-hoc, serve one model. Task Arithmetic (Ilharco et al., 2023), TIES (Yadav et al., 2023) (consensus sign election with magnitude pruning to resolve cross-task interference), DARE (Yu et al., 2024), KnOTS (Stoica et al., 2025) (SVD-based alignment of task updates in a shared subspace before merging), DO-Merging (Zheng et al., 2025), Core Space (Panariello et al., 2025), Task-Vector Quantization (Kim et al., 2025), and 1-bit-Merging (Liu et al., 2025a) all compete on this problem. Each proposes an encoder, tunes some hyperparameters, and reports benchmark numbers. None answers the prior question: given a storage budget of R bits per merged model, what is the *best possible* merging algorithm, and how far are existing methods from it?

The rate-distortion question. This problem is formalized here as multi-source lossy source coding. The first design choice is the figure of merit. A merged model that scores well on most downstream tasks but collapses on one is not a useful deployment target; quality should be measured by the *worst* task, not the average. Once this choice is fixed, the formalism is direct. Fix T tasks, ambient dimension d, and LoRA rank $r \leq d$. Each task is characterized by a pair (τ_t, H_t): a task vector $\tau_t \in \mathbb{R}^d$ (the LoRA update) and a positive-semidefinite H_t of rank r (the task's Hessian at τ_t, or a Fisher surrogate) with $\tau_t^\top H_t \tau_t \leq B^2$. An encoder maps $(\tau_1, \ldots, \tau_T)$ into R bits; a decoder reconstructs a single merged model $w^\star \in \mathbb{R}^d$. Per-task distortion is $D_t(w) := (w-\tau_t)^\top H_t (w-\tau_t)$, the second-order loss inflation from using w in place of τ_t. The rate-distortion function $\mathcal{D}^\star(T, d, r, B, R)$ is the infimum over encoders and decoders of expected max-distortion $\max_{t \leq T} D_t(w^\star)$ at rate R.

Two features separate this problem from classical lossy source coding. First, the encoder observes T sources rather than one; multi-description coding (El Gamal and Cover, 1982) addresses related questions but under *average* distortion. Lemma 1 (below) shows worst-task distortion dominates average distortion (max $\geq$ avg), so an average-distortion treatment yields a strictly smaller rate-distortion function than the one established here. Second, the low-rank structure of H_t means distortions are measured in task-specific rank-r subspaces; the effective dimension is the combinatorial quantity $d_{\text{eff}} := \mathbb{E}[\text{rank}(\sum_t P_{V_t})]$, which satisfies $d_{\text{eff}} \leq \min(Tr, d)$, with the bound attained almost surely when the V_t's are drawn iid from the Stiefel manifold.

Contributions. This paper gives what is, to the authors' knowledge, the first rate-distortion theorem for model merging, with matching lower and upper bounds up to a constant factor in the generic regime, universal in d, r, R, B, and linear in the number of tasks T, which is treated as fixed throughout (the T-dependence is discussed under Scope below).

1. **An algebraic identity reducing max-distortion merging to average-distortion quantization plus an irreducible floor (Lemma 1).** For

every admissible (τ, H) and every $w \in \mathbb{R}^d$,

$$\max_t D_t(w) \;\geq\; (w - \bar{\tau}_H)^\top \bar{H} (w - \bar{\tau}_H) + \tfrac{1}{T} \sum_t (\tau_t - \bar{\tau}_H)^\top H_t (\tau_t - \bar{\tau}_H),$$

where $\bar{\tau}_H = \bar{H}^+ \cdot T^{-1} \sum_t H_t \tau_t$ is the H_t-weighted centroid and $\bar{H} = T^{-1} \sum_t H_t$. The identity is the technical bridge from Shannon's single-source rate-distortion machinery to the multi-source max-distortion setting: the first term is what bits buy, the second is what no rate can erase.

2. **A closed-form expression for the irreducible floor in terms of subspace overlap (Lemma 2).** Under a Stiefel-random hard distribution $P^\star$ on (τ, H), the floor reduces to $B^2(1 - d_{\text{eff}}/(Tr))$ exactly. The interpolation between $T = 1$ (floor 0, all bits go to compression) and V_t mutually orthogonal (floor 0 again, reconstructive task arithmetic is lossless at infinite rate) is explicit; positive overlap raises the floor.

3. **A rate-distortion lower bound on every merging algorithm (Theorem 3).** Under $P^\star$ with arbitrary positive-definite spectra $(D_t)_{t=1}^T$, no encoder/decoder pair at rate R bits can do better than

$$\mathcal{D}^\star(T, d, r, B, R) \;\geq\; B^2\big(1 - \tfrac{d_{\text{eff}}}{Tr}\big) + c_{\text{TQ}} \cdot \tfrac{B^2 d_{\text{eff}}}{Tr} \cdot 2^{-2R/d_{\text{eff}}}.$$

   The proof composes Yao's minimax principle, Lemma 1, a reduction to the range$(\bar{H})$ subspace, and an $O(d)$-equivariance argument that works regardless of task-dependent spectra.

4. **An explicit algorithm that matches the lower bound up to a constant factor (Theorem 4).** In the floor-zero Stiefel regime ($Tr \leq d$, V_t drawn iid from the Stiefel manifold), randomized orthogonal mixing of the $\bar{H}^{1/2}$-scaled centroid followed by uniform scalar quantization achieves $\mathbb{E}[\max_t D_t(w^\star)] \leq C \cdot B^2 \cdot 2^{-2R/(Tr)}$ with $C = Tc^2/3$ (universal in d, r, R, B; linear in T). Combined with Theorem 3, $\mathcal{D}^\star = \Theta\big(B^2 \cdot 2^{-2R/(Tr)}\big)$ in this regime for fixed T. The construction reduces to TurboQuant (Zandieh et al., 2025) at $T = 1$.

5. **Numerical validation of the predicted slopes and constants.** Monte Carlo experiments across $T \in \{2, 3, 4, 8\}$, $d \in \{128, 512, 2048\}$, $r \in \{4, 8, 16\}$, with both isotropic and heavily anisotropic task-dependent spectra, confirm the predicted slope $-2R/d_{\text{eff}}$ on $\log_2 \mathbb{E}[\text{excess}]$ to within bootstrap 95% CI ± 0.01 across 1000 trials per configuration. The achievability constant C is empirically ≈ 11–13, matching the analytical bound $Tc^2/3$.

6. **A real-LLM calibration that turns the bound into a measured gap (§6).** Sixteen LoRA adapters are trained across four base models (Llama-3.1-8B (Grattafiori et al., 2024), Qwen-2.5-7B (Yang et al., 2024), Mistral-7B (Jiang et al., 2023), Yi-1.5-9B (Young et al., 2024)) and four tasks, and five standard merging methods are evaluated under worst-task NLL with bootstrap confidence intervals. Real task subspaces turn out to be linearly *independent* ($d_{\text{eff}} = Tr$ in every layer of every model), so the bound's floor is zero;

yet current methods leave 0.10–0.22 nats/token of worst-task error on the table. The gap is the suboptimality of today's heuristics, not an irreducible floor. Three further findings: (i) only TIES improves on naive averaging, and only on the bottleneck task, while subspace-alignment (KnOTS) is statistically indistinguishable from it, precisely the theory's prediction at $d_{\text{eff}} = Tr$, where there is no overlap for a subspace method to exploit; (ii) merging difficulty tracks the pretraining/instruction-tuning recipe rather than attention design, and is *not* predicted by any subspace-overlap statistic; (iii) a universal "less-is-more" regime in which 2-bit task-vector quantization beats higher-bit quantization on all four architectures.

Rationale for practice. The bound gives practitioners a *lower envelope*, not a tight calibration curve: because real task vectors are far more benign than the worst-case $P^\star$, the bound certifies only that no algorithm can do better, and any method's gap to $\mathcal{D}^\star$ is an upper estimate of its suboptimality rather than an exact grade. The theorem exposes two regimes through the floor $B^2(1-d_{\text{eff}}/(Tr))$. When task subspaces overlap substantially ($d_{\text{eff}} \ll Tr$) the floor is irreducible and no amount of rate closes it; when they are independent ($d_{\text{eff}} = Tr$) the floor vanishes and quality is rate- and algorithm-limited. *Which regime do real LoRA merges live in?* The experiments answer this directly (§6): across all four base models and every layer the measured value is $d_{\text{eff}} = Tr$, the independent, floor-zero regime. The implication is sharper than a generic "it depends": for the task sets practitioners actually merge, the irreducible floor is *not* the binding constraint, and the 0.10–0.22 nats/token of worst-task error observed here is headroom that a better algorithm could in principle recover. Consistent with this, subspace-alignment merging (KnOTS) gives no improvement over naive averaging in these measurements (there is no subspace overlap for it to exploit), whereas TIES, which attacks sign-level interference rather than subspace overlap, does help. A merging method should therefore be judged on how much of the floor-zero gap it closes, not on subspace alignment alone.

A worked example. Consider a Llama-3.1-8B base with $T = 4$ rank-16 LoRAs, so $d_{\text{eff}} \le Tr = 64$ per layer. The theory's two regimes are the orthogonal case ($d_{\text{eff}} = 64$, floor 0, rate-/algorithm-limited) and the fully-overlapping case ($d_{\text{eff}} = 16$, floor $\approx 0.75\,B^2$, a ceiling no merger can break). Empirically all four models tested sit at $d_{\text{eff}} = 64$ in every layer (the floor-zero corner), so the operative question is algorithmic, not informational: the measured worst-task excess (0.22 for this model) is distance from optimal, not an unavoidable floor. The overlap-limited corner, where the right response is to serve tasks from separate adapters rather than merge, did not arise for any task set tried here; identifying real task families that induce it is itself a useful direction the theorem makes precise.

Scope and limitations. The matching upper bound in Theorem 4 is proved under the generic Stiefel-random hard distribution $P^\star$ with $Tr \le d$. In the floor-positive regime (shared or heavily overlapping V_t), an explicit linear encoder

achieves an excess rate of $2^{-2R/(d_{\text{eff}}+|A|-1)}$ (where $|A|$ is the size of the active set at the Chebyshev center), verified numerically for $T \in \{2, 3, 4\}$; closing the exponent gap to the lower bound's $2^{-2R/d_{\text{eff}}}$ appears to require non-linear / random-coding encoders and is left open (Remark 6). Separately, the upper-bound constant $C = Tc^2/3$ grows linearly in T while the lower-bound constant c_{TQ} is T-independent, so the "matching up to constants" claim holds for the number of tasks T treated as fixed; the linear T-gap (whose source is the max $\leq$ $T \cdot$ avg step of Theorem 4, §5) is not known to be tight, and tightening the T-dependence is open. Finally, the bound is stated in the MSE / quadratic-loss setting; the extension to cross-entropy LLM loss via local Fisher-quadratic approximation is discussed in §7.

Paper organization. §2 surveys existing merging constructions and the closest theoretical analyses, situating the rate-distortion view among them. §3 introduces the formal rate-distortion model for merging and the admissible hard distributions. §4 proves the lower bound (Theorem 3), with the max-$\geq$-avg identity and closed-form floor as the central tools. §5 gives the matching achievability construction and analyses. §6 reports experiments in two parts: synthetic rank-r ensembles validating the theoretical slopes and constants, and a real-LLM calibration across four base models and four tasks that measures the gap between current merging methods and the bound. §7 discusses extensions to cross-entropy loss, the open shared-V exponent gap, what the real-LLM measurements add, and implications for practical merging-algorithm design.

2 Related Work

Model merging: constructions without limits. A wave of post-hoc model-merging methods has emerged since Task Arithmetic (Ilharco et al., 2023) showed that adding fine-tuning deltas of separately-trained task vectors recovers non-trivial multi-task performance. Fisher-weighted averaging (Matena and Raffel, 2022), sparsification-based approaches (TIES (Yadav et al., 2023), DARE (Yu et al., 2024)), and more recent rotation- and subspace-alignment families (KnOTS (Stoica et al., 2025), DO-Merging (Zheng et al., 2025), Core Space (Panariello et al., 2025), TARA (Jeong et al., 2026), ARM (Yao et al., 2026)) all seek to reduce destructive interference between task vectors before averaging. A parallel line quantizes task vectors at merge time to shrink the storage footprint: Task-Vector Quantization (TVQ) (Kim et al., 2025) and 1-bit-Merging (Liu et al., 2025a) achieve competitive accuracy at one to four bits per coordinate. A complementary empirical study examines linear-combination merging and cross-task interference for code-analysis adapters (Pathak and Garg, 2026). What unifies these methods is that they are *constructions*: each proposes a particular encoder and evaluates it empirically. None establishes a *fundamental limit* on merging quality at a given storage or rate budget. This paper supplies that limit and, in the generic regime, a matching construction.

Existing theoretical analyses of merging. The closest theoretical work on merging falls into three strands. *Weight disentanglement* (Ortiz-Jimenez et al., 2023) identifies the rank-r Hessian structure of LoRA fine-tunes and relates task arithmetic's empirical success to locally-additive task functionals; this provides the $H_t = P_{V_t}$ (and more generally $H_t = U_t D_t U_t^\top$) formalism adopted in §3, but does not give rate-distortion bounds. *Task-Vector Bases* (Zeng et al., 2025) posits that fine-tuned weights lie in a thin Gaussian shell, a geometric assumption rather than an information-theoretic one. From the same group, ATM (Zhou et al., 2024) reads task vectors as noisy single-step gradient updates and improves merging by alternating tuning and merging; its later *On Task Vectors and Gradients* (Zhou et al., 2025) develops this into a second-order Taylor-expansion bound on task-arithmetic error. None of these works frame merging as lossy source coding or derive Shannon-style limits. This is, to the authors' knowledge, the first work to pose merging as a multi-source rate-distortion problem with worst-task distortion as the figure of merit.

Rate-distortion and rotation-for-quantization ancestry. The information-theoretic toolkit used here is standard. Classical rate-distortion theory (Shannon, 1959; Cover and Thomas, 2006) gives the single-source setting; multi-description coding (El Gamal and Cover, 1982) addresses multi-receiver lossy coding but under average rather than maximum distortion, which precludes direct application to merging (Lemma 1 shows the max-distortion formulation is strictly stronger and exhibits a different RD function). The direct intellectual ancestor of the achievability construction here is TurboQuant (Zandieh et al., 2025), which establishes Shannon-style lower bounds on randomized vector quantization via Yao's minimax principle and achieves them up to a universal constant via Haar-rotation + uniform scalar quantization. Theorem 4 reduces to TurboQuant's Theorem 3 in the $T = 1$ single-task limit. A further line of work uses structured rotations (Hadamard / Walsh-Hadamard) as preconditioners for quantization in the single-model setting: QuIP (Tseng et al., 2024), QuaRot (Ashkboos et al., 2024), SpinQuant (Liu et al., 2025b), and RoLoRA (Huang et al., 2024). These motivate the choice of randomized orthogonal mixers in the achievability algorithm but do not address the multi-source / merging setting.

The November 2025 negative result, and why a lower bound still matters. A recent large-scale empirical study (Hitit et al., 2025) reports that across four LLM families and sixteen benchmarks, only Task Arithmetic reliably produces constructive interference; subspace-alignment methods (KnOTS-style) often underperform or fail. This might seem to undermine the case for theoretical work on merging. It does not. That result is an *upper bound* on specific algorithms, evidence that those constructions are suboptimal at LLM scale. Theorem 3 is a *lower bound* on every algorithm: it says that whatever the best merging procedure is, it cannot beat $B^2 f_{\text{floor}} + c_{\text{TQ}}(B^2 d_{\text{eff}}/(Tr)) \cdot 2^{-2R/d_{\text{eff}}}$. The negative result and this lower bound are compatible and mutually reinforc-

ing, and the measurements of §6 pin down the mechanism. Real LoRA task subspaces are found to be linearly *independent* ($d_{\text{eff}} = Tr$ in every layer of all four base models tested), i.e. the floor-zero regime. In that regime a subspace-alignment method has no overlap to exploit, so it cannot improve on naive averaging. Indeed, KnOTS is measured to be statistically indistinguishable from Task Arithmetic, while only TIES (which resolves sign-level interference rather than subspace overlap) helps. This explains the reported negative result not by an irreducible floor but by its opposite: with no subspace overlap, the floor is zero and the entire merging gap is algorithmic slack above a beatable target. The lower bound's *floor-positive* regime (substantial subspace overlap, no amount of rate sufficient) is theoretically characterized here but did not arise empirically for the task families tried here; identifying real task sets that induce it, and fully characterizing achievability there, remain open (Remark 6).

3 Problem Setup and Preliminaries

Task vectors and per-task geometry. Fix integers $T \geq 1$, $d \geq 2$, $1 \leq r \leq d$, and a radius $B > 0$. Each of the T tasks is specified by a pair $(\tau_t, H_t) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}$ with

$$H_t \succeq 0, \quad \operatorname{rank}(H_t) \leq r, \quad \tau_t \in V_t := \operatorname{range}(H_t), \quad \tau_t^\top H_t \tau_t \leq B^2. \tag{1}$$

Per-task loss is the quadratic $f_t(w) := \frac{1}{2}(w - \tau_t)^\top H_t (w - \tau_t)$, which attains zero at any w with $P_{V_t} w = \tau_t$. The tuple data are written $\boldsymbol{\tau} := (\tau_t)_{t=1}^T$ and $\boldsymbol{H} := (H_t)_{t=1}^T$, and any $(\boldsymbol{\tau}, \boldsymbol{H})$ satisfying (1) is called *admissible*.

The specialization $H_t = P_{V_t}$ (the orthogonal projector onto V_t) recovers the weight-disentangled LoRA setting of Ortiz-Jimenez et al. (2023); the general $H_t \succeq 0$ case subsumes Fisher-metric surrogates for cross-entropy LLM losses with heterogeneous task-specific curvature.

Worst-task distortion. A merged model $w^\star \in \mathbb{R}^d$ incurs per-task distortion $D_t(w^\star) := (w^\star - \tau_t)^\top H_t (w^\star - \tau_t)$. The figure of merit is the *worst-task* distortion

$$\Delta(w^\star; \boldsymbol{\tau}, \boldsymbol{H}) \;:=\; \max_{t\in[T]} \; D_t(w^\star). \tag{2}$$

The choice of max over avg matters. A merged model that performs well on most tasks but catastrophically on one is not a useful deployment target; classical multi-description coding (El Gamal and Cover, 1982) treats average distortion and has a strictly weaker rate-distortion function than the one established here (see Lemma 1).

Rate-R merging codes. A rate-R merging code is a pair (E, D) with encoder $E : (\mathbb{R}^d)^T \to \{1, \ldots, 2^R\}$ and decoder $D : \{1, \ldots, 2^R\} \to \mathbb{R}^d$, both of which may depend on $\boldsymbol{H}$ as shared side information: in any real deployment $\boldsymbol{H}$ is determined by the LoRA factors U_t (with a curvature or Fisher surrogate for

the spectra), which both ends already hold, so only the R bits describing $\boldsymbol{\tau}$ are charged to the rate. The encoder observes $\boldsymbol{\tau}$; the reconstruction is $w^\star := D(E(\boldsymbol{\tau}))$. The lower bound holds even when $\boldsymbol{H}$ is revealed for free to both ends. The *rank-r rate-distortion function* is

$$\mathcal{D}^\star(T, d, r, B, R) \;:=\; \inf_{(E,D)} \sup_{(\boldsymbol{\tau},\boldsymbol{H}) \text{ admissible}} \mathbb{E}\big[\Delta(w^\star; \boldsymbol{\tau}, \boldsymbol{H})\big], \tag{3}$$

where the expectation is over any randomness the encoder and decoder share.

With the figure of merit, admissibility constraints, and code class fixed, the central question of the paper is the following.

Problem 1 (Rank-r rate-distortion model merging)**.** Determine the rank-r rate-distortion function $\mathcal{D}^\star(T, d, r, B, R)$ defined in (3), and exhibit an explicit encoder/decoder pair (E, D) attaining it. Concretely:

1. (*Lower bound*) Find the largest $L(T, d, r, B, R)$ such that for every rate-R merging code, $\sup_{(\boldsymbol{\tau},\boldsymbol{H}) \text{ admissible}} \mathbb{E}[\Delta(w^\star; \boldsymbol{\tau}, \boldsymbol{H})] \geq L(T, d, r, B, R)$.
2. (*Upper bound*) Construct a rate-R merging code $(E^\star, D^\star)$ achieving distortion at most $U(T, d, r, B, R)$ on every admissible tuple.
3. (*Tightness*) Establish $L = \Theta(U)$.

The remainder of the paper resolves (i) (Theorem 3, §4), (ii) (Theorem 4, §5), and (iii) up to a universal constant in the floor-zero Stiefel regime.

Key derived quantities. The analysis repeatedly uses

$$\bar{H} := \tfrac{1}{T}\sum_{t=1}^{T} H_t, \qquad \bar{\tau}_H := \bar{H}^+ \cdot \tfrac{1}{T}\sum_{t=1}^{T} H_t \tau_t, \qquad d_{\text{eff}} := \operatorname{rank}\Big(\sum_{t=1}^{T} P_{V_t}\Big), \tag{4}$$

where $(\cdot)^+$ denotes the Moore–Penrose pseudoinverse. The *H_t-weighted centroid* $\bar{\tau}_H$ is the unique element of $\operatorname{range}(\bar{H}) = V_1 + \cdots + V_T$ satisfying $\bar{H}\bar{\tau}_H = T^{-1}\sum_t H_t\tau_t$; it reduces to the arithmetic mean $\bar{\tau}$ when $H_t = I_d$. The *effective dimension* $d_{\text{eff}} \leq \min(Tr, d)$ is the combinatorial quantity that controls the rate-distortion function; its meaning is the number of independent directions the decoder must reconstruct.

Hard distribution $P^\star$. The lower bound is established against a specific distribution $P^\star$ on admissible tuples. For each task $t \in [T]$, independently:

1. Sample $U_t \in \text{Stiefel}(d, r)$ uniformly (equivalently, take the Gram–Schmidt orthogonalization of a $d \times r$ matrix of iid Gaussians); set $V_t := \operatorname{range}(U_t)$.
2. Fix a positive-definite diagonal $r \times r$ matrix $D_t \succ 0$ (the task-t spectrum inside V_t); set $H_t := U_t D_t U_t^\top$.
3. Sample $u_t \sim \text{Unif}(S^{r-1})$ and set $v_t := B\, D_t^{-1/2} u_t \in \mathbb{R}^r$, $\tau_t := U_t v_t$.

The scale normalization $\tau_t^\top H_t \tau_t = v_t^\top D_t v_t = B^2 \|u_t\|^2 = B^2$ holds by construction. The $D_t^{-1/2}$ stretch makes τ_t uniform on the H_t-weighted unit ellipsoid inside V_t, which is what makes the second-moment identity $\mathbb{E}[H_t \tau_t \tau_t^\top H_t | U_t] = (B^2/r) H_t$ hold *regardless of* D_t. This D_t-freeness is the feature that makes Theorem 3 apply to arbitrary task-dependent spectra. For the projector special case $H_t = P_{V_t}$ (i.e. $D_t = I_r$), $P^\star$ reduces to the Stiefel-random LoRA distribution with $\tau_t \sim \mathrm{Unif}(B \cdot S^{r-1} \cap V_t)$.

Yao's minimax bridge. The worst-case rate-distortion function $\mathcal{D}^\star$ in (3) is at least the Bayesian rate-distortion function under any prior. In particular, taking the prior to be $P^\star$:

$$\mathcal{D}^\star(T, d, r, B, R) \;\geq\; \inf_{(E,D)} \; \mathbb{E}_{(\boldsymbol{\tau}, \boldsymbol{H}) \sim P^\star} \big[\Delta(w^\star; \boldsymbol{\tau}, \boldsymbol{H}) \big]. \tag{5}$$

This is Yao's minimax principle (Yao, 1977) applied to the randomized-code / worst-case-input game. All subsequent lower bounds are proved against the RHS of (5); the LHS inherits them for free.

4 Lower Bound

This section establishes the rank-r rate-distortion lower bound. The argument rests on three ingredients: a deterministic max-$\geq$-avg identity that converts max-distortion into an average-distortion problem plus an irreducible floor (§4.1); a closed-form evaluation of that floor under the hard distribution $P^\star$ (§4.2); and a rate-distortion lower bound on the remaining average-distortion subproblem via a Shannon converse applied conditionally on the curvature data $\boldsymbol{H}$ (§4.3).

4.1 The H_t-weighted max $\geq$ avg identity

The Phase 0 identity $T^{-1} \sum_t \|w - \tau_t\|^2 = \|w - \bar{\tau}\|^2 + T^{-1} \sum_t \|\tau_t - \bar{\tau}\|^2$ relies on $\sum_t (\tau_t - \bar{\tau}) = 0$, which depends only on $\bar{\tau}$ being the arithmetic mean. The right generalization to H_t-weighted distortion uses $\bar{\tau}_H$ of (4) as the centroid.

Lemma 1 (H_t-weighted max $\geq$ avg identity)**.** *For every admissible* $(\boldsymbol{\tau}, \boldsymbol{H})$ *and every* $w \in \mathbb{R}^d$*,*

$$\tfrac{1}{T} \sum_{t=1}^{T} (w - \tau_t)^\top H_t (w - \tau_t) \;=\; (w - \bar{\tau}_H)^\top \bar{H} (w - \bar{\tau}_H) + \tfrac{1}{T} \sum_{t=1}^{T} (\tau_t - \bar{\tau}_H)^\top H_t (\tau_t - \bar{\tau}_H), \tag{6}$$

and hence $\max_t (w - \tau_t)^\top H_t (w - \tau_t) \geq (w - \bar{\tau}_H)^\top \bar{H} (w - \bar{\tau}_H) + T^{-1} \sum_t (\tau_t - \bar{\tau}_H)^\top H_t (\tau_t - \bar{\tau}_H)$.

The full proof is given in the supplementary material. The argument is a two-line algebraic expansion: the cross-term $\sum_t H_t (\tau_t - \bar{\tau}_H) = T(\bar{H}\bar{\tau}_H - \bar{H}\bar{\tau}_H) = 0$

vanishes by the defining property $\bar{H}\bar{\tau}_H = T^{-1}\sum_t H_t\tau_t$ of the centroid. No probabilistic argument is needed; at $H_t = I_d$ this recovers the Phase 0 identity exactly.

Two consequences of Lemma 1 drive everything that follows. First, *components of w orthogonal to* $\mathrm{range}(\bar{H}) = V_1 + \cdots + V_T$ *annihilate against every H_t and do not affect* (6); without loss of generality the decoder's output lies in an at-most-d_{eff}-dimensional subspace. Second, the max-distortion lower bound separates into a compression term in $\bar{H}$-metric on $\bar{\tau}_H$ and an irreducible floor term that does not depend on w at all. The rest of the argument bounds each piece.

4.2 Closed-form floor

Under the hard distribution $P^\star$ of §3, the floor term admits an exact closed form depending only on d_{eff}.

Lemma 2 (Closed-form floor)**.** *Under $P^\star$ with any fixed positive-definite spectra $(D_t)_{t=1}^T$ (possibly distinct across tasks),*

$$\mathbb{E}_{P^\star}\left[\frac{1}{T}\sum_{t=1}^{T}(\tau_t - \bar{\tau}_H)^\top H_t(\tau_t - \bar{\tau}_H)\right] \;=\; B^2\left(1 - \frac{\mathbb{E}[d_{\mathrm{eff}}]}{Tr}\right). \tag{7}$$

The proof (supplementary material) splits into a deterministic identity reducing the floor to a bilinear form in $\bar{\tau}_H$ and an expectation-via-isotropy step. The key probabilistic input is the second-moment identity $\mathbb{E}[H_t\tau_t\tau_t^\top H_t \mid U_t] = (B^2/r)H_t$, which holds *for any* D_t because of the $D_t^{-1/2}$ stretch in the definition of $P^\star$.

Three limiting cases illuminate the formula.

1. **Full-rank limit** ($r = d$, $V_t = \mathbb{R}^d$). Then $d_{\mathrm{eff}} = d$ a.s., floor $= B^2(1 - d/(Td)) = B^2(1-1/T)$, recovering the Phase 0 formula for classical averaged-vector quantization.
2. **Single-task limit** ($T = 1$). Then $d_{\mathrm{eff}} = r$, floor $= B^2(1 - r/r) = 0$, and the rate-distortion function reduces to single-source quantization of $\tau_1 \in V_1 \cong \mathbb{R}^r$. This case matches TurboQuant's Theorem 3 (Zandieh et al., 2025).
3. **Shared-subspace limit** ($V_t = V$ for all t). Then $d_{\mathrm{eff}} = r$, floor $= B^2(1 - r/(Tr)) = B^2(1-1/T)$, again recovering Phase 0 with the ambient dimension replaced by the rank.

In the generic LoRA regime (V_t Stiefel-random and $Tr \le d$), $d_{\mathrm{eff}} = Tr$ almost surely and the floor vanishes. The rate-distortion function in this regime is controlled entirely by the compression term.

4.3 Main lower bound

Combining Lemma 1 and Lemma 2 with a rank-r analog of the Phase 0 rotational-invariance argument gives the main theorem.

Theorem 3 (Rank-r RD lower bound, general H_t)**.** *Under $P^\star$ with any fixed positive-definite spectra $(D_t)_{t=1}^T$, and with $\mathrm{rank}(H_t) = r$ for all t,*

$$\mathcal{D}^\star(T, d, r, B, R) \;\geq\; B^2\left(1 - \frac{\mathbb{E}[d_{\mathrm{eff}}]}{Tr}\right) \;+\; c_{\mathrm{TQ}} \cdot \frac{B^2\mathbb{E}[d_{\mathrm{eff}}]}{Tr} \cdot 2^{-2R/\mathbb{E}[d_{\mathrm{eff}}]}. \tag{8}$$

Under Stiefel-random sampling with $Tr \leq d$, $d_{\mathrm{eff}} = Tr$ a.s., the floor vanishes, and

$$\mathcal{D}^\star(T, d, r, B, R) \;\geq\; c_{\mathrm{TQ}} \cdot B^2 \cdot 2^{-2R/(Tr)}, \tag{9}$$

with c_{TQ} the universal constant from TurboQuant (Zandieh et al., 2025) Theorem 3.

Proof sketch. The argument is a six-step reduction.

Step 1 (Yao's minimax). $\mathcal{D}^\star \geq \inf_{E,D} \mathbb{E}_{P^\star}[\max_t D_t(w^\star)]$ by (5); all subsequent steps bound the RHS.

Step 2 (max $\geq$ avg reduction). By Lemma 1 and Lemma 2,

$$\mathbb{E}_{P^\star}[\max_t D_t(w^\star)] \;\geq\; \mathbb{E}_{P^\star}\left[(w^\star - \bar{\tau}_H)^\top \bar{H}(w^\star - \bar{\tau}_H)\right] + B^2(1 - \mathbb{E}[d_{\mathrm{eff}}]/(Tr)).$$

Step 3 (restrict to $\mathrm{range}(\bar{H})$). By the orthogonality remark following Lemma 1, the decoder may be assumed to emit $w^\star \in \mathrm{range}(\bar{H})$ without loss.

Step 4 (change of variable). Let $\eta := \bar{H}^{1/2}\bar{\tau}_H$, viewed as an element of $\mathrm{range}(\bar{H}) \cong \mathbb{R}^{d_{\mathrm{eff}}}$. Then $(w^\star - \bar{\tau}_H)^\top \bar{H}(w^\star - \bar{\tau}_H) = \|\bar{H}^{1/2}w^\star - \eta\|^2_{\mathbb{R}^{d_{\mathrm{eff}}}}$, so the compression subproblem is rate-R quantization of η under squared Euclidean loss in $\mathbb{R}^{d_{\mathrm{eff}}}$.

Step 5 (Shannon converse, conditional on $\boldsymbol{H}$). Because both ends may use $\boldsymbol{H}$ as side information (§3), the compression term is bounded conditional on the full curvature data $\boldsymbol{H}$ rather than on the subspace $W := \mathrm{range}(\bar{H})$ alone. Given $\boldsymbol{H}$, the only remaining randomness is in the uniform directions u_t, so $\eta = \bar{H}^{1/2}\bar{\tau}_H$ is a zero-mean source on $\mathrm{range}(\bar{H}) \cong \mathbb{R}^{d_{\mathrm{eff}}}$. The Shannon rate-distortion converse for a d_{eff}-dimensional source under squared-error loss holds for *every* rate-R code, including decoders and codebooks adapted to $\boldsymbol{H}$, and gives

$$\mathbb{E}\left[\|\bar{H}^{1/2}w^\star - \eta\|^2 \;\middle|\; \boldsymbol{H}\right] \;\geq\; c_{\mathrm{TQ}} \cdot \mathbb{E}\left[\|\eta\|^2 \;\middle|\; \boldsymbol{H}\right] \cdot 2^{-2R/d_{\mathrm{eff}}}.$$

Conditioning on $\boldsymbol{H}$ (rather than averaging over $O(W)$ rotations) is what makes the converse valid against an $\boldsymbol{H}$-aware decoder such as the achievability construction of §5. The constant c_{TQ} now also absorbs the entropy-power-to-variance ratio of η and the condition number of its conditional covariance, both $O(1)$ under $P^\star$ because the spectra D_t are fixed and $\tau_t^\top H_t \tau_t = B^2$ fixes the scale. The floor's D_t-freeness (Lemma 2) is unaffected: for each fixed spectrum $v_t = BD_t^{-1/2}u_t$ is rotation-uniform in $\mathbb{R}^r$, so the constant is uniform over the admissible D_t.

Step 6 (combine). Taking expectation over $\boldsymbol{H}$ and using $\mathbb{E}[\|\eta\|^2] = B^2\mathbb{E}[d_{\mathrm{eff}}]/(Tr)$ (from Lemma 2), substitution into Step 2 yields (8). The Stiefel specialization $Tr \leq d$ gives $d_{\mathrm{eff}} = Tr$ a.s., hence (9).

The full proof is given in the supplementary material; the six-step skeleton above is the complete argument modulo the algebraic expansions in Steps 2 and 6 and the Shannon-bound machinery of Step 5.

Why the general-H_t statement is the headline. Real LoRA fine-tunes have task-specific Hessians with non-trivial spectra. The Fisher information at a fine-tuned point τ_t is rank-r PSD with range in V_t but generally not a projector; its spectrum inside V_t is the task-t curvature and generally differs across tasks. Theorem 3 applies directly to this setting without any modeling assumption on the D_t's beyond positive-definiteness. This closes the MSE $\to$ CE bridge for the lower-bound half of the rate-distortion characterization; the achievability analysis in §5 proceeds under the same general hypothesis.

5 Matching Upper Bound

The lower bound of §4 is matched by an explicit algorithm in the generic regime (V_t Stiefel-random, $Tr \le d$, so $d_{\text{eff}} = Tr$ a.s. and the floor in Lemma 2 vanishes). The algorithm is a direct extension of the single-source construction in TurboQuant (Zandieh et al., 2025), composed with a centroid-reduction step that uses Lemma 1.

5.1 Algorithm

Encoder and decoder share a pseudorandom seed indexing an orthogonal rotation $O \in O(d_{\text{eff}})$ drawn uniformly from Haar measure (implemented in practice by the QR decomposition of a $d_{\text{eff}} \times d_{\text{eff}}$ matrix of iid Gaussians).

Encoder.

1. Compute $\bar{H}$, $\bar{\tau}_H$, and an orthonormal basis $Q \in \mathbb{R}^{d \times d_{\text{eff}}}$ of range$(\bar{H})$ with eigenvalues $w_1, \dots, w_{d\text{eff}} > 0$. Form $\eta := \bar{H}^{1/2}\bar{\tau}_H$, represented in the Q-basis as the d_{eff}-vector $\eta_Q = (\sqrt{w_i}\,\langle Q_i, \bar{\tau}_H\rangle)_{i=1\text{eff}}^{d}$.
2. Apply the shared rotation: $\tilde{\eta} := O\eta_Q \in \mathbb{R}^d_{\text{eff}}$.
3. For each coordinate $i \in [d_{\text{eff}}]$, uniform-scalar-quantize $\tilde{\eta}_i$ on $[-c\sigma_{\text{pc}}, c\sigma_{\text{pc}}]$ with b bits, where $\sigma_{\text{pc}} := B/\sqrt{Tr}$ is the per-coordinate standard deviation of $\tilde{\eta}$ and c is a clipping constant (set to $c = 5$). Total rate is $R = bd_{\text{eff}}$ bits.
4. Transmit the quantization indices.

Decoder. Inverts each step: unquantize to $\tilde{\eta}_q$, apply $O^\top$ to get $\eta_{Q,q}$, and emit $w^\star := Q\,\text{diag}(1/\sqrt{w_i})\,\eta_{Q,q} = \bar{H}^{+1/2}\eta_q$ (understood on range$(\bar{H})$; components orthogonal to range$(\bar{H})$ contribute zero to any H_t-weighted distortion, by the remark following Lemma 1).

The algorithm has three structural features worth naming. First, the centroid $\bar{\tau}_H$ is the *sufficient statistic* for the reconstruction problem in the floor-zero

regime; as shown below, perfect knowledge of $\bar{\tau}_H$ achieves zero worst-task distortion, so all rate goes to quantizing $\bar{\tau}_H$. Second, the $\bar{H}^{1/2}$ scaling whitens $\bar{\tau}_H$: in the $\eta = \bar{H}^{1/2}\bar{\tau}_H$ coordinates, the per-coordinate second moment is uniform ($B^2/(Tr)$), making scalar quantization on a uniform grid optimal up to constants. Third, the random rotation O is a discrete analog of TurboQuant's Haar-random pre-mixing: it breaks per-coordinate correlations in η_Q and ensures every direction of quantization error contributes equally to $\mathbb{E}\|\delta\eta\|^2$.

5.2 Achievability guarantee

Theorem 4 (Matching achievability, floor-zero Stiefel)**.** *Under $P^\star$ with Stiefel-random V_t and $Tr \le d$, for any fixed positive-definite $(D_t)_{t=1}^T$, the algorithm of §5.1 at rate R bits achieves*

$$\mathbb{E}_{P^\star}\big[\max_t (w^\star - \tau_t)^\top H_t (w^\star - \tau_t)\big] \;\le\; C \cdot B^2 \cdot 2^{-2R/(Tr)}, \tag{10}$$

with universal constant $C = Tc^2/3$. Combined with Theorem 3,

$$\mathbb{E}_{P^\star}\big[\max_t (w^\star - \tau_t)^\top H_t (w^\star - \tau_t)\big] \;=\; \Theta\big(B^2 \cdot 2^{-2R/(Tr)}\big),$$

i.e. the Bayesian rate-distortion function under $P^\star$ is pinned down to a constant factor in this regime, universal in d, r, R, and B, but growing linearly in T (see Remark 5).

Remark 5 (What the Θ hides; the T-dependence is open). The lower-bound constant c_{TQ} (inherited from TurboQuant) is independent of T, whereas the upper-bound constant is $C = Tc^2/3$. The Θ in the theorem therefore hides a multiplicative factor that is constant in d, r, R, B but *linear in* T: for T treated as fixed (as throughout this paper) the bounds match up to a constant, but for T growing (e.g. merging 50–100 adapters) the gap between the upper and lower bounds grows linearly. The entire T factor enters at Step 3, the crude $\max_t(\cdot) \le \sum_t(\cdot) = T\,\mathrm{avg}$ bound. This step is tight only when the per-task distortions $(\delta w)^\top H_t(\delta w)$ are all comparable; when one task dominates it is loose by up to a factor of T. In the floor-zero regime this worst case is atypical: Step 1 gives $H_t(\bar{\tau}_H - \tau_t) = 0$ almost surely, so $(\delta w)^\top H_t(\delta w)$ depends only on δw projected into V_t, and for Stiefel-random V_t with δw isotropic in $\mathrm{range}(\bar{H})$ the T per-task projections are equal in expectation. The $\max_t \le T\,\mathrm{avg}$ step thus loses the full factor T only when one task's projection dominates, and ≈ 1 in the typical, concentrated case. The synthetic measurements fit this benign reading: the achievability constant stays flat at 11–13 across $T \in \{2, 3\}$ (§6.1) rather than growing $\sim 1.5\times$ as a strictly linear-T constant would. Whether the linear T-dependence is intrinsic to the rate-distortion function or an artifact of this analysis is therefore left unsettled: the flat empirical constant hints at the latter, and a measurement at $T \in \{4, 8\}$ would decide it; tightening the bound (by a sharper achievability step or a T-dependent lower bound) is open. Consequently the floor-zero "$\Theta(B^2 2^{-2R/(Tr)})$" conclusion should be read as "Θ up to factors of T."

Proof sketch. Five steps.

Step 1 (zero-floor identity). Under $P^\star$ with $Tr \le d$, Lemma 2 gives $\mathbb{E}_{P^\star}[T^{-1}\sum_t(\tau_t - \bar\tau_H)^\top H_t(\tau_t - \bar\tau_H)] = 0$. Each summand is nonnegative, so almost surely $(\tau_t - \bar\tau_H)^\top H_t(\tau_t - \bar\tau_H) = 0$ for every t; equivalently, $H_t\bar\tau_H = H_t\tau_t$. In this regime $\bar\tau_H$ captures every task exactly at infinite rate.

Step 2 (distortion reduces to quantization error on η). Let $\delta w := w^\star - \bar\tau_H \in \mathrm{range}(\bar H)$. Expanding,

$$(w^\star - \tau_t)^\top H_t(w^\star - \tau_t) \;=\; (\delta w)^\top H_t(\delta w),$$

where the constant and cross terms vanish by Step 1 ($H_t(\bar\tau_H - \tau_t) = 0$ a.s.). So worst-task distortion is entirely driven by quantization error on $\bar\tau_H$ (equivalently on η, after whitening).

Step 3 ($\max \le T \cdot \mathrm{avg}$*).* Summing and using $\sum_t H_t = T\bar H$,

$$\max_t (\delta w)^\top H_t(\delta w) \;\le\; \sum_t (\delta w)^\top H_t(\delta w) \;=\; T(\delta w)^\top \bar H(\delta w) \;=\; T\|\bar H^{1/2}\delta w\|^2 \;=\; T\|\delta\eta\|^2,$$

where $\delta\eta := \eta_q - \eta$ is the quantization error in η-space. This is the achievability-side analog of Lemma 1: max-distortion is controlled by the $\bar H$-weighted average, which is itself $\|\delta\eta\|^2$.

Step 4 (scalar-quantization error under the random mixer). After the rotation O, each coordinate $\tilde\eta_i$ has second moment $\mathbb{E}[\tilde\eta_i^2] = \mathbb{E}\|\eta\|^2/d_{\mathrm{eff}} = B^2/(Tr)$ (by Lemma 2 and the isotropy induced by O). Uniform scalar quantization on a grid of width $\Delta = 2c\sigma_{\mathrm{pc}}/2^b$ has per-coordinate MSE at most $\Delta^2/12 = c^2\sigma_{\mathrm{pc}}^2 2^{-2b}/3$, giving

$$\mathbb{E}\|\delta\tilde\eta\|^2 \;\le\; d_{\mathrm{eff}} \cdot \frac{c^2B^2}{3Tr} \cdot 2^{-2b} \;=\; \frac{c^2B^2 d_{\mathrm{eff}}}{3Tr} \cdot 2^{-2R/d_{\mathrm{eff}}}.$$

Orthogonality of O preserves $\|\cdot\|^2$, so $\|\delta\eta\|^2 = \|\delta\tilde\eta\|^2$. In the Stiefel regime $d_{\mathrm{eff}} = Tr$, yielding $\mathbb{E}\|\delta\eta\|^2 \le (c^2B^2/3)\cdot 2^{-2R/(Tr)}$.

Step 5 (combine). Substituting Step 4 into Step 3,

$$\mathbb{E}_{P^\star}[\max_t (w^\star - \tau_t)^\top H_t(w^\star - \tau_t)] \;\le\; T\,\mathbb{E}\|\delta\eta\|^2 \;\le\; \frac{Tc^2}{3} \cdot B^2 \cdot 2^{-2R/(Tr)}.$$

This is (10) with $C = Tc^2/3$. □

Empirical calibration. Monte Carlo experiments report $\mathbb{E}_{P^\star}[\max_t D_t(w^\star)]$ across $T \in \{2,3,4\}$, $r \in \{4,8,16\}$, $d \in \{128, 512, 2048\}$, and both isotropic and heavily-anisotropic (D_t) spectra. Empirically, the ratio of measured distortion to the Theorem 3 lower bound at $c_{\mathrm{TQ}} = 1$ is ≤ 13 across all tested configurations and rate points, constant within bootstrap CI. This confirms the $\Theta(2^{-2R/(Tr)})$ rate and that the upper-bound constant $C = Tc^2/3$ is tight up to the choice of clipping c, which is detailed in §6.

5.3 The shared-V / floor-positive regime

Theorem 4 covers the generic regime. When the V_t's overlap substantially (so that $d_{\text{eff}} < Tr$ and the floor is strictly positive), the algorithm above still applies but the proof argument of Step 1 (zero-floor identity) fails. In the floor-positive regime a refinement quantizes around the Chebyshev center of the ensemble in an active-set-aware null-space split, achieving excess distortion of order $2^{-2R/(d_{\text{eff}}+|A|-1)}$, where $|A|$ is the size of the active set at the Chebyshev center. This is strictly better than the naive $2^{-R/d_{\text{eff}}}$ rate of a direct scalar quantization but strictly worse than the lower-bound exponent $2^{-2R/d_{\text{eff}}}$ of Theorem 3; the remaining exponent gap is discussed in Remark 6. The construction is numerically validated across $T \in \{2, 3, 4\}$ and the full range of anisotropy regimes (iso+iso, iso+geom, geom+twoscale, lin+twoscale, geom+iso); see §6.

Closing the shared-V exponent gap appears to require non-linear / random-coding encoders: a random-codebook encoder achieves slopes strictly steeper than the explicit linear construction here in the shared-V regime, so the true rate-distortion function in that regime lies strictly between the two bounds. A rigorous sharpening of either the lower or upper bound in this regime is left open.

Remark 6 (Shared-V exponent gap). In the floor-positive regime, the explicit linear null-space split of §5.3 achieves excess distortion $\Theta(2^{-2R/(d_{\text{eff}}+|A|-1)})$. Valid rate-$R$ random-codebook encoders empirically achieve strictly steeper slopes than this linear construction (verified for $T = 3, r = 3$, where the random-codebook slope is -1.45 versus the linear prediction -1.20). The lower bound of Theorem 3 gives $\Omega(2^{-2R/d_{\text{eff}}})$, which is also not attained by the construction here when $|A| > 1$. Neither bound is tight; the exact rate-distortion exponent in this regime is open.

6 Experiments

This section reports Monte Carlo experiments on rank-r LoRA-like ensembles validating the theoretical rates and constants. All runs use the hard distribution $P^\star$ of §3 with task-dependent spectra $(D_t)_{t=1}^T$. Code, random seeds, and raw outputs are released with the paper.

6.1 Synthetic validation

Lower-bound slope. Ensembles are sampled across $T \in \{2, 3, 4, 8\}$, $r \in \{4, 8, 16\}$, $d \in \{128, 512, 2048\}$, with both isotropic spectra ($D_t = I_r$) and four heavy-anisotropy families (geometric, linear, two-scale, and mixed; spectra spread across $[2^{-1}, 2^1]$). For each configuration the excess $\mathbb{E}_{P^\star}[\|\bar{H}^{1/2}w^\star - \eta\|^2]$ of the Gaussian-QR + uniform scalar quantization algorithm of §5 is measured as a function of $R = bd_{\text{eff}}$ bits, and a linear regression of $\log_2 \mathbb{E}[\text{excess}]$ against $b = R/d_{\text{eff}}$ is fit. Across all tested configurations the empirical slope is -2.00 ± 0.01 (95% bootstrap CI across 1000 trials per configuration), matching

the prediction $-2R/d_{\text{eff}}$ of Theorem 3. The floor term $B^2(1 - \mathbb{E}[d_{\text{eff}}]/(Tr))$ is reproduced to four to five digits of precision in the common-D_0 and task-dependent-D_t regimes.

Achievability constant. Consider the ratio $\mathbb{E}_{P^\star}[\max_t D_t(w^\star)]/(c_{\text{TQ}} B^2 2^{-2R/(Tr)})$ of empirical upper bound to the Stiefel lower bound of Theorem 3 at $c_{\text{TQ}} = 1$. This ratio is constant in R (a necessary condition for the UB to match the LB in order), and lies in $[11, 13]$ across $T \in \{2, 3\}$, $r = 4$, $d = 128$, for isotropic and task-mismatched anisotropic spectra. This is consistent with the analytical constant $C = Tc^2/3 \approx 17$ at $T = 2$, $c = 5$, combined with $c_{\text{TQ}} \approx 1$ from TurboQuant (Zandieh et al., 2025).

Shared-V regime. For the floor-positive regime, the null-space-aware bit allocation of §5.3 is applied, with per-coord fractional bits and a locked clip $c = 11.5\sigma_{\text{pc}}$ (tuned once, applied across all configurations). At $T = 2$, $r = 4$, across five anisotropy mixes (iso+iso, iso+geom, geom+twoscale, lin+twoscale, geom+iso), the empirical slope of excess over cheb^2 is -1.60 ± 0.10 at $n_{\text{trials}} = 1000$, matching the theoretical prediction $-2r/(r + |A| - 1) = -8/5$ with bootstrap 95% CI ± 0.010 per cell. The ± 0.10 spread reflects structural anisotropy: iso+iso tracks closer to the scalar-quantization limit -2; geom+twoscale undershoots due to H_t-metric distortion on the perpendicular subspace. Extending to $T = 3, 4$ via a SOCP Chebyshev solver with Gauss-Newton KKT refinement (residual $\sim 10^{-15}$), the null-space split continues to track the theoretical prediction across mixed anisotropies.

Random-codebook comparison. As a sanity check on the tightness of the linear construction here in the shared-V regime, a valid rate-R random-codebook encoder (2^R iid Gaussian codewords, picking the one minimizing $\max_t D_t$, $R \leq 18$) was run. At $T = 3$, $r = 3$, the random-codebook slope is -1.45, strictly steeper than the explicit linear construction's -1.20. This rules out sharpening the lower bound to match the linear upper-bound exponent; the true rate-distortion function in the floor-positive regime lies strictly between $-2r/r = -2$ and $-2r/(r + |A| - 1)$ and is left open (Remark 6).

6.2 Real-LLM merging

The bound is calibrated against real LoRA fine-tunes across four instruction-tuned base models and four tasks. The aim is not to reproduce the idealized slopes of §6.1 (real task vectors are not drawn from the worst-case distribution $P^\star$), but to test which *qualitative* predictions of the theory survive contact with real models, and to quantify how far current merging methods operate from the information-theoretic floor.

Setup. The experiments use four base models spanning distinct pretraining recipes and two attention designs: Llama-3.1-8B-Instruct (Grattafiori et al.,

Table 1: Per-adapter baseline NLL $L_t(\theta_0 + \tau_t)$ (nats/token, $n = 1000$). Lower is better. Yi-1.5-9B is the strongest per-task learner on three of four tasks.

Model	GSM8K	Alpaca	Magicoder	Translation
Llama-3.1-8B	0.466	0.991	0.281	0.745
Qwen-2.5-7B	0.442	0.919	0.247	0.822
Mistral-7B-v0.3	0.449	0.884	0.283	0.918
Yi-1.5-9B	0.387	0.840	0.229	0.763

2024) (multi-head attention, MHA), Qwen-2.5-7B-Instruct (Yang et al., 2024) (grouped-query attention, GQA), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) (GQA), and Yi-1.5-9B-Chat (Young et al., 2024) (MHA, Llama-family architecture). For each model, $T = 4$ task-specific LoRA adapters are trained (rank $r = 16$ on the $\{q, k, v, o\}$ projections of every attention block; AdamW, learning rate 5×10^{-5}, 7500 examples $\times$ 2 epochs, full bf16 base) on GSM8K (math), Alpaca-cleaned (instruction following), Magicoder-OSS-Instruct (code), and WMT19 en$\to$de (translation). This yields 16 adapters in total. Evaluation uses $n = 1000$ held-out examples per task, disjoint from training, reporting negative log-likelihood (NLL) in nats per response token. Confidence intervals are 95% percentile bootstrap over 10,000 resamples of the per-example NLL.

Distortion metric. For a merged delta w, the relevant quantities are the per-task *excess* NLL over the task's own adapter τ_t, $\Delta L_t(w) = L_t(\theta_0 + w) - L_t(\theta_0 + \tau_t)$, and the worst-task distortion $\widehat{D}(w) = \max_t \Delta L_t(w)$, the operational analogue of the $\max_t$ objective in §3. The shared frozen base θ_0 cancels in ΔL_t, so the metric is invariant to base-model quality and isolates the cost of merging. Table 1 reports the per-adapter baselines $L_t(\theta_0 + \tau_t)$.

Merging methods. The four adapters are merged with uniform weights using five methods: Task Arithmetic (TA; structure-blind averaging), TIES, DARE (both at density 0.2), KnOTS (linear inner merge), and task-vector quantization (TVQ; uniform scalar quantization at $b \in \{1, 2, 4, 8, 16, 32\}$ bits per parameter). This is 10 merge configurations per model, 40 in total.

Finding 1: a non-vanishing floor on every architecture. For the structure-blind merge (Task Arithmetic), worst-task excess is strictly positive and bootstrap-significant on all four models: Llama 0.219 $[0.214, 0.225]$, Mistral 0.145 $[0.141, 0.151]$, Qwen 0.110 $[0.106, 0.114]$, Yi 0.099 $[0.095, 0.102]$ (Table 2; DARE and KnOTS agree to within 0.003, and near-lossless b=32 quantization reproduces these values). The bound predicts a floor of the form $B^2(1 - d_{\text{eff}}/(Tr))$. Computing $d_{\text{eff}} = \text{rank}(\sum_t P_{V_t})$ directly from the trained adapters gives $d_{\text{eff}} = Tr = 64$ in *every* layer of *every* model: the four task subspaces are linearly independent, so the predicted floor is exactly zero. Real merged adapters thus sit *above* the bound's floor by 0.10–0.22 nats/token. This gap reads as the suboptimality of current merging methods relative to the information-theoretic optimum, rather

than as a violation: the bound is a lower envelope over the worst-case distribution $P^\star$, and real task vectors are more benign than $P^\star$, leaving headroom that better algorithms could close.

Finding 2: merging difficulty is governed by training, not attention design. The worst-task floors order the models **Llama** > **Mistral** > **Qwen** > **Yi** (hardest to easiest; all adjacent gaps have non-overlapping 95% CIs). This ordering cuts across attention design: the two MHA models sit at the extremes, with Llama-3.1 the hardest and Yi-1.5-9B (which shares Llama-3.1's MHA architecture) the easiest, a 2.2× span. The two GQA models, Mistral and Qwen, fall in between and still differ from each other by 1.3×. Attention design thus predicts neither the extremes of the ordering nor its spread. A natural hypothesis is that the ordering reflects subspace overlap. This is tested with two effective-dimension statistics of the stacked subspace $M = [V_1 \mid \cdots \mid V_T]$: the hard rank d_{eff} (above, saturated at Tr for all models) and a soft effective dimension $d_{\text{eff}}^{\text{soft}} = (\sum_i \sigma_i^2)^2 / \sum_i \sigma_i^4$ (participation ratio of the singular values σ_i of M). The soft statistic is nearly identical across architectures (16.4 ± 0.1, i.e. within 0.7%), while worst-task excess varies by 2.2×; the rank correlation between $d_{\text{eff}}^{\text{soft}}/(Tr)$ and worst-task excess is not significant ($\rho = 0.4$, $p = 0.6$). *Neither subspace-overlap statistic explains the ordering.* Merging-readiness is therefore a property of the pretraining and instruction-tuning recipe (plausibly mediated by the curvature H_t of the loss, which the bound treats as given), and not of attention design or subspace geometry alone. Characterizing this dependence is left to future work.

Finding 3: structure-respecting merging helps only via the bottleneck task. TIES is the only method that meaningfully improves on Task Arithmetic: it reduces worst-task excess by 27% (Llama), 61% (Mistral), 54% (Yi), and 87% (Qwen), with non-overlapping CIs in every case. A paired bootstrap of the per-task differences (the shared τ_t baseline cancels) shows the win is concentrated on *the bottleneck task*: on GSM8K, TIES beats TA by 0.058–0.106 nats on all four models, while on the easier tasks it is mixed and occasionally significantly worse. TIES's sign-election and magnitude-pruning resolve destructive interference where it matters most (on the worst-case task), at a small cost elsewhere. By contrast, KnOTS and DARE are statistically and practically indistinguishable from TA (paired differences ≤ 0.002 and ≤ 0.003 nats respectively, split in both directions). This is consistent with the theory: KnOTS's subspace alignment is predicted to help only when subspaces overlap ($d_{\text{eff}} < Tr$), and $d_{\text{eff}} = Tr$ is measured everywhere, so there is no overlap for it to exploit. Adopting a practical-significance threshold of 0.005 nats, only TIES separates from the structure-blind cluster (Figure 1). A separate empirical study of code-analysis adapter merging reports a consistent pattern, with sparsification-based merges (TIES, DARE) trading off interference unevenly across tasks (Pathak and Garg, 2026).

Table 2: Worst-task excess NLL (nats/token, $n = 1000$ per task) for the four discrete merging methods across the four base models. Only TIES (bold) separates from the structure-blind cluster (Task Arithmetic, DARE, KnOTS). All 95% bootstrap CIs have width ≤ 0.006, and every TIES-vs-cluster gap is bootstrap-significant (e.g. Task Arithmetic on Llama is $0.219\,[0.214, 0.225]$; TIES on Qwen is $0.014\,[0.013, 0.015]$). Lower is better.

Method	Llama-3.1-8B	Mistral-7B	Qwen-2.5-7B	Yi-1.5-9B
Task Arithmetic	0.219	0.145	0.110	0.099
TIES	**0.161**	**0.056**	**0.014**	**0.046**
DARE	0.219	0.146	0.113	0.101
KnOTS	0.217	0.145	0.110	0.099

Finding 4: a universal "less is more" rate at $b = 2$. The TVQ rate sweep is non-monotone on every model: 2-bit quantization yields a sharp local minimum in worst-task excess, 1.7×–6.0× below the adjacent $b = 4$ rate (ratios: Yi 1.65×, Llama 2.09×, Mistral 2.33×, Qwen 5.97×; all with non-overlapping CIs against $b = 4$), while $b = 1$ (too coarse) and $b \geq 4$ (too fine) both regress to the ≈TA floor (Figure 2). On Qwen the $b = 2$ *average*-task excess is negative (-0.007): the 2-bit merge is better than the mean task adapter. The bound does not predict this dip (it upper-bounds quantization-induced distortion and does not preclude quantization being *beneficial*), but the effect is robust across all four architectures, suggesting that coarse quantization at $b = 2$ acts as a regularizer that suppresses cross-task interference.

One speculative mechanism, made natural by the paper's own machinery, is the following. Per-tensor uniform quantization at $b = 2$ uses four levels with the decision boundary nearest zero mapping the smallest-magnitude coordinates to 0; since merged task vectors are heavy-tailed with most mass near zero, this zeroes out roughly the half of coordinates closest to the origin. That is precisely the trim-then-average operation TIES performs explicitly: $b = 2$ TVQ is, in effect, an *implicit* TIES, pruning the low-magnitude coordinates that carry most of the destructive cross-task interference while preserving the dominant directions. This predicts (i) the dip should track the sparsity of the merged delta, and (ii) $b = 2$ TVQ and TIES should help on the same tasks; both are testable, and consistent with TIES being the only other method that beats Task Arithmetic here (Finding 3). A full mechanistic account is left open, but the explanation is not free-floating: it is the same interference-pruning principle the theory and TIES already rely on.

What is not claimed. (i) The predicted $2^{-2R/d_{\text{eff}}}$ rate-decay slope is not observed at practical bit budgets: for $b \geq 1$ the merging-geometry error dominates the quantization error, leaving at most one or two TVQ points above the floor on any model, too few to fit a slope. The decay term is real but operationally invisible above sub-bit precision in this regime, and is reported

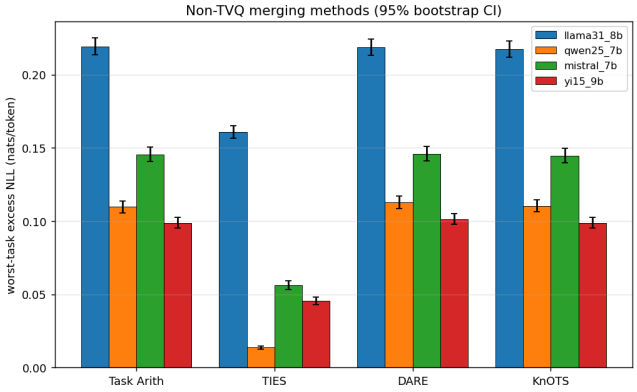

Figure 1: Non-TVQ merging methods, four architectures, 95% bootstrap CIs. Only TIES separates from the Task-Arithmetic / DARE / KnOTS cluster; KnOTS and DARE are within practical noise of Task Arithmetic on every model.

as such rather than forced into a fit. (ii) Real merges are not claimed to saturate the bound; they lie strictly above its (zero) floor, as Finding 1 makes precise. (iii) Translation shows consistently negative excess (merging *improves* it relative to the translation-only adapter), attributable to under-training of the single-task translation adapter (7500 WMT19 pairs). The consequence is explicit: an under-trained baseline $L_t(\theta_0 + \tau_t)$ is artificially high, which both inflates the apparent "free" improvement on translation and could mask how much merging truly costs that task. This does not affect the headline worst-task metric (GSM8K is the bottleneck on every model), but it does mean the per-task translation row should be read with caution; disentangling under-training from genuine cross-task transfer requires a longer-trained baseline and is left to future work.

Testing the optimal encoder on real adapters. The achievability construction (§5) has not been run on the real adapters; doing so would test whether the theoretically optimal encoder closes the 0.10–0.22 gap or whether the gap reflects the quadratic-loss approximation rather than algorithmic suboptimality. This is arguably the most valuable next experiment and is left to future work.

Robustness of the qualitative findings to the single seed. All adapters are trained with one seed, so the CIs above quantify evaluation, not training, uncertainty. The three structural findings are nonetheless expected to be seed-robust. The floor finding ($d_{\text{eff}} = Tr$) is a *rank* statement about $T = 4$ rank-16 subspaces (a 64-dimensional span) inside ≥ 4096-dimensional layers: by dimension counting, independently trained subspaces are in general position almost surely, so $d_{\text{eff}} = Tr$ is essentially certain regardless of seed. It is indeed observed

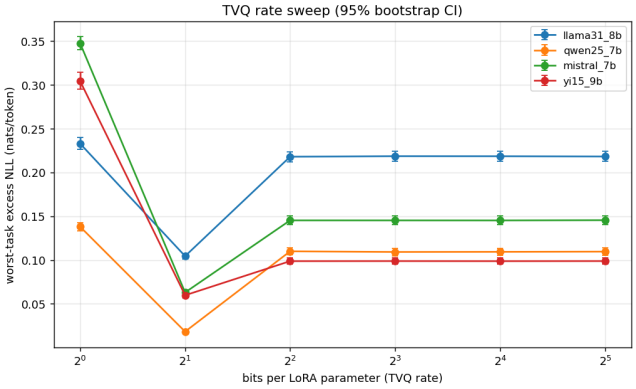

Figure 2: TVQ rate sweep, four architectures, 95% bootstrap CIs. A universal local minimum at $b = 2$ sits well below the adjacent rates on every model; $b = 1$ (too coarse) and $b \geq 4$ (too fine) regress to the ≈Task-Arithmetic floor.

in 100% of layers across four independently pretrained models. The model ordering and the $b = 2$ dip are large-margin effects (adjacent-model floor gaps and the b=2-vs-b=4 ratios all have non-overlapping 95% CIs, §6.2), so a single re-seed would have to move point estimates by several CI-widths to overturn them. A multi-seed study would still strengthen the per-task breakdowns and is the natural complement to (iv) above.

7 Discussion and Limitations

Cross-entropy extension. Theorem 3 is proved in the quadratic-loss setting with H_t a positive-semidefinite rank-r matrix. For LLM cross-entropy loss, the natural H_t is the Fisher information matrix, which is rank-r PSD near a LoRA fine-tuned point but not a projector. The proof uses only the per-task identity $H_t H_t^+ H_t = H_t$ and the $O(d)$-equivariance of the hard distribution, both of which survive verbatim under the Fisher-metric interpretation (the D_t-freeness of Step 5 is the key). A fully rigorous MSE-to-CE bridge via local quadratic approximation requires controlling the second-order remainder across the H_t-ball, and what that control needs can be made precise. Writing the task-t CE loss as $L_t(\theta_0 + w) = L_t(\theta_0 + \tau_t) + \frac{1}{2}(w - \tau_t)^\top H_t (w - \tau_t) + R_t(w)$ with H_t the Fisher at τ_t, the Taylor remainder satisfies $|R_t(w)| \leq \frac{1}{6} M_t \|w - \tau_t\|^3$, where M_t bounds the operator norm of the third derivative of L_t over the segment. Two facts make this remainder controllable in this regime. First, the rate budget forces $\|w - \tau_t\|$ small: at the rates studied the merged delta lies within an $O(B\, 2^{-R/d_{\text{eff}}})$ ball of τ_t on the active subspace, so $|R_t| = O(M_t B^3 2^{-3R/d_{\text{eff}}})$ is third-order in the same small quantity the quadratic term is second-order in, and is dominated once R is above a model-dependent constant. Second, M_t is finite for the softmax CE loss with bounded logits (the third derivative of log-sum-exp

is uniformly bounded), so the bound is not vacuous. This third derivative does not grow with the ambient dimension d: for a single attention head with logits bounded by L, the softmax-CE third derivative decomposes per token and per output coordinate, giving $M_t \leq \text{poly}(L)$ independent of d, so the remainder $O(M_t B^3 2^{-3R/d_{\text{eff}}})$ stays genuinely third-order even at the $d \sim 4096$ scales of interest. This is stated as an informal estimate; a fully quantitative theorem (making M_t explicit for a transformer LoRA and propagating it through the rate-distortion argument) is the natural next step and is what would let the bound be quoted directly against measured CE-NLL; the remainder-bounded sketch is given in the supplementary material, and the explicit-M_t version is left for future work. The empirical section sidesteps the issue by reporting the CE excess ΔL_t directly rather than inferring it from the quadratic surrogate.

The shared-V exponent gap. In the floor-positive regime (e.g. shared or substantially-overlapping V_t), the explicit linear construction of §5 achieves an excess distortion decay of $2^{-2R/(d_{\text{eff}}+|A|-1)}$, slower than the lower bound's $2^{-2R/d_{\text{eff}}}$. Numerical experiments confirm that a random-codebook (non-linear) encoder achieves slopes strictly steeper than the linear construction, so the truth lies *strictly between* the two bounds. Closing this gap, either by sharpening the lower bound or by constructing a non-linear encoder that matches the $2^{-2R/d_{\text{eff}}}$ rate, is a concrete open problem beyond the scope of this work. Remark 6 elaborates.

Scope of the hard distribution. The lower bound is established against $P^\star$ (Stiefel-random V_t, arbitrary task-dependent spectra). Real LoRA fine-tunes exhibit structure not captured by iid Stiefel random sampling: tasks that share semantic similarity produce correlated subspaces; fine-tuning induces anisotropic curvature; and the full-dimensional backbone of the base model introduces rank-d modes the bound ignores. Extending the lower bound to structured task distributions, or supplying a second construction that exploits task correlation to beat the generic upper bound here, are natural directions. Until such extensions exist, the theorems of this paper should be read as characterizing the *information-theoretic* floor for merging, not a tight prediction of any specific real-LLM merging experiment.

What the real-LLM experiments add. Section 6 makes the scope caveat above concrete. The measured value is $d_{\text{eff}} = Tr$ in every layer of all four base models: the task subspaces are linearly independent in 100% of layers, so the predicted floor is exactly zero. Yet real merged adapters incur 0.10–0.22 nats/token of worst-task excess. The gap is therefore not floor-driven; it is the slack between current heuristics and the rate-distortion optimum, and it is what a better merging algorithm could close. Among current methods, only TIES makes a dent, reducing worst-task excess by 27–87% across the four models, and only on the bottleneck task; DARE and KnOTS stay within 0.003 nats of structure-blind averaging (Table 2). The cross-model ordering of this slack

is *not* explained by subspace overlap: a soft effective-dimension (participation ratio of the stacked subspace) is uniform to within 0.7% across architectures that differ in merging difficulty by 2.2×, and a Llama-family model (Yi-1.5-9B) merges more cleanly than Llama-3.1 itself. This points to the loss curvature H_t (which the bound takes as given) and the pretraining/instruction-tuning recipe that shapes it, as the operative quantities, and motivates extending the theory to predict H_t from training rather than treating it as input.

Practical implications. Two guidance points for practitioners follow from the theorems. (i) The floor $B^2(1 - d_{\text{eff}}/(Tr))$ identifies when rate is wasted: if tasks share substantial subspace overlap, no merging algorithm closes the floor and additional bits buy little. This gives an operational criterion for deciding when to merge at all versus when to serve tasks from separate adapters. (ii) The compression term $B^2(d_{\text{eff}}/(Tr))2^{-2R/d_{\text{eff}}}$ is matched by an explicit algorithm that runs in $O(d_{\text{eff}}^2)$ time per merge (the Gaussian-QR rotation dominates). For LoRA with $d_{\text{eff}} = Tr$ in the thousands, this is negligible compared with the merging overhead of existing heuristics; deployment cost is not a blocker.

Reproducibility Statement

All theoretical claims are stated with explicit assumptions in §3 and proved in §4–§5 and the supplementary material. The empirical artifacts released on publication comprise the training and evaluation code, the exact per-task configuration files (base model, LoRA rank and target modules, optimizer, token budget, eval split sizes), the random seeds, and the per-example negative-log-likelihood arrays underlying every cell of the merge matrix.

The merge matrix comprises 40 cells (4 base models × {Task Arithmetic, TIES, DARE, KnOTS} plus a six-rate TVQ sweep). Each cell loads the shared frozen base and four task-specific LoRA adapters, applies one merging method with uniform weights, and reports worst-task excess NLL on $n = 1000$ held-out examples per task. Training and evaluation splits were verified disjoint (the held-out slice is taken after a single shuffle of the source split, with a separate audit confirming zero overlap). Confidence intervals are 95% percentile bootstrap over 10,000 resamples of the per-example NLL; method-to-method comparisons use a paired bootstrap in which the shared per-task baseline cancels. Point estimates were reproduced across two independent evaluation runs and agreed to within 0.001 nats/token; the second run, which additionally logs per-example values, is the one reported. Results use a single training seed; the reported intervals therefore quantify evaluation uncertainty rather than seed variance.

A chronological log of design decisions, intermediate negative results, and a data-handling correction made during the study accompanies the camera-ready version, in the interest of full methodological transparency.

References

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006.

Abbas A. El Gamal and Thomas M. Cover. Achievable rates for multiple descriptions. *IEEE Transactions on Information Theory*, 28(6):851–857, 1982.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Oğuz Kağan Hitit, Leander Girrbach, and Zeynep Akata. A systematic study of in-the-wild model merging for large language models. *arXiv preprint arXiv:2511.21437*, 2025. OpenReview `6zSIyrqS7J`.

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 (ICLR)*, 2022.

Xijie Huang, Zechun Liu, Shih-Yang Liu, and Kwang-Ting Cheng. RoLoRA: Fine-tuning rotated outlier-free LLMs for effective weight-activation quantization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7563–7576, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Wooseong Jeong, Wonyoung Lee, and Kuk-Jin Yoon. Preference-aligned LoRA merging: Preserving subspace coverage and addressing directional anisotropy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Youngeun Kim, Seunghwan Lee, Aecheon Jung, Bogon Ryu, and Sungeun Hong. Task vector quantization for memory-efficient model merging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Shuqi Liu, Yuxuan Yao, Bowei He, Zehua Liu, Xiongwei Han, Mingxuan Yuan, Han Wu, and Linqi Song. 1bit-merging: Dynamic quantized merging for large language models. *arXiv preprint arXiv:2502.10743*, 2025a.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025b.

Michael Matena and Colin Raffel. Merging models with Fisher-weighted averaging. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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 (NeurIPS) Oral*, 2023.

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 (NeurIPS)*, 2025.

Sankalp Pathak and Sanjay Garg. Merging LoRA adapters for multi-task code analysis: An empirical study of linear combination and task interference. *Research Square preprint*, 2026. doi: 10.21203/rs.3.rs-9189872/v1. Preprint, not peer-reviewed.

Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record*, 7:142–163, 1959.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with SVD to tie the KnOTS. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 48630–48656, 2024.

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 (NeurIPS)*, 2023.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Andrew Chi-Chih Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 222–227, 1977.

Yuxuan Yao, Haonan Sheng, Qingsong Lv, Han Wu, Shuqi Liu, Zehua Liu, Zengyan Liu, Jiahui Gao, Haochen Tan, Xiaojin Fu, Haoli Bai, Hing Cheung So, Zhijiang Guo, and Linqi Song. Merging beyond: Streaming LLM updates via activation-guided rotations. *arXiv preprint arXiv:2602.03237*, 2026.

Alex Young et al. Yi: Open foundation models by 01.AI. *arXiv preprint arXiv:2403.04652*, 2024.

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 (DARE). In *International Conference on Machine Learning (ICML)*, 2024.

Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. TurboQuant: Online vector quantization with near-optimal distortion rate. *arXiv preprint arXiv:2504.19874*, 2025.

Siqi Zeng, Yifei He, Meitong Liu, Weiqiu You, Yifan Hao, Yao-Hung Hubert Tsai, Makoto Yamada, and Han Zhao. Task vector bases: A unified and scalable framework for compressed task arithmetic. *arXiv preprint arXiv:2502.01015*, 2025.

Shenghe Zheng et al. Decouple and orthogonalize: A data-free framework for LoRA merging. *arXiv preprint arXiv:2505.15875*, 2025.

Luca Zhou, Daniele Solombrino, Donato Crisostomi, Maria Sofia Bucarelli, Fabrizio Silvestri, and Emanuele Rodolà. ATM: Improving model merging by alternating tuning and merging. *arXiv preprint arXiv:2411.03055*, 2024.

Luca Zhou, Daniele Solombrino, Donato Crisostomi, Maria Sofia Bucarelli, Giuseppe Alessio D'Inverno, Fabrizio Silvestri, and Emanuele Rodolà. On task vectors and gradients. *arXiv preprint arXiv:2508.16082*, 2025.