SEARCHING THE QUANTIZED WEIGHT SPACE FOR ADAPTATION: TASK-AWARE JOINT QUANTIZATION AND LOW-RANK INITIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantized low-rank fine-tuning enables memory-efficient adaptation of large language models by training low-rank adapters on a frozen low-bit backbone. Since the backbone remains fixed during fine-tuning, the quantized weights selected at initialization define the starting point from which the adapter must learn. Existing methods mainly select quantized weights to reconstruct the pretrained model's weights or activations. However, reconstruction quality before fine-tuning does not necessarily identify the best starting point for adaptation. We therefore view initialization as a search for quantized weights that perform well after adaptation. However, this poses a major challenge: the adapted model is not yet available during the calculation of the quantized initialization. To address this challenge, we propose *Task-Aware Quantized Initialization* (TAQI), which uses target-task output gradients measured before fine-tuning to guide this search. TAQI weights layer-wise reconstruction errors by gradient energy and mixes the resulting task-sensitive curvature with general calibration curvature. We show that the task-weighted error bounds the energy of local first-order task-loss perturbations. Using IBM Granite-4.0-H-350M at 4 bits, TAQI improves the primary metric by 0.45–1.33 points over GPTQ-intrinsic across medical, financial, and scientific question-answering benchmarks. These results highlight the promise of adaptation-oriented quantized initialization for quantized low-rank fine-tuning.

1 INTRODUCTION

Quantized low-rank fine-tuning enables memory-efficient adaptation of large language models by training high-precision low-rank adapters on a frozen low-bit backbone (Hu et al., 2022; Dettmers et al., 2023). Freezing the backbone reduces the memory required for model parameters, gradients, and optimizer states, but also makes its initialization consequential. The selected quantized weights define the fixed base for adaptation and, together with the adapter rank and training budget, shape the set of models attainable through fine-tuning. Quantization in this setting is therefore not only a reconstruction problem, but also an initialization-design problem for subsequent adaptation.

Most existing methods select quantized weights by preserving the pretrained model's weights or activations. Joint initialization methods such as LQ-LoRA and GPTQ-intrinsic additionally construct a low-rank component to compensate for quantization error (Guo et al., 2024; Zhang & Saab, 2026), while sensitivity-aware methods such as GuidedQuant use end-loss gradients to prioritize important reconstruction errors (Kim et al., 2025). Despite these differences, candidate initializations are still primarily evaluated by their fidelity to the pretrained model before adaptation. The practical objective, however, is the task performance attained after fine-tuning. A starting point that best reconstructs the pretrained model need not be the one that produces the best adapted model under a fixed training budget.

Figure 1 provides motivating evidence for this distinction. Under matched within-task training budgets, 4-bit+LoRA outperforms BF16+LoRA on eight of 14 benchmarks (Figure 1(a)), while Figure 1(b) relates these post-adaptation precision gaps to the gains from adapting the BF16 backbone. The full results further reveal five reversals in the relative ranking of BF16 and 4-bit models before and after fine-tuning (Appendix C.1). These observations show that pre-adaptation task scores alone

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.

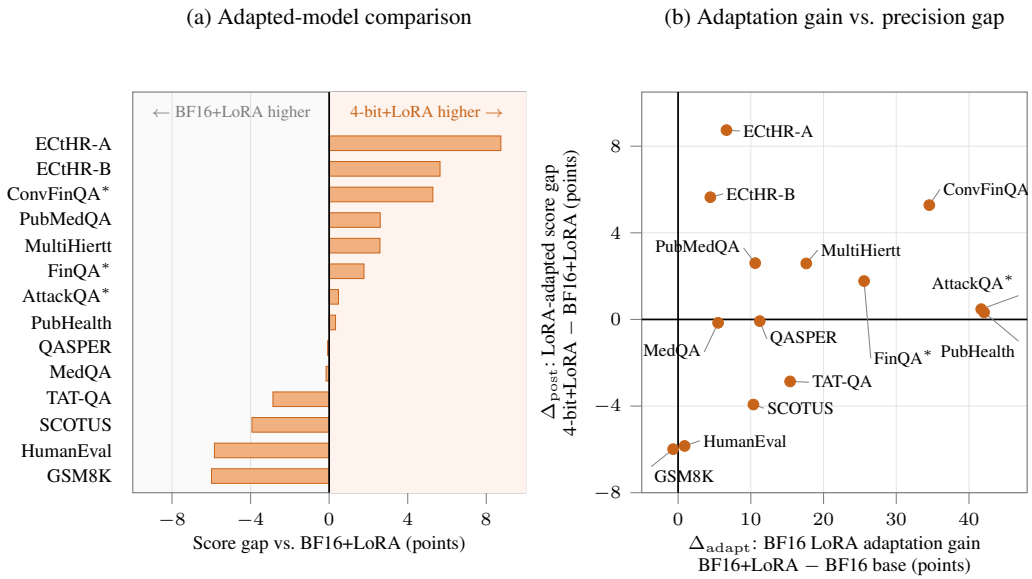


Figure 1: Adaptation gains and precision gaps on 14 benchmarks with Qwen2.5-0.5B-Instruct (Qwen et al., 2024). (a) After LoRA fine-tuning: 4-bit+LoRA minus BF16+LoRA. (b) BF16+LoRA minus BF16 base on the horizontal axis, versus the adapted precision gap from (a) on the vertical axis. Positive horizontal values indicate gains from BF16 LoRA adaptation; positive vertical values favor 4-bit+LoRA. Each difference is in the task's primary-metric points, not a common cross-task scale. Protocols, marker conventions, and full results are in Appendix C.

do not determine post-adaptation rankings. They do not measure reconstruction error, nor do they imply that quantization is inherently beneficial.

These observations motivate selecting initializations for their post-adaptation utility (Section 3.1), but evaluating each candidate through fine-tuning is computationally expensive. We therefore propose *Task-Aware Quantized Initialization* (TAQI), which uses task sensitivity measured before adaptation to guide joint quantized–low-rank construction. TAQI collects task-loss output gradients once on the unchanged BF16 model and uses their squared norms to weight input-activation second moments (Section 3.2). This retains the pretrained reconstruction target while directing reconstruction accuracy toward task-sensitive responses (Section 3.3).

During evaluation, we re-evaluate LLM capabilities when selecting downstream task datasets, rather than directly adopting evaluation datasets from prior work (Li et al., 2024; Meng et al., 2024). This is because contemporary LLMs have updated knowledge that already covers these tasks; as shown in Figure 1 (b), GSM8K and HumanEval benefit little from BF16 LoRA. Based on the detailed evaluation in Figure 1 (b), we select MedQA, TAT-QA, and QASPER. At 4 bits, TAQI improves over GPTQ-intrinsic on IBM Granite-4.0-H-350M by 0.71 points in MedQA accuracy, 1.33 points in TAT-QA F1, and 0.45 points in QASPER answer F1. On Qwen3.5-0.8B at the same precision, it improves TAT-QA and QASPER F1 by 0.67 and 0.77 points, respectively, but reduces MedQA accuracy by 2.99 points. Experiments across three backbones and multiple bit widths further characterize the task- and precision-dependent gains.

Our contributions are:

- We reformulate quantized low-rank initialization from a reconstruction-oriented problem into an adaptation-oriented joint search over quantized weights and low-rank initial values, evaluated by target-domain risk after a fixed adaptation budget. We show analytically that reconstruction optimality does not in general imply optimality under this objective, and empirically observe that pre-adaptation model quality does not consistently predict post-adaptation performance.

- Based on this formulation, we propose *Task-Aware Quantized Initialization* (TAQI), a static initialization procedure that uses target-task sensitivity measured before adaptation to guide joint quantized–low-rank construction. Its task-weighted reconstruction surrogate reallocates reconstruction accuracy toward task-sensitive responses, with a corresponding bound on the energy of local first-order task-loss perturbation contributions.
- We empirically characterize task-, backbone-, and precision-dependent behavior of task-aware joint initialization across two backbones, three domain QA tasks, and 4-, 3-, and 2-bit quantization. Single runs per configuration reveal both gains and regressions relative to GPTQ-intrinsic, delineating the observed benefits and limitations of TAQI. A supporting 14-benchmark adaptation study motivates the choice of evaluation tasks.

2 Related Work

Quantized low-rank adaptation. QLoRA enables efficient specialization of pretrained models by training high-precision low-rank adapters on a frozen low-bit backbone (Hu et al., 2022; Dettmers et al., 2023). PEQA adapts quantization scales (Kim et al., 2023); QA-LoRA uses group-wise operators to balance quantization and adaptation (Xu et al., 2024). Q-BLoRA simplifies adapter inputs and outputs while increasing rank; its QA-BLoRA extension aligns adaptation with block-wise quantization (Shen et al., 2025). IR-QLoRA calibrates low-bit representations to preserve pretrained information (Qin et al., 2024). Joint initialization makes the backbone and adapter complementary from the outset. LoftQ, LQ-LoRA, ApiQ, and quantized PiSSA reconstruct pretrained weights or activations using the combined components (Li et al., 2024; Guo et al., 2024; Liao et al., 2024; Meng et al., 2024). GPTQ-intrinsic incorporates continuous low-rank coordinates directly into the GPTQ scan (Zhang & Saab, 2026). We build on joint construction while distinguishing its reconstruction criterion from target-domain risk after finite-budget adaptation.

Initialization and sensitivity-aware quantization. Parameter-efficient methods offer several ways to specialize pretrained models (Houlsby et al., 2019; Li & Liang, 2021; Lester et al., 2021; Zhang et al., 2023; Liu et al., 2024). For low-rank initialization, PiSSA uses principal weight components (Meng et al., 2024), LoRA-GA uses task-loss gradients (Wang et al., 2024), and CorDA uses task-context covariances (Yang et al., 2024). Quantization also depends on calibration data (Williams & Aletras, 2024) and the geometry used to allocate error. GPTQ uses activation second moments; SqueezeLLM uses sensitivity information; data-aware LQ-LoRA uses Fisher-weighted reconstruction (Frantar et al., 2023; Kim et al., 2024; Guo et al., 2024). GuidedQuant incorporates end-loss gradients into the quantization criterion (Kim et al., 2025). Sensitivity weighting and coupled quantized–low-rank scanning are thus established ingredients. TAQI's methodological distinction is where task conditioning acts: a common mixed reconstruction geometry determines the construction basis and governs both discrete codes and continuous low-rank coordinates in the same augmented scan. This differs from initializing adapters on an already fixed backbone, weighting residuals only after codes have been selected, and joint scanning under general activation geometry alone. The pretrained reconstruction target is retained; task information changes the geometry of the coupled decisions. Appendix F discusses additional quantization methods.

3 Method

Overview. TAQI uses a static task-sensitive reconstruction geometry to guide adaptation-oriented initialization. Figure 2 contrasts independent quantization with joint backbone–adapter construction.

3.1 Problem Formulation: Adaptation-Oriented Quantized Initialization

Domain adaptation on a fixed quantized backbone. Let $W_l, Q_l \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ be input-first weights, so $z^l = W_l^\top x^l$. As an unrestricted baseline setting, the base adaptation model $f_{Q,\Phi}$ has effective weights $Q_l + cA_l^\top B_l^\top$, with $A_l \in \mathbb{R}^{r \times d_{\text{in}}}$, $B_l \in \mathbb{R}^{d_{\text{out}} \times r}$, $c = \alpha/r$, and $\Phi = \{(A_l, B_l)\}_l$. For fixed Q and target-domain training data $\mathcal{D}_{\text{tr}} = \{u_n\}_{n=1}^N$, adaptation trains both factors:

$$\mathcal{L}_{\text{tr}}(\Phi; Q) = \frac{1}{N} \sum_n \ell_{\text{tr}}(f_{Q,\Phi}; u_n), \qquad \Phi_K = \mathcal{U}_K(\Phi_0; Q, \mathcal{D}_{\text{tr}}, \xi). \tag{1}$$

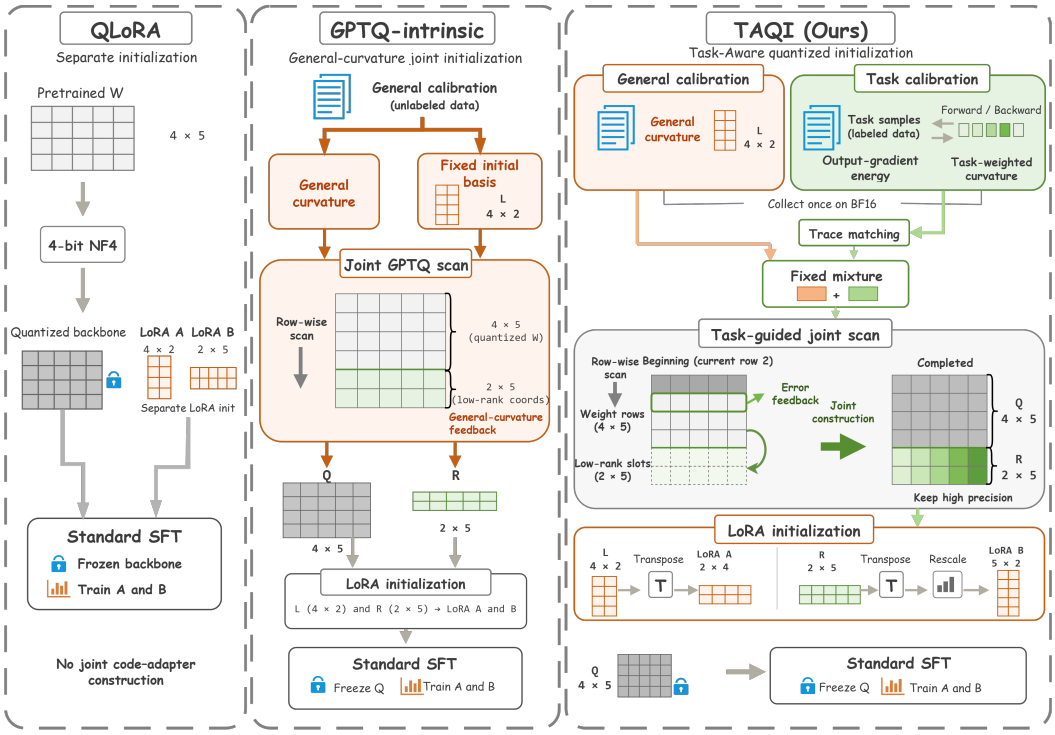


Figure 2: Schematic initialization pathways contrasting independent quantization with the general-basis joint-construction core and its low-rank mapping. The joint pathway constructs quantized weights and continuous correction coordinates together. Toy dimensions follow the input-first convention; runtime weights use its transpose.

Here $\mathcal{U}_K$ comprises $K < \infty$ optimizer steps, their schedule, and optimizer states; ξ collects randomness. Evaluation uses a separate target-domain risk $\mathcal{R}_T(f) = \mathbb{E}_{u\sim P_T}[\ell_{\text{eval}}(f;u)]$. Token cross-entropy supplies the training loss, whereas downstream accuracy and F1 measure prediction quality. Thus $\mathcal{L}_{\text{tr}}$ determines the updates, while $\mathcal{R}_T$ evaluates held-out answer quality.

Expressible models and finite-budget trajectories. For one layer the base family is $\mathcal{M}_r(Q) = \{Q + D : \text{rank}(D) \le r\}$. Changing Q translates this family; higher backbone precision does not by itself establish inclusion between two such families. In a linear squared-loss model with $H \succ 0$, the best rank-r fit to a task target $W_T^\star$ leaves the squared singular-value tail of $H^{1/2}(W_T^\star - Q)$ beyond rank r (Appendix D.2). The relevant residual is therefore $(W_T^\star - W) + (W - Q)$: task and quantization shifts can reinforce or offset one another in directions that a rank-limited update cannot remove. Expressibility is distinct from reaching a model in K steps. For $M = Q + cA^\top B^\top$ and $G = \nabla_M \mathcal{L}$, one gradient-descent step gives $M^+ - M = -\eta c^2(A^\top AG + GBB^\top) + O(\eta^2)$. Rescaling A by t and B by $1/t$ preserves the initial weight but changes this update (Appendix D.5). Thus both the backbone and the individual initial factors influence adaptation.

Adaptation-oriented initialization objective. These roles lead to the joint design criterion

$$\min_{(Q,\Phi_0)\in\mathcal{I}_{b,r}} \mathcal{J}_K(Q,\Phi_0), \qquad \mathcal{J}_K(Q,\Phi_0) = \mathbb{E}_\xi\left[\mathcal{R}_T\left(f_{Q,\mathcal{U}_K(\Phi_0;Q,\mathcal{D}_{\text{tr}},\xi)}\right)\right]. \tag{2}$$

The feasible set fixes the quantization format, bit width, rank, and adapter scale. TAQI searches integer codes on the source GPTQ grid $\mathcal{Q}_b$, retaining scales, zeros, and input-group assignments. Evaluating $\mathcal{J}_K$ requires a complete adaptation trajectory for each candidate and repeated trajectories to estimate the expectation, making direct search computationally impractical. Finite K retains the role of initialization, connecting this viewpoint to meta-learning and bilevel optimization (Finn et al., 2017; Franceschi et al., 2018) (Appendix D.1). The criterion specifies what makes an initialization useful; the static construction below proposes candidates for empirical evaluation without running adaptation during construction.

Reconstruction as a distinct criterion. Calibration on the unchanged BF16 reference gives

$$H_G^l = \frac{1}{N_G} \sum_i x_i^l (x_i^l)^\top. \tag{3}$$

For coordinates $R_l \in \mathbb{R}^{r \times d_{\text{out}}}$ in a fixed construction basis L_l, a reconstruction objective is

$$\mathcal{J}_{\text{rec}}(Q, R) = \sum_l \|W_l - Q_l - L_l R_l\|_{H_G^l}^2, \qquad \|E\|_H^2 := \text{tr}(E^\top H E). \tag{4}$$

GPTQ-intrinsic uses $L_l = V_r(H_G^l)$, where V_r contains leading orthonormal eigenvectors. TAQI uses the mixed geometry defined below. The two criteria need not agree.

Proposition 1 *Reconstruction-optimal joint initialization need not minimize finite-budget adaptation risk: there exist two feasible initializations with identical bit width, rank, basis, training-loss function, optimizer, and budget such that the one with zero reconstruction error has strictly higher risk after adaptation.*

For a short calculation, take $W = I_2$, task target $\frac{2}{3} I_2$, rank one, $c = 1$, basis $L = e_1$, and grid $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$. The choices $Q_a = I_2$ and $Q_b = \frac{2}{3} I_2$ have optimal reconstruction coordinates $R_a = 0$ and $R_b = \frac{1}{3} e_1^\top$, with errors 0 and $1/9$. One simultaneous gradient step of size one on both factors for $\mathcal{R}(M) = \frac{1}{2}\|M - \frac{2}{3} I_2\|_F^2$ gives $M_{a,1} = \text{diag}(\frac{2}{3}, 1)$ and $M_{b,1} = \frac{2}{3} I_2$: the corresponding risks are $1/18$ and 0. Appendix D.4 gives the factor updates. This example separates reconstruction optimality from adaptation optimality.

Task-sensitive initialization hypothesis. We replace candidate-by-candidate adaptation during construction with a local sensitivity heuristic: measure teacher-forced token-cross-entropy output gradients once on the fixed BF16 reference, then prioritize reconstruction at positions with greater gradient energy. Our hypothesis is that, at fixed bit width and rank, this allocation provides more useful starting points than general activation energy alone when the measured sensitivities remain informative during adaptation. It is tested through the downstream comparisons in Section 4.2; the local diagnostic in Section 4.4 examines the proposed error allocation mechanism. The heuristic is not a Taylor or theoretical approximation to $\mathcal{J}_K$: its mathematical justification concerns local reference-loss perturbations, not a guarantee of post-adaptation accuracy or F1.

3.2 Task-Sensitive Surrogate and Local Perturbation Bound

Evaluating $\mathcal{J}_K$ requires adaptation; we therefore use static output-gradient measurements as a tractable proxy for task sensitivity.

Target-task output gradients. A static task-sensitive criterion needs to distinguish output perturbations by their local effect on the target loss. Output gradients provide this first-order signal without adapting candidate initializations. On the unchanged BF16 reference, let ℓ_n be the per-example teacher-forced token-cross-entropy loss, retaining its sample-local normalization. For sampled supervised positions $p \in \mathcal{S}_n^l$ at target layer l, define

$$g_{np}^l = \frac{\partial \ell_n}{\partial z_{np}^l}, \qquad \omega_{np}^l = \|g_{np}^l\|_2^2. \tag{5}$$

Task-sensitive curvature. To allocate more reconstruction accuracy to positions with larger local loss sensitivity while retaining a quadratic reconstruction criterion, we weight the input second moment by gradient energy,

$$H_T^l = \frac{1}{Z_s^l} \sum_{n,p \in \mathcal{S}_n^l} \omega_{np}^l x_{np}^l (x_{np}^l)^\top, \qquad Z_s^l = \max\left\{ \sum_{n,p \in \mathcal{S}_n^l} \omega_{np}^l, 1 \right\}. \tag{6}$$

This gradient-weighted input second moment defines the reconstruction geometry. Appendix E specifies sampling and normalization.

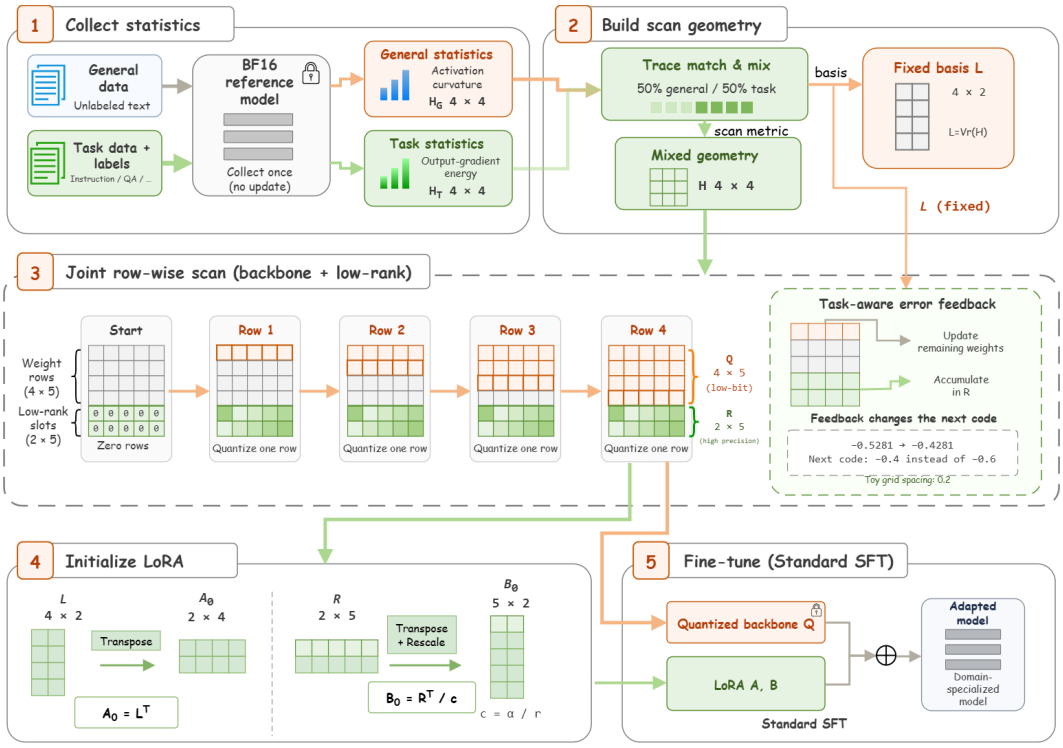


Figure 3: Task-general statistics mixing and coupled reconstruction. General (H_G) and task-sensitive (H_T) statistics, collected once on the unchanged BF16 model, are trace-matched and mixed with $\beta = 0.5$ into H. The same H determines the fixed basis $L = V_r(H)$ and governs the joint row-wise scan, which propagates rounding errors to unprocessed weight rows and the continuous low-rank coordinates R. The resulting Q is the frozen backbone, with LoRA factors initialized as $A_0 = L^\top$ and $B_0 = R^\top / c$. Output-group selection and the persistent adapter support during SFT, omitted here, are specified in Section 4.1. Toy dimensions: 4×5, $r = 2$, grid spacing 0.2.

Local first-order perturbation bound. To relate this weighted reconstruction criterion to task loss, we examine local first-order perturbations at fixed reference inputs. For $E_l = W_l - Q_l - L_l R_l$, the local output perturbation at fixed reference input is $-E_l^\top x_{np}^l$. With $N_s^l = \sum_n |\mathcal{S}_n^l| > 0$,

$$\Gamma_{\mathcal{S}}^l(E_l) := \frac{1}{N_s^l} \sum_{n,p \in \mathcal{S}_n^l} \left[(g_{np}^l)^\top E_l^\top x_{np}^l \right]^2 \tag{7}$$

$$\leq \frac{Z_s^l}{N_s^l} \|E_l\|_{H_T^l}^2. \tag{8}$$

The quantity $\Gamma_{\mathcal{S}}^l$ is the mean squared first-order contribution at selected positions of one layer, with inputs and gradients fixed at the BF16 reference. Cauchy–Schwarz bounds each term by $\|g\|_2^2 \|E^\top x\|_2^2$ (Appendix D.6). Higher-gradient positions therefore incur a larger penalty for a fixed output-error norm. Scalar gradient energy weights all output directions equally, including those orthogonal to g, yielding an efficiently accumulated upper envelope of local sensitivity.

3.3 TAQI: Task-Aware Quantized Initialization

Mixed task-general geometry. To balance task-sensitive and general reconstruction, we trace-match their curvature matrices before mixing. Suppressing layer indices, we form

$$\bar{H}_T = \frac{\mathrm{tr}(H_G)}{\mathrm{tr}(H_T)} H_T, \qquad H = (1-\beta) H_G + \beta \bar{H}_T, \qquad L = V_r(H). \tag{9}$$

The statistics must be finite and symmetric with positive traces. Unlike GPTQ-intrinsic, TAQI uses the mixed geometry for both the basis and scan; L is fixed only during construction. Trace matching

makes $\mathrm{tr}(\bar{H}_T) = \mathrm{tr}(H_G)$, so β changes the directional allocation rather than the total trace of H. The leading mixed eigenspace allocates the rank budget, while the same matrix weights errors during code construction. Task information therefore enters both continuous correction directions and the discrete rounding decisions.

Task-guided joint construction. We adopt GPTQ-intrinsic's augmented scan (Zhang & Saab, 2026) and condition both its basis and reconstruction metric on the mixed H. The scan illustrated in Figure 3 constructs codes and continuous coordinates jointly, allowing their errors to compensate. With $C = [I\ L]$, $W_{\mathrm{aug}} = [W; 0]$, and $\Theta = [Q; R]$, augmenting the curvature as $C^\top HC$ and damping it by λI gives the constrained surrogate

$$\min_{Q \in \mathcal{Q}_b, R} \|W - Q - LR\|_H^2 + \lambda \|W - Q\|_F^2 + \lambda \|R\|_F^2. \tag{10}$$

Here $\lambda > 0$ is set from the mean diagonal of the augmented curvature, with the implemented coefficient given in Appendix E.3. A greedy augmented GPTQ scan rounds weight rows on the source grid and propagates errors to future rows, including continuous coordinates R. It returns the greedy candidate $(Q_{\mathrm{cand}}, R_{\mathrm{cand}})$. To see the coupling, let $K_R = L^\top HL + \lambda I_r$. At fixed Q, the minimizing coordinates are $R^\star = K_R^{-1} L^\top H(W - Q)$, giving the effective quadratic

$$\min_R \mathcal{S}(Q, R) = \|W - Q\|_{H_{\mathrm{eff}}}^2, \qquad H_{\mathrm{eff}} = H + \lambda I - HLK_R^{-1} L^\top H. \tag{11}$$

Here $\mathcal{S}$ denotes the objective in (10). The subtracted positive semidefinite term discounts residual directions that the continuous coordinates can compensate. Thus a code's cost depends on the available correction, not only its distance to W under H. This Schur-complement identity characterizes the exact minimum over continuous coordinates at fixed codes. Appendices D.7 and E.3 give the derivation and greedy scan mechanics.

4 Experiments

4.1 Experimental setup

Tasks and backbones. We evaluate medical-exam multiple-choice QA on MedQA (Jin et al., 2021) (accuracy), numerical QA over financial tables and text on TAT-QA (Zhu et al., 2021) (answer F1), and QA over scientific papers on QASPER (Dasigi et al., 2021) (answer F1). These tasks span complementary adaptation demands: specialist medical knowledge, structured financial evidence, and evidence-based reasoning over long scientific documents. In our motivation study, BF16+LoRA improves these tasks by 5.50, 15.40, and 11.23 points, respectively, versus -0.68 on GSM8K (Cobbe et al., 2021) and $+0.89$ on HumanEval (Chen et al., 2021a). These gains motivate the three domain benchmarks. Parentheses give primary metrics; macro-F1, exact match, scale accuracy, and evidence F1 are reported in Appendix A. We evaluate two backbones: Granite-4.0-H-350M (IBM Granite Team, 2025) and Qwen3.5-0.8B (Qwen Team, 2026); comparisons are within-backbone.

Configuration. We evaluate targeted projections at $b \in \{4, 3, 2\}$ bits; architecture-specific targets, construction settings, and training details are in Appendix E. Baselines are QA-LoRA, QBLoRA, GPTQ followed by LoRA (GPTQ+LoRA), and GPTQ-intrinsic. GPTQ-intrinsic provides the general-geometry joint-construction baseline. TAQI uses the mixed-geometry construction in Section 3.3. All methods freeze the quantized backbone during adapter SFT. Each configuration reports a single run, so score differences are descriptive.

Implemented TAQI protocol. After the coupled scan, output-channel groups are retained only if their raw task-weighted reconstruction error improves over the source GPTQ backbone Q_0 while the general reconstruction error remains within the specified tolerance. Rejected groups restore Q_0 and receive zero low-rank correction. The accepted output support persists during SFT through a fixed output mask, while both low-rank factors train; this is not unrestricted ordinary LoRA. Appendix E.4 gives the exact selection rule, factor mapping, and algorithm, and Appendix E provides the statistics, numerical settings, and training protocol.

Table 1: Primary metrics on Granite-4.0-H-350M and Qwen3.5-0.8B at 4, 3, and 2 bits (MedQA accuracy, TAT-QA F1, and QASPER answer F1). Within each bit width and model/metric column, the best distinct value among the compared methods is bold and the second-best is underlined (ties share the same emphasis). The Δ rows report TAQI minus GPTQ-intrinsic, which uses the same joint scan with general calibration curvature; positive differences are bold. Diagnostic metrics are reported in Appendix A.

		Granite-4.0-H-350M			Qwen3.5-0.8B		
Bit	Method	MedQA	TAT-QA	QASPER	MedQA	TAT-QA	QASPER
4	QA-LoRA	26.787	16.715	16.844	37.078	41.027	12.464
	QBLoRA	25.137	27.746	16.218	40.613	49.481	17.511
	GPTQ+LoRA	30.715	42.413	18.797	**45.719**	<u>63.432</u>	22.976
	GPTQ-intrinsic	<u>33.386</u>	<u>44.395</u>	<u>22.772</u>	<u>45.090</u>	63.289	<u>23.179</u>
	TAQI (ours)	**34.093**	**45.721**	**23.217**	42.105	**63.957**	**23.945**
	Δ vs. GPTQ-intr.	$\mathbf{+0.707}$	$\mathbf{+1.326}$	$\mathbf{+0.445}$	-2.985	$\mathbf{+0.668}$	$\mathbf{+0.766}$
3	QA-LoRA	26.002	9.316	14.608	28.751	34.648	6.930
	QBLoRA	27.101	20.542	11.946	31.658	43.764	17.251
	GPTQ+LoRA	27.808	38.633	19.127	35.428	<u>59.233</u>	21.063
	GPTQ-intrinsic	**29.851**	<u>41.430</u>	<u>22.588</u>	<u>36.371</u>	59.130	<u>21.305</u>
	TAQI (ours)	<u>29.694</u>	**42.155**	**22.632**	**37.863**	**60.618**	**23.865**
	Δ vs. GPTQ-intr.	-0.157	$\mathbf{+0.725}$	$\mathbf{+0.044}$	$\mathbf{+1.492}$	$\mathbf{+1.488}$	$\mathbf{+2.560}$
2	QA-LoRA	21.524	0.048	2.174	<u>25.295</u>	0.265	1.301
	QBLoRA	25.216	0.117	1.835	23.488	0.676	0.923
	GPTQ+LoRA	24.823	21.056	15.320	24.352	<u>34.737</u>	<u>8.323</u>
	GPTQ-intrinsic	**26.551**	<u>31.124</u>	<u>18.418</u>	24.038	29.081	4.494
	TAQI (ours)	<u>26.316</u>	**33.278**	**21.695**	**27.730**	**48.190**	**21.457**
	Δ vs. GPTQ-intr.	-0.235	$\mathbf{+2.154}$	$\mathbf{+3.277}$	$\mathbf{+3.692}$	$\mathbf{+19.109}$	$\mathbf{+16.963}$

4.2 Main results

Granite-4.0-H-350M. At 4 bits, TAQI is best on all three primary metrics and all four diagnostics among the ranked methods, improving over GPTQ-intrinsic, a general-geometry joint baseline, by 0.707, 1.326, and 0.445 points. At 3 and 2 bits, it remains best on TAT-QA F1 and QASPER answer F1 (2-bit QASPER: 21.695 versus 18.418) but trails GPTQ-intrinsic on MedQA (-0.157, -0.235).

Qwen3.5-0.8B. TAQI leads all three primary metrics at 3 and 2 bits. At 3 bits, its QASPER F1 (23.865) is close to its 4-bit score (23.945), while GPTQ-intrinsic falls from 23.179 to 21.305. At 2 bits, TAQI reaches 48.190 TAT-QA F1 and 21.457 QASPER F1, versus 29.081 and 4.494 for GPTQ-intrinsic; it also exceeds GPTQ+LoRA (34.737, 8.323). Granite shows the same 2-bit QASPER advantage with a smaller margin ($+3.277$ versus $+16.963$). MedQA is less consistent: TAQI trails GPTQ-intrinsic at 4 bits but leads at 3 and 2 bits. Its 2-bit macro-F1 is 10.855, indicating weaker class-balanced performance despite its leading accuracy.

Overall. TAQI scores higher than GPTQ-intrinsic in 15 of the 18 model–bit-width–task comparisons in Table 1: two backbones, three bit widths, and three primary metrics: 7 of 9 comparisons on Granite and 8 of 9 on Qwen. At 4 bits, the corresponding counts are 3 of 3 and 2 of 3. TAT-QA F1 and QASPER answer F1 are higher in all six backbone–precision settings, whereas all three regressions occur on MedQA. These counts summarize single-run configuration comparisons, not independent repeated trials; we retain each task's metric rather than average accuracy and F1.

4.3 Ablation of TAQI components

We ablate TAQI on Granite-4.0-H-350M at 4 bits along its two construction components (Table 2). TAQI-core keeps the mixed task-general geometry in the basis and the coupled scan but omits the output-group selection protocol; GPTQ-intrinsic additionally removes task curvature. The GPTQ-intrinsic–TAQI-core comparison therefore isolates TAQI's central design choice, task-conditioned

Table 2: Component ablation on Granite-4.0-H-350M at 4 bits. Task curv. denotes task curvature in the basis and coupled scan; Select. denotes the output-group selection protocol of Section 4.1. For each benchmark, the primary metric is rightmost; the best value in each column is bold. Appendix A.2 reports 3- and 2-bit results.

			MedQA		TAT-QA			QASPER	
Method	Task curv.	Select.	Macro-F1	Acc.	EM	Scale Acc.	F1	Evidence F1	Answer F1
GPTQ-intrinsic	–	–	33.226	33.386	35.839	93.145	44.395	6.616	22.772
TAQI-core	✓	–	33.158	33.621	36.621	93.085	45.188	6.754	22.826
TAQI	✓	✓	**33.417**	**34.093**	**37.162**	**93.626**	**45.721**	**6.892**	**23.217**

geometry in the coupled scan, and the TAQI-core–TAQI comparison isolates the selection protocol of Section 4.1, including its persistent adapter support during SFT. Appendix A.2 extends the ablation to 3 and 2 bits.

Task curvature. Adding task curvature to the coupled scan (GPTQ-intrinsic $\rightarrow$ TAQI-core) raises all three primary metrics, by 0.235 MedQA accuracy, 0.793 TAT-QA F1, and 0.054 QASPER answer F1. The largest gain is on TAT-QA, where EM also rises by 0.782, and QASPER evidence F1 rises by 0.138. MedQA macro-F1 and TAT-QA scale accuracy stay within 0.07 points of GPTQ-intrinsic (-0.068 and -0.060).

Output-group selection. The selection protocol then raises all seven metrics, including 0.472 MedQA accuracy, 0.533 TAT-QA F1, and 0.391 QASPER answer F1. TAQI is therefore best on every metric at 4 bits and exceeds GPTQ-intrinsic by 0.707, 1.326, and 0.445 points on the primary metrics and by 1.323 TAT-QA EM. Both components contribute at this precision: task curvature raises all three primary metrics, and the selection protocol builds on that gain. At lower precision, task curvature continues to raise TAT-QA and QASPER but lowers MedQA accuracy; Appendix A.2 analyzes these cases.

4.4 Initialization-level analysis of task-sensitive error allocation

A static diagnostic probes the mechanism behind task curvature before any fine-tuning. On three preselected Granite-4.0-H-350M shared-MLP output projections (layers 6, 15, 25) at 4 bits, it compares joint construction under general curvature (GJ), unweighted target activations (XJ), and task-gradient-weighted curvature (TJ); all three share pretrained targets and a fixed general basis, so the XJ–TJ comparison isolates gradient weighting. On common BF16 probes with TAT-QA output gradients, TJ increases the general error D_G by 7.71–9.20% relative to GJ but reduces the task-weighted error D_T by 80.95–81.11% and the first-order energy Γ by 77.40–81.19%. Relative to XJ, gradient weighting further reduces D_T by 52.16–60.40% and Γ by 32.87–70.56%: task curvature reallocates error toward task-sensitive responses, and the direction-aware Γ responds more variably than D_T across layers. Appendix B gives the figure, protocol, and design differences from TAQI.

5 Conclusion

Quantized low-rank initialization sets the backbone and initial factors for finite-budget adaptation. We formalized its evaluation through target-domain risk after adaptation and established that reconstruction and adaptation criteria can favor different initializations. TAQI uses static task sensitivity to shape the geometry of joint quantized–low-rank construction. It scores higher than GPTQ-intrinsic on all three Granite tasks at 4 bits and in 15 of 18 primary-metric comparisons, and at 4 bits the ablation shows that both task curvature and the output-group selection protocol contribute. Its largest gains appear at 2 bits, up to 19.109 TAT-QA F1 on Qwen3.5, where reconstruction capacity is most constrained and allocating it to task-sensitive responses matters most.

Scope and future work. The experiments cover 0.35–0.8B-parameter backbones, with gains varying by task and precision; larger models and repeated runs would characterize this variation more fully. TAQI requires labeled target examples and BF16 backward statistics. When static sensitivities remain informative through adaptation is a central question for extending it.

AI USE STATEMENT

An LLM-based assistant was used during manuscript preparation only for: (i) organizing/formatting part of the data; (ii) assistance with generating the LaTeX structure; and (iii) minor language polishing of selected portions of the manuscript.

REPRODUCIBILITY STATEMENT

Section 4.1 describes the experimental setup, and Appendix E details statistics collection, construction, selection, factor mapping, and training. Appendix E.2 states the numerical assumptions; Appendix D.4 and Appendix D.6 give the counterexample and local-bound derivation. Full downstream results appear in Appendix A, diagnostic protocols in Appendix B, and the supporting adaptation study in Appendix C.

ETHICS STATEMENT

The medical, financial, and scientific QA evaluations measure research benchmark performance, not deployment safety or fitness for professional advice. Incorrect or unsupported answers, including numerical errors and misleading evidence, can cause harm in these settings. Any deployment would require domain-expert verification, evaluation of bias and reliability under the intended use conditions, and appropriate safeguards for sensitive data and consequential decisions.

REFERENCES

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4310–4330, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.297. URL https://aclanthology.org/2022.acl-long.297/.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021a. doi: 10.48550/arXiv.2107.03374. URL https://arxiv.org/abs/2107.03374.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711, Online and Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.300. URL https://aclanthology.org/2021.emnlp-main.300/.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6279–6292, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.421. URL https://aclanthology.org/2022.emnlp-main.421/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, 2018.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

Han Guo, Philip Greengard, Eric P. Xing, and Yoon Kim. LQ-LoRA: Low-rank plus quantized matrix decomposition for efficient language model finetuning. In *International Conference on Learning Representations*, 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2019.

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.

IBM Granite Team. Granite-4.0-H-350M: Official model card, 2025. URL https://huggingface.co/ibm-granite/granite-4.0-h-350m.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421. URL https://doi.org/10.3390/app11146421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259/.

Jeonghoon Kim, Jung Hyun Lee, Sungdong Kim, Joonsuk Park, Kang Min Yoo, Se Jung Kwon, and Dongsoo Lee. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. In *Advances in Neural Information Processing Systems*, 2023.

Jinuk Kim, Marwa El Halabi, Wonpyo Park, Clemens JS Schaefer, Deokjae Lee, Yeonhong Park, Jae W. Lee, and Hyun Oh Song. GuidedQuant: Large language model quantization via exploiting end loss guidance. In *International Conference on Machine Learning*, pp. 30011–30037, 2025.

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. SqueezeLLM: Dense-and-sparse quantization. In *International Conference on Machine Learning*, 2024.

Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7740–7754, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.623. URL https://aclanthology.org/2020.emnlp-main.623/.

Varun Badrinath Krishna. AttackQA: Development and adoption of a dataset for assisting cybersecurity operations using fine-tuned and open-source LLMs. *arXiv preprint arXiv:2411.01073*, 2024. doi: 10.48550/arXiv.2411.01073. URL https://arxiv.org/abs/2411.01073.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing*, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Annual Meeting of the Association for Computational Linguistics*, 2021.

Yixiao Li, Yifan Yu, Chen Liang, Nikos Karampatziakis, Pengcheng He, Weizhu Chen, and Tuo Zhao. LoftQ: LoRA-fine-tuning-aware quantization for large language models. In *International Conference on Learning Representations*, 2024.

Baohao Liao, Christian Herold, Shahram Khadivi, and Christof Monz. ApiQ: Finetuning of 2-bit quantized large language model. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, 2024.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning*, 2024.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, 2024.

Haotong Qin, Xudong Ma, Xingyu Zheng, Xiaoyang Li, Yang Zhang, Shouda Liu, Jie Luo, Xianglong Liu, and Michele Magno. Accurate LoRA-finetuning quantization of LLMs via information retention. In *International Conference on Machine Learning*, 2024.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115. Checkpoint: https://huggingface.co/Qwen/Qwen2.5-0.5B-Instruct.

Qwen Team. Qwen3.5-0.8B: Official model card, 2026. URL https://huggingface.co/Qwen/Qwen3.5-0.8B.

Ao Shen, Zhiquan Lai, Qiang Wang, Xionglve Li, Lizhi Zhang, Dongsheng Li, and Jiaxin Li. Accurate and efficient fine-tuning of quantized large language models through optimal balance in adaptation. *Transactions of the Association for Computational Linguistics*, 13:861–877, 2025. doi: 10.1162/tacl.a.23. URL https://aclanthology.org/2025.tacl-1.40/.

Shaowen Wang, Linxi Yu, and Jian Li. LoRA-GA: Low-rank adaptation with gradient approximation. In *Advances in Neural Information Processing Systems*, 2024.

Miles Williams and Nikolaos Aletras. On the impact of calibration data in post-training quantization and pruning. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, 2023.

Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, Xiaopeng Zhang, and Qi Tian. QA-LoRA: Quantization-aware low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2024.

Yibo Yang, Xiaojie Li, Zhongzhu Zhou, Shuaiwen Leon Song, Jianlong Wu, Liqiang Nie, and Bernard Ghanem. CorDA: Context-oriented decomposition adaptation of large language models for task-aware parameter-efficient fine-tuning. In *Advances in Neural Information Processing Systems*, 2024.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023.

Shihao Zhang and Rayan Saab. GPTQ-intrinsic LoRA: A near-optimal algorithm for low-precision quantization with low-rank adaptation. *arXiv preprint arXiv:2606.01412*, 2026.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6588–6600, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.454. URL https://aclanthology.org/2022.acl-long.454/.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3277–3287. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.254. URL https://aclanthology.org/2021.acl-long.254/.

A FULL RESULTS

This appendix reports all metrics and bit widths for the two backbones. Methods are compared within each backbone and bit width.

A.1 IBM GRANITE-4.0-H-350M

Table 3: IBM Granite-4.0-H-350M results on three question-answering tasks. Within each bit width and metric, best distinct values among the compared methods are bold and second-best values are underlined; ties share the same emphasis. For each benchmark, the primary metric is rightmost.

		MedQA		TAT-QA			QASPER	
Bit	**Method**	Macro-F1	**Acc.**	EM	Scale Acc.	**F1**	Evidence F1	**Answer F1**
4	GPTQ+LoRA	27.476	30.715	33.434	92.002	42.413	6.823	18.797
4	QA-LoRA	26.000	26.787	7.396	63.019	16.715	6.478	16.844
4	QBLoRA	23.150	25.137	18.641	76.969	27.746	5.720	16.218
4	GPTQ-intrinsic	33.225	33.386	35.839	93.145	44.395	6.616	22.772
4	TAQI	**33.417**	**34.093**	**37.162**	**93.626**	**45.721**	**6.892**	**23.217**
3	GPTQ+LoRA	22.639	27.808	30.247	89.537	38.633	**6.823**	19.127
3	QA-LoRA	25.320	26.002	4.570	53.698	9.316	**6.823**	14.608
3	QBLoRA	24.298	27.101	12.387	67.348	20.542	6.065	11.946
3	GPTQ-intrinsic	**28.593**	**29.851**	32.892	**91.942**	41.430	6.713	22.588
3	TAQI	26.766	29.694	**33.434**	91.882	**42.155**	6.754	**22.632**
2	GPTQ+LoRA	20.067	24.823	14.071	79.254	21.056	**6.823**	15.320
2	QA-LoRA	14.442	21.524	0.000	46.482	0.048	**6.823**	2.174
2	QBLoRA	**24.060**	25.216	0.000	49.970	0.117	**6.823**	1.835
2	GPTQ-intrinsic	18.832	**26.551**	23.452	88.815	31.124	6.754	18.418
2	TAQI	18.775	26.316	**25.316**	**90.138**	**33.278**	6.685	**21.695**

Table 3 reports the Granite results. At 4 bits, TAQI leads every metric among the ranked methods. Relative to GPTQ-intrinsic, the primary-metric differences are $+0.707$, $+1.326$, and $+0.445$ at 4 bits, -0.157, $+0.725$, and $+0.044$ at 3 bits, and -0.235, $+2.154$, and $+3.277$ at 2 bits (MedQA, TAT-QA, QASPER order).

A.2 COMPONENT ABLATION ACROSS BIT WIDTHS

Table 4 extends the 4-bit ablation of Section 4.3 to 3 and 2 bits on the primary metrics, and Table 5 reports all metrics. The GPTQ-intrinsic and TAQI rows are the runs reported in Table 3; the ablation adds TAQI-core. MedQA invalid-answer rates are 0.000 and TAT-QA operator accuracy is 0.000 for all nine configurations, so both are omitted.

Table 4: Component ablation on Granite-4.0-H-350M across bit widths (primary metrics: MedQA accuracy, TAT-QA F1, QASPER answer F1). Task curv. and Select. follow Table 2. Within each bit width and column, the best value is bold; ties are all bold.

			4-bit			3-bit			2-bit		
Method	Task curv.	Select.	MedQA	TAT-QA	QASPER	MedQA	TAT-QA	QASPER	MedQA	TAT-QA	QASPER
GPTQ-intrinsic	–	–	33.386	44.395	22.772	**29.851**	41.430	22.588	**26.551**	31.124	18.418
TAQI-core	✓	–	33.621	45.188	22.826	29.694	41.551	**23.143**	26.316	**33.278**	21.520
TAQI	✓	✓	**34.093**	**45.721**	**23.217**	29.694	**42.155**	22.632	26.316	**33.278**	**21.695**

Task curvature across bit widths. Adding task curvature to the coupled scan (GPTQ-intrinsic → TAQI-core) raises TAT-QA F1 at every precision, by 0.793, 0.121, and 2.154 points at 4, 3, and 2

bits, and QASPER answer F1 by 0.054, 0.555, and 3.102. Its gains are largest at 2 bits, where it lifts QASPER answer F1 from 18.418 to 21.520. MedQA is the exception: accuracy rises by 0.235 at 4 bits but decreases by 0.157 and 0.235 at 3 and 2 bits.

Selection protocol across bit widths. At 3 bits, the selection protocol adds 0.604 TAT-QA F1, offsets most of the QASPER gain from task curvature (-0.511), and leaves MedQA unchanged. At 2 bits, it leaves MedQA and TAT-QA unchanged and adds 0.175 QASPER answer F1, for a total gain of 3.277 over GPTQ-intrinsic. Below 4 bits, the gains of TAQI therefore come mainly from task curvature. The MedQA regressions relative to GPTQ-intrinsic at 3 and 2 bits are already present in TAQI-core, at the step where pre-adaptation sensitivity guides construction.

Table 5: Full component ablation on Granite-4.0-H-350M. Task curv. and Select. follow Table 2. Within each bit width and metric, the best value is bold; ties are all bold. For each benchmark, the primary metric is rightmost.

				MedQA		**TAT-QA**			**QASPER**	
Bit	**Method**	Task curv.	Select.	Macro-F1	**Acc.**	EM	Scale Acc.	**F1**	Evidence F1	**Answer F1**
4	GPTQ-intrinsic	–	–	33.226	33.386	35.839	93.145	44.395	6.616	22.772
4	TAQI-core	✓	–	33.158	33.621	36.621	93.085	45.188	6.754	22.826
4	TAQI	✓	✓	**33.417**	**34.093**	**37.162**	**93.626**	**45.721**	**6.892**	**23.217**
3	GPTQ-intrinsic	–	–	**28.593**	**29.851**	32.892	**91.942**	41.430	6.713	22.588
3	TAQI-core	✓	–	26.766	29.694	33.253	91.882	41.551	6.713	**23.143**
3	TAQI	✓	✓	26.766	29.694	**33.434**	91.882	**42.155**	**6.754**	22.632
2	GPTQ-intrinsic	–	–	**18.832**	**26.551**	23.452	88.815	31.124	**6.754**	18.418
2	TAQI-core	✓	–	18.775	26.316	**25.316**	**90.138**	**33.278**	6.685	21.520
2	TAQI	✓	✓	18.775	26.316	**25.316**	**90.138**	**33.278**	6.685	**21.695**

The diagnostic metrics follow the primary ones. The 4-bit pattern in Table 2 holds for all metrics. At 3 bits, task curvature lowers MedQA macro-F1 from 28.593 to 26.766 alongside the small accuracy decrease, while the selection protocol leaves the MedQA metrics unchanged. At 2 bits, task curvature raises TAT-QA EM by 1.864 and scale accuracy by 1.323 points. On 2-bit QASPER, it raises answer F1 by 3.102 while evidence F1 stays nearly unchanged (6.685 versus 6.754); the selection protocol adds a further 0.175 answer F1 and leaves evidence F1 unchanged.

A.3 QWEN3.5-0.8B

Table 6: Qwen3.5-0.8B downstream results. Within each bit width, best distinct values among the compared methods are bold and second-best values are underlined for every metric; ties share the same emphasis. For each benchmark, the primary metric is rightmost.

		MedQA		TAT-QA			QASPER	
Bit	**Method**	Macro-F1	**Acc.**	EM	Scale Acc.	**F1**	Evidence F1	**Answer F1**
4	GPTQ+LoRA	**45.356**	**45.719**	55.382	93.867	63.432	19.596	22.976
4	QA-LoRA	36.023	37.078	31.630	79.735	41.027	17.161	12.464
4	QBLoRA	39.276	40.613	40.168	90.559	49.481	17.032	17.511
4	GPTQ-intrinsic	44.740	45.090	55.081	93.746	63.289	**19.876**	23.179
4	TAQI	41.844	42.105	**55.562**	**94.227**	**63.957**	19.734	**23.945**
3	GPTQ+LoRA	34.790	35.428	50.511	93.806	59.233	19.300	21.063
3	QA-LoRA	27.771	28.751	24.594	74.564	34.648	19.848	6.930
3	QBLoRA	29.401	31.658	34.336	88.094	43.764	19.228	17.251
3	GPTQ-intrinsic	35.925	36.371	50.391	93.626	59.130	19.698	21.305
3	TAQI	**36.797**	**37.863**	**52.255**	**94.588**	**60.618**	**19.906**	**23.865**
2	GPTQ+LoRA	18.277	24.352	25.376	77.210	34.737	19.711	8.323
2	QA-LoRA	21.623	25.295	0.060	47.805	0.265	**19.848**	1.301
2	QBLoRA	17.220	23.488	0.241	47.865	0.676	**19.848**	0.923
2	GPTQ-intrinsic	**21.880**	24.038	20.024	69.573	29.081	19.779	4.494
2	TAQI	10.855	**27.730**	**38.845**	**93.265**	**48.190**	**19.848**	**21.457**

Table 6 shows that TAQI has the highest MedQA accuracy, TAT-QA F1, and QASPER answer F1 at both two and three bits. Relative to GPTQ-intrinsic, these three metrics differ by $+3.692$, $+19.109$, and $+16.963$ points at two bits, and $+1.492$, $+1.488$, and $+2.560$ points at three bits. At 2 bits, TAQI's macro-F1 is 10.855 alongside its 27.730 accuracy, reflecting the distinction between accuracy and class-balanced performance. At four bits, among the ranked methods, TAQI leads TAT-QA F1 (63.957) and QASPER answer F1, while GPTQ+LoRA leads MedQA accuracy. Relative to GPTQ-intrinsic at four bits, the three primary metrics differ by -2.985, $+0.668$, and $+0.766$ points. The four-bit QBLoRA TAT-QA run additionally reports operator accuracy of 0.000, a diagnostic not included in the table.

A.4 TRAINING-LOSS TRAJECTORIES ON GRANITE

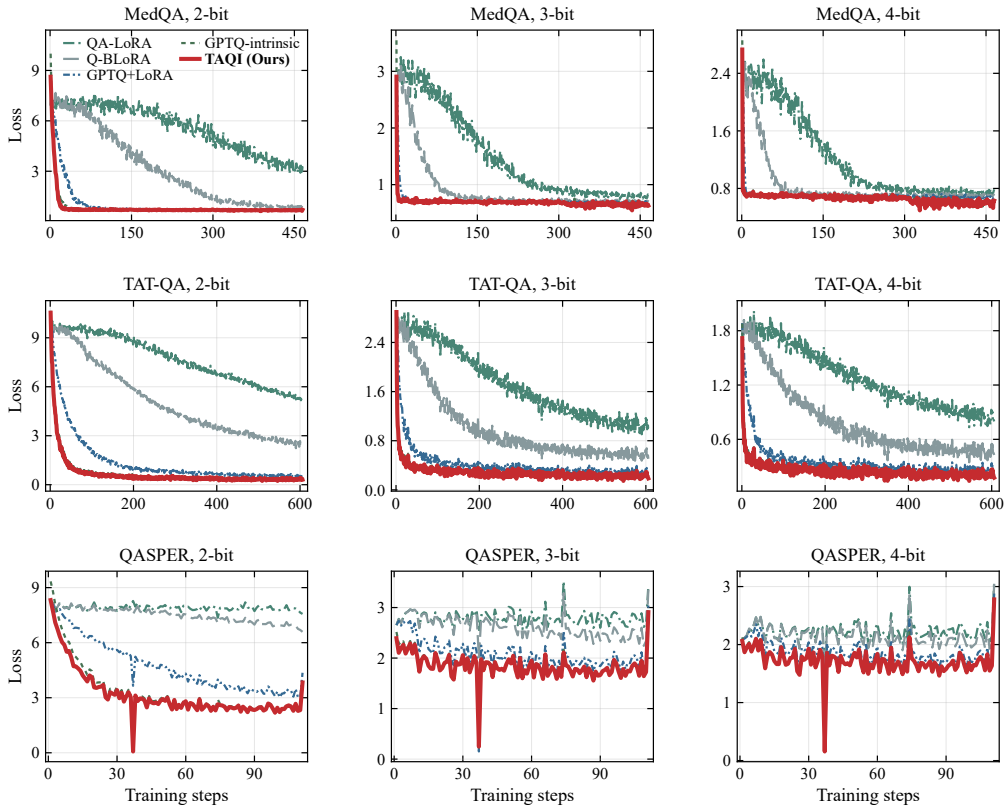


Figure 4: Training-loss trajectories on Granite-4.0-H-350M (seed 42). Rows show MedQA, TAT-QA, and QASPER; columns show 2-, 3-, and 4-bit quantization. The five compared methods are shown at each bit width. Loss is the sample mean of each example's mean cross-entropy over supervised target tokens, measured before the optimizer update indexed on the horizontal axis. All 45 native-step traces are unsmoothed traces. Loss normalization and axis ranges are task-specific.

Figure 4 complements the downstream results with optimization trajectories over 465, 606, and 111 steps for MedQA, TAT-QA, and QASPER, respectively. TAQI and GPTQ-intrinsic show similarly rapid early loss decreases on MedQA and TAT-QA, reaching lower losses than QA-LoRA and Q-BLoRA. Their trajectories remain close. QASPER at 3 and 4 bits shows pronounced step-to-step fluctuations, including a shared dip at step 37 and a terminal spike.

The trajectories complement held-out task metrics. At 2 bits on QASPER, TAQI's training loss is below GPTQ-intrinsic's at 110 of 111 steps, and its answer F1 is also higher, 21.695 versus 18.418 (Table 1).

B INITIALIZATION-LEVEL DIAGNOSTIC DETAILS

Figure 5 visualizes the reallocation summarized in Section 4.4. The relative changes of D_T and Γ reflect different quantities: D_T uses scalar gradient energy, whereas Γ also depends on output-error alignment with the task-gradient direction. Layer 25 illustrates the distinction: relative to XJ, TJ lowers D_T by 52.16% but Γ by 32.87%; layer 6 instead lowers them by 60.40% and 70.56%.

Diagnostic design. The diagnostic examines three shared-MLP projections before fine-tuning, using all nonpadding positions and common sequence-length gradient normalization. TAQI instead

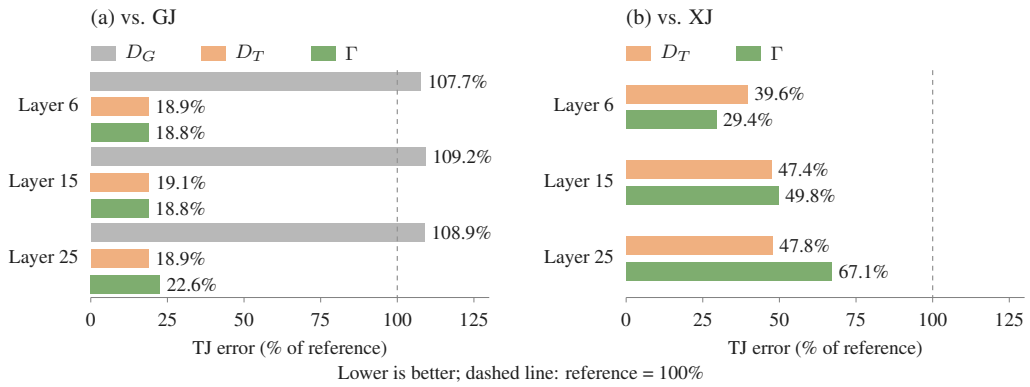


Figure 5: Initialization-level error reallocation on three preselected Granite layers under the common diagnostic protocol. Reference error is 100%; lower is better. TJ trades higher D_G for lower D_T and Γ. Table 7 reports the relative changes at the reported precision.

uses the main adapter modules, a mixed basis, selected shifted-response positions, sample-local mean losses, and the protocol in Section 4.1. The observed reductions therefore provide local evidence for gradient-weighted error allocation in this simplified setting; the downstream component ablation in Section 4.3 evaluates the full pipeline.

Table 7: Initialization-level error reallocation on three preselected Granite-4.0-H-350M projections for TAT-QA. Values are relative changes $100(\mathrm{TJ}/\mathrm{reference} - 1)$ of the static task-guided initialization (TJ); negative values indicate lower error.

Layer	ΔD_G vs. GJ	ΔD_T vs. GJ	$\Delta\Gamma$ vs. GJ	ΔD_T vs. XJ	$\Delta\Gamma$ vs. XJ
6	$+7.71\%$	-81.08%	-81.19%	-60.40%	-70.56%
15	$+9.20\%$	-80.95%	-81.19%	-52.57%	-50.23%
25	$+8.94\%$	-81.11%	-77.40%	-52.16%	-32.87%

Methods and quantities. We use the GJ/XJ/TJ controls defined in Appendix E.6 and the design in Section 4.4. The diagnostic includes prompt positions and assigns zero gradients to final positions. Its gradients g_i^{diag} use supervised cross-entropy sums with common normalization by T_{all}. Explicitly, $g_i^{\mathrm{diag}} = T_{\mathrm{all}}^{-1}\partial\mathcal{L}_{\mathrm{CE,sum}}/\partial z_i$. For N_T nonpadding positions, define

$$s_i^{\mathrm{diag}} = \|g_i^{\mathrm{diag}}\|_2^2/d_{\mathrm{out}}, \qquad H_T^{\mathrm{diag}} = \frac{\sum_i s_i^{\mathrm{diag}} x_i x_i^\top}{\sum_i s_i^{\mathrm{diag}} + N_T\epsilon}, \tag{12}$$

$$\Gamma^{\mathrm{diag}}(E) = \frac{1}{N_T}\sum_i [(g_i^{\mathrm{diag}})^\top E^\top x_i]^2, \tag{13}$$

$$U^{\mathrm{diag}}(E) = d_{\mathrm{out}}(\bar{s}^{\mathrm{diag}} + \epsilon)\|E\|^2_{H_T^{\mathrm{diag}}}. \tag{14}$$

where $\bar{s}^{\mathrm{diag}} = N_T^{-1}\sum_i s_i^{\mathrm{diag}}$. Cauchy–Schwarz gives $\Gamma^{\mathrm{diag}} \le U^{\mathrm{diag}}$. In Table 7, D_T and Γ abbreviate $\|E\|^2_{H_T^{\mathrm{diag}}}$ and Γ^{diag}. $D_G = \|E\|^2_{H_G}$. All comparisons are within one layer on common probe statistics.

Protocol. Construction uses 128 general and 128 task examples, giving 262,144 general and 72,013 task activation positions per layer. The probe uses 32 general and 32 task examples with 17,196 valid task activation positions; probe sequences are padded to $T_{\mathrm{all}} = 961$ positions, prompt positions are included, and the 32 final positions have zero gradient. Construction/probe separation is defined at the shared table/text-component level, the granularity of the available identifiers. A single pre-adaptation construction is analyzed per method.

Normalization and relative comparisons. For a common probe layer, $\epsilon \geq 0$ in the diagnostic convention rescales H_T^{diag} by a common positive scalar whenever total gradient energy is positive. This scalar cancels in the within-layer ratios in Table 7 and in trace matching during construction. Entries report within-layer relative errors; Γ^{diag} uses the specified gradients directly.

Concentration of gradient weights. The task-gradient effective sample sizes over $17{,}196$ valid activation positions are 28.519, 459.447, and 273.461 for layers 6, 15, and 25, respectively. These are token-weight concentration measures, with the strongest concentration in the early layer.

C Full results of the adaptation-precision study

Benchmark references. Medical: MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), and PubHealth (Kotonya & Toni, 2020); scientific: QASPER (Dasigi et al., 2021); financial: TAT-QA (Zhu et al., 2021), MultiHiertt (Zhao et al., 2022), FinQA (Chen et al., 2021b), and ConvFinQA (Chen et al., 2022); security: AttackQA (Krishna, 2024); code: HumanEval (Chen et al., 2021a); math: GSM8K (Cobbe et al., 2021). The legal task definitions for ECtHR-A, ECtHR-B, and SCOTUS are described in LexGLUE (Chalkidis et al., 2022).

Table 8 lists the configurations for the precision comparison and an additional KD control (Section C.2). All LoRA configurations share the following within-task training protocols. On the 12 domain tasks, fine-tuning uses 3 epochs, learning rate 10^{-5}, and an effective batch size of 64, and checkpoint selection uses at most 128 validation examples. HumanEval and GSM8K use training sources that differ from those of the domain tasks. All values are test-set results. Ratio metrics are percentages; Invalid is lower-is-better, and all other metrics are higher-is-better. Values without $\pm$ are single runs with seed 42; $\pm$ denotes the mean and sample standard deviation over seeds 42, 43, and 44. In primary-metric rows, the best value among the three fine-tuned configurations is bold; the unadapted backbones are not ranked.

Figure conventions. Figure 1 compares the non-KD configurations, with NF4 for all 4-bit backbones. Figure 1 (b) uses $\Delta_{\mathrm{adapt}} = \mathrm{Score}(\text{BF16+LoRA}) - \mathrm{Score}(\text{BF16 base})$ and $\Delta_{\mathrm{post}} = \mathrm{Score}(\text{4-bit+LoRA}) - \mathrm{Score}(\text{BF16+LoRA})$, computed from the four-decimal primary scores in Tables 10–12. BF16+LoRA is low-rank adaptation, not full-parameter fine-tuning. Stars on FinQA, ConvFinQA, and AttackQA indicate adapted-score means over seeds 42, 43, and 44; other adapted scores and all unadapted scores use seed 42. All scatter markers are filled; no region denotes rank reversals. MultiHiertt, FinQA, and ConvFinQA have mostly invalid unadapted outputs, so their large adaptation gains can include learning valid output formats.

Table 8: Configurations for the adaptation-precision comparison and additional KD control. The KD teacher is high-precision and fully fine-tuned on the corresponding task; BF16+LoRA is only a comparator.

Configuration	Backbone	Task fine-tuning	Teacher
BF16 base	BF16	None	None
4-bit base	NF4	None	None
BF16+LoRA	BF16	LoRA ($r = 16$, $\alpha = 32$)	None
4-bit+LoRA	NF4	LoRA ($r = 16$, $\alpha = 32$)	None
4-bit+LoRA+KD	NF4	LoRA + CE + KD	High-precision full-parameter FT

Table 9: Adaptation-precision study on Qwen2.5-0.5B-Instruct (one primary metric per benchmark, in %). BF16+LoRA, 4-bit+LoRA, and 4-bit+LoRA+KD are fine-tuned with rank 16 and $\alpha = 32$. The backbone-only gap Δ_{pre} is the unadapted 4-bit backbone minus the unadapted BF16 backbone; the LoRA-adapted gap Δ_{post} is 4-bit+LoRA minus BF16+LoRA, in bold when positive. All 4-bit backbones use NF4. Values with $\pm$ are means and sample standard deviations over seeds 42/43/44; other fine-tuned results are single runs with seed 42. Figure 1 uses only the non-KD configurations; the KD control is analyzed separately in Section C.2. Tables 10–12 report all metrics.

Domain	Benchmark	Primary metric	BF16+LoRA	4-bit+LoRA	4-bit+LoRA+KD	Δ_{pre}	Δ_{post}
Medical	MedQA	Accuracy	38.26	38.10	38.49	+0.16	−0.16
	PubMedQA	Accuracy	51.00	53.60	53.60	−2.60	**+2.60**
	PubHealth	Accuracy	58.00	58.33	57.68	+26.89	**+0.32**
Scientific	QASPER	Answer F1	28.44	28.36	28.15	−2.72	−0.08
Financial	TAT-QA	F1	33.90	31.04	33.01	−10.57	−2.86
	MultiHiertt	Answer Acc.	18.68	21.26	19.16	−1.05	**+2.59**
	FinQA	Answer/Exec. Acc.	27.40±0.66	29.18±1.29	27.67±1.08	−1.31	**+1.77**
	ConvFinQA	Answer/Exec. Acc.	36.96±0.54	42.24±0.57	36.17±1.00	−2.28	**+5.28**
Legal	ECtHR-A	Macro-F1	18.41	27.15	21.23	+3.61	**+8.74**
	ECtHR-B	Macro-F1	18.59	24.24	23.95	+5.77	**+5.65**
	SCOTUS	Accuracy	15.50	11.57	16.14	−3.00	−3.93
Security	AttackQA	Token/Answer F1	81.86±0.12	82.34±0.20	82.02±0.12	−12.87	**+0.47**
Code	HumanEval	Pass@1	24.76	18.91	20.10	−6.91	−5.84
Math	GSM8K	Accuracy	34.50	28.51	26.84	−13.04	−5.99
Benchmarks on which 4-bit+LoRA outperforms BF16+LoRA							8/14

Table 10: Full results on medical, scientific, and financial benchmarks (single runs, seed 42). TAT-QA operator accuracy is 0 for all configurations and is omitted. In this summary, QASPER answer F1 equals its token F1 and F1, and MultiHiertt answer accuracy equals its execution accuracy.

Benchmark	Metric	BF16 base	4-bit base	BF16+LoRA	4-bit+LoRA	4-bit+LoRA+KD
MedQA	**Accuracy** ↑	32.7573	32.9144	38.2561	38.0990	**38.4918**
	Macro-F1 ↑	31.4725	30.9895	37.3555	37.1294	37.4260
	Invalid ↓	0.7855	2.1995	0	0	0
PubMedQA	**Accuracy** ↑	40.4000	37.8000	51.0000	**53.6000**	**53.6000**
	Macro-F1 ↑	27.3641	25.9812	27.6382	28.2292	30.3777
	Invalid ↓	0	0	0	0	0
PubHealth	**Accuracy** ↑	15.9220	42.8107	58.0016	**58.3266**	57.6767
	Macro-F1 ↑	15.4139	28.0228	32.4259	34.5150	32.6558
	Invalid ↓	0	0	0	0	0
QASPER	**Answer F1** ↑	17.2084	14.4859	**28.4360**	28.3591	28.1500
	EM ↑	0.4135	0	13.4390	13.7147	14.3349
	Evidence F1 ↑	0	0	1.9516	0.4683	2.0294
TAT-QA	**F1** ↑	18.5009	7.9266	**33.9032**	31.0409	33.0102
	EM ↑	10.7035	1.6236	24.9549	22.5496	24.5340
	Scale Acc. ↑	50.8719	49.9098	78.4726	69.3927	80.7577
MultiHiertt	**Answer Acc.** ↑	1.0536	0	18.6782	**21.2644**	19.1571
	Program Acc. ↑	0	0	18.4866	21.0728	18.6782
	Invalid ↓	99.9042	99.9042	9.0038	2.9693	10.3448

Table 11: Benchmarks with three seeds. Unadapted backbones are single runs with seed 42; fine-tuned configurations report the mean ± sample standard deviation over seeds 42/43/44. AttackQA evidence F1 is 0 for all configurations.

Benchmark	Metric	BF16 base	4-bit base	BF16+LoRA	4-bit+LoRA	4-bit+LoRA+KD
FinQA	**Answer/Exec. Acc.** ↑	1.8309	0.5231	27.4048±0.6601	**29.1776±1.2853**	27.6664±1.0831
	Program Acc. ↑	0	0	24.6731±0.6103	26.4458±1.2523	25.3124±0.8601
	Invalid ↓	99.9128	99.8256	1.8599±0.7863	3.4583±0.1815	2.4993±0.4301
ConvFinQA	**Answer/Exec. Acc.** ↑	2.4161	0.1342	36.9575±0.5425	**42.2371±0.5708**	36.1745±1.0022
	Program Acc. ↑	0	0	34.2058±0.2357	39.7763±0.4367	33.7136±0.5629
	Invalid ↓	95.5034	99.5302	9.1499±0.6379	1.4318±0.6379	10.3132±0.4714
AttackQA	**Token/Answer F1** ↑	40.1852	27.3124	81.8647±0.1187	**82.3396±0.2044**	82.0163±0.1178
	EM ↑	3.2604	1.5109	38.2638±0.3694	39.6952±0.3564	38.5288±0.1590
	BERTScore F1 ↑	83.7680	79.8784	95.1895±0.0127	95.3691±0.0237	95.2246±0.0328
	Technical-ID F1 ↑	73.6478	66.8967	85.1784±0.0768	85.7183±0.1047	85.4141±0.0831
	Evidence F1 ↑	0	0	0	0	0

Table 12: Full results on legal, code, and math benchmarks (single runs, seed 42). GSM8K #Correct and #Invalid are counts, not percentages. HumanEval uses 164 problems with 100 samples each, so its Pass@1 is a sampling estimate rather than a single greedy-decoding result. HumanEval and GSM8K use training sources that differ from the domain tasks.

Benchmark	Metric	BF16 base	4-bit base	BF16+LoRA	4-bit+LoRA	4-bit+LoRA+KD
ECtHR-A	**Macro-F1** ↑	11.7567	15.3672	18.4102	**27.1526**	21.2254
	Micro-F1 ↑	13.1238	16.2355	40.0373	45.2009	40.7390
	EM ↑	2.9000	2.9000	26.3000	32.3000	27.4000
ECtHR-B	**Macro-F1** ↑	14.1613	19.9263	18.5942	**24.2400**	23.9507
	Micro-F1 ↑	16.0121	21.3471	41.9405	44.6775	45.4002
	EM ↑	2.5000	1.9000	23.2000	26.4000	26.2000
SCOTUS	**Accuracy** ↑	5.1429	2.1429	15.5000	11.5714	**16.1429**
	Macro-F1 ↑	3.3473	2.3547	5.7181	5.1285	6.0748
	Invalid ↓	68.1429	67.8571	23.5000	26.0714	16.2143
HumanEval	**Pass@1** ↑	23.8659	16.9512	**24.7561**	18.9146	20.1037
	Pass@10 ↑	38.3958	30.0950	40.6500	31.3929	32.0417
	Pass@100 ↑	51.2195	39.0244	54.8780	45.1220	45.7317
GSM8K	**Accuracy** ↑	35.1782	22.1380	**34.4958**	28.5064	26.8385
	#Correct ↑	464	292	455	376	354
	#Invalid ↓	0	2	1	0	0

C.1 Analysis of the adaptation-precision study

Study design. These results examine finite-budget adaptation with independent NF4 quantization and LoRA on Qwen2.5-0.5B-Instruct. The horizontal axis of Figure 1(b) measures the observed BF16 LoRA adaptation gain under each task's training protocol. The two axes share the BF16+LoRA score with opposite signs and are mathematically coupled; the scatter summarizes task-wise outcomes on their respective primary-metric scales. BF16+LoRA and quantized LoRA both train rank-r updates on frozen backbones, whose expressible families are described in Appendix D.2. The rank-reversal analysis below also identifies tasks with mostly invalid base outputs.

Post-adaptation precision rankings. 4-bit+LoRA exceeds BF16+LoRA on 8 of the 14 primary metrics. The largest gains appear on legal classification (ECtHR-A $+8.74$ and ECtHR-B $+5.65$

macro-F1) and financial numerical reasoning (ConvFinQA $+5.28$ and MultiHiertt $+2.59$). Among the three benchmarks with three seeds, the ConvFinQA and AttackQA gains exceed the sum of the two sample standard deviations, whereas the FinQA gain ($+1.77$) does not. BF16+LoRA leads on the other six benchmarks, although by less than 0.2 points on MedQA and QASPER. Its clearest advantages are on HumanEval (5.84 Pass@1 points) and GSM8K (5.99 accuracy points). On these two benchmarks, BF16+LoRA changes the score of the unadapted BF16 backbone by only $+0.89$ and -0.68 points, respectively, compared with $+5.50$ on MedQA, $+15.40$ on TAT-QA, and $+11.23$ on QASPER. These within-task adaptation gains motivate the selected domain benchmarks.

Pre- and post-adaptation ranking reversals. The backbone-only gap Δ_{pre} (4-bit base minus BF16 base) in Table 9 measures the unadapted task-score change; Figure 1(b) places the adapted precision gap alongside the BF16 LoRA adaptation gain. Quantization lowers the unadapted score on 10 of the 14 benchmarks, yet 4-bit+LoRA still leads after adaptation on five of them; we call these cases rank reversals. On AttackQA, for example, quantization lowers token F1 by 12.87 points before adaptation, whereas 4-bit+LoRA exceeds BF16+LoRA by 0.47 points after adaptation; on PubMedQA, a 2.60-point loss becomes a 2.60-point gain. MultiHiertt, FinQA, and ConvFinQA have mostly invalid unadapted outputs, making their base-score gaps primarily measures of output validity. In contrast, PubMedQA and AttackQA exhibit reversals with informative base predictions.

Implication. The task-dependent rankings motivate evaluating initializations through finite-budget adaptation. Proposition 1 supplies the corresponding distinction between reconstruction and adaptation criteria.

C.2 Additional Control: Quantized Fine-Tuning with Knowledge Distillation

Question and setup. We examine how task-adapted teacher supervision changes quantized LoRA performance. The student uses an NF4 backbone with trainable LoRA adapters; the high-precision teacher is obtained by full-parameter fine-tuning on the corresponding target task. BF16+LoRA serves as the evaluation comparator.

We supplement task supervision with answer-token KL distillation at temperature 2:

$$\mathcal{L}_{\text{student}} = 0.5\,\mathcal{L}_{\text{CE}} + 0.5\,\mathcal{L}_{\text{KD}}. \tag{15}$$

The student uses the same LoRA rank $r = 16$ and scale $\alpha = 32$ as the non-distilled quantized configuration. The teacher's standalone scores and training hyperparameters are unreported, limiting assessment of teacher quality and reproducibility of this control.

Table 13: Additional KD control: primary-metric gaps in points. $\Delta_Q = \text{Score}(\text{4-bit+LoRA}) - \text{Score}(\text{BF16+LoRA})$; $\Delta_K = \text{Score}(\text{4-bit+LoRA+KD}) - \text{Score}(\text{BF16+LoRA})$; $\Delta_{\text{KD}} = \Delta_K - \Delta_Q$ is the student improvement from KD. Positive values favor the first configuration in each difference. Values are rounded from four decimals to two; stars mark means over seeds 42/43/44, with seed 42 elsewhere. Raw scores remain in Tables 10–12.

Task	Δ_Q vs. BF16+LoRA	Δ_K vs. BF16+LoRA	Δ_{KD}
GSM8K	-5.99	-7.66	-1.67
HumanEval	-5.84	-4.65	$+1.19$
SCOTUS	-3.93	$+0.64$	$+4.57$
TAT-QA	-2.86	-0.89	$+1.97$
MedQA	-0.16	$+0.24$	$+0.39$
QASPER	-0.08	-0.29	-0.21
PubHealth	$+0.33$	-0.32	-0.65
AttackQA*	$+0.47$	$+0.15$	-0.32
FinQA*	$+1.77$	$+0.26$	-1.51
MultiHiertt	$+2.59$	$+0.48$	-2.11
PubMedQA	$+2.60$	$+2.60$	0.00
ConvFinQA*	$+5.28$	-0.78	-6.06
ECtHR-B	$+5.65$	$+5.36$	-0.29
ECtHR-A	$+8.74$	$+2.82$	-5.93

Results. Table 13 summarizes the task-wise effects. KD improves the primary metric over ordinary 4-bit+LoRA on 4 of 14 benchmarks, ties on 1, and lowers it on 9. Among the six benchmarks on which ordinary 4-bit+LoRA trails BF16+LoRA, KD improves MedQA, TAT-QA, SCOTUS, and HumanEval. It closes the observed deficit on MedQA and SCOTUS, while TAT-QA and HumanEval remain below BF16+LoRA. The deficits on QASPER and GSM8K instead increase. Some changes are small, including the 0.39-point improvement on MedQA and the 0.21-point decrease on QASPER. Conversely, on the eight benchmarks where ordinary 4-bit+LoRA already leads, KD lowers the primary metric on seven and ties on PubMedQA. ConvFinQA shows the largest decrease relative to the non-distilled student, at 6.06 points. Under this KD recipe, additional supervision closes two of the six observed deficits relative to BF16+LoRA, while the remaining four persist.

Interpretation. This control examines training-time supervision, complementing the study of the quantized backbone and low-rank initialization from which adaptation begins. Its task-wise outcomes characterize answer-token KL distillation with the stated teacher, temperature, and loss mixture. Because KD changes the training objective, Figure 1 uses the non-KD configurations for its precision comparison. FinQA, ConvFinQA, and AttackQA report means over seeds 42/43/44; the remaining fine-tuned results use seed 42. Differences are descriptive, evaluated separately on each task's metric.

D Derivations for the method section

D.1 Details of the design objective

Initialization variables. The base objective uses $I = (Q, \Phi_0)$; $\mathcal{I}_{b,r}$ fixes the quantization format, bit width, group size, rank, and scale α. The risk on P_T evaluates generalization, while $\mathcal{L}_{\mathrm{tr}}$ is empirical token cross-entropy. Calibration examples come from the training side, separately from the test sets used for reporting. Appendix E.4 specifies the implemented extension $I = (Q, \Phi_0, m)$ and its restricted expressible family.

Why the budget must be finite. Replacing $\mathcal{U}_K$ by an exact minimizer $\Phi^\star(Q) \in \arg\min_\Phi \mathcal{L}_{\mathrm{tr}}(\Phi; Q)$ with initialization-independent minimizer selection would remove Φ_0 from the objective. Finite-budget evaluation retains the optimizer, its states, the initial factors, and the number of updates. This connects the design problem to optimization-based meta-learning and bilevel optimization (Finn et al., 2017; Franceschi et al., 2018). Both factors train inside $\mathcal{U}_K$; the construction basis specifies their initial values.

D.2 Reconstruction versus adaptation

The backbone changes the reachable model set. For the base adaptation task with fixed Q, rank-r updates express

$$\mathcal{M}_r(Q) = \{Q + D : \mathrm{rank}(D) \le r\}. \tag{16}$$

Changing Q therefore translates the entire set of reachable models, not merely the initial error. BF16+LoRA corresponds to $\mathcal{M}_r(W)$ and quantized adaptation to $\mathcal{M}_r(Q)$; numerical precision alone does not establish an inclusion between them. In a linear squared-loss model with $H \succ 0$ and a hypothetical task target $W_T^\star$,

$$\inf_{\mathrm{rank}(D)\le r} \|W_T^\star - Q - D\|_H^2 = \sum_{j>r} \sigma_j^2\big(H^{1/2}(W_T^\star - Q)\big) \tag{17}$$

(Appendix D.3). The attainable fit is governed by $W_T^\star - Q = (W_T^\star - W) + (W - Q)$, the sum of the task shift and the quantization shift. The hypothetical target $W_T^\star$ describes this geometry; TAQI's reconstruction target remains the pretrained W.

A finite-step counterexample (proof of Proposition 1). Let $W = I_2$, $W_T^\star = \frac{2}{3} I_2$, and $\mathcal{R}(M) = \frac{1}{2}\|M - W_T^\star\|_F^2$, with the 2-bit grid $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$, $r = 1$, $c = 1$, and $L = e_1$. We compare $Q_a = I_2$ and $Q_b = \frac{2}{3} I_2$, give each its optimal fixed-basis reconstruction coordinates, and apply the same single gradient-descent step with learning rate 1 to both LoRA factors. Table 14 reports the result (Appendix D.4): after the step, the effective weights are $\mathrm{diag}(\frac{2}{3}, 1)$ and $\frac{2}{3} I_2$, respectively. With

identical bit width, rank, basis, training loss, optimizer, and budget, the initialization with zero reconstruction error has higher post-adaptation risk. This establishes the claimed separation between the two criteria. □

Table 14: Two-dimensional counterexample: reconstruction-optimal initialization is not adaptation-optimal under an identical one-step budget.

Initialization	R	Reconstruction error $\|W - Q - LR\|_F^2$	Risk after one step
$Q_a = I_2$	0	0	$1/18$
$Q_b = \frac{2}{3} I_2$	$\frac{1}{3} e_1^\top$	$1/9$	0

Factor initialization changes the training path. Even with the same initial effective weight, the individual factors matter. For a single layer $M = Q + cA^\top B^\top$ with $G = \nabla_M \mathcal{L}$, one gradient-descent step with step size η on both factors gives

$$M^+ - M = -\eta c^2 \big(A^\top A\, G + G\, BB^\top\big) + O(\eta^2) \tag{18}$$

(Appendix D.5). The update depends on the Gram matrices of the two factors, not only on their product: the rescaling $A \mapsto tA$, $B \mapsto B/t$ leaves the initial effective weight unchanged but alters the update. This is why (2) retains $\Phi_0 = (A_0, B_0)$ rather than scoring only the reconstruction error of the initialized model; LoRA-GA studies adapter initialization from the same initial-update perspective (Wang et al., 2024). This calculation describes plain gradient descent.

D.3 Proof of Equation (17)

For $H \succ 0$, $\|E\|_H^2 = \mathrm{tr}(E^\top H E) = \|H^{1/2} E\|_F^2$. Hence $\|W_T^\star - Q - D\|_H^2 = \|H^{1/2}(W_T^\star - Q) - Z\|_F^2$ with $Z = H^{1/2} D$. Because $H^{1/2}$ is invertible, $D \mapsto Z$ is a bijection that preserves rank, so the infimum over $\mathrm{rank}(D) \le r$ equals the infimum over $\mathrm{rank}(Z) \le r$. By the Eckart–Young–Mirsky theorem, the latter equals $\sum_{j>r} \sigma_j^2\big(H^{1/2}(W_T^\star - Q)\big)$. □

D.4 Step-by-step computation of the counterexample

Setting. $W = I_2$, $W_T^\star = \frac{2}{3} I_2$, $\mathcal{R}(M) = \frac{1}{2}\|M - W_T^\star\|_F^2$, grid $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$, $r = 1$, $c = 1$, $L = e_1$. Both $Q_a = I_2$ and $Q_b = \frac{2}{3} I_2$ lie on the grid. With $c = 1$, the LoRA factors are $A = L^\top$ and $B = R^\top$. Writing $a := A^\top \in \mathbb{R}^{2\times 1}$ and $b := B^\top \in \mathbb{R}^{1\times 2}$, the effective weight is $M = Q + ab$, and with $G = \nabla_M \mathcal{R} = M - W_T^\star$ the gradients are $\nabla_a \mathcal{R} = G b^\top$ and $\nabla_b \mathcal{R} = a^\top G$. Both factors are updated simultaneously with learning rate 1.

Reconstruction. For Q_a, $W - Q_a = 0$, so $R_a = 0$ and the reconstruction error is 0. For Q_b, $W - Q_b = \frac{1}{3} I_2$. The term $e_1 R$ can only change the first row, so the optimal coordinates are $R_b = \frac{1}{3} e_1^\top$, leaving the residual $\frac{1}{3} e_2 e_2^\top$ with reconstruction error $\frac{1}{9}$.

One step from Q_a. $a_0 = e_1$, $b_0 = 0$, $M_{a,0} = I_2$, and $G = \frac{1}{3} I_2$. Then $\nabla_a \mathcal{R} = G b_0^\top = 0$ and $\nabla_b \mathcal{R} = e_1^\top G = \frac{1}{3} e_1^\top$, so $a_1 = e_1$ and $b_1 = -\frac{1}{3} e_1^\top$. Hence $M_{a,1} = I_2 - \frac{1}{3} e_1 e_1^\top = \mathrm{diag}(\frac{2}{3}, 1)$ and $\mathcal{R}(M_{a,1}) = \frac{1}{2}\big(\frac{1}{3}\big)^2 = \frac{1}{18}$.

One step from Q_b. $a_0 = e_1$, $b_0 = \frac{1}{3} e_1^\top$, $M_{b,0} = \frac{2}{3} I_2 + \frac{1}{3} e_1 e_1^\top = \mathrm{diag}(1, \frac{2}{3})$, and $G = \frac{1}{3} e_1 e_1^\top$. Then $\nabla_a \mathcal{R} = G b_0^\top = \frac{1}{9} e_1$ and $\nabla_b \mathcal{R} = e_1^\top G = \frac{1}{3} e_1^\top$, so $a_1 = \frac{8}{9} e_1$ and $b_1 = 0$. Hence $M_{b,1} = \frac{2}{3} I_2$ and $\mathcal{R}(M_{b,1}) = 0$.

Remark. Before the step, the task risks are $\mathcal{R}(M_{a,0}) = \frac{1}{9}$ and $\mathcal{R}(M_{b,0}) = \frac{1}{18}$, so Q_b is already better on the task at initialization. Reconstruction fidelity and task risk thus favor different candidates both before and after the same adaptation step.

D.5 Derivation of Equation (18)

For $M = Q + cA^\top B^\top$, we have $\mathrm{d}M = c\,(\mathrm{d}A)^\top B^\top + c\,A^\top(\mathrm{d}B)^\top$, so $\nabla_A \mathcal{L} = c\,B^\top G^\top$ and $\nabla_B \mathcal{L} = c\,G^\top A^\top$. After one step, $(A^+)^\top = A^\top - \eta c G B$ and $(B^+)^\top = B^\top - \eta c A G$. Therefore

$$M^+ - M = c\big[(A^+)^\top (B^+)^\top - A^\top B^\top\big] = -\eta c^2\big(A^\top A\,G + G\,BB^\top\big) + \eta^2 c^3\,GBAG.$$

Under $A \mapsto tA$ and $B \mapsto B/t$, the product $A^\top B^\top$ is unchanged, whereas $A^\top A \mapsto t^2 A^\top A$ and $BB^\top \mapsto t^{-2} BB^\top$, so the first-order update changes. □

D.6 Proof of Equation (8)

For each selected position, Cauchy–Schwarz gives $[(g^l_{np})^\top E_l^\top x^l_{np}]^2 \le \omega^l_{np} \|E_l^\top x^l_{np}\|_2^2$. Summing over the selected sets,

$$\Gamma^l_{\mathcal{S}}(E_l) \le \frac{1}{N^l_s}\,\mathrm{tr}\left(E_l^\top \left[\sum_{n,p\in\mathcal{S}^l_n} \omega^l_{np} x^l_{np} (x^l_{np})^\top\right] E_l\right) = \frac{Z^l_s}{N^l_s}\|E_l\|^2_{H^l_T}.$$

The gradient-energy convention $\omega = \|g\|_2^2$ gives the coefficient Z^l_s/N^l_s directly. □

D.7 Profiling out the low-rank coordinates

Condition on the collected statistics. Fix one layer, its mixed basis $L = V_r(H)$, a positive semidefinite scan matrix H, and $\lambda > 0$, and let $D = W - Q$. The regularized surrogate is $\mathcal{S}(Q,R) = \|D - LR\|_H^2 + \lambda\|D\|_F^2 + \lambda\|R\|_F^2$. With $K_R = L^\top H L + \lambda I_r \succ 0$, $\mathcal{S}$ is strictly convex in R, and setting its gradient $2K_R R - 2L^\top H D$ to zero gives $R^\star = K_R^{-1} L^\top H D$. Completing the square,

$$\mathcal{S}(Q,R) = \mathrm{tr}\big(D^\top H_{\mathrm{eff}} D\big) + \mathrm{tr}\big((R-R^\star)^\top K_R (R-R^\star)\big), \qquad H_{\mathrm{eff}} = H + \lambda I - HLK_R^{-1}L^\top H,$$

so $\min_R \mathcal{S}(Q,R) = \|W-Q\|^2_{H_{\mathrm{eff}}}$. □

E Implementation details

E.1 Statistics collection

Per-example teacher-forced model losses use sample-local normalization. At each target Linear, $g_{np} = \partial \ell_n / \partial z_{np}$ and $\omega_{np} = \|g_{np}\|_2^2$. Eligible positions satisfy $\mathrm{labels}[n, p+1] \neq -100$ and $\omega_{np} > 0$; the final position is excluded. If more than 32 qualify, 32 are selected deterministically, approximately evenly spaced in positional order. Sets $\mathcal{S}_n$ are layer-specific. Accumulation uses $Z_s = \max\{\sum \omega_{np}, 1\}$ with summed output-gradient energy. H_G is the full uncentered second moment. All statistics come from the dense BF16 reference and remain fixed.

Trace matching uses the ratio of positive traces in (9). The implementation's relative weight $\beta_{\mathrm{code}} = 1$ forms $(H_G + \beta_{\mathrm{code}}\bar{H}_T)/(1+\beta_{\mathrm{code}})$; the normalized manuscript convention is therefore $\beta = \beta_{\mathrm{code}}/(1+\beta_{\mathrm{code}}) = 0.5$.

E.2 Numerical domain and reproducibility

The construction assumes finite symmetric positive semidefinite H_G, H_T with strictly positive traces, finite source weights and grids, nonzero source scales, $1 \le r \le d_{\mathrm{in}}$, and $0 \le \beta \le 1$. Selected-position diagnostics additionally require $N_s > 0$. The normalizer $Z_s = \max\{\sum \omega, 1\}$ sets the accumulation scale; trace matching requires a finite positive task trace independently.

Positive trace gives $\lambda > 0$ in (19) and $K \succ 0$ in exact arithmetic. Numerical evaluation requires finite inverse factorizations and well-defined scan divisions. The notation $V_r(H)$ permits any orthonormal leading eigenspace: repeated eigenvalues allow basis rotations, and a tie at the rank cutoff allows different retained subspaces. Reproducibility requires the eigensolver and returned basis, or an explicit tie-breaking convention, together with any policy used for out-of-domain statistics or failed numerical factorizations.

E.3 AUGMENTED SCAN, DAMPING, AND EFFECTIVE GEOMETRY

Let $d = d_{\text{in}}$, $L = V_r(H)$, $C = [I_d\ L]$, $W_{\text{aug}} = [W; 0]$, and $\Theta = [Q; R]$. Define

$$H_{\text{aug}} = C^\top H C, \qquad \lambda = 0.01\frac{\operatorname{tr}(H_{\text{aug}})}{d+r}, \qquad K = H_{\text{aug}} + \lambda I_{d+r}. \tag{19}$$

H_{aug} is singular for $r > 0$; damping makes it invertible and gives

$$\|W_{\text{aug}} - \Theta\|_K^2 = \|W - Q - LR\|_H^2 + \lambda\|W - Q\|_F^2 + \lambda\|R\|_F^2. \tag{20}$$

For source input group $g(i)$ and output channel j, feasible entries are $Q_{ij} = s_{g(i),j}(k_{ij} - z_{g(i),j})$ with $k_{ij} \in \{0, \dots, 2^b - 1\}$. Source scales, zeros, and group assignments remain fixed.

Take upper triangular U with $U^\top U = K^{-1}$ and initialize $Z = [W; 0]$. In input-row order, round and clamp $Z_i/s_{g(i)} + z_{g(i)}$ to obtain the dequantized row q_i. Compute $e_i = (Z_i - q_i)/U_{ii}$ and update each future row by $Z_j \leftarrow Z_j - U_{ij}e_i$, $j > i$. The blocked implementation uses 128-row blocks. Only the first d rows are quantized; the final r continuous rows become R_{cand} through the same greedy error-feedback process. The illustrative scan in Figure 3 changes an unprocessed entry from -0.5281 to -0.4281 and its next code from -0.6 to -0.4 on a toy grid of spacing 0.2.

Conditioned on H, L, profiling gives $R^\star = (L^\top H L + \lambda I_r)^{-1} L^\top H (W - Q)$ and

$$H_{\text{eff}} = H + \lambda I_d - HL(L^\top H L + \lambda I_r)^{-1} L^\top H. \tag{21}$$

The exact minimum over R at fixed Q is $\|W - Q\|_{H_{\text{eff}}}^2$ (Appendix D.7).

E.4 SELECTION AND LORA MAPPING

Output-group selection. Output-group screening requires improved task-weighted reconstruction while limiting the increase in general reconstruction error relative to the source GPTQ backbone. For each output-channel group J, compare the candidate error $E_{1,J} = (W - Q_{\text{cand}} - LR_{\text{cand}})_{:,J}$ to the original GPTQ error $E_{0,J} = (W - Q_0)_{:,J}$, without a baseline low-rank correction. Writing $D_G(E) = \|E\|_{H_G}^2$ and $D_T(E) = \|E\|_{H_T}^2$, accept finite candidates precisely when

$$D_G(E_{1,J}) \le (1+\tau) D_G(E_{0,J}), \qquad D_T(E_{1,J}) < D_T(E_{0,J}). \tag{22}$$

Accepted groups retain their codes and coordinates; rejected groups restore Q_0 and set their coordinates to zero. With the resulting groupwise mask $M = \operatorname{diag}(m)$,

$$\widehat{Q} = Q_0 + (Q_{\text{cand}} - Q_0)M, \qquad \widehat{R} = R_{\text{cand}} M. \tag{23}$$

Selection evaluates the raw reference quadratics using the collected BF16 statistics and constructed basis, codes, and coordinates.

Masked low-rank adaptation. The selected support is enforced throughout adaptation by a persistent mask. Set $A_0 = L^\top$, $B_0 = \widehat{R}^\top/c$, and use the row-input forward map $x\widehat{Q} + c(xA^\top)B^\top M$. Both factors train by SFT, but the mask persists: rejected adapter outputs stay zero, while the shared A learns through accepted outputs. The expressible family is therefore $\{\widehat{Q} + DM : \operatorname{rank}(D) \le r\}$.

Source reference and design criterion. The source Q_0 with zero adapter supplies the groupwise reference in (22) and the retained weights for rejected groups. For the implemented configuration $I = (Q, \Phi_0, m)$, the design criterion is $\mathcal{J}_K(Q, \Phi_0; m) = \mathbb{E}_\xi[\mathcal{R}_T(f_{Q,\mathcal{U}_K(\Phi_0;Q,m,\mathcal{D}_{\text{tr}},\xi),m})]$, with m fixed throughout training.

Input quantization groups contain 64 channels. Separately, output selection uses groups of size 64 and tolerance $\tau = 0.05$ in (22).

With $r = 16$, $\alpha = 32$, and $c = 2$, the masked mapping is

$$A_0 = L^\top, \quad B_0 = \widehat{R}^\top/c, \quad x\widehat{Q} + c(xA_0^\top)B_0^\top M = x(\widehat{Q} + L\widehat{R}). \tag{24}$$

It follows from $\widehat{R}M = \widehat{R}$. A layerwise FP32 mapping check with tolerance 10^{-6} observed maximum error 3.95×10^{-7}. SFT optimizes A, B with the stated task loss; codes, scales, zeros, input groups, and output masks remain fixed.

Algorithm 1 TAQI construction and adaptation

Require: BF16 weights, source GPTQ grids and Q_0, general and task data
1: Collect H_G and selected-position H_T on the BF16 reference
2: **for** each target layer **do**
3: Trace-match and mix statistics; set $L \leftarrow V_r(H)$
4: Form $K = C^\top HC + \lambda I$ and run the fixed-grid augmented scan
5: Apply the output-group gate to $(Q_{\mathrm{cand}}, R_{\mathrm{cand}})$
6: Construct $(\widehat{Q}, \widehat{R}, M)$ by (23)
7: Set $A_0 \leftarrow L^\top$, $B_0 \leftarrow \widehat{R}^\top / c$
8: **end for**
9: Freeze backbone, quantization parameters, and masks; train A, B by SFT

E.5 Main-experiment training and protocol scope

Hardware environment. All experiments were run on a single NVIDIA A100 PCIe GPU with 40 GiB of device memory per run; no experiment used multi-GPU execution. The host server is equipped with eight such GPUs, two Intel Xeon Gold 6348 CPUs at 2.60 GHz (56 physical cores and 112 hardware threads in total), and approximately 503.5 GiB of system RAM. It runs Ubuntu 20.04.6 LTS with Linux kernel 5.15.0-86-generic (x86_64).

Training protocol. TAQI SFT uses AdamW, learning rate 2×10^{-5}, weight decay 0.01, three epochs, effective batch size 64, and gradient clipping at 1. The final epoch checkpoint is evaluated. Microbatch sizes may vary with memory needs. Qwen3.5 has 150 adapter targets across hybrid linear/full-attention and MLP projections. Granite has 72: 56 Mamba input/output projections across 28 blocks and 16 query/key/value/output projections across four attention blocks. Granite quantizes 136 modules.

Formal training excludes fold-0 calibration/support/query examples. The 128 fold-1 task-construction examples overlap SFT training data. Granite uses base-model text prompts; Qwen training and generation use model templates. Candidate curvature remains tied to its construction inputs, whose prompt representation can differ from SFT. QASPER results are compared within each backbone: Qwen scoring and Granite inline-gold scoring are distinct evaluation pipelines.

E.6 Initialization diagnostic controls

The GJ, XJ, and TJ controls compare static reconstruction geometries under the design in Section 4.4. All reconstruct the pretrained W_l with the fixed general basis $L = V_{16}(H_G)$. GJ uses the general scan curvature H_G. XJ uses the unweighted target-activation covariance $H_X^l = N_T^{-1} \sum_i x_i^l (x_i^l)^\top$ in the mixed scan curvature $(1-\beta)H_G + \beta \widetilde{H}_X$. TJ instead uses task-gradient-weighted curvature in $(1-\beta)H_G + \beta \widetilde{H}_T$. XJ and TJ use identical task examples, trace matching, β, basis, and scanner, isolating the effect of gradient weighting. Statistics are collected on the unchanged BF16 reference; Appendix B details the diagnostic gradients and probes.

F Additional related work

Quantization replaces high-precision weights or activations with low-bit representations to reduce storage, memory traffic, and, with appropriate kernels, computation. LLM.int8() isolates activation outliers in higher precision to preserve accuracy at eight bits (Dettmers et al., 2022), while SmoothQuant migrates activation difficulty into weights through an equivalent rescaling to enable efficient W8A8 inference (Xiao et al., 2023). At lower weight precision, GPTQ uses calibration activations and approximate second-order information to compensate errors while quantizing weights sequentially (Frantar et al., 2023), and AWQ identifies salient channels from activation magnitude and rescales them before weight-only quantization (Lin et al., 2024). These methods balance low-bit efficiency against preservation of the pretrained function. Our setting treats the quantized model as an initialization for domain adaptation and evaluates it by post-adaptation performance.