GRADIENT SUBSPACE DYNAMICS DURING SUPERVISED FINE-TUNING OF LLMS: INPUT–OUTPUT ASYMMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

The gradient dynamics of supervised fine-tuning (SFT) in large language models (LLMs) remain poorly understood. For a layer $y = Wx$, we call the side of the weight gradient corresponding to the input features x the **input side**, and the side corresponding to the output features y the **output side**. To unveil the gradient dynamics of SFT, we conduct seventeen full-parameter SFT runs using AdamW, spanning eight datasets and models from the Qwen3, Llama, and MiniCPM families. Fifteen runs use diagnostic rank $k = 8$, and two additional runs use $k = 16$. We identify four main findings: input-side energy overlap generally remains high during fine-tuning; output-side energy overlap is generally low, especially outside query/key projections; input-side energy overlap is consistently higher than output-side overlap in model averages and is higher in the vast majority of individual-matrix comparisons; and reversals are concentrated in query/key projections and embedding tables. Across the fifteen $k = 8$ runs, input-side energy overlap exceeds output-side overlap in 93.58% of individual-matrix comparisons with the initial subspace and 92.98% of comparisons with the subspace from the previous measurement. As an application of our findings, we propose a memory-efficient optimizer, Overlap-Guided Gradient Low-Rank Projection (OG-GaLore), which always projects gradients onto input-side subspaces. It achieves the lowest or near-lowest validation loss among the compared projection rules across the different model–rank configurations. Our code is available at `https://anonymous.4open.science/r/gradient_subspace_dynamics-3DD6/`.

1 INTRODUCTION

Post-training improves pretrained large language models' ability to follow instructions and solve reasoning tasks (Ouyang et al., 2022; Shao et al., 2024). Supervised fine-tuning (SFT) updates parameters using gradients, yet how these gradients evolve during training remains poorly understood.

We want to understand gradient dynamics during SFT. For a layer $y = Wx$ with $W \in \mathbb{R}^{m\times n}$, let $G_t \in \mathbb{R}^{m\times n}$ denote its weight gradient. We call the coordinate space $\mathbb{R}^n$, corresponding to x, the **input side**, and the coordinate space $\mathbb{R}^m$, corresponding to y, the **output side**. If $G_t = U_t\Sigma_t V_t^\top$, their top-$k$ subspaces are $\mathrm{span}(V_t[:, 1:k])$ and $\mathrm{span}(U_t[:, 1:k])$, respectively.

To characterize gradient dynamics, we use **energy overlap**, which measures the fraction of the current gradient's top-k energy captured by a historical top-k subspace on the input or output side. We compare the current gradient with subspaces from both the first recorded step and the previous recorded step, characterizing changes relative to the initial reference and between neighboring measurements. This allows us to examine how gradient energy overlap changes over training and whether the input and output sides exhibit different temporal patterns.

We conduct seventeen full-parameter SFT runs using original AdamW (Loshchilov & Hutter, 2019), spanning eight datasets and models from the Qwen3, Llama, and MiniCPM families. Our main analysis uses fifteen runs at $k = 8$; two additional runs at $k = 16$ on Llama-3.2-1B-Instruct and Qwen3-1.7B provide complementary evidence (Appendix H). Our empirical study yields four main findings. **First,** input-side energy overlap generally remains high during fine-tuning. **Second,** output-side en-

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.

ergy overlap is generally low, particularly outside query/key projections. **Third,** input-side energy overlap is consistently higher than output-side energy overlap in model averages and is higher in the vast majority of individual-matrix comparisons. **Fourth,** reversals of the input-side energy-overlap advantage are concentrated in query/key projections and embedding tables. After excluding query/key projections and embedding tables, input-side energy overlap exceeds output-side energy overlap in 99.90–100% of individual-matrix comparisons across five OpenMathReasoning runs, under both initial-reference and local comparisons. These module-specific exceptions do not alter the overall finding that input-side energy overlap is higher than output-side energy overlap.

Standard GaLore selects the projection side according to matrix shape (Zhao et al., 2024). Motivated by the higher input-side energy overlap observed in our empirical findings, we introduce a memory-efficient optimizer, **Overlap-Guided Gradient Low-Rank Projection (OG-GaLore)**, which selects input-side projection based on the observed energy-overlap asymmetry.

Our contributions are threefold:

- **Gradient dynamics during SFT.** Across seventeen runs on eight datasets, we study gradient subspaces at $k = 8$ and $k = 16$. The fifteen $k = 8$ runs reveal high input-side energy overlap, generally low output-side overlap, and input–output asymmetry in model averages and most individual matrices.
- **Module-specific exceptions.** Overlap reversals occur primarily in query/key projections and embedding tables. We discuss why query/key projections can be exceptions: both sides involve input-derived representations. We examine the structure of embedding lookup.
- **An application of our findings.** We propose Overlap-Guided Gradient Low-Rank Projection (OG-GaLore), which always projects onto input-side subspaces. It achieves the lowest or near-lowest validation loss across the evaluated models and ranks.

2 Related Work

Gradient subspace dynamics. Gradient spectra and subspaces provide complementary descriptions of learning structure. Li et al. (2026) use gradient nuclear norms and effective ranks to study instruction and reasoning data quality. WeLore examines gradient subspace similarity across training steps and its relation to low-rank weights, identifying module- and depth-dependent stabilization (Jaiswal et al., 2025). Q-GaLore uses layer-dependent projection stability to adapt subspace refresh intervals (Zhang et al., 2024). More recently, Thomas (2026) examines the sensitivity of estimated gradient subspaces to minibatch sampling and its consequences for optimizer states. For cumulative parameter changes, Cai et al. (2026b) report rank-1 dominance in RL, while Cai et al. (2026a) find low-rank concentration and early subspace alignment in on-policy distillation. We study how historical gradient subspaces retain current dominant energy on the input and output sides.

Input–output asymmetry. LoRA represents parameter changes through two low-rank factors (Hu et al., 2022). LoRA-FA freezes the input projection and trains the output mapping to reduce activation memory (Zhang et al., 2023). Zhu et al. (2024) analyze the distinct feature-extraction and output-adaptation roles of these factors. In an exploratory appendix analysis of Qwen3-4B trained on DeepMath, Ding et al. (2026) compare the cumulative changes $\Delta W_{\mathrm{GRPO}}(t) = W_{\mathrm{GRPO}}(t) - W_0$ and $\Delta W_{\mathrm{OPD}}(t) = W_{\mathrm{OPD}}(t) - W_0$ for corresponding parameter matrices at the same training step t. They report higher cross-method overlap between the top-k input-side subspaces than between the output-side subspaces, using the mean cosine of principal angles with $k = 20$. We examine a different comparison: how much of the current gradient's dominant energy is retained by historical input- and output-side gradient subspaces during supervised fine-tuning. We characterize the persistent input–output difference across training steps, together with its prevalence and exceptions among individual parameter matrices.

Memory-efficient gradient projection. GaLore maintains optimizer states in low-rank gradient coordinates (Zhao et al., 2024), and GaLore 2 reduces projection-update cost through randomized SVD (Su et al., 2025). GoLore analyzes stochastic convergence and introduces random projections (He et al., 2025). LDAdam accounts for changing subspaces through projection-aware state updates and error feedback (Robert et al., 2025); SubTrack++ combines subspace tracking with

state adaptation and gradient recovery (Rajabi et al., 2025). CompAct compresses input activations before forming gradients and optimizer states (Shamshoum et al., 2025), while OASIS learns the activation subspace online (Choudhary et al., 2026). These approaches connect projection design to memory-efficient training. Our findings motivate **Overlap-Guided Gradient Low-Rank Projection (OG-GaLore)**, which selects input-side projection in GaLore to reuse a basis on the side with greater observed energy retention.

3 PRELIMINARIES

3.1 MATRIX PRELIMINARIES

Singular-value decomposition (SVD). Let ℓ index layers, m index individual weight matrices within each layer, and t denote the training step. For each gradient matrix $\mathbf{G}_{\ell,m,t} \in \mathbb{R}^{d_{\ell,m,\mathsf{out}} \times d_{\ell,m,\mathsf{in}}}$, let $Q_{\ell,m} = \min(d_{\ell,m,\mathsf{out}}, d_{\ell,m,\mathsf{in}})$ and write its thin singular-value decomposition (SVD) (Golub & Van Loan, 2013) as

$$\mathbf{G}_{\ell,m,t} = \mathbf{U}_{\ell,m,t}\Sigma_{\ell,m,t}\mathbf{V}_{\ell,m,t}^{\top}, \qquad \Sigma_{\ell,m,t} = \mathrm{diag}(\sigma_{\ell,m,t,1}, \ldots, \sigma_{\ell,m,t,Q_{\ell,m}}), \tag{1}$$

where $\mathbf{U}_{\ell,m,t} \in \mathbb{R}^{d_{\ell,m,\mathsf{out}} \times Q_{\ell,m}}$ and $\mathbf{V}_{\ell,m,t} \in \mathbb{R}^{d_{\ell,m,\mathsf{in}} \times Q_{\ell,m}}$ have orthonormal columns representing, respectively, the output-side (left) and input-side (right) singular directions, and $\sigma_{\ell,m,t,1} \geq \cdots \geq \sigma_{\ell,m,t,Q_{\ell,m}} \geq 0$ are the singular values. For embedding tables, the right and left sides correspond to embedding-feature and vocabulary coordinates, respectively (Section 4.5).

Top-k energy. For a nonzero matrix, top-k energy is the fraction of squared Frobenius norm captured by the top-k singular components:

$$\mathcal{E}_{\ell,m}^{(k)}(t) = \frac{\sum_{i=1}^{k} \sigma_{\ell,m,t,i}^2}{\sum_{i \geq 1} \sigma_{\ell,m,t,i}^2}. \tag{2}$$

Top-k input- and output-side subspaces. For a given rank k ($1 \leq k \leq Q_{\ell,m}$), the top-k output-side (left) and input-side (right) singular subspaces of this matrix have orthonormal bases

$$\mathbf{Q}_{\ell,m,t}^{\mathsf{out}} = \mathbf{U}_{\ell,m,t}[:, 1\!:\!k] \in \mathbb{R}^{d_{\ell,m,\mathsf{out}} \times k}, \qquad \mathbf{Q}_{\ell,m,t}^{\mathsf{in}} = \mathbf{V}_{\ell,m,t}[:, 1\!:\!k] \in \mathbb{R}^{d_{\ell,m,\mathsf{in}} \times k}. \tag{3}$$

3.2 ENERGY OVERLAP

Let $t_0 < t_1 < \cdots$ denote the training steps with subspace measurements. For a current measurement $t = t_j$ with $j \geq 1$, let $s < t$ denote a historical measurement: we use $s = t_0$ for **initial-reference comparisons** and $s = t_{j-1}$ for **local comparisons**, where t_{j-1} is the previous recorded step.

Energy overlap. Energy overlap measures the fraction of the current gradient's top-k energy retained by projection onto a previously recorded top-k subspace on the corresponding input or output side. For either side $z \in \{\mathsf{out}, \mathsf{in}\}$, we define

$$\boxed{\mathcal{O}_{\ell,m,z}^{E}(s,t) = \left\| (\mathbf{Q}_{\ell,m,s}^{z})^{\top} \mathbf{Q}_{\ell,m,t}^{z} D_{\ell,m,t}^{1/2} \right\|_F^2.} \tag{4}$$

Here, $D_{\ell,m,t}$ is a diagonal matrix containing the normalized top-k squared singular values.

$$D_{\ell,m,t} = \mathrm{diag}(a_{\ell,m,t,1}, \ldots, a_{\ell,m,t,k}), \qquad a_{\ell,m,t,i} = \frac{\sigma_{\ell,m,t,i}^2}{\sum_{j=1}^{k} \sigma_{\ell,m,t,j}^2}.$$

Initial-reference and local energy overlap. For measurement steps $t_0 < t_1 < \cdots$, the historical subspace is either the first recorded subspace (initial reference) or the most recently recorded subspace before t_j (local):

Comparison	**Input-side energy overlap**	**Output-side energy overlap**
Initial reference	$\mathcal{O}_{\ell,m,\mathsf{in}}^{E}(t_0, t_j)$	$\mathcal{O}_{\ell,m,\mathsf{out}}^{E}(t_0, t_j)$
Local	$\mathcal{O}_{\ell,m,\mathsf{in}}^{E}(t_{j-1}, t_j)$	$\mathcal{O}_{\ell,m,\mathsf{out}}^{E}(t_{j-1}, t_j)$

Here t_{j-1} is the previous recorded measurement step, not the directly preceding optimizer step $t_j - 1$.

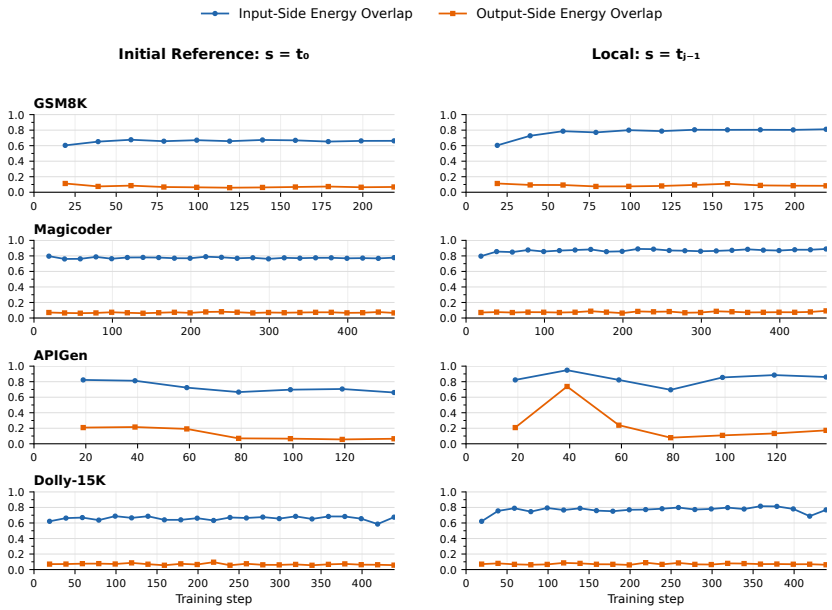


Figure 1: **Across datasets, Qwen3-1.7B exhibits higher weighted mean energy overlap on the input side than on the output side.** The four datasets are GSM8K, Magicoder, APIGen, and Dolly-15K. Each point is the weighted mean across all measured parameter matrices at that step. Blue and orange denote input- and output-side overlap; left and right panels use the initial and previously measured subspaces as references, respectively. All runs use original AdamW and $k = 8$. Appendix Figure 4 shows all eight datasets.

4 Main Findings

Section 4.1 describes the experimental setup; Sections 4.2–4.5 present our main empirical findings.

4.1 Experimental Setup

Tasks and Models. We study gradient dynamics in supervised fine-tuning across eight datasets covering six task categories: (1) **Mathematical Reasoning**, including **GSM8K** (Cobbe et al., 2021) and **OpenMathReasoning** (Moshkov et al., 2025); (2) **Multi-Domain Reasoning**, using **OpenThoughts** (Guha et al., 2026); (3) **Code Generation**, using **Magicoder-OSS-Instruct-75K** (Wei et al., 2024), abbreviated as **Magicoder**; (4) **General Instruction Following**, using **Dolly-15K** (Conover et al., 2023); (5) **Multi-Turn Tool Use**, using **APIGen-MT-5k** (Prabhakar et al., 2025); and (6) **Domain-Specific Question Answering**, including medical questions from **MedQA** (Jin et al., 2021) and science questions from **SciQ** (Welbl et al., 2017). We evaluate **Qwen3-0.6B/1.7B/4B/8B** (Yang et al., 2025), **Llama-3.2-1B-Instruct/3B-Instruct** (Meta, 2024), and **MiniCPM5-2B** (OpenBMB, 2026). In total, we conduct seventeen experiments: fifteen at diagnostic rank $k = 8$ and two at $k = 16$. Sections 4.2–4.5 report the fifteen-run $k = 8$ analysis; Appendix H presents the two $k = 16$ experiments.

Implementation Details. All runs update the full model with AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019). We use learning rate 2×10^{-5}, weight decay of 0.01, $(\beta_1, \beta_2) = (0.9, 0.999)$, and $\epsilon = 10^{-8}$. Training uses a cosine learning-rate schedule with 3% warmup, gradient clipping at norm 1.0, and bfloat16 mixed precision. Training horizons and data sizes vary across configurations; measurements characterize each evaluated trajectory. We measure subspaces at rank $k = 8$ before optimizer steps $1, T_{\text{meas}}, 2T_{\text{meas}}, \ldots$, with $T_{\text{meas}} = 20$.

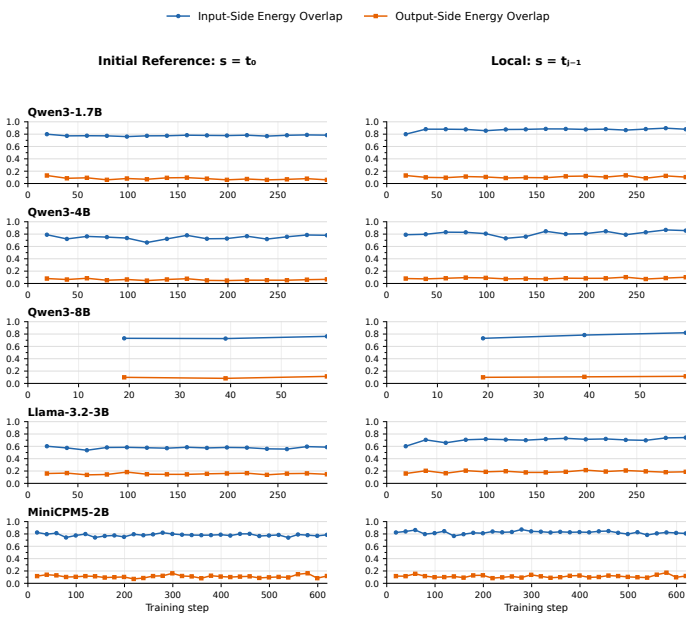


Figure 2: **Across models, input-side weighted mean energy overlap exceeds output-side overlap on OpenMathReasoning.** Each point is the weighted mean across all measured parameter matrices at that step. Blue and orange denote input- and output-side overlap; left and right panels show initial-reference and local comparisons, respectively. All runs use original AdamW and $k = 8$.

4.2 Persistently High Input-Side Gradient Energy Overlap

Finding 1 ***Input-side energy overlap*** *generally remains high during fine-tuning.*

Table 1: **Prevalence of high weighted mean input-side energy overlap.** Counts cover 223 measured steps with valid overlap values across the fifteen SFT runs at $k = 8$; weights follow Equation 5.

Comparison	Steps with weighted mean overlap > 0.7	Proportion
Initial reference	113/223	50.7%
Local	212/223	95.1%

Table 2: **Persistence of high input-side energy overlap in individual matrices.** f is the percentage of measurements with input-side overlap above 0.7. Statistics cover 2,999 matrices across the fifteen $k = 8$ runs, counting each matrix once per run. Q25 and Q75 are the quartiles of f.

Comparison	Median f (%)	Q25–Q75 (%)	$f \geq 80\%$	$f = 100\%$
Initial reference	42.9	[0.0, 100.0]	1320/2999 (44.0%)	1127/2999 (37.6%)
Local	93.3	[45.5, 100.0]	1960/2999 (65.4%)	1345/2999 (44.8%)

Let $\mathcal{M}$ contain all measured matrices of the model, and $N_{\ell,m}$ denote the number of parameters in matrix (ℓ, m). At each recorded step, we define the **weighted mean energy overlap** as

$$\overline{\mathcal{O}}_z^E(s,t) = \frac{\sum_{(\ell,m)\in\mathcal{M}} w_{\ell,m}(t)\mathcal{O}^E_{\ell,m,z}(s,t)}{\sum_{(\ell,m)\in\mathcal{M}} w_{\ell,m}(t)}, \qquad w_{\ell,m}(t) = N_{\ell,m}\mathcal{E}^{(k)}_{\ell,m}(t). \tag{5}$$

Here $z \in \{\mathsf{in}, \mathsf{out}\}$, and $\mathcal{E}^{(k)}_{\ell,m}(t)$ is the current top-k energy fraction defined in Equation 2.

Across the fifteen runs, time-averaged input-side energy overlap ranges from 0.704 to 0.880 for local comparisons, with a mean of 0.808, and from 0.479 to 0.783 for initial-reference comparisons, with a mean of 0.682. Across the fifteen runs, weighted mean input-side energy overlap exceeds 0.7 at 95.1% (212/223) of measured steps for local comparisons and 50.7% (113/223) for initial-reference comparisons, as shown in Table 1.

For each matrix, f is the percentage of measurements with input-side energy overlap above 0.7. Table 2 summarizes f for initial-reference and local comparisons across 2,999 matrices from fifteen runs, counting each matrix once per run. For local comparisons, the median f is 93.3%, and 1,960 matrices (65.4%) exceed 0.7 at least 80% of the time. The initial-reference median is 42.9%, with variation across matrices. Results locate the strongest persistence in neighboring measurements.

4.3 Low Output-Side Gradient Energy Overlap

Finding 2 ***Output-side energy overlap*** *is usually low, especially outside* ***query/key projections****.*

Table 3: **Distribution of output-side energy overlap in individual matrices.** For each comparison type, we collect 45,827 output-side energy overlap values across measured matrices and recorded training steps in the fifteen $k = 8$ experiments; 32,807 values remain after excluding query and key projections. The final two columns report the percentages of overlap values below 0.3 and 0.5, respectively.

Matrices	Comparison	Median overlap	Percentage below 0.3	Percentage below 0.5
All	Initial reference	0.031	85.67%	92.42%
All	Local	0.042	82.46%	89.77%
Excluding Q/K	Initial reference	0.018	99.15%	99.78%
Excluding Q/K	Local	0.025	98.31%	99.45%

We evaluate output-side energy overlap separately for each parameter matrix at each recorded measurement. Across all fifteen runs, its median is 0.031 for initial-reference comparisons and 0.042 for local comparisons (Table 3). Overlap is below 0.3 in 85.67% and 82.46% of the respective comparisons. Excluding query and key projections, output-side energy overlap is below 0.3 in 99.15% of initial-reference comparisons and 98.31% of local comparisons.

The low values also persist within individual matrices. For each matrix, we compute the fraction of its valid measurements with output-side overlap below 0.5. In initial-reference comparisons, 2,765 of 2,999 matrices (92.20%) remain below this threshold at least 80% of the time; for local comparisons, the count is 2,666 (88.90%). These frequency statistics assign equal weight to each matrix within each run. The thresholds 0.3 and 0.5 summarize the observed distribution.

4.4 Input–Output Asymmetry in Gradient Subspace Dynamics

Finding 3 ***Input-side energy overlap*** *is consistently higher than* ***output-side energy overlap*** *in model averages and is higher in the vast majority of individual-matrix comparisons.*

Findings 1 and 2 describe the energy retained on each side separately. We now compare the two sides of the same gradient against the same historical measurement. This comparison reveals the central pattern: historical input-side subspaces retain more of the current gradient's dominant energy than historical output-side subspaces.

Model-level evidence. *Across datasets.* Figure 1 shows this separation on Qwen3-1.7B fine-tuned on GSM8K, Magicoder, APIGen, and Dolly-15K: the input-side curves remain above the output-side curves throughout the measured training steps, for both initial-reference and local comparisons.

The Qwen3-1.7B results across GSM8K, OpenMathReasoning, OpenThoughts, Magicoder, Dolly-15K, APIGen, MedQA, and SciQ contain 126 valid measurement points; all favor the input side.

Across models. Figure 2 shows the same ordering on OpenMathReasoning with Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Llama-3.2-3B-Instruct, and MiniCPM5-2B. Appendix E compares the overlap magnitudes and query/key reversal rates across these families. The five-model OpenMathReasoning comparison contains 79 valid points, which also all favor the input side in initial-reference and local comparisons. Across the full set of model–dataset configurations described in Section 4.1, this ordering holds at all 223 valid measurement points:

$$\overline{\mathcal{O}}^{E}_{\mathsf{in}}(s,t) > \overline{\mathcal{O}}^{E}_{\mathsf{out}}(s,t), \qquad s \in \{t_0, t_{j-1}\}, \quad t = t_j.$$

Matrix-level evidence. Directly comparing the two sides of each parameter matrix also reveals a widespread input-side advantage. Among the 45,827 matrix–step comparisons, input-side energy overlap is higher in 93.58% of initial-reference comparisons and 92.98% of local comparisons. The remaining 2,940 initial-reference comparisons (6.42%) and 3,216 local comparisons (7.02%) favor the output side. Section 4.5 examines the reversals by module type. Additional experiments at $k = 16$ on Llama-3.2-1B-Instruct and Qwen3-1.7B show patterns consistent with Findings 1–4 (Appendix H).

4.5 Module-Specific Exceptions

Finding 4 *Reversals of the input-side energy-overlap advantage are concentrated in* ***query/key projections*** *and* ***embedding tables****.*

The weighted mean curves establish an input-side advantage at the model level. We examine individual parameter matrices for exceptions to this pattern. Across fifteen runs, 246 of 2,999 matrices (8.20%) exhibit higher output-side overlap at least once under either initial-reference or local comparisons. Excluding query/key projections, 14 of 2,147 matrices (0.65%) show at least one reversal. Further excluding embedding tables, 12 of 2,132 matrices (0.56%) show at least one reversal.

Table 4: **Prevalence of higher input-side energy overlap.** "Other matrices" excludes query/key projections and the stored embedding table. Denominators are the number of eligible matrices times valid records; observations from the same training trajectory are correlated. All five models are fine-tuned on OpenMathReasoning, with subspaces measured at diagnostic rank $k = 8$.

			All matrices (%)		Other matrices (%)	
Model	Matrices	Records	Ref.	Local	Ref.	Local
Llama-3.2-3B	197	15	77.46	76.99	99.95	100.00
MiniCPM5-2B	296	31	88.24	86.87	100.00	100.00
Qwen3-1.7B	197	15	98.54	98.07	99.90	99.90
Qwen3-4B	253	15	99.92	99.76	100.00	100.00
Qwen3-8B	254	3	99.48	99.21	100.00	100.00

Matrix-level evidence. Table 4 compares the prevalence of higher input-side energy overlap across five models fine-tuned on OpenMathReasoning: Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Llama-3.2-3B-Instruct, and MiniCPM5-2B. At each measured step, we compare input- and output-side energy overlap for each matrix. The step-1 reference is excluded.

Input-side energy overlap exceeds output-side overlap in 98.07–99.92% of matrix–step comparisons for Qwen3, 76.99–77.46% for Llama, and 86.87–88.24% for MiniCPM across the two reference choices (Table 4). Excluding query/key projections and embeddings raises these proportions to 99.90–100% across all five runs; query/key projections account for most reversals. For embeddings, the input-side advantage holds in all 15 comparisons of each type for Qwen3-1.7B, Qwen3-4B, and Llama-3.2-3B, and in 14/31 initial-reference and 24/31 local comparisons for MiniCPM. Qwen3-8B instead shows higher output-side overlap in all three comparisons of each type.

Why embedding tables differ. For embedding lookup, the input is a one-hot token encoding and the output is a dense embedding vector, whereas ordinary projection layers map dense inputs to dense outputs. This structural difference may lead to exceptions to the input-side energy-overlap advantage.

Why query and key can differ. In self-attention, $Q = XW_Q^\top$ and $K = XW_K^\top$ share the same input X. Thus,

$$S = \frac{QW_K X^\top}{\sqrt{d_k}}, \qquad S^\top = \frac{KW_Q X^\top}{\sqrt{d_k}}. \tag{6}$$

Here, W_K connects Q and X, while W_Q connects K and X. Both sides involve input-derived representations, which may help explain why query/key projections are the main exceptions to the input-side energy-overlap advantage.

5 An Application: Overlap-Guided Gradient Low-Rank Projection

5.1 Input-Side Projection for GaLore

GaLore (Zhao et al., 2024) reduces optimizer-state memory by maintaining optimizer states in a low-rank gradient subspace. For $G \in \mathbb{R}^{m\times n}$ with $G = U\Sigma V^\top$, we describe the official implementation's `std` mode, which selects the projection side according to the matrix shape:

$$R = \begin{cases} GV_r \in \mathbb{R}^{m\times r}, & m \geq n \quad \text{input side (right)}, \\ U_r^\top G \in \mathbb{R}^{r\times n}, & m < n \quad \text{output side (left)}. \end{cases} \tag{7}$$

Appendix Table 7 compares optimizer-state storage costs with standard AdamW. Updates are mapped back to the parameter space, and the projection basis is reused between periodic refreshes.

Guided by the higher input-side energy overlap, we propose **Overlap-Guided Gradient Low-Rank Projection (OG-GaLore)**, which always selects input-side projection. At each refresh, we set $P = V_r$ from $G = U\Sigma V^\top$ and reuse P until the next refresh. Each gradient is compressed as $R = GP$; AdamW operates in these compact coordinates, and its update is mapped back through $P^\top$. Here overlap guides the choice of side; the refresh interval is fixed.

We compare this rule with Output-side Projection, standard GaLore (input side for $m \geq n$, output side otherwise), Reverse GaLore (the opposite projection-side choice to GaLore). For either projection side z, let P denote the basis selected at the most recent refresh: $P = V_r \in \mathbb{R}^{n\times r}$ for $z = \mathsf{in}$ and $P = U_r \in \mathbb{R}^{m\times r}$ for $z = \mathsf{out}$. The compressed gradient R and reconstructed update direction H are

$$R = \begin{cases} GP, & z = \mathsf{in}, \\ P^\top G, & z = \mathsf{out}, \end{cases} \qquad H = \begin{cases} DP^\top, & z = \mathsf{in}, \\ PD, & z = \mathsf{out}, \end{cases} \tag{8}$$

where D is the Adam update direction computed from R and has the same shape as R: $m \times r$ for input-side projection and $r \times n$ for output-side projection. In both cases, $H \in \mathbb{R}^{m\times n}$ matches the weight matrix's shape. Appendices B and C give the algorithm and experimental settings.

Table 5: **Final validation loss for four GaLore projection rules.** Bold marks the lowest loss in each row. Table 8 reports results for all evaluated ranks.

Model	Rank	OG-GaLore	GaLore	Output-side	Reverse GaLore
0.6B	4	1.0589	**1.0570**	1.3316	1.2368
	8	1.0589	**1.0564**	1.1848	1.0908
1.7B	4	**0.8451**	0.9970	1.2090	0.8527
	8	**0.8404**	0.8875	1.1372	0.8479
	16	**0.8359**	0.8562	0.9778	0.8418
4B	8	**0.6996**	0.7057	0.9304	0.6997
	16	0.6973	0.6993	0.7355	**0.6969**

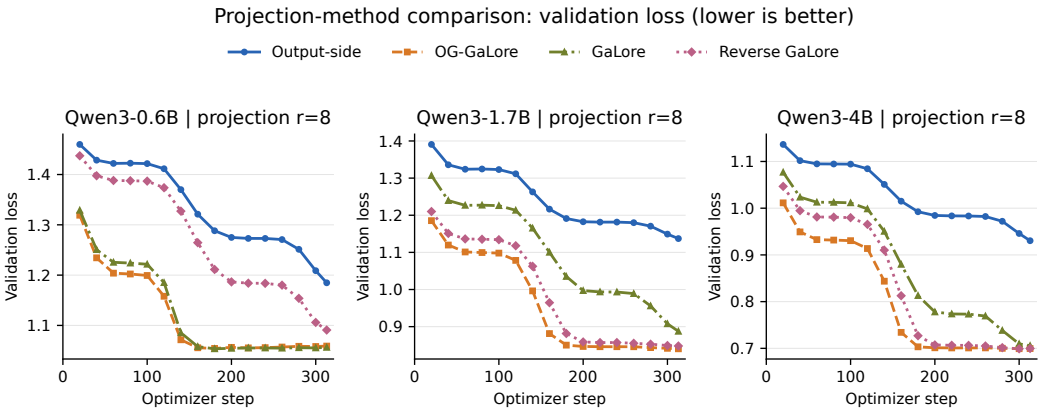


Figure 3: **Validation-loss curves for four GaLore projection rules.** Panels show Qwen3-0.6B, Qwen3-1.7B, and Qwen3-4B, all at rank 8. OG-GaLore reaches low validation loss earlier, while some other rules attain similar final losses. Lower is better.

5.2 Validation Loss and Rank Sensitivity

Table 5 compares final validation losses, with all ranks reported in Appendix Table 8. OG-GaLore outperforms standard GaLore in eight of ten available configurations, including all six comparable Qwen3-1.7B ranks and both Qwen3-4B ranks. It does not always achieve the lowest loss: GaLore leads on Qwen3-0.6B by 0.0019 at rank 4 and 0.0025 at rank 8, while Reverse GaLore leads on Qwen3-4B at rank 16 by 0.00032.

Figure 3 compares the validation-loss trajectories. OG-GaLore reaches low validation loss in fewer steps on all three models, with a clear early advantage on Qwen3-1.7B and Qwen3-4B. Although GaLore on 0.6B and Reverse GaLore on 4B eventually attain similar losses, their early curves lie above OG-GaLore. Output-side Projection has the highest final loss in all three panels.

Matrix shapes help interpret the performance of the projection rules. On Qwen3-0.6B, GaLore selects the input side for 71.57% of target matrices and achieves validation loss close to OG-GaLore. On Qwen3-4B, Reverse GaLore selects the input side for 56.92% of target matrices and likewise achieves similar loss. In both cases, competitive performance accompanies input-side projection on a majority of matrices. Their agreement with OG-GaLore depends on matrix shapes, and close results in some configurations coexist with substantial gaps in others. At rank 4, GaLore's loss on 1.7B is 0.9970, compared with 0.8451 for OG-GaLore; Reverse GaLore's loss on 0.6B is 1.2368, compared with 1.0589 for OG-GaLore. By always selecting the input side based on the observed energy-overlap pattern, OG-GaLore achieves the lowest or near-lowest loss across different configurations. These results support using energy retention to guide projection-side selection. Appendix C.2 details the module assignments and discusses their relation to validation loss.

6 Conclusion

Our experiments reveal a general pattern in SFT gradient dynamics: input-side energy overlap is generally high and output-side overlap usually low. The input-side advantage holds in model averages and most individual-matrix comparisons; reversals concentrate in query/key projections and embedding tables. Guided by these findings, we propose OG-GaLore, which always selects input-side projection and achieves the lowest or near-lowest validation loss across different configurations.

Limitations and Future Work. Our empirical study finds the same pattern across diverse datasets and model families. Ding et al. (2026) examine related input–output differences in cumulative parameter changes; we study gradient subspaces over training. Time and computational resources limit model coverage. We hope these findings motivate further investigation into this asymmetry and inform future LLM training. Future work will investigate whether this input–output asymmetry in gradient energy overlap also holds in reinforcement learning and on-policy distillation (OPD).

REFERENCES

Yuchen Cai, Ding Cao, Liang Lin, Chunxi Luo, Xin Xu, Kai Yang, Weijie Liu, Saiyong Yang, Tianxiang Zhao, Guangzhong Sun, Guiquan Liu, and Junfeng Fang. Learning to foresee: Unveiling the unlocking efficiency of on-policy distillation, 2026a. URL https://arxiv.org/abs/2605.11739.

Yuchen Cai, Ding Cao, Xin Xu, Zijun Yao, Yuqing Huang, Benyi Zhang, Zhenyu Tan, Guiquan Liu, and Junfeng Fang. On predictability of reinforcement learning dynamics for large language models. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=SdHmA6BYVJ.

Sakshi Choudhary, Utkarsh Saxena, and Kaushik Roy. OASIS: Online activation subspace learning for memory-efficient training, 2026. URL https://arxiv.org/abs/2604.09406.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free Dolly: Introducing the world's first truly open instruction-tuned LLM. Databricks Blog, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Yifan Ding, Xincheng Wei, Yoshua Y. Li, Ziheng Li, Yuquan Lu, Siyu Zhang, Dongsheng Ma, Rongxiang Weng, Xunliang Cai, and Yun Chen. SAF-OPD: Stable advantage fusion for on-policy distillation, 2026. URL https://arxiv.org/abs/2607.29209.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, fourth edition, 2013.

Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Rea Sprague, Ashima Suvarna, Benjamin Feuer, Leon Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah M Pratt, Vivek Ramanujan, Jon Saad-Falcon, Stutee Acharya, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alex Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7xjoTuaNmN.

Yutong He, Pengrui Li, Yipeng Hu, Chuyan Chen, and Kun Yuan. Subspace optimization for large language models with convergence guarantees. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=4BARFrWnTI.

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. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Ajay Kumar Jaiswal, Yifan Wang, Lu Yin, Shiwei Liu, Runjin Chen, Jiawei Zhao, Ananth Grama, Yuandong Tian, and Zhangyang Wang. From low rank gradient subspace stabilization to low-rank weights: Observations, theories, and applications. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=fsf7LhbYdf.

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. ISSN 2076-3417. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Ming Li, Yanhong Li, Ziyue Li, and Tianyi Zhou. How instruction and reasoning data shape post-training: Data quality through the lens of layer-wise gradients. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33249–33299, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1536. URL https://aclanthology.org/2026.acl-long.1536/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Meta. Llama 3.2 model card, 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. AIMO-2 winning solution: Building state-of-the-art mathematical reasoning models with OpenMathReasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025. URL https://arxiv.org/abs/2504.16891.

OpenBMB. MiniCPM5-2B model card, 2026. URL https://huggingface.co/openbmb/MiniCPM5-2B.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Akshara Prabhakar, Zuxin Liu, Ming Zhu, Jianguo Zhang, Tulika Manoj Awalgaonkar, Shiyu Wang, Zhiwei Liu, Haolin Chen, Thai Hoang, Juan Carlos Niebles, Shelby Heinecke, Weiran Yao, Huan Wang, Silvio Savarese, and Caiming Xiong. APIGen-MT: Agentic Pipeline for Multi-Turn Data Generation via Simulated Agent-Human Interplay. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-2184. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5e3661f7fe4c8ac5652d62eb3d3c96ea-Paper-Datasets_and_Benchmarks_Track.pdf.

Sahar Rajabi, Nayeema Nonta, and Sirisha Rambhatla. Subtrack++ : Gradient subspace tracking for scalable LLM training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=6geRIdlFWJ.

Thomas Robert, Mher Safaryan, Ionut-Vlad Modoranu, and Dan Alistarh. LDAdam: Adaptive optimization from low-dimensional gradient statistics. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Zkp1GuHerF.

Yara Shamshoum, Nitzan Hodos, Yuval Sieradzki, and Assaf Schuster. CompAct: Compressed activations for memory-efficient LLM training. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1511–1524, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.71. URL https://aclanthology.org/2025.naacl-long.71/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

DiJia Su, Andrew Gu, Jane Xu, Yuandong Tian, and Jiawei Zhao. GaLore 2: Large-scale LLM pre-training by gradient low-rank projection, 2025. URL https://arxiv.org/abs/2504.20437.

Noel Thomas. No subspace to track: Non-identifiability and optimizer state in low-rank training, 2026. URL https://arxiv.org/abs/2607.05872.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Empowering code generation with OSS-instruct. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=XUeoOBid3x.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413/.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Longteng Zhang, Lin Zhang, Shaohuai Shi, Xiaowen Chu, and Bo Li. LoRA-FA: Memory-efficient low-rank adaptation for large language models fine-tuning, 2023. URL https://arxiv.org/abs/2308.03303v1.

Zhenyu Zhang, Ajay Jaiswal, Lu Yin, Shiwei Liu, Jiawei Zhao, Yuandong Tian, and Zhangyang Wang. Q-GaLore: Quantized GaLore with INT4 projection and layer-adaptive low-rank gradients, 2024. URL https://arxiv.org/abs/2407.08296.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient LLM training by gradient low-rank projection. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=hYHsrKDiX7.

Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Sáez de Ocáriz Borde, Rickard Brüel Gabrielsson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. Asymmetry in low-rank adapters of foundation models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=txRZBD8tBV.

A Subspace-Dynamics Training Configurations

Table 6 lists the batch and sequence settings for each of the fifteen supervised fine-tuning runs used in the subspace analysis at diagnostic rank $k = 8$. Batch size is per device, and maximum sequence length is measured in tokens.

Table 6: Batch and maximum sequence settings for the fifteen subspace-dynamics runs at diagnostic rank $k = 8$.

Dataset	Model	Batch size	Accum. steps	Max. length
APIGen	Qwen3-0.6B	1	8	8,192
APIGen	Qwen3-1.7B	1	8	16,384
Dolly-15K	Qwen3-1.7B	4	2	2,048
GSM8K	Llama-3.2-1B-Instruct	1	8	2,048
GSM8K	Qwen3-1.7B	4	2	2,048
Magicoder	Qwen3-1.7B	4	2	2,048
MedQA	Qwen3-1.7B	4	2	2,048
OpenMathReasoning	Llama-3.2-3B-Instruct	1	8	16,384
OpenMathReasoning	MiniCPM5-2B	1	4	16,384
OpenMathReasoning	Qwen3-1.7B	1	8	16,384
OpenMathReasoning	Qwen3-4B	1	8	16,384
OpenMathReasoning	Qwen3-8B	1	4	8,192
OpenThoughts	Llama-3.2-1B-Instruct	4	2	2,048
OpenThoughts	Qwen3-1.7B	4	2	2,048
SciQ	Qwen3-1.7B	4	2	2,048

B Gradient Projection Algorithm

Algorithm 1 Gradient projection with different projection rules.

Input: weights $W \in \mathbb{R}^{m\times n}$, rank r, refresh interval τ, rule q, learning rate η, weight decay λ, scale α, Adam coefficients β_1, β_2, and numerical constant ϵ.

Select the projection side for each matrix:

$$z_q(m,n) = \begin{cases} \mathsf{in}, & q = \texttt{right} \quad \text{(OG-GaLore)}, \\ \mathsf{out}, & q = \texttt{left} \quad \text{(Output-side Projection)}, \\ \mathsf{in} \text{ if } m \geq n, \ \mathsf{out} \text{ otherwise}, & q = \texttt{std} \quad \text{(GaLore)}, \\ \mathsf{out} \text{ if } m \geq n, \ \mathsf{in} \text{ otherwise}, & q = \texttt{reverse_std} \quad \text{(Reverse GaLore)}. \end{cases}$$

Set $z = z_q(m,n)$; initialize $M_0 = S_0 = 0$ with shape $m \times r$ if $z = \mathsf{in}$, else $r \times n$.
for each training step t and weight matrix W **do**
  Compute $G = \nabla_W \mathcal{L}_t$.
  if $t = 1$ or $t \equiv 0 \pmod{\tau}$ **then**
    Compute $G = U\Sigma V^\top$; set $P = V_r$ if $z = \mathsf{in}$, else $P = U_r$.
  end if
  Project: $R = GP$ if $z = \mathsf{in}$, else $R = P^\top G$.
  Update first moment: $M_t \leftarrow \beta_1 M_{t-1} + (1-\beta_1)R$.
  Update second moment: $S_t \leftarrow \beta_2 S_{t-1} + (1-\beta_2)(R \odot R)$.
  Correct bias: $\widehat{M}_t \leftarrow M_t/(1-\beta_1^t)$, $\widehat{S}_t \leftarrow S_t/(1-\beta_2^t)$.
  Compute $D \leftarrow \widehat{M}_t/(\sqrt{\widehat{S}_t} + \epsilon)$ (elementwise).
  Reconstruct: $H = DP^\top$ if $z = \mathsf{in}$, else $H = PD$.
  Apply decoupled weight decay: $W \leftarrow (1-\eta\lambda)W$.
  Update: $W \leftarrow W - \eta\alpha H$.
end for

C Experimental Setup for Projection Comparisons

We fine-tune **Qwen3-0.6B, Qwen3-1.7B, and Qwen3-4B** (Yang et al., 2025) on up to 10,000 **OpenThoughts** examples (Guha et al., 2026). These projection experiments are separate from the seventeen runs using original AdamW used to characterize subspace dynamics. Training uses AdamW, learning rate 2×10^{-5}, weight decay 0.01, maximum sequence length 2,048, one epoch, and seed 42. The accumulated batch size per device is 8; all reported endpoints are evaluated at step 313. Projectors are refreshed every 100 steps. We use Adam parameters $(\beta_1, \beta_2) = (0.9, 0.999)$ and $\epsilon = 10^{-8}$, a cosine learning-rate schedule with a 3% warmup ratio, and gradient clipping at norm 1.0. Training uses BF16 mixed precision and gradient checkpointing.

Projection settings. All four rules use a projection scale $\alpha = 1$ and the same configured rank across target matrices. Projection targets the attention matrices (`q_proj`, `k_proj`, `v_proj`, `o_proj`), the feed-forward matrices (`gate_proj`, `up_proj`, `down_proj`), the token embedding (`embed_tokens`), and the language-model head (`lm_head`). Standard GaLore uses `std`, choosing input-side projection for $m \geq n$ and output-side projection otherwise; Reverse GaLore uses the complementary rule. OG-GaLore always uses input-side projection, and Output-side Projection always uses output-side projection.

The four projection rules use the same rank within each comparison. For a weight matrix of shape $m \times n$, the two AdamW moment states contain $2mr$ elements under input-side projection and $2rn$ elements under output-side projection. Figure 3 compares all three models at rank 8. The 0.6B and 1.7B configurations otherwise coincide. For 4B, OG-GaLore uses microbatch size 1 with eight accumulation steps, whereas the other rules use microbatch size 2 with four accumulation steps; evaluation microbatch sizes also differ. Table 8 reports all available projection ranks.

Table 7: **Moment-state and projection-basis storage** for an $m \times n$ parameter matrix and projection rank r, measured in elements. The totals include the two Adam moment states and the projection basis; model weights, gradients, and activations are excluded.

Stored elements	AdamW	GaLore (`std`)
Two moment states	$2mn$	$2r \max(m, n)$
Projection basis	0	$r \min(m, n)$
Total	$2mn$	$2r \max(m, n) + r \min(m, n)$

C.1 Complete Validation-Loss Results

Table 8 includes all evaluated ranks, extending the Qwen3-1.7B rows in Table 5 to ranks 32, 64, and 128. The Qwen3-0.6B and Qwen3-4B rows are unchanged.

Table 8: **Complete validation-loss results for four projection rules.** All methods use the indicated projection rank, AdamW, and evaluation step 313. Bold marks the lowest loss in each row.

Model	Rank	OG-GaLore	GaLore	Output-side	Reverse GaLore
0.6B	4	1.0589	**1.0570**	1.3316	1.2368
	8	1.0589	**1.0564**	1.1848	1.0908
1.7B	4	**0.8451**	0.9970	1.2090	0.8527
	8	**0.8404**	0.8875	1.1372	0.8479
	16	**0.8359**	0.8562	0.9778	0.8418
	32	**0.8343**	0.8487	0.8691	0.8381
	64	**0.8334**	0.8432	0.8574	0.8350
	128	**0.8315**	0.8377	0.8499	0.8330
4B	8	**0.6996**	0.7057	0.9304	0.6997
	16	0.6973	0.6993	0.7355	**0.6969**

Validation loss and rank sensitivity. The complete results in Table 8 show that OG-GaLore achieves lower validation loss than standard GaLore in eight of ten available model–rank configurations. On Qwen3-1.7B, it improves over GaLore at all six evaluated ranks. At rank 8, their losses are 0.8404 and 0.8875, respectively. On Qwen3-0.6B, OG-GaLore is within 0.0019 and 0.0025 of the lowest validation loss, achieved by standard GaLore at ranks 4 and 8, respectively. On Qwen3-4B, OG-GaLore improves over standard GaLore at ranks 8 and 16, while Reverse GaLore attains the lowest loss at rank 16.

In the three fixed-rank panels of Figure 3, Output-side Projection has the highest final loss. Among the other rules, GaLore achieves the lowest final loss on 0.6B, while OG-GaLore achieves the lowest on 1.7B and 4B. On 4B, the final losses of OG-GaLore and Reverse GaLore are 0.6996 and 0.6997, respectively.

Across different configurations, OG-GaLore achieves the lowest or near-lowest validation loss among the compared projection rules. Appendix C.2 details how matrix shapes cause the shape-based rules to share input-side choices with OG-GaLore.

C.2 Matrix Shapes and Projection-Side Agreement

The shape-based rules can coincide with input-side projection on a substantial subset of target matrices. We derive their assignments from the archived Qwen3 configurations, cross-checked against the official configurations for 0.6B, 1.7B, and 4B. All shapes below use the stored weight convention $m \times n$ (output dimension $\times$ input dimension). For attention, query width is the number of query heads times head dimension; key/value width is computed from the number of key/value heads. Consequently, query and output projections need not be square.

Table 9: **Shapes of projection targets.** Each layer contains seven listed attention and feed-forward matrices. The embedding and language-model head share one parameter matrix and are counted once.

Target matrix	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B
Query	2048×1024	2048×2048	4096×2560
Key/value (each)	1024×1024	1024×2048	1024×2560
Attention output	1024×2048	2048×2048	2560×4096
Gate/up (each)	3072×1024	6144×2048	9728×2560
Down	1024×3072	2048×6144	2560×9728
Embedding/head	151936×1024	151936×2048	151936×2560
Transformer layers	28	28	36

Under GaLore's $m \geq n$ rule, the input-side targets are query, key, value, gate, up, and embedding on 0.6B; query, attention output, gate, up, and embedding on 1.7B; and query, gate, up, and embedding on 4B. Reverse GaLore assigns the complementary side to every matrix, including square matrices. Table 10 reports the resulting agreement with OG-GaLore, measured by the number and percentage of unique target matrices assigned to the input side. Each matrix is counted equally, and the shared embedding/head matrix is counted once.

Table 10: **Agreement with always-input projection.** Matrix counts and percentages follow deterministically from Table 9 and the projection-side rules.

	GaLore: input-side targets		Reverse GaLore: input-side targets	
Model	Matrices	Percentage	Matrices	Percentage
Qwen3-0.6B	141/197	71.57%	56/197	28.43%
Qwen3-1.7B	113/197	57.36%	84/197	42.64%
Qwen3-4B	109/253	43.08%	144/253	56.92%

For 0.6B, GaLore already selects the input side for five of seven matrices per layer and the shared embedding. Its substantial agreement with OG-GaLore provides a structural explanation for their

similar validation losses. For 1.7B, the key/value matrices become wide and GaLore selects their output side, whereas the square attention-output matrix selects the input side. Thus both the set of differing modules and the fraction of input-side assignments change. For 4B, Reverse GaLore selects the input side for key, value, attention output, and down projections—four of seven matrices per layer. This shared choice is consistent with its close performance to OG-GaLore. The assignments therefore help interpret small loss differences, but matrix counts alone do not establish their cause or predict the ordering of validation losses.

Table 11: **Input-side matrix percentages and validation loss.** Input % is the percentage of unique target matrices assigned to input-side projection and is constant across ranks within each model. OG-GaLore uses 100% input-side projection; Output-side Projection uses 0%. Losses are the step-313 validation losses from Table 8 on OpenThoughts, with the same experimental settings. Bold marks the lowest loss in each row.

		OG-GaLore	GaLore		Reverse GaLore		Output-side
Model	Rank	Loss	Input %	Loss	Input %	Loss	Loss
0.6B	4	1.0589	71.57%	**1.0570**	28.43%	1.2368	1.3316
	8	1.0589		**1.0564**		1.0908	1.1848
1.7B	4	**0.8451**	57.36%	0.9970	42.64%	0.8527	1.2090
	8	**0.8404**		0.8875		0.8479	1.1372
	16	**0.8359**		0.8562		0.8418	0.9778
	32	**0.8343**		0.8487		0.8381	0.8691
	64	**0.8334**		0.8432		0.8350	0.8574
	128	**0.8315**		0.8377		0.8330	0.8499
4B	8	**0.6996**	43.08%	0.7057	56.92%	0.6997	0.9304
	16	0.6973		0.6993		**0.6969**	0.7355

Interpreting GaLore and Reverse GaLore together. Table 11 places these assignments alongside validation loss at every evaluated rank. The results should be read in terms of the module-wise assignments of both shape-based rules. On 0.6B, GaLore closely matches OG-GaLore while already using input-side projection for most target matrices. On 4B, Reverse GaLore closely matches OG-GaLore at rank 8 and slightly outperforms it at rank 16; it too uses input-side projection for most target matrices, although on a different set of modules. These results are consistent with the usefulness of input-side projection: the competitive shape-based rule in each case shares many of its input-side choices with OG-GaLore. They also show why a small loss gap need not indicate that projection direction is unimportant. The comparison changes the direction only on the matrices where the rules disagree. This interpretation supports input-side projection as a simple, consistently competitive choice, without claiming that the number of input-side assignments alone determines loss or that always-input projection must outperform every mixed assignment.

D Parameter Matrices with Reversed Energy-Overlap Ordering

We inspect the fifteen $k = 8$ runs at the level of individual weight matrices. A reversal means that output-side energy overlap strictly exceeds input-side energy overlap at a measured step. No equalities occur among the valid comparisons. Table 13 identifies every query/key matrix with at least one reversal under either initial-reference or local comparisons; Table 14 lists all remaining exceptions. A listed matrix need not reverse the ordering at every step. Layer indices are zero-based. Llama models use the Instruct checkpoints.

Table 12 reports both orderings for initial-reference and local comparisons separately. For brevity, write $E_z^{\mathrm{ref}} = \mathcal{O}^E_{\ell,m,z}(t_0, t_j)$ and $E_z^{\mathrm{loc}} = \mathcal{O}^E_{\ell,m,z}(t_{j-1}, t_j)$. Each percentage gives the fraction of valid matrix–step comparisons satisfying the indicated inequality, with equal weight per comparison.

Table 12: Energy-overlap ordering across the fifteen runs at diagnostic rank $k = 8$. N is the number of valid comparisons for each reference choice. No ties occur. The total includes all matrices other than query/key projections and embedding tables; `lm_head` counts separately measured heads.

		Initial reference		Local	
Matrix	N	$E_{\text{in}}^{\text{ref}} > E_{\text{out}}^{\text{ref}}$	$E_{\text{out}}^{\text{ref}} > E_{\text{in}}^{\text{ref}}$	$E_{\text{in}}^{\text{loc}} > E_{\text{out}}^{\text{loc}}$	$E_{\text{out}}^{\text{loc}} > E_{\text{in}}^{\text{loc}}$
`o_proj`	6,510	99.8925%	0.1075%	99.9386%	0.0614%
`up_proj`	6,510	99.9693%	0.0307%	100.0000%	0.0000%
`down_proj`	6,510	99.9693%	0.0307%	99.9846%	0.0154%
`gate_proj`	6,510	100.0000%	0.0000%	100.0000%	0.0000%
`lm_head`	34	100.0000%	0.0000%	100.0000%	0.0000%
All except Q/K/embeddings	32,584	99.9662%	0.0338%	99.9847%	0.0153%

Table 13: **Query/key matrices with at least one reversal across the fifteen $k = 8$ runs.** Each layer index ℓ identifies `model.layers.`ℓ`.self_attn.q_proj.weight` or `k_proj.weight`, respectively. A dash indicates no reversal. The listed layers are the union of initial-reference and successive-measurement exceptions.

Dataset	Model	Query projection layers	Key projection layers
APIGen	Qwen3-0.6B	20, 23, 26–27	1
APIGen	Qwen3-1.7B	20, 23–24, 26–27	—
Dolly-15K	Qwen3-1.7B	3, 6, 18–27	—
GSM8K	Llama-3.2-1B	0–6, 8, 10–13, 15	0–15
GSM8K	Qwen3-1.7B	3, 6, 8, 19–20, 23–27	—
Magicoder	Qwen3-1.7B	20, 23, 26–27	—
MedQA	Qwen3-1.7B	3, 6, 8, 19–27	—
OpenMathReasoning	Llama-3.2-3B	0–7, 9, 13, 16–17, 19–25, 27	0–27
OpenMathReasoning	MiniCPM5-2B	8–9, 26, 30, 33–34, 40–41	1–12, 14–41
OpenMathReasoning	Qwen3-1.7B	3, 20, 23, 25–27	0
OpenMathReasoning	Qwen3-4B	3, 5, 32, 35	—
OpenMathReasoning	Qwen3-8B	34–35	—
OpenThoughts	Llama-3.2-1B	0–6, 11–13, 15	0–15
OpenThoughts	Qwen3-1.7B	1, 3, 6, 20, 23, 26–27	1
SciQ	Qwen3-1.7B	3, 18–27	—

E Comparison across Model Families

We compare Qwen3, Llama, and MiniCPM on OpenMathReasoning, the dataset shared by all three families in our experiments. Table 15 summarizes the five trajectories in Figure 2. All use AdamW, diagnostic rank $k = 8$, and measurements every 20 optimizer steps. The Qwen3-1.7B, Qwen3-4B, and Llama-3.2-3B-Instruct runs also share batch size 1, eight accumulation steps, and maximum sequence length 16,384. MiniCPM5-2B uses four accumulation steps; Qwen3-8B uses four and a maximum length of 8,192. We report each run over its available measurements: steps 20–300 for Qwen3-1.7B/4B and Llama, 20–620 for MiniCPM, and 20–60 for Qwen3-8B. These comparisons describe the evaluated models and training trajectories; they do not isolate architecture from model size, checkpoint training history, or the settings in Table 6.

Input and output energy retention. All five runs exhibit higher input-side than output-side weighted mean overlap at every measured step, under both reference choices. The input-side averages range from 0.5763 to 0.7827 for the initial reference and from 0.7039 to 0.8732 for local comparisons; the corresponding output-side ranges are 0.0610–0.1544 and 0.0842–0.1892. Llama has lower input-side and higher output-side averages than the Qwen3 and MiniCPM runs, while retaining a clear separation between the two sides. Within Qwen3, the averages do not increase with model size: Qwen3-1.7B has higher input-side averages than Qwen3-4B under their shared settings. The short Qwen3-8B trajectory provides additional evidence of the same ordering, without establishing its behavior over longer training.

Table 14: **All reversals outside query/key projections across the fifteen $k = 8$ runs.** Projection names refer to their weight matrices: `o_proj` is the attention output projection, and `up_proj`/`down_proj` are MLP projections. The embedding matrix is `model.embed_tokens.weight`; its entries report reversal counts out of the available measurements (31 for MiniCPM5-2B and 3 for Qwen3-8B) rather than step numbers. All other entries give optimizer steps. A dash indicates no reversal.

Dataset	Model	Matrix	Reference steps	Local steps
APIGen	Qwen3-1.7B	Layer 18: `down_proj`	40	—
APIGen	Qwen3-1.7B	Layer 26: `o_proj`	40	—
Dolly-15K	Qwen3-1.7B	Layer 27: `o_proj`	200	—
GSM8K	Llama-3.2-1B	Layer 15: `up_proj`	180	—
GSM8K	Qwen3-1.7B	Layer 26: `o_proj`	20	20
GSM8K	Qwen3-1.7B	Layer 27: `o_proj`	20	20
OpenMathReasoning	Llama-3.2-3B	Layer 27: `up_proj`	40	—
OpenMathReasoning	MiniCPM5-2B	`embed_tokens`	17/31	7/31
OpenMathReasoning	Qwen3-8B	`embed_tokens`	3/3	3/3
OpenMathReasoning	Qwen3-1.7B	Layer 26: `o_proj`	20	20
OpenMathReasoning	Qwen3-1.7B	Layer 27: `o_proj`	20	20
OpenThoughts	Qwen3-1.7B	Layer 0: `down_proj`	—	240
SciQ	Qwen3-1.7B	Layer 23: `down_proj`	140	—
SciQ	Qwen3-1.7B	Layer 27: `o_proj`	160	—

Table 15: Energy overlap on OpenMathReasoning across three model families. Each value averages the per-step weighted mean energy overlap over the available measurements in that run. The initial step is excluded. Llama uses the Instruct checkpoint.

		Initial reference		Local	
Model	Records	Input	Output	Input	Output
Qwen3-1.7B	15	0.7781	0.0786	0.8732	0.1072
Qwen3-4B	15	0.7458	0.0610	0.8127	0.0842
Qwen3-8B	3	0.7401	0.0981	0.7786	0.1068
Llama-3.2-3B	15	0.5763	0.1544	0.7039	0.1892
MiniCPM5-2B	31	0.7827	0.1108	0.8241	0.1129

Where the matrix-level differences occur. The lower prevalence of input-side dominance in Llama and MiniCPM (Table 4) is concentrated in query/key projections. For Llama, output-side overlap exceeds input-side overlap in every key-projection comparison and in 58.33%/61.90% of query-projection comparisons (initial reference/local). For MiniCPM, the corresponding reversal rates are 79.19%/86.10% for key projections and 2.38%/5.91% for query projections. In the three Qwen3 runs, key-projection reversal rates are at most 0.24%, and query-projection rates range from 0.56% to 12.86%. These percentages count matrix–step comparisons equally. After excluding query/key projections and embedding tables, input-side overlap is higher in 99.90–100% of comparisons in every run. Embedding reversals occur in MiniCPM5-2B and Qwen3-8B, as detailed in Table 14. Thus, the three families share the input–output asymmetry described by Findings 1–3, while the frequency and location of the exceptions in Finding 4 distinguish their measured trajectories.

F QWEN3-1.7B RESULTS ON EIGHT DATASETS

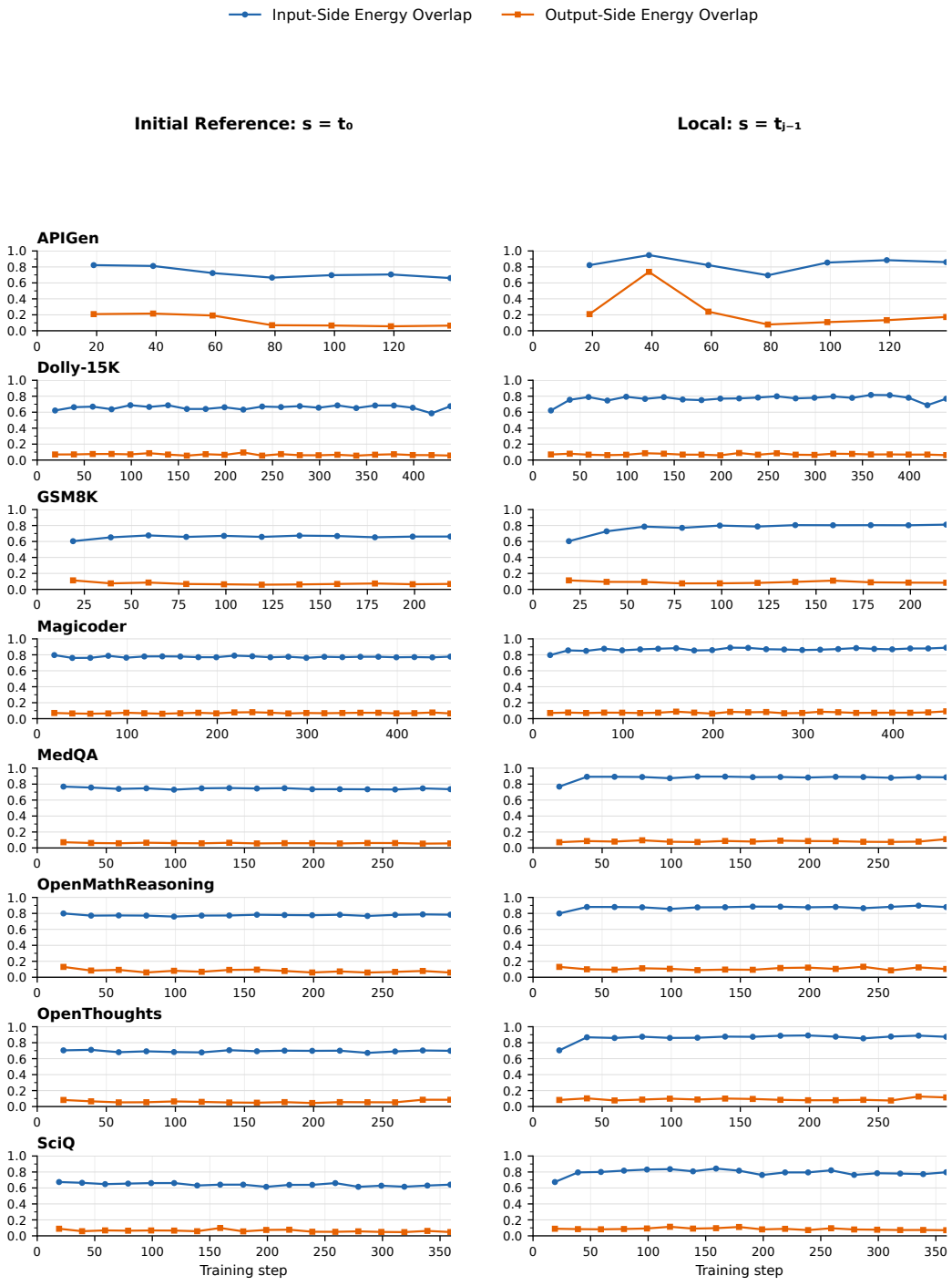


Figure 4: **Energy overlap across training steps on all eight datasets for Qwen3-1.7B.** Each row shows one dataset; blue circles denote the input side and orange squares the output side. Left: $\overline{\mathcal{O}}_z^E(t_0, t_j)$; right: $\overline{\mathcal{O}}_z^E(t_{j-1}, t_j)$. All panels use the same $[0, 1]$ vertical scale; horizontal ranges reflect each run's training horizon. Points are unsmoothed measurements, with lines connecting successive points. The initial step is excluded because its reference overlap is trivially one. All runs use original AdamW and diagnostic rank $k = 8$.

Table 16: **Energy overlap on eight datasets for Qwen3-1.7B.** Columns report averages over recorded training steps of $\overline{\mathcal{O}}^{E}_{\mathsf{in}}(s, t_j)$ and $\overline{\mathcal{O}}^{E}_{\mathsf{out}}(s, t_j)$, with $s = t_0$ for initial reference and $s = t_{j-1}$ for local comparisons. All eight runs use original AdamW and diagnostic rank $k = 8$.

	Initial reference		Local	
Dataset	Input	Output	Input	Output
APIGen	0.7267	0.1247	0.8412	0.2399
Dolly-15K	0.6588	0.0684	0.7680	0.0717
GSM8K	0.6579	0.0726	0.7730	0.0901
Magicoder	0.7744	0.0700	0.8679	0.0765
MedQA	0.7439	0.0606	0.8798	0.0840
OpenMathReasoning	0.7781	0.0786	0.8732	0.1072
OpenThoughts	0.6940	0.0605	0.8617	0.0915
SciQ	0.6422	0.0632	0.7940	0.0863

G TOP-8 ENERGY COVERAGE

Energy overlap in Equation 4 is normalized by the current gradient's top-k energy. To quantify how much of the full gradient this represents, Table 17 reports the logged top-8 energy fraction $E^{(8)}_{\ell,m}(t)$ from Equation 2 for the same fifteen AdamW runs. We use the 223 post-reference measurement steps and 45,827 matrix–step observations underlying the temporal comparisons; the step-1 reference is excluded. Each matrix–step observation is counted once, since the input and output sides share the same singular values. All measured modules are included, including query/key projections and embedding tables.

Table 17: **Fraction of full gradient energy captured by the top eight singular components.** Statistics assign equal weight to each matrix–step observation within a run; the last row pools all observations. N is the number of observations, and Q25–Q75 gives their interquartile range. These are distributions of per-matrix energy fractions, not fractions of the model's summed gradient energy. Llama uses the Instruct checkpoints.

Dataset	Model	N	Mean	Median	Q25–Q75
APIGen	Qwen3-0.6B	1,379	0.661	0.684	[0.542, 0.818]
APIGen	Qwen3-1.7B	1,379	0.675	0.695	[0.541, 0.825]
Dolly-15K	Qwen3-1.7B	4,334	0.637	0.611	[0.491, 0.820]
GSM8K	Llama-3.2-1B	1,243	0.555	0.542	[0.403, 0.706]
GSM8K	Qwen3-1.7B	2,167	0.609	0.587	[0.438, 0.806]
Magicoder	Qwen3-1.7B	4,531	0.623	0.601	[0.472, 0.818]
MedQA	Qwen3-1.7B	2,955	0.671	0.655	[0.550, 0.843]
OpenMathReasoning	Llama-3.2-3B	2,955	0.491	0.470	[0.320, 0.649]
OpenMathReasoning	MiniCPM5-2B	9,176	0.688	0.695	[0.585, 0.820]
OpenMathReasoning	Qwen3-1.7B	2,955	0.658	0.674	[0.505, 0.839]
OpenMathReasoning	Qwen3-4B	3,795	0.611	0.607	[0.431, 0.818]
OpenMathReasoning	Qwen3-8B	762	0.605	0.582	[0.403, 0.842]
OpenThoughts	Llama-3.2-1B	1,695	0.533	0.515	[0.369, 0.688]
OpenThoughts	Qwen3-1.7B	2,955	0.622	0.619	[0.452, 0.820]
SciQ	Qwen3-1.7B	3,546	0.734	0.729	[0.611, 0.888]
All runs		45,827	0.638	0.651	[0.488, 0.815]

Across all observations, the mean top-8 energy fraction is 0.638 and the median is 0.651, with an interquartile range of $[0.488, 0.815]$; 73.47% of observations exceed 0.5. Within-run means range from 0.491 for Llama-3.2-3B on OpenMathReasoning to 0.734 for Qwen3-1.7B on SciQ. Thus, the eight leading components capture a substantial but incomplete fraction of gradient energy, with variation across matrices and runs. These measurements calibrate the fixed-k analysis; they do not establish that $k = 8$ captures nearly all energy or that the overlap ordering is invariant to the choice of k.

Relation to full-gradient energy retention. Suppressing the layer and matrix indices, let $G_t^{(k)}$ denote the rank-k truncated SVD of G_t, and define the historical projection operators $\mathcal{P}_{\mathsf{in},s}(G) = GQ_s^{\mathsf{in}}(Q_s^{\mathsf{in}})^\top$ and $\mathcal{P}_{\mathsf{out},s}(G) = Q_s^{\mathsf{out}}(Q_s^{\mathsf{out}})^\top G$. For either side z, the fraction of *full* gradient energy retained from the current top-k component is exactly

$$\frac{\|\mathcal{P}_{z,s}(G_t^{(k)})\|_F^2}{\|G_t\|_F^2} = E^{(k)}(t)\,\mathcal{O}_z^E(s,t). \tag{9}$$

Orthogonality of the retained and residual singular components gives

$$E^{(k)}(t)\,\mathcal{O}_z^E(s,t) \le \frac{\|\mathcal{P}_{z,s}(G_t)\|_F^2}{\|G_t\|_F^2} \le E^{(k)}(t)\,\mathcal{O}_z^E(s,t) + 1 - E^{(k)}(t). \tag{10}$$

The additional contribution from the residual spectrum is not measured by Equation 4. In particular, a top-k overlap of 0.8 is not itself an 80% full-gradient retention rate. The relation above holds per matrix and measurement; multiplying separately averaged energy fractions and overlaps does not in general recover the average product.

H Additional Experiments at Diagnostic Rank 16

H.1 Llama-3.2-1B-Instruct

We examine whether the input–output pattern also occurs at diagnostic rank $k = 16$ in an additional full-parameter AdamW run of Llama-3.2-1B-Instruct on OpenMathReasoning. This run is separate from the fifteen $k = 8$ runs summarized in the main text. It uses up to 10,000 training examples, seed 42, a maximum sequence length of 16,384, per-device batch size 1, and eight gradient-accumulation steps. The learning rate is 2×10^{-5}, with weight decay 0.01, Adam coefficients $(0.9, 0.999)$, $\epsilon = 10^{-8}$, a cosine schedule with 3% warmup, and gradient clipping at norm 1.0. Training uses BF16, gradient checkpointing, and no gradient projection, and ends at step 313.

Measurements and aggregation. The logs contain both sides of 113 parameter matrices at step 1 and every 20 steps through step 300. Excluding the step-1 reference gives 15 measurement points and 1,695 paired matrix–step observations. We use the logged current-top-k energy capture for each reference choice, corresponding to the current-energy weighting in Equation 4. The separately logged symmetric average of current- and historical-energy capture is not used here. Table 18 assigns equal weight to each matrix–step observation; these means differ in aggregation from the weighted model means in Equation 5.

Table 18: **Input–output energy overlap at diagnostic rank 16.** All observations are from the additional Llama-3.2-1B-Instruct run on OpenMathReasoning. Counts compare the two sides of the same matrix at the same measurement; they are correlated observations within one run.

Statistic	Initial reference	Local
Mean input-side overlap	0.6045	0.6840
Mean output-side overlap	0.2661	0.2856
Input-side advantage, all matrices	1,308/1,695 (77.17%)	1,301/1,695 (76.76%)
Input-side advantage, excluding Q/K/embedding	1,200/1,200 (100%)	1,200/1,200 (100%)
Query reversals	147/240 (61.25%)	154/240 (64.17%)
Key reversals	240/240 (100%)	240/240 (100%)

Results. The equal-weight mean input-side overlap exceeds its output-side counterpart at all 15 post-reference measurement points under both reference choices. All reversals occur in query/key projections; the embedding table favors the input side in all 15 comparisons of each type. Across the 1,695 observations, the top-16 fraction of full gradient energy has mean 0.6241, median 0.6130, and interquartile range $[0.4538, 0.7909]$. This is a distribution of per-matrix energy fractions, not a fraction of the model's summed gradient energy.

This additional experiment reproduces the input-side advantage and the concentration of reversals in query/key projections at $k = 16$, providing evidence that the pattern also occurs beyond the main diagnostic choice of $k = 8$. It is not a paired rank sweep on the same training trajectory: the fifteen main runs do not include this model–dataset combination. Accordingly, it does not isolate the effect of changing k or establish rank-independent ordering.

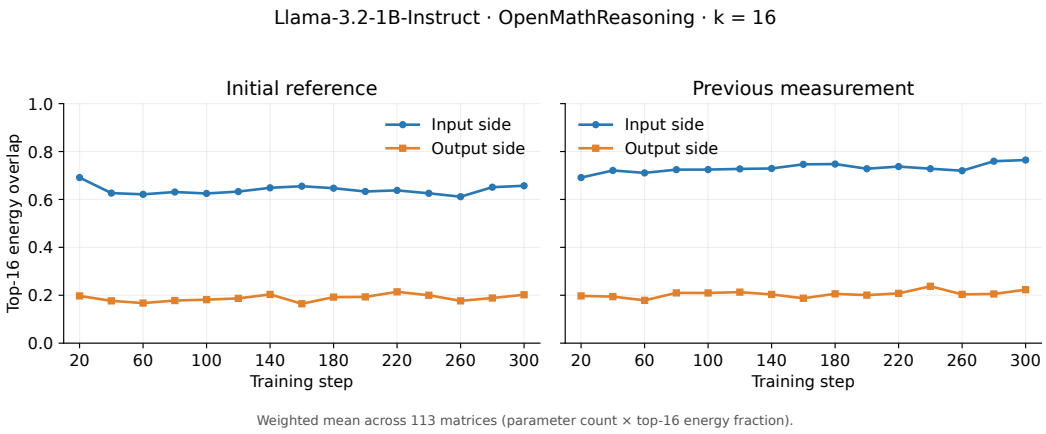


Figure 5: **Top-k energy overlap for Llama-3.2-1B-Instruct at $k = 16$.** Curves show weighted means across 113 matrices using Equation 5, for initial-reference (left) and local (right) comparisons. The initial step is excluded. These curves use the current-top-k energy overlap from Equation 4; the table above instead reports equal-weight means across matrix–step observations.

H.2 QWEN3-1.7B

We also evaluate top-k energy overlap at $k = 16$ in the additional Qwen3-1.7B run on OpenMathReasoning described in Appendix I. The 197 matrices and 15 post-reference measurements yield 2,955 matrix–step comparisons for each reference choice. Figure 6 uses the current-top-k energy overlap from Equation 4, with the parameter-count-times-top-k-energy weights from Equation 5.

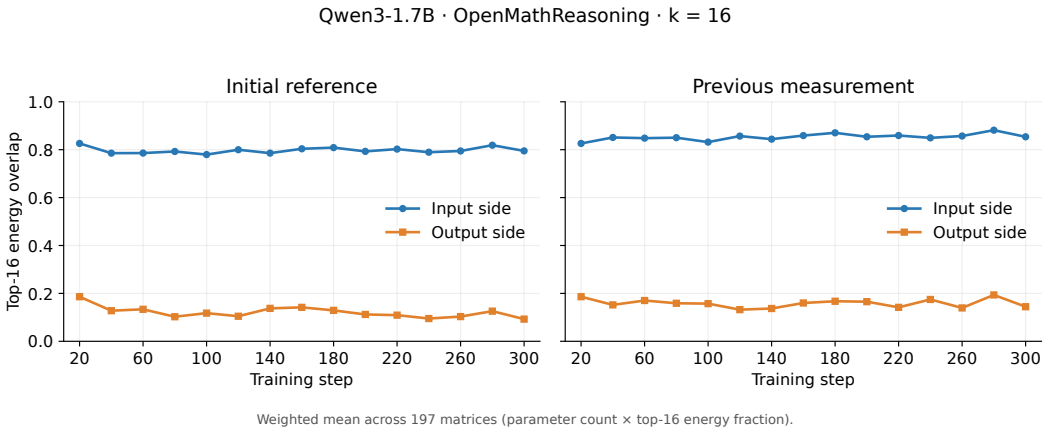


Figure 6: **Top-k energy overlap for Qwen3-1.7B at $k = 16$.** Input-side weighted mean overlap exceeds output-side overlap at every measured step under initial-reference (left) and local (right) comparisons. The initial step is excluded. This metric measures retention of the current top-k energy, rather than full-gradient energy retention.

The time-averaged weighted mean overlap is 0.7974 on the input side versus 0.1213 on the output side for initial-reference comparisons, and 0.8528 versus 0.1587 for local comparisons. At the individual-matrix level, input-side overlap is higher in 2,899/2,955 (98.10%) initial-reference and 2,886/2,955 (97.66%) local comparisons. Excluding query/key projections and embeddings, the proportion is 2,098/2,100 (99.90%) for both. The embedding favors the input side in all 15 comparisons of each type. Across matrix–step observations, the top-16 energy fraction has mean 0.7216 and median 0.7445. These results provide additional evidence of input–output asymmetry at $k = 16$ in Qwen3; they do not constitute a paired rank sweep on a fixed training trajectory.

I FULL-GRADIENT ENERGY RETENTION

We additionally measure how much of the full current gradient is retained by a historical rank-k subspace, including contributions outside the current top-k components. For a gradient G_t and historical orthonormal bases Q_s^{in} and Q_s^{out}, define

$$C_{\mathsf{in}}(s,t) = \frac{\|G_t Q_s^{\mathsf{in}}\|_F^2}{\|G_t\|_F^2}, \qquad C_{\mathsf{out}}(s,t) = \frac{\|(Q_s^{\mathsf{out}})^\top G_t\|_F^2}{\|G_t\|_F^2}. \tag{11}$$

Unlike the top-k energy overlap in Equation 4, these quantities directly measure the fraction of full gradient energy retained by projection onto each historical subspace.

We evaluate this metric in an additional full-parameter AdamW run of Qwen3-1.7B on OpenMath-Reasoning with diagnostic rank $k = 16$, without LoRA or gradient projection. The run uses up to 10,000 training examples, a maximum sequence length of 16,384, batch size 1, eight gradient-accumulation steps, and seed 42. The configured learning rate is 2×10^{-5}, with weight decay 0.01 and Adam coefficients $(0.9, 0.999)$. Training ends at step 313. We measure 197 parameter matrices at step 1 and every 20 steps through step 300, yielding 15 post-reference measurements. This run is separate from the fifteen $k = 8$ runs in the main text and the Llama $k = 16$ run in Appendix H.

At each measurement, we average the per-matrix retention fractions using parameter counts as weights:

$$\overline{C}_z(s,t) = \frac{\sum_{(\ell,m)\in\mathcal{M}} N_{\ell,m} C_{\ell,m,z}(s,t)}{\sum_{(\ell,m)\in\mathcal{M}} N_{\ell,m}}, \qquad z \in \{\mathsf{in}, \mathsf{out}\}. \tag{12}$$

These weights do not include the top-k energy fraction used in Equation 5. The resulting statistic is a weighted mean of per-matrix fractions, not a fraction of summed model gradient energy.

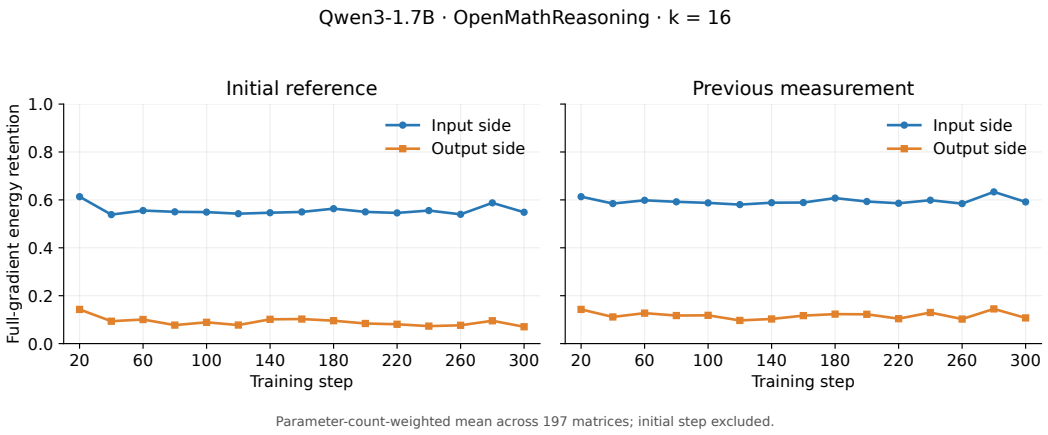


Figure 7: **Full-gradient energy retention on Qwen3-1.7B at $k = 16$.** The input side retains more energy than the output side at every measured step under both initial-reference (left) and local (right) comparisons. Curves show parameter-count-weighted means across 197 matrices. The initial step is excluded.

Table 19: **Time-averaged full-gradient energy retention.** Values average $\overline{C}_z(s,t)$ over the 15 post-reference measurements.

Comparison	Input side	Output side
Initial reference	0.5557	0.0906
Local	0.5952	0.1178

Input-side retention exceeds output-side retention at all 15 measurements under both reference choices (Figure 7). The time-averaged values are 55.57% versus 9.06% for the initial reference and 59.52% versus 11.78% for local comparisons (Table 19). The input-side advantage therefore extends to full-gradient energy retention in this additional trajectory at $k = 16$.

AI USE STATEMENT

We used generative AI tools to assist with translation, manuscript drafting and editing, literature search and reference formatting, interpretation of experimental results, and the creation and revision of analysis scripts and LaTeX source. Numerical summaries were checked against recorded experimental results, and bibliographic information was checked against available primary sources. The manuscript was compiled and its layout inspected after revisions. The authors are responsible for the accuracy and integrity of the final manuscript, code, and reported results.