GIST: Targeted Data Selection for Instruction Tuning via Coupled Optimization Geometry

Guanghui Min[1], Tianhao Huang[1], Ke Wan[1], Chen Chen[1]

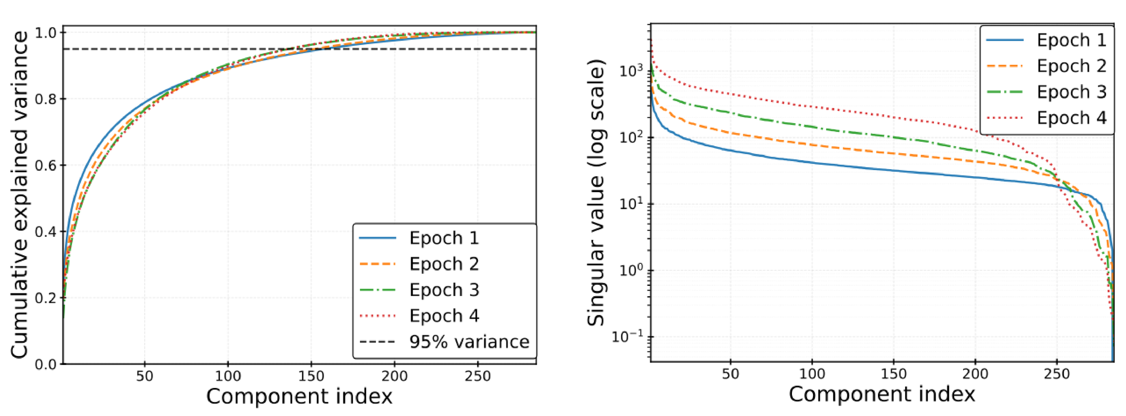
[1]University of Virginia

Paper Code

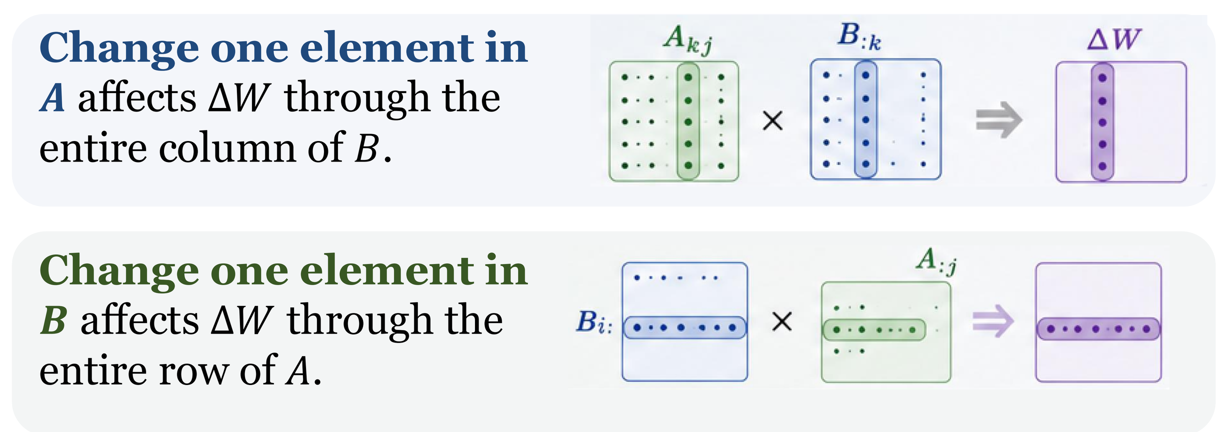
Motivation

- Targeted data selection aims to identify a small yet influential subset of instruction data that improves performance on a specific target task.
- Existing scalable methods like LESS are largely optimizer-oriented — they estimate influence through gradient alignment and diagonal optimizer statistics, such as Adam states.
- Optimization geometry — especially the coupled, low-rank geometry induced by LoRA-style fine-tuning — is mostly overlooked when measuring data influence.

Gradient Spectral Analysis: Low-rank → Not Axis-aligned



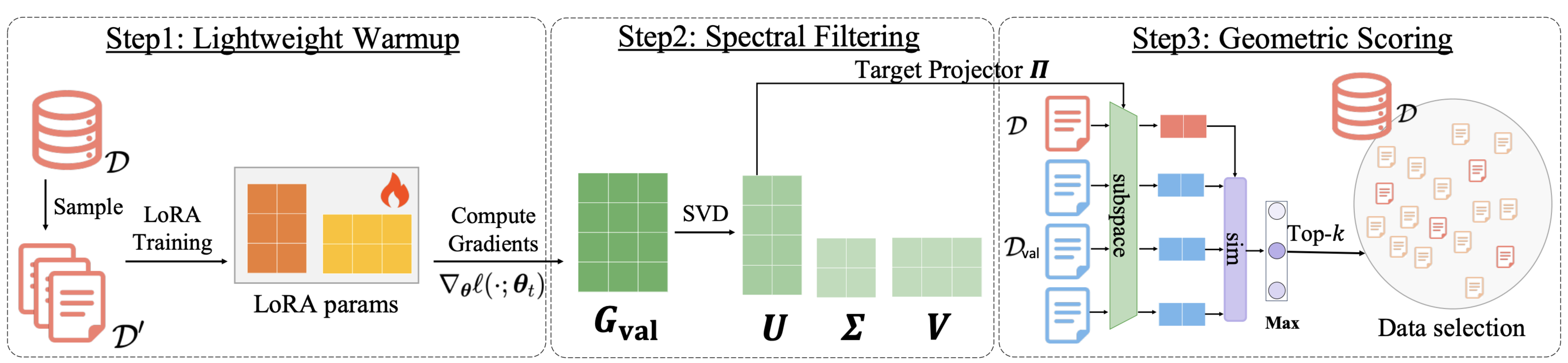
LoRA Parameterization Induces Coupling



- How should we design a geometry-aware data selection framework to effectively select target-relevant instruction data under coupled PEFT geometry?

Key takeaway: Good influence scores require good geometry; diagonal surrogates miss the coupled low-rank directions that matter in LoRA fine-tuning.

Proposed Framework: Gradient Isometric Subspace Transformation (GIST)



① **Warmup & Gradient Collection:** Run a short LoRA warmup and compute validation / candidate gradients in the LoRA parameter space.

② **Target Subspace Recovery:** Apply SVD to validation gradients to recover a low-rank task subspace that captures coupled target geometry.

③ **Projected Alignment Scoring:** Project candidate gradients into the target subspace and score them by alignment with target gradients:

$$\mathbf{Score}(z_i) = \max_j \cos(\Pi g_i, \Pi g_{\text{val}}^{(j)})$$

④ **Top-k Selection:** Select the highest-scoring examples for final instruction tuning.

Core idea: Replace diagonal rescaling with low-rank, non-diagonal task geometry.

Contributions

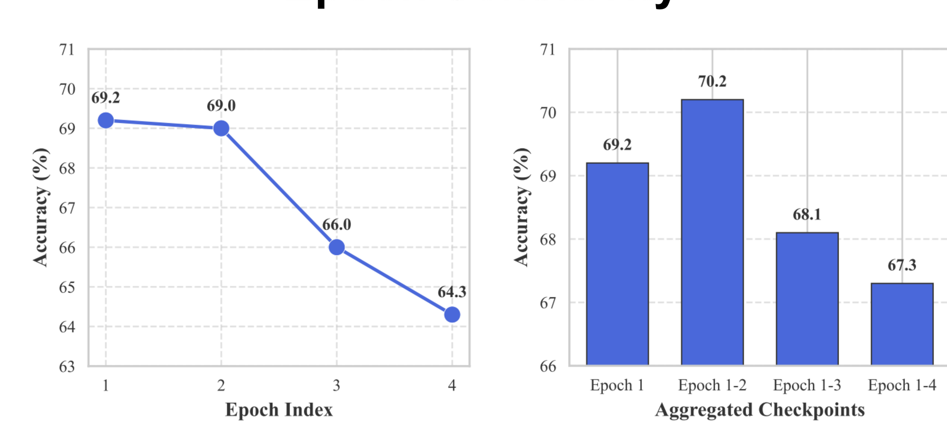
- **Theoretical Unification & Analysis.** We formulate targeted data selection as geometry-aware optimization and show that existing methods approximate the same objective under different surrogate geometries. This exposes a key limitation of diagonal optimizer-based preconditioners: **they miss rotated low-rank coupling induced by LoRA-style PEFT.**
- **GIST Algorithm.** We propose GIST, which recovers a low-rank task subspace from validation gradients via SVD and ranks candidates by projected target alignment.
- **Empirical Effectiveness.** Experiments on MMLU, TYDIQA, and BBH show that GIST matches or outperforms the state-of-the-art baseline under the same 5% selection budget, while using only **0.29% of the storage and 25% of the computational time**.

Experiments

Performance comparison

Model	Dataset	Base (0%)	Full (100%)	Rand.	Length	PPL.	Embed.	RDS+	LESS	GIST
Llama2-7B	MMLU	*45.6*	*51.6*	46.5±0.5	50.5±0.3	38.9±0.9	47.3±0.3	46.1±0.6	50.2±0.5	**51.2**±0.7
	TYDIQA	*46.4*	*54*	52.7±0.4	51.1±0.3	32.9±0.4	49.8±0.8	51.0±0.3	**56.2**±0.7	55.8±0.6
	BBH	*38.3*	*43.2*	38.9±0.5	39.8±0.7	34.5±0.2	38.6±0.6	39.0±0.7	41.5±0.6	**41.8**±0.8
	Avg. Δ	–	*+6.2*	+2.6	+3.7	-8.0	+1.8	+1.9	+5.9	**+6.2**
Llama3.2-3B	MMLU	*53.9*	*55.2*	53.2±0.6	**57.6**±0.9	51.8±0.3	53.9±0.6	51.9±0.2	56.5±0.6	56.1±0.4
	TYDIQA	*60.4*	*66.6*	64.1±0.4	63.9±1.3	57.1±0.7	62.8±0.4	62.6±0.5	67.1±0.8	**69.2**±0.3
	BBH	*45.5*	*47.8*	45.1±0.2	45.3±0.3	43.0±0.4	44.6±0.5	45.1±0.4	46.1±0.7	**48.0**±0.5
	Avg. Δ	–	*+3.3*	+0.9	+2.3	-2.6	+0.5	-0.1	+3.3	**+4.5**
Qwen2.5-1.5B	MMLU	*62.0*	*62.3*	61.5±0.3	62.8±0.1	61.7±0.2	62.3±0.2	62.4±0.2	62.2±0.2	**62.9**±0.6
	TYDIQA	*57.8*	*59.9*	55.6±0.3	56.7±0.3	58.4±0.4	54.1±0.3	56.2±0.3	**61.4**±0.8	61.2±0.5
	BBH	*43.8*	*44.0*	43.0±0.4	42.8±0.2	43.1±0.2	43.6±0.2	39.0±0.6	43.2±0.5	**44.1**±0.4
	Avg. Δ	–	*+0.9*	-1.2	-0.4	-0.1	-1.2	-2	+1.1	**+1.5**

Epoch Sensitivity



(a) Single Epoch Performance (b) Cumulative Performance

Efficiency Comparison

Method	Metric	Warmup	Target SVD	Grad. Feats.
LESS	Time	6.0 h	–	48.0 h
	Compl.	$\mathcal{O}(\lvert\mathcal{D}'\rvert \cdot N)$	–	$\mathcal{O}(\lvert\mathcal{D}\rvert \cdot N)$
GIST	**Time**	**1.5 h**	**< 1 m**	**12.0 h**
	Compl.	$\mathcal{O}(\lvert\mathcal{D}'\rvert)$	$\mathcal{O}(\lvert\mathcal{D}_{\text{val}}\rvert^2 d)$	$\mathcal{O}(\lvert\mathcal{D}\rvert)$

Observations

- GIST achieves the best Avg. Δ on all three backbones; 5% selected data can match or beat full fine-tuning and improves over LESS by +0.3/+1.2/+0.4.
- A short warmup with **one early checkpoint** is enough to recover useful target geometry; later or aggregated checkpoints add little.
- By avoiding multi-checkpoint feature extraction, GIST cuts per-task runtime **from 54.0h to 13.5h**, adds negligible SVD overhead, and **reduces storage from 75GB to 217MB (350× smaller)**.

Let's Connect!

I welcome research collaborations and am actively seeking Summer 2027 internships and academic-year internship opportunities. **Interests:** efficient LLM post-training, data selection & curation and long-horizon agents.

Email: jjm8vr@virginia.edu