PHLoRA: Data-Free Post-Hoc Low-Rank Adapter Extraction from Full-Rank Checkpoint

Bhoomit Vasani
Amazon AGI
vbhoomit@amazon.com

Jack FitzGerald
EdgeRunner AI
jack@edgerunnerai.com

Anjie Fang
Amazon AGI
njfn@amazon.com

Sushmit Vaish
Amazon AGI
sushvai@amazon.com

Abstract

We introduce **PHLoRA**[1] (Post-hoc LoRA), a simple yet powerful method to extract low-rank adapters from full-rank fine-tuned models without requiring access to training data or gradients. By computing the low-rank decomposition of weight differences between a base model and its fine-tuned counterpart, our method reconstructs adapter modules that can be merged, dynamically routed at inference time via S-LoRA, or served in scalable industry settings using platforms like NVIDIA NIM. This approach amortizes latency overhead across requests and yields substantial cost savings. Unlike prior work that trains each adapter explicitly, our approach decouples fine-tuning from adapter generation, allowing adapter extraction from existing full-rank models or third-party checkpoints. Experiments on text, image, and video benchmarks using the Amazon Nova model family demonstrate that extracted adapters preserve high energy from the full weight delta, can be pruned safely, and yield negligible degradation in downstream task performance when re-merged. Overall, PHLoRA provides a practical path for making all existing full-rank checkpoints adapter-ready, democratizing scalable inference for all models.

1 Introduction

The Low-Rank Adapters (LoRA) technique [Hu et al., 2022] is a popular way to reduce memory during training, and it offers an additional advantage at inference: it allows a single server to host adapters for hundreds or thousands of users in a shared inference API, as in S-LoRA [Sheng et al., 2024]. Modern industry platforms such as NVIDIA NIM support scalable, low-latency serving of LoRA-based adapters in production. However, many practitioners have existing models trained with full-rank fine-tuning, including through the use of other training methods beyond standard fine-tuning like DPO [Rafailov et al., 2024] or PPO [Schulman et al., 2017]. To serve these users, we introduce and evaluate a method for compressing full-rank updates into low-rank adapters compatible with dynamic serving frameworks, called **Post-hoc Low-Rank Adapter Extraction** (PHLoRA). Our contributions include the following:

- **Post-hoc LoRA formulation:** We pose adapter extraction as a low-rank decomposition solved with truncated SVD over the checkpoint's weight delta; it does not require any gradients or data.
- **LoRA Rank compression:** PHLoRA can also be used to compress rank of existing LoRA adapters (e.g., convert LoRA trained with rank 128 to rank 32)
- **Flexible deployment and fast start-up:** Compact adapters cut model-load latency by over 10× compared to full-rank checkpoints and can be merged for static inference or dynamically routed via shared-adapter execution (e.g., S-LoRA), and are compatible with scalable industry platforms such as NVIDIA NIM, to minimize run-time cost.
- **Multimodal results:** We evaluate on three text, one image, and one video understanding benchmark, showing that PHLoRA preserves performance while reducing inference cost by up to 4×.

2 Background and Related Work

PHLoRA uniquely provides constant-cost, post-hoc adapter generation that is fully LoRA-inference-compatible for both text and multimodal settings [Sung et al., 2022], with further comparisons in Table 1.

[1] Pronounced "flora".

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1992–1999
December 20-24, 2025 ©2025 Association for Computational Linguistics

Method	Stage	Input	SVD On	Output	LoRA-comp.	Task-spec. Train?	Needs Data?
PHLoRA (ours)	Post-hoc	ΔW	ΔW	LoRA A, B	✓	✗	✗
SLiM	Post-hoc	W	W	LR+Q weights	✗	✗	✗
SVD-LLM	Post-hoc	W	W	Trunc. LR model	✗	✗	✗
SVDQuant	Post-hoc	W	W	LR+Q weights	✗	✗	✗
Dobi-SVD	Post-hoc+Grad	W	Diff. SVD	Compressed model	✗	✓	✓
SORSA	PEFT init	W	W	Struct. adapter	△	✓	✓
PiSSA	PEFT init	W	W	Init. adapter	✓	✓	✓

Table 1: Qualitative comparison of PHLoRA and related approaches. ✓: yes; ✗: no; △: partially.

LoRA inserts rank-r matrices in parallel with linear layers and trains only these additions, reducing memory and compute [Hu et al., 2022]. LoRA+ further re-balances the optimizer by raising the learning rate on the B matrix [Hayou et al., 2024]. Other variants explore dynamic rank schedules (AdaLoRA [Zhang et al., 2023]), quantized training (QLoRA [Dettmers et al., 2023]), and selective layer targeting. Soft prompt-tuning [Lester et al., 2021], BitFit [Zaken et al., 2021], AdapterFusion [Pfeiffer et al., 2021], and VL-Adapter [Sung et al., 2022] trade different portions of trainable parameters for efficiency, but all require task-specific optimization. Recent methods extend parameter-efficient transfer to vision-language models [Sung et al., 2022]. PiSSA initializes LoRA adapters with principal singular vectors *before* adapter training, accelerating convergence but not eliminating the need for training [Meng et al., 2025]. SLiM [Mozaffari et al., 2025], SVD-LLM [Wang et al., 2025b], and SVDQuant [Li et al., 2025] apply low-rank decomposition (often combined with quantization) directly to pretrained weights W for inference compression and acceleration. GPTQ [Frantar et al., 2023] is another widely-used post-hoc quantization approach. However, these methods do not expose LoRA-compatible factors nor leverage the fine-tuning delta. Dobi-SVD [Wang et al., 2025a] makes SVD differentiable and tunes the factors with task supervision, achieving lower reconstruction error at the cost of additional gradient steps. SORSA [Cao, 2024] proposes a structured low-rank adaptation that replaces dense LoRA matrices but still requires full adapter training.

While prior works have explored low-rank approximation techniques for fine-tuning (e.g., Hu et al., 2022, Zhang et al., 2023), we also found a recent GitHub implementation, LoRD [Gauthier-Caron, 2024], that performs similar post-hoc low-rank extraction, though without an associated peer-reviewed manuscript.

3 Methodology

3.1 Problem Setup

Given a pretrained model and a fine-tuned model, each consisting of weights, we define the weight delta as

$$\Delta W = W_{\text{ft}} - W_{\text{base}}, \text{ where } W \in \mathbb{R}^{d\times k} \tag{1}$$

Our objective is to approximate each ΔW with a rank-r factorization in the LoRA form:

$$\Delta W \approx BA, \text{where } A \in \mathbb{R}^{r\times k}, B \in \mathbb{R}^{d\times r} \tag{2}$$

Once A and B are obtained, they can be deployed as standard LoRA adapters (for dynamic or conditional routing) or merged back into the backbone via $W_{\text{base}} \leftarrow W_{\text{base}} + BA$. This process is repeated for all target components (typically attention and MLP submodules).

3.2 Post-hoc LoRA Extraction

We perform a truncated singular value decomposition (SVD) on ΔW:

$$\begin{gathered} U\Sigma V^{\top} = \text{SVD}(\Delta W), \text{where} \\ U \in \mathbb{R}^{d\times d}, \quad \Sigma \in \mathbb{R}^{d\times k}, \quad V \in \mathbb{R}^{k\times k} \end{gathered} \tag{3}$$

The low-rank LoRA factorization is then:

$$\begin{aligned} B &= U_{[:,:r]}\Sigma^{\frac{1}{2}}_{[:r,:r]} \\ A &= \Sigma^{\frac{1}{2}}_{[:r,:r]} V^{\top}_{[:r,:]} \end{aligned} \tag{4}$$

where the first r columns of U, the first r rows of V, and the first r rows and columns of Σ are taken, and the $\frac{1}{2}$ exponent represents the element-wise square root. This SVD-based decomposition ensures that BA is the best rank-r approximation of ΔW [Eckart and Young, 1936]. All computations are performed independently for each target weight matrix (e.g., q_{proj}, k_{proj}, mlp_{fc1}).

Merged inference computes $W_{\text{base}} + BA$ once, fully restoring the original fine-tuned model up to

Dataset	**Metric(s)**
TAT–QA [Zhu et al., 2021]	Accuracy, Exact Match (EM)
MKFE [Owkin, 2024]	Key Overlap Mean, Value Overlap Mean (%)
MedMCQA [Pal et al., 2022]	Accuracy, F1
VQA–RAD [Lau et al., 2018]	Avg. Normalized Similarity
CaptionGen [Chen and Dolan, 2011]	ROUGE_L, CIDEr

Table 2: Primary evaluation metrics and citations for each benchmark.

truncation error (no runtime adapter overhead). Dynamic routing, as in S-LoRA [Sheng et al., 2024], loads A and B as lightweight adapters and activates them on demand, enabling low-cost serving of multiple adapters in a single process.

PHLoRA is compatible with the HuggingFace PEFT library [Hugging Face, 2023], PyTorch LoRA implementations, and multi-adapter serving frameworks. No access to gradients or training data is needed, but only the base and fine-tuned checkpoints.

3.3 Energy-Based Analysis

In low-rank matrix approximation, the *energy* of a matrix refers to the sum of the squares of its singular values, quantifying the total information content or signal present in the matrix. For a weight delta ΔW with singular values $\sigma_1, \sigma_2, \ldots \sigma_d$, we define the preserved energy at rank r as

$$E_r = \frac{\sum_{i=1}^{r} \sigma_i^2}{\sum_{i=1}^{d} \sigma_i^2}. \qquad (5)$$

Intuitively, E_r measures what fraction of the “important” weight update is retained by the top-r singular directions. High preserved energy typically correlates with the adapter’s ability to recover full-rank performance.

In Figure 1, we present the values of E_r as the rank parameter varies across three modalities and three Nova model sizes. We observe a consistent pattern: energy preservation improves with higher rank values, independent of modality or model size. Instead of exploring all possible ranks, this paper focuses on three representative settings, 32, 64, and 512, which correspond to low, medium, and high energy levels, respectively. This selection enables us to analyze the approach’s performance under different energy conditions.

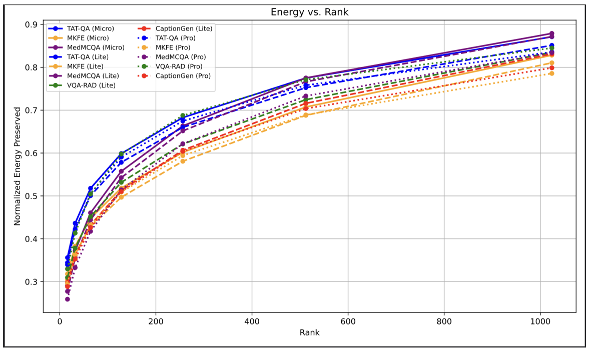

Figure 1: Preserved energy vs rank.

4 Experiments

4.1 Experimental Setup

We benchmark **PHLoRA** on three text only datasets, *TAT-QA* [Zhu et al., 2021], *Medical Knowledge from Extracts (MKFE)* [Owkin, 2024], *MedMCQA* [Pal et al., 2022]; one image and text dataset, *VQA-RAD* [Lau et al., 2018]; and one video and text dataset, *CaptionGen* [Chen and Dolan, 2011]. We sub-sample and reformat the datasets. See Table 2 for primary metrics, and Appendix A for detailed statistics. All experiments use the Nova model family.

We compare: (i) the base model, (ii) the full-rank fine-tuned model, (iii) LoRA+ with rank 32, and (iv) PHLoRA with rank 32 (default) and 64 (see Section 4.3 for larger r). We use an evaluation prompt that does not specify formatting, which results in many formatting errors in the base model. Though the base model could be improved substantially with prompt optimization, our choice of prompts accentuates the effect of fine-tuning.

4.2 Results and Analysis

Table 3 reports test-set performance across all model sizes and benchmarks. For each task, we include one or two evaluation metrics (e.g., Accuracy / Exact Match), with the best score for each metric shown in **bold**. PHLoRA demonstrates consistency with full-rank fine-tuning across the Nova Micro, Lite, and Pro model families, often coming within 1% performance while occasionally even surpassing full-rank results.

On Nova Micro, full-rank leads on TAT-QA, but PHLoRA remains close and even surpasses it on MedMCQA, with only minor gaps on MKFE. On

Table 3: Test set results for Nova Micro (text-only), Lite, and Pro. Where two metrics are shown, the best per metric is **bolded**. The evaluation prompts do not provide formatting instructions, substantially increasing the difficulty of the task for the base model prior to fine-tuning.

Dataset (Metric)	**Base Model**	**Full-Rank**	**LoRA+ (r32)**	**PHLoRA (r32)**	**PHLoRA (r64)**
Nova Micro					
TAT–QA (Acc / Exact Match)	0 / 0	**84.95 / 51.68**	82.47 / 48.49	83.54 / 48.14	84.07 / 49.02
MKFE (Key Overlap / Val Overlap)	50.0 / 22.0	**100.0 / 28.0**	**100.0** / 26.75	99.0 / 27.75	99.0 / **28.0**
MedMCQA (Accuracy / F1)	0.03 / 0.05	60.49 / 60.52	60.52 / 60.52	**60.79 / 60.85**	60.60 / 60.63
Nova Lite					
TAT–QA (Acc / Exact Match)	0 / 0	83.89 / 48.85	82.48 / 48.50	85.84 / 52.39	**86.02 / 53.45**
MKFE (Key Overlap / Val Overlap)	49.50 / 19.50	**99.50** / 22.75	99.0 / 24.75	**99.50 / 26.50**	**99.50** / 26.0
MedMCQA (Accuracy / F1)	0.19 / 0.38	63.40 / 63.33	59.52 / 59.52	64.11 / 64.07	**64.30 / 64.25**
VQA–RAD (Avg Norm Similarity)	23.22	54.87	57.15	**58.56**	57.56
CaptionGen (ROUGE_L / CIDEr)	31.37 / 0.81	49.40 / 1.43	48.26 / 1.46	49.19 / **1.51**	**49.43** / 1.50
Nova Pro					
TAT–QA (Acc / Exact Match)	0 / 0	**89.38 / 62.48**	87.79 / 53.98	87.96 / 55.22	89.00 / 58.00
MKFE (Key Overlap / Val Overlap)	50.0 / 17.0	99.50 / 24.75	99.50 / **25.50**	**100.0** / 24.0	**100.0** / 25.0
MedMCQA (Accuracy / F1)	0 / 0	69.40 / 69.42	**71.14 / 71.16**	70.0 / 70.0	70.0 / 70.0
VQA–RAD (Avg Norm Similarity)	27.03	56.20	56.57	**57.58**	56.92
CaptionGen (ROUGE_L / CIDEr)	37.63 / 1.13	48.94 / 1.37	**50.12 / 1.55**	48.55 / 1.45	48.85 / 1.48

Nova Lite, PHLoRA (r64) delivers the best scores on TAT-QA, MedMCQA, VQA-RAD, and CaptionGen, which highlights its strength in reasoning and multimodal tasks. Nova Pro further demonstrates scalability: PHLoRA nearly matches full-rank on TAT-QA and outperforms it on MedMCQA and VQA-RAD, while it also remains competitive on CaptionGen. Overall, the margin between PHLoRA and full-rank shrinks as the Nova model scales, with PHLoRA often taking the lead.

Inference Cost and Latency. PHLoRA, when merged into the backbone ("m-packed" as in S-LoRA [Sheng et al., 2024]), is computationally equivalent to full-rank and merged LoRA inference for a single adapter or task. All three approaches require only a single matrix multiplication per layer. For scalable multi-adapter deployment, we estimate cost and throughput improvements using S-LoRA-like dynamic routing [Sheng et al., 2024], which achieves up to 4× higher throughput and cost efficiency than naive dynamic LoRA serving (e.g., PEFT or vLLM) in multi-tenant settings, as shown in Table 3 and Figure 4 of S-LoRA. These reference results provide a strong indication that PHLoRA, when paired with S-LoRA-like serving, is highly cost-effective for scalable, multi-user inference scenarios.[2]

4.3 Ablation: Rank and Energy Preservation

We vary the PHLoRA rank (from 32 to 512) and report preserved energy E_r (as defined in Equation 5). The results across all three Nova model sizes (Micro, Lite, Pro) are presented in Tables 4, 5, and 6, where each table reports test-set scores alongside preserved energy (E_r, %) for different PHLoRA ranks and the full-rank reference.

Across all three Nova model scales (Micro, Lite, Pro), PHLoRA rank shows a clear correlation between preserved energy (E_r) and downstream task performance. Higher ranks consistently recover full-rank accuracy, while lower ranks maintain strong results with considerable efficiency gains. For Nova Micro (text-only), performance is stable across ranks, with r512 closely matching or slightly exceeding full-rank metrics on MedMCQA. In Nova Lite (multimodal), intermediate ranks such as r64 achieve performance comparable to or better than full-rank on tasks like VQA-RAD and CaptionGen. Similarly, in Nova Pro, r32 and r64 occasionally surpass full-rank scores, particularly in multimodal settings, though MKFE value overlap metrics appear more sensitive to rank and do not always improve with higher E_r. Overall, higher PHLoRA ranks reliably recover accuracy, while intermediate ranks can offer a strong balance between efficiency and performance across different model sizes and tasks.

5 Conclusion

We presented PHLoRA, a practical post-hoc method for deriving LoRA-compatible adapters directly from fully fine-tuned models, without requiring access to training data or gradients. Our experiments focused on three modalities—text, image,

[2] We use "S-LoRA-like" to refer to any scalable, dynamic multi-adapter LoRA serving implementation; S-LoRA [Sheng et al., 2024] is used as a reference.

Table 4: Ablation study for Nova Micro (text-only). We show one or two evaluation metrics, with the best score for each metric shown in **bold** and E_r (preserved energy, %) for PHLoRA in parentheses.

Dataset (Metric)	**Full-Rank**	**PHLoRA (r32)**	**PHLoRA (r64)**	**PHLoRA (r512)**
TAT–QA (Accuracy / EM)	**84.96 / 51.68**	83.54 / 48.14 (44)	84.07 / 49.03 (52)	**84.96** / 51.15 (77)
MKFE (Key Overlap / Value Overlap)	**100.0** / 28.0	99.0 / 27.75 (38)	99.0 / 28.0 (45)	**100.0 / 28.75** (71)
MedMCQA (Accuracy / F1)	60.49 / 60.52	60.79 / 60.85 (37)	60.60 / 60.63 (46)	**60.93 / 60.94** (78)

Table 5: Ablation study for Nova Lite (text, image, video). We show one or two evaluation metrics, with the best score for each metric shown in **bold** and E_r (preserved energy, %) for PHLoRA in parentheses.

Dataset (Metric)	**Full-Rank**	**PHLoRA (r32)**	**PHLoRA (r64)**	**PHLoRA (r512)**
TAT–QA (Accuracy / EM)	83.89 / 48.45	85.84 / 52.39 (42)	86.02 / 53.45 (49)	**86.90 / 55.58** (74)
MKFE (Key Overlap / Value Overlap)	**99.50** / 22.75	**99.50 / 26.50** (36)	**99.50** / 26.00 (42)	99.0 / 24.75 (68)
MedMCQA (Accuracy / F1)	63.40 / 63.33	64.11 / 64.07 (35)	**64.30 / 64.25** (44)	63.83 / 63.77 (76)
VQA–RAD (Similarity)	54.87	58.57 (37)	**57.56** (45)	55.01 (71)
CaptionGen (ROUGE_L / CIDEr)	49.40 / 1.43	49.19 / **1.51** (36)	49.43 / 1.50 (43)	**49.84** / 1.50 (71)

and video—using three Amazon Nova [AGI, 2024] models and five moderate-sized benchmarks, all in the supervised fine-tuning (SFT) setting. PHLoRA maintains competitive task accuracy while reducing inference GPU-hour costs by up to 4-fold compared to merged adapter inference, and by a similar or greater margin compared to full-rank model inference, in dynamic multi-adapter routing scenarios such as S-LoRA. This cost reduction reflects improvements in inference throughput, i.e., the number of tokens or requests processed per unit time, as demonstrated in S-LoRA [Sheng et al., 2024].

PHLoRA provides a practical path for making all existing full-rank checkpoints adapter-ready, democratizing scalable inference for legacy models.

6 Future Work

Several avenues remain for future research:

- **Scaling to Larger and More Diverse Tasks:** Our current experiments are limited to moderate-sized SFT datasets. Future work should evaluate PHLoRA on larger-scale, more challenging benchmarks and additional modalities.

- **Advanced Tuning Strategies:** Extending PHLoRA to support advanced fine-tuning techniques such as DPO, PPO, or reward-based learning.

- **Extending Beyond Linear Layers:** While LoRA has been generalized to convolutions [Zhong et al., 2024], post-hoc SVD-based extraction for higher-order tensors requires further research, potentially leveraging advanced tensor decompositions [Kolda and Bader, 2009] or alternative adapter parametrizations [Chen et al., 2023].

- **Rank Selection and Usability:** Further developing practical methods for adaptive, data-free, or black-box rank selection, and enabling adapter extraction even when the base model is unavailable.

- **Empirical data displaying inference efficiency improvements:** Our current experiments are limited to generation of LoRA adapters. Future work should evaluate the empirical data displaying inference efficiency improvement with the generated adapters.

Limitations

While **PHLoRA** offers a simple and effective post-hoc mechanism for adapter extraction, it comes with several important limitations.

PHLoRA is currently designed for standard linear (matrix-shaped) layers, as it relies on singular value decomposition (SVD) to extract low-rank adapters from weight differences. While LoRA and similar adapters have been extended to convolutional layers — either via kernel reshaping or structured convolutional approximations (e.g., [Zhong et al., 2024]), post-hoc SVD extraction for convolutions or other higher-order tensors is non-trivial and depends on the decomposition or flattening strategy, which may lose spatial structure or interpretability. More generally, advanced tensor decompositions [Kolda and Bader, 2009] or alternative adapter parametrizations [Chen et al., 2023] would be required for such modules, which we leave to future work. Note also that attention

Table 6: Ablation study for Nova Pro (text, image, video). We show one or two evaluation metrics, with the best score for each metric shown in **bold** and E_r (preserved energy, %) for PHLoRA in parentheses.

Dataset (Metric)	Full-Rank	PHLoRA (r32)	PHLoRA (r64)	PHLoRA (r512)
TAT–QA (Accuracy / EM)	**89.38 / 62.48**	87.96 / 55.22 (42)	89.0 / 58.0 (50)	89.0 / 61.0 (75)
MKFE (Key Overlap / Value Overlap)	99.50 / 24.75	**100.0** / 24.00 (36)	**100.0 / 25.0** (43)	50.0 / 23.0 (68)
MedMCQA (Accuracy / F1)	69.4 / 69.42	**70.0 / 70.0** (35)	**70.0 / 70.0** (41)	**70.0** / 69.70 (73)
VQA–RAD (Similarity)	56.20	**57.58** (41)	56.92 (50)	57.07 (77)
CaptionGen (ROUGE_L / CIDEr)	**48.94** / 1.37	48.55 / 1.45 (35)	48.85 / **1.48** (42)	48.87 / 1.39 (70)

"caches" refer to runtime data, not persistent parameters, and so are out of scope for PHLoRA.

In this paper, we fix the adapter rank r globally for all layers. Although we analyze energy-based rank selection, adaptive or per-layer rank scheduling—which could further improve the efficiency/accuracy tradeoff—remains for future work. Furthermore, while the preserved energy metric (E_r) is a useful indicator of information retention at a given rank, model quality on the target task does not always correlate perfectly with energy preservation. Thus, optimal adapter rank cannot be reliably selected solely from energy curves; empirical evaluation remains necessary.

PHLoRA assumes access to both base and fully fine-tuned weights. In settings where only the fine-tuned model is available (e.g., closed-source vendors), post-hoc adapter extraction is not directly possible.

The principal benefits of PHLoRA are realized in dynamic inference scenarios (e.g., S-LoRA or multi-adapter routing), where multiple adapters are loaded or swapped at runtime. In conventional merged-inference pipelines—where a single adapter is fused into the model for all requests—the practical advantage of post-hoc extraction is diminished, as cost and latency resemble standard LoRA or full-rank fine-tuning.

Our evaluation is limited to a set of public text, image, and video benchmarks. Results may differ for larger, more diverse real-world applications. While PHLoRA enables substantial inference cost reductions with dynamic adapter routing, there remains a modest runtime latency penalty versus full-rank merging; practical savings will depend on system-level batch sizes and workload characteristics.

We encourage future work to address these limitations by extending PHLoRA to non-linear modules, developing robust energy-aware or data-free rank selection strategies, enabling black-box or partial-weight extraction, and improving dynamic adapter composition schemes.

References

Amazon AGI. 2024. The amazon nova family of models: Technical report and model card.

Yang Cao. 2024. Sorsa: Singular values and orthonormal regularized singular vectors adaptation of large language models. *Preprint*, arXiv:2409.00055.

David Chen and Bill Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Collecting Highly Parallel Data for Paraphrase Evaluation*. Association for Computational Linguistics.

Jiaao Chen, Aston Zhang, Xingjian Shi, Mu Li, Alex Smola, and Diyi Yang. 2023. Parameter-efficient fine-tuning design spaces. *Preprint*, arXiv:2301.01821.

DAMO-NLP-SG. Multi-source video captioning dataset. `https://huggingface.co/datasets/DAMO-NLP-SG/Multi-Source-Video-Captioning`. Accessed: 2025-05-19.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, Kai Tai, Arthur Szlam, Ari S. Tillman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*.

Carl Eckart and G. Marion Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. Gptq: Accurate post-training quantization for generative pre-trained transformers. *Preprint*, arXiv:2210.17323.

Thomas Gauthier-Caron. 2024. Lord: Low-rank decomposition from full-rank fine-tuning. `https://github.com/thomasgauthier/LoRD`. Accessed: 2025-06-15.

Gene H. Golub and Charles F. Van Loan. 2013. *Matrix Computations*, 4th edition. Johns Hopkins University Press, Baltimore, MD, USA.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. 2024. Lora+: Efficient low rank adaptation of large models. *Preprint*, arXiv:2402.12354.

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

Hugging Face. 2023. Peft: Parameter efficient fine-tuning library. https://github.com/huggingface/peft. Accessed: 2025-05-18.

Tamara G. Kolda and Brett W. Bader. 2009. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *Preprint*, arXiv:2104.08691.

Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. 2025. Svdquant: Absorbing outliers by low-rank components for 4-bit diffusion models. *Preprint*, arXiv:2411.05007.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2025. Pissa: Principal singular values and singular vectors adaptation of large language models. *arXiv preprint arXiv:2404.02948*.

Mohammad Mozaffari, Amir Yazdanbakhsh, and Maryam Mehri Dehnavi. 2025. Slim: One-shot quantization and sparsity with low-rank approximation for llm weight compression. *Preprint*, arXiv:2410.09615.

Owkin. 2024. Medical knowledge from extracts (mkfe). https://huggingface.co/datasets/owkin/medical_knowledge_from_extracts. Accessed: 2025-05-18.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347.

Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, Joseph E. Gonzalez, and Ion Stoica. 2024. S-lora: Serving thousands of concurrent lora adapters. *Preprint*, arXiv:2311.03285.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. *Preprint*, arXiv:2112.06825.

Qinsi Wang, Jinghan Ke, Masayoshi Tomizuka, Yiran Chen, Kurt Keutzer, and Chenfeng Xu. 2025a. Dobi-svd: Differentiable svd for llm compression and some new perspectives. *Preprint*, arXiv:2502.02723.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. 2025b. Svd-llm: Truncation-aware singular value decomposition for large language model compression. *Preprint*, arXiv:2403.07378.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1–16.

Xiaohui Zhang, Xiang Lisa Li, and 1 others. 2023. Adaptive low-rank adapter (adalora): Compressing lora further via rank allocation. *arXiv preprint arXiv:2303.10512*.

Zihan Zhong, Zhiqiang Tang, Tong He, Haoyang Fang, and Chun Yuan. 2024. Convolution meets lora: Parameter efficient finetuning for segment anything model. *Preprint*, arXiv:2401.17868.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *Preprint*, arXiv:2105.07624.

A Dataset Descriptions

We provide summary statistics and descriptions for each benchmark used in this study.

- **TAT-QA** [Zhu et al., 2021]: A table-augmented question answering dataset in the financial domain, requiring models to reason over both natural language and tabular data. *Train:* 2,830; *Test:* 565. License: MIT. Evaluated using Accuracy and Exact Match.

- **MKFE** [Owkin, 2024]: Medical Knowledge from Extracts. Evaluates the ability to extract structured key-value medical facts from unstructured text. *Key Overlap* measures the proportion of gold-standard keys correctly predicted; *Value Overlap* measures the fraction of correct values among matched keys. *Train:* 1,000; *Test:* 200. License: Apache 2.0.

- **MedMCQA** [Pal et al., 2022]: A large-scale medical multiple-choice question answering dataset. *Train:* 20,000; *Test:* 3,683. License: MIT. Evaluated using Accuracy and F1.

- **VQA-RAD** [Lau et al., 2018]: Visual question answering over radiology images, requiring both visual and textual understanding. *Train:* 1,793; *Test:* 451. License: CC0 1.0 Universal. Evaluated by average normalized similarity.

- **CaptionGen**: A video captioning benchmark with 2,000 training and 500 test examples. Videos are sourced from MSVD [Chen and Dolan, 2011]; captions are from the Multi-Source Video Captioning dataset [DAMO-NLP-SG]. License: MIT. Evaluated using ROUGE_L and CIDEr.

B Implementation Details

Hardware. All experiments were performed on AWS P5.48xlarge instances, each equipped with 8×NVIDIA H100 80GB GPUs. Post-hoc LoRA adapter extraction and energy analysis steps were also executed on the same hardware.

Fine-tuning Hyperparameters. We used the AdamW optimizer (β_1=0.9, β_2=0.999), a learning rate of 1×10^{-5}, batch size 32, and trained for 2 epochs.

LoRA+ Training Hyperparameters.

- **Nova Micro:** Learning rate 1×10^{-5}, `loraplus_lr_ratio` 16.0, rank $r = 32$, $\alpha = 128$, `lora_dropout` 0.01, `target_modules = [attention_qkv, attention_dense, mlp_fc1, mlp_fc2]`.

- **Nova Lite/Pro:** Learning rate 1×10^{-5}, `loraplus_lr_ratio` 8.0, rank $r = 32$, $\alpha = 32$, `lora_dropout` 0.01, `target_modules = [attention_qkv, attention_dense, mlp_fc1, mlp_fc2]`.

PHLoRA Extraction. SVD was performed per linear layer using PyTorch's `torch.linalg.svd`. The default low-rank approximation used rank $r = 32$, with ablations at ranks $r = 64$ and $r = 512$.

Energy Plots. Energy preserved at rank r, E_r, was calculated as in Equation 5.

C Reproducibility Checklist

- **Random seeds:** Fixed to 42.

D Optimality of SVD for Low-Rank Adapter Extraction

Given any real matrix $\Delta W \in \mathbb{R}^{m\times n}$, the Eckart–Young–Mirsky theorem [Eckart and Young, 1936] states that the rank-r matrix $\hat{W}_r = U_r \Sigma_r V_r^\top$ (where U_r, Σ_r, V_r are the top r components from the SVD of ΔW) uniquely minimizes the Frobenius norm $\|\Delta W - \hat{W}_r\|_F$ over all matrices of rank at most r.

To see this, let $\Delta W = U\Sigma V^\top$ be the full SVD, with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)}$. The truncated approximation is

$$\hat{W}_r = \sum_{i=1}^{r} \sigma_i u_i v_i^\top,$$

and satisfies

$$\|\Delta W - \hat{W}_r\|_F^2 = \sum_{i=r+1}^{\min(m,n)} \sigma_i^2.$$

Therefore, by setting $A = U_r \operatorname{diag}(\sqrt{\Sigma_r})$ and $B = \operatorname{diag}(\sqrt{\Sigma_r})V_r^\top$, as in PHLoRA, $AB = \hat{W}_r$ is the best rank-r LoRA update (minimizing Frobenius error).

For more details, see [Eckart and Young, 1936, Golub and Loan, 2013].