Compress then Serve: Serving Thousands of LoRA Adapters with Little Overhead

Rickard Brüel Gabrielsson [1] **Jiacheng Zhu** [1] **Onkar Bhardwaj** [2] **Leshem Choshen** [1 2] **Kristjan Greenewald** [2] **Mikhail Yurochkin** [2] **Justin Solomon** [1]

Abstract

Fine-tuning large language models (LLMs) with low-rank adapters (LoRAs) has become common practice, often yielding numerous copies of the same LLM differing only in their LoRA updates. This paradigm presents challenges for systems that serve real-time responses to queries that each involve a different LoRA. Prior works optimize the design of such systems but still require continuous loading and offloading of LoRAs, as it is infeasible to store thousands of LoRAs in GPU memory. To mitigate this issue, we investigate the efficacy of compression when serving LoRA adapters. We consider compressing adapters individually via SVD and propose a method for joint compression of LoRAs into a shared basis paired with LoRA-specific scaling matrices. Our experiments with up to 500 LoRAs[1] demonstrate that compressed LoRAs preserve performance while offering major throughput gains in realistic serving scenarios with over a thousand LoRAs, maintaining 75% of the throughput of serving a *single* LoRA.

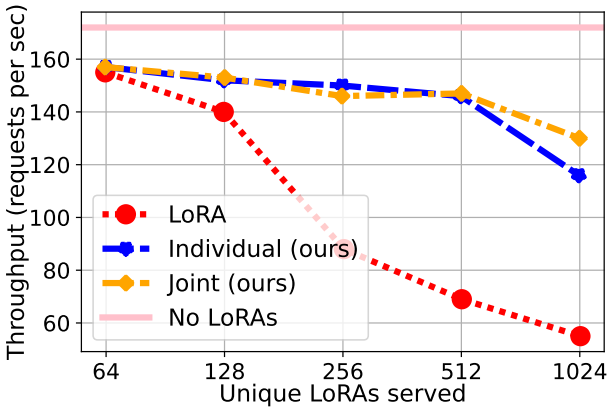


Figure 1. Throughput gains when serving 1000s of compressed LoRAs with vLLM.

1. Introduction

The myriad uses for foundation models (FMs) has led to a proliferation of specialized models, each fine-tuned to perform downstream tasks. The growing number of parameters of these models, however, incurs significant costs both for the fine-tuning and for serving these models.

To avoid fine-tuning giant foundation models, parameter-efficient fine-tuning (PEFT) algorithms update a smaller set of values that define an edit to the original model. An especially successful PEFT method is low-rank adaptation (LoRA) [8], which achieves parameter-efficient fine-tuning by learning low-rank additive changes to neural network matrices. Because of the low-rank parameterization, these matrices (called adapter weights) contain orders-of-magnitude fewer parameters than the base model. LoRA can achieve performance on par with full fine-tuning [8]. In addition, by merging the adapter weights into the base model, the fine-tuned model does not suffer from increased latency.

LoRA's popularity has triggered a growing need to serve large collections of LoRA adapters at the scale where serving multiple copies of the base model becomes infeasible [25]. In the extreme, one may wish to serve hundreds to thousands of LoRAs, where each incoming request requires a different LoRA adapter, e.g., for a personalized chat assistant. To this end, S-LoRA [25] separates the base model computation from individual LoRA computations and optimizes the system's inner workings via custom CUDA kernels and memory management to increase the throughput when serving multiple LoRAs. Despite the optimized system design, serving LoRAs still has a fundamental limitation: when the number of adapters is large, they need to be constantly loaded and offloaded from GPU memory to accommodate incoming requests, degrading throughput. We present extended related work discussion in Appendix A.

In this work, we consider the problem of compressing a

[1]MIT CSAIL [2]MIT-IBM Watson AI Lab. Correspondence to: Rickard Brüel Gabrielsson <brg@mit.edu>.

Work presented at the ES-FoMo Workshop at ICML 2024, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]Available at: https://huggingface.co/Lots-of-LoRAs

collection of LoRAs. We have two key objectives: (1) preserving the performance of the original LoRAs and (2) improving the throughput of serving many LoRAs. We formulate LoRA compression as a reconstruction problem, where the goal is to approximate the original adapters via collections of matrices of a smaller total size. We investigate two approaches: compressing each LoRA *individually* by lowering its rank via SVD and compressing LoRAs *jointly* by finding a shared basis and LoRA-specific scaling matrices. The former is inspired by [11], who demonstrated that lowering LoRA ranks is beneficial for multi-task learning and model merging. For the latter, we propose a joint compression algorithm inspired by models for joint diagonalization.

Our individual and joint compression methods present interesting trade-offs. For example, joint compression can achieve extreme parameter count reduction when representing any specific LoRA by using a shared basis that can be pre-loaded onto the GPU. The shared basis, however, will typically increase the rank of LoRAs. On the other hand, individual compression can reduce the rank, but is limited in terms of its compression efficiency. We investigate how these trade-offs affect throughput when serving 1000s of LoRAs using vLLM [12], a state-of-the-art LLM serving system. In Figure 1 we illustrate the benefits of compression and the effect of these trade-offs. When serving over 1000 LoRAs, compression increases throughput over 2 times and maintains 75% of the throughput of serving the base LLM (or a single LoRA merged into the LLM).

We summarize our main contributions below:

- We formulate the problem of compressing a collection of LoRAs and propose two solutions: individual and joint compression.
- We establish theoretical guarantees for the reconstruction error central to our compression formulation and verify the relation between reconstruction loss and performance empirically.
- We train a collection of 500 high-quality LoRAs for `Mistral-7B-Instruct-v0.2` [10] on 500 natural instruction tasks [27] and demonstrate that our compression techniques can preserve the performance of the original LoRAs. We will release the 500 LoRAs to facilitate future work on LoRA compression as well as the code to compress new models.
- We incorporate LoRA compression into a state-of-the-art LLM serving system and demonstrate that it is possible to serve over 1000 LoRAs across thousands of asynchronous requests with throughput comparable to serving a *single LoRA*.
- We analyze the trade-offs between individual and joint compression, presenting opportunities for further improvement of LLM serving systems.

2. Rank-Based LoRA Compression

LoRA updates are parameterized by pairs of matrices A, B, whose product BA updates the fixed weight matrices $W_0 \in \mathbb{R}^{d_B \times d_A}$ of a neural network foundation model. Given an input x to a layer, the output of the LoRA-updated model at this layer is $(W_0 + BA)x$.

In formulating our compression algorithms, we consider a collection of given LoRA adapters $\{(A_i, B_i)\}_{i=1}^n$ that we would like to serve. We let r_i refer to the rank of the LoRA adapter-pair (A_i, B_i), i.e., $B_i \in \mathbb{R}^{d_B \times r_i}$, $A_i \in \mathbb{R}^{r_i \times d_A}$. We consider a simple baseline parameter reduction approach §2.1 as well as a more aggressive joint diagonalization method in §2.2.

2.1. Independent Compression via SVD

The simplest approach is to replace each rank-r_i LoRA adapter B_iA_i with a reduced rank-r approximation, where typically $r < \frac{1}{n}\sum_{i=1}^n r_i$:

$$\begin{aligned} \mathrm{SVD}_r(B_iA_i) &= U_i\Sigma_iV_i^\top, \\ &\forall i = 1, \ldots, n \qquad (\mathsf{r-SVD}) \end{aligned} \tag{1}$$

Thanks to the Eckart-Young Theorem, this strategy finds the rank-r approximation that best fits B_iA_i in terms of Frobenius norm. As $\Sigma_iV_i^\top$ can be saved as a single matrix, this approach reduces the number of parameters used from $\sum_i(d_A + d_B)r_i$ to $rn(d_A + d_B)$. While this approximation is effective, it is limited in its compression abilities as n increases, since it does not share information across LoRAs. This shortcoming motivates our next approach.

2.2. Joint Diagonalization

Next, we suggest a Joint Diagonalization (JD) method, which optimizes a shared basis onto which we can project the set of n LoRAs. This will allow structure to be shared, implicitly clustering the collection of LoRAs.

In this model, each LoRA-product B_iA_i is factorized into the form $U\Sigma_iV$, where U and V are shared across all LoRAs and Σ_i is specific to each LoRA. In this formulation, every Σ_i shares the same rank r. This allows U and V to be pre-loaded onto the GPU, with Σ_i loaded when necessary for each batch. The matrices Σ_i can be either diagonal or small square matrices, accelerating the forward pass compared to conventional multi-LoRA serving configurations.

Objective function. Motivated by the relationship of singular value decomposition to minimizing the Frobenius norm of the reconstruction error, we also propose to minimize the Frobenius norm of the adapter matrix approximation error.

Specifically, we use the following objective function:

$$\min_{\{\Sigma_i\}_{i=1}^n, U, V} \sum_{i=1}^{n} \|B_i A_i - U \Sigma_i V^\top\|_{\text{Fro}}^2. \tag{2}$$

Using the Frobenius norm has the added benefit of making the objective convex in each argument separately, suggesting the possibility of efficient optimization. This objective function is underdetermined, however, so we consider two constrained regimes below.

Full Σ_i approximation. The first method we call $\mathsf{JD-Full}$. Without loss of generality, U and V can be constrained to be orthogonal, so long as Σ_i remains an unconstrained full matrix. $\mathsf{JD-Full}$ adopts this restriction to make the optimization better posed, but note it does not restrict the expressiveness of the objective (2). This setting yields the following optimization problem:

$$\begin{aligned} &\text{JD-Full}_r(\{B_i A_i\}_{i=1}^n) \\ &= \underset{\substack{\{\Sigma_i\}_{i=1}^n \\ U^\top U = V^\top V = I_r}}{\operatorname{argmin}} \sum_{i=1}^{n} \|B_i A_i - U \Sigma_i V^\top\|_{\text{Fro}}^2 \\ &(\mathsf{JD-Full}) \end{aligned} \tag{3}$$

An efficient alternating algorithm to solve this objective function can be found in Appendix C.

Diagonal Σ_i approximation. As an alternative, we can leave U, V unconstrained (other than to have r columns) and instead constrain the matrices Σ_i to be diagonal (but not necessarily positive). This formulation yields the following optimization problem:

$$\begin{aligned} &\text{JD-Diag}_r(\{B_i A_i\}_{i=1}^n) \\ &= \underset{\{\Sigma_i\}_{i=1}^n, U, V}{\operatorname{argmin}} \sum_{i=1}^{n} \|B_i A_i - U \operatorname{diag}(\Sigma_i) V^\top\|_{\text{Fro}}^2 \\ &(\mathsf{JD-Diag}) \end{aligned} \tag{4}$$

An efficient alternating least squares algorithm to optimize this objective can be found in Appendix C. This diagonal version has some per-LoRA parameter savings when compared to $\mathsf{JD-Full}$, since the diagonal Σ_i only needs r parameters instead of r^2.

In Appendix D we establish theoretical bounds for the reconstruction loss of the $\mathsf{JD-Full}$ algorithm and discuss its connections to merging LoRAs [23; 9].

3. Training LoRAs & Evaluating Task Performance

3.1. Training

We trained LoRA adapters on 500 natural instruction tasks [27] using `Mistral-7B-Instruct-v0.2` [10] as the base model. All LoRA adapters were configured with a rank of 16, i.e., $\forall i, r_i = 16$.

We selected 10 diverse tasks (Table 2 in Appendix F) manually for consistent evaluation across experiments and randomly sampled an additional 490 tasks, resulting in a total of 500 tasks. These tasks were exclusively in English (both input and output), ensuring higher quality and thorough review [27]. Each task dataset was divided into training, validation, and test sets. Hyperparameters, such as early stopping, were tuned using the validation sets. Evaluation on the test sets demonstrated that LoRA consistently outperformed the base model in terms of both Rouge scores and loss metrics, as shown in Table 1. Details are provided in Appendix F.

Table 1. Comparison of metrics before and after LoRA training across 500 tasks.

Metric	Base Model	LoRA
Loss	4.99 ± 3.11	0.43 ± 0.57
Exact Match	2.28 ± 7.89	66.66 ± 34.34
Rouge-1	20.38 ± 18.90	76.74 ± 24.89
Rouge-L	19.66 ± 18.16	76.22 ± 25.27

3.2. Evaluation

We evaluated multiple metrics for the natural instruction tasks, including cross-entropy loss, Rouge-1, Rouge-L [15], exact match, and *agreement* between uncompressed and compressed LoRA. Here, *agreement* measures the exact match in task-generations between the uncompressed LoRA model and the compressed LoRA model, rather than comparing to ground truth data. While detailed results and discussions for all metrics are provided in the Appendix, our primary focus in the main text is on Rouge-L. We find that all metrics correlate, but Rouge-L correlates most strongly with downstream utility. This finding aligns with prior work [27], which demonstrates that Rouge-L correlates well with classification accuracy.

Joint diagonalization optimizes reconstruction error measured by the Frobenius norm, and our theoretical analysis in §D bounds this reconstruction error. Empirically, reconstruction error and downstream Rouge-L performance correlate.

Instead of listing the absolute performance of different meth-

ods, we compute the performance difference between the base model and the LoRA model for each task. We present the ratio

$$\text{Performance relative to LoRA} := \frac{\text{method-performance}}{\text{LoRA-performance}}$$

for the specific method in question, highlighting relative improvement with respect to the uncompressed LoRA and the base model.

4. Experiments

4.1. Task Performance

For each method, we vary the number of n LoRAs that are compressed and the compression rank r. We run each experiment three times with different random seeds and report the mean and standard deviation. See Table 3 for results where we evaluate on the same ten manually selected tasks (Table 2) across settings. Every compressed collection of LoRAs contains these 10 tasks (i.e., in-distribution tasks), and each collection contains the smaller collections as subsets.

Figure 2a plots the mean performance improvement relative to uncompressed LoRAs (y-axis) against the GPU parameter saved ratio, i.e., the reduction in per-LoRA parameters (x-axis). Similarly, Figure 2b relates mean performance improvement to the total parameter saved ratio. Both the independent compression via SVD and joint diagonalization methods significantly compress the LoRAs while preserving—and occasionally enhancing—performance. Notably, the JD methods are unique in their ability to approach the compression efficacy of a single LoRA, although this aggressive reduction in size may decrease performance in larger LoRA collections. As shown in Figure 4, the JD methods achieve the highest throughput. These results provide practitioners with a spectrum of methods and configurations, enabling them to optimize the balance between throughput, compression, and performance according to their specific requirements.

4.2. Performance and Reconstruction Error

Figure 3 relates reconstruction error and performance. The y-axis measures mean performance improvement of Rouge-L relative to uncompressed LoRA, and the x-axis quantifies the mean reconstruction error between the compressed reconstruction of the product BA and the original uncompressed product BA. Although the relationship between performance and reconstruction error is nonlinear, it demonstrates a generally decreasing, somewhat exponential trend. Notably, the minimal reconstruction error does not correlate with optimal performance, indicating that a degree of lossy reconstruction may be advantageous for enhancing generalization.

4.3. Throughput of Serving Compressed LoRAs

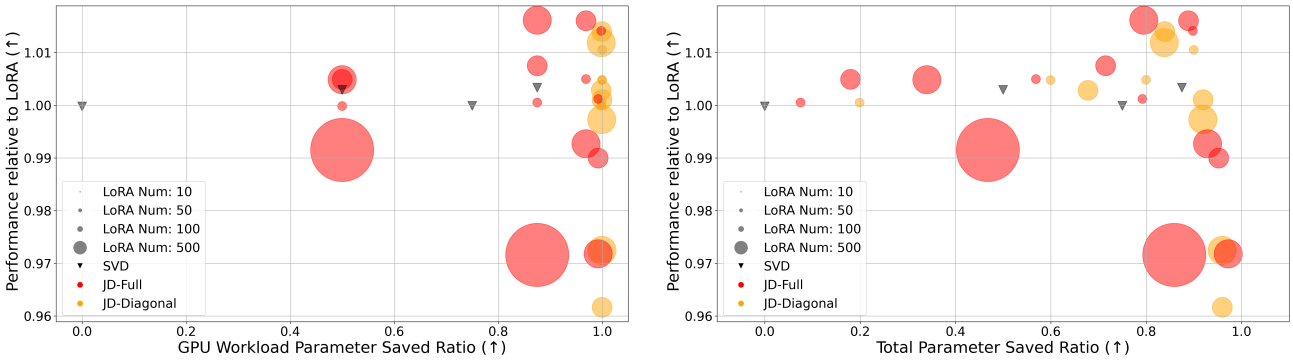
Figure 4 studies the efficacy of compression in a real-world serving scenario. We consider a varying number of rank-16 LoRAs, using a dataset of Shakespeare sonnets as inputs[2] arriving *asynchronously*. We measure throughput, i.e., the number of requests served per second when generating ten tokens per request. In these experiments, the base model is Mistral 7B Instruct as in the previous experiments; we simulate random LoRAs and assign inputs to LoRAs at random. Experiments were conducted on A100 80GB GPU capped at 30% memory consumption. This is done to reflect cost concerns in practical situations where a service provider might want to serve many LoRAs from cheaper hardware with lower memory than higher-end GPUs. This setting also takes into account the scenario where the LLM is large compared to the size of GPU and yet a provider may want to serve many LoRAs efficiently using the same device.

Although LoRAs are meant as parameter-efficient adapters, when serving over 200 LoRAs, the throughput starts to degrade rapidly, illustrating the challenge of serving many LoRAs. When serving over 500 LoRAs, our compression methods provide over $2\times$ improvement in throughput, and at over 1000 LoRAs, JD-Full-64 serves at 75% of the throughput of the base LLM (equivalent to serving a single LoRA after merging it with the LLM).

Next, we analyze trade-offs between the compression methods. Each compression method except JD-Full-64 and JD-Diag-128 starts to degrade faster at a certain point as the number of LoRAs increases; this point is the threshold at which compressed LoRAs no longer fit into the GPU memory and the throughput degrades due to scheduling. The JD LoRAs are the most efficient in terms of memory consumption, and we can simultaneously load all 1024 LoRAs. Due to their increased rank (e.g., each LoRA is of rank 128 for JD-Diag-128), however, the throughput starts lower, but degrades at a lower rate. We conclude that for the current vLLM design, when serving a moderate number of LoRAs, SVD compression provides favorable throughput as long as it can fit all LoRAs into memory. When serving a large number of LoRAs that cannot fit into memory even with SVD compression, JD is the better option, as it can fit all LoRAs onto the GPU at the cost of a slight throughput reduction due to its larger rank.

Fast LoRA [28] is a recent work that aims to alleviate the batched matrix multiplication (BMM) bottleneck when serving many LoRAs. They propose an adapter parameterization that replaces addition with elementwise multiplication, avoiding BMM and improving LoRA throughput at lower ranks. Our JD LoRA formulation also circumvents or heav-

[2] https://www.kaggle.com/datasets/shivamshinde123/william-shakespeares-sonnet



(a) Performance with GPU parameter saved ratio

(b) Performance with total parameter saved rate

Figure 2. We qualitatively illustrate the performance of JD-LoRA. The GPU workload parameter saved ratio indicates the number of parameters saved compared to LoRA during inference when switching between different LoRAs. For JD-Diagonal methods, it is computed as $r_{pergpu} := 1 - \frac{r_{JD}}{r_{LoRA} * d * 2}$. The total parameter saved ratio depicts the number of parameters saved for a system with a large number n of LoRAs for different tasks. It is computed as: $r_{total} := 1 - \frac{2 * r_{JD} * d + K * r_{JD}}{n * r_{LoRA} * d * 2}$, where r_{JD} and r_{LoRA} are the rank for JD-LoRA and LoRA respectively, d is the weight dimension, and n is the number of LoRAs served.

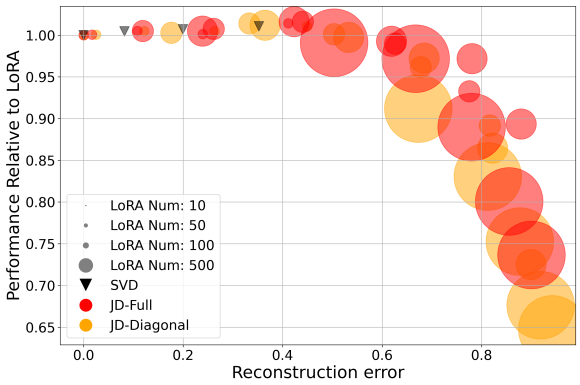


Figure 3. Reconstruction error vs. performance.

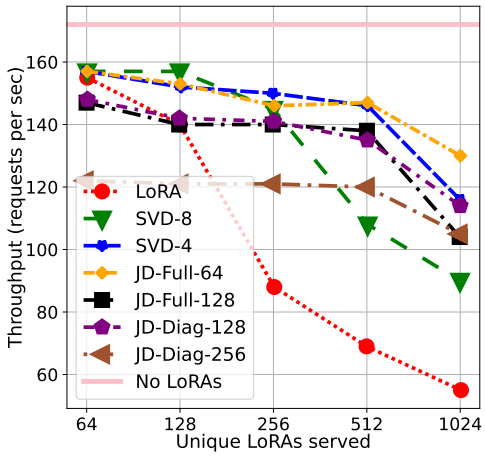


Figure 4. Throughput when serving varying number of LoRAs with vLLM.

ily reduces the impact of BMM (see Appendix H), and both individual and joint compression methods can be applied to Fast LoRAs.

5. Discussion

This study introduces approaches to LoRA compression, addressing significant challenges facing foundation models and large language models. Our contributions include theoretical formulations, empirical validation, and practical implementations that enhance understanding and application of LLMs in scalable environments. The implications of our findings are manifold. Our theoretical guarantees for reconstruction error not only increase confidence in the use of compressed models but also lay a groundwork for future explorations in this area. Demonstrating that our compression techniques can preserve up to 100% of the original LoRAs' performance highlights the effectiveness of our methods. Furthermore, integrating LoRA compression into state-of-the-art LLM serving systems demonstrates the potential for resource optimization, with throughput for thousands of LoRAs nearing that of a single LoRA. In conclusion, our research significantly advances the deployment of LLMs by providing robust, scalable, and efficient compression solutions. The ability of compressed LoRAs to maintain high performance while facilitating substantial resource savings opens new avenues for the broader application and adoption of LLMs across various industries. We encourage the community to build upon our findings and the shared LoRAs to further explore and enhance the utility of these technologies.

6. Acknowledgments

The MIT Geometric Data Processing Group acknowledges the generous support of Army Research Office grants W911NF2010168 and W911NF2110293, from the CSAIL Systems that Learn program, from the MIT–IBM Watson AI Laboratory, from the Toyota–CSAIL Joint Research Center,

and from an Amazon Research Award.

References

[1] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017.

[2] Leshem Choshen, Elad Venezian, Noam Slonim, and Yoav Katz. Fusing finetuned models for better pre-training. *ArXiv*, abs/2204.03044, 2022.

[3] Leshem Choshen, Elad Venezian, Shachar Don-Yehiya, Noam Slonim, and Yoav Katz. Where to start? analyzing the potential value of intermediate models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1446–1470, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.90. URL https://aclanthology.org/2023.emnlp-main.90.

[4] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*, pages 291–326. Chapman and Hall/CRC, 2022.

[6] Moshik Hershcovitch, Leshem Choshen, Andrew Wood, Ilias Enmouri, Peter Chin, Swaminathan Sundararaman, and Danny Harnik. Lossless and near-lossless compression for foundation models. *arXiv preprint arXiv:2404.15198*, 2024.

[7] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[8] 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. *arXiv preprint arXiv:2106.09685*, 2021.

[9] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition, 2024.

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

[11] Weisen Jiang, Baijiong Lin, Han Shi, Yu Zhang, and James T Kwok. Byom: Building your own multi-task model for free. *arXiv preprint arXiv:2310.01886*, 2023.

[12] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.

[13] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes, 2018.

[14] Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*, 2023.

[15] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

[16] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

[17] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation, 2024.

[18] Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. *arXiv preprint arXiv:2111.09832*, 2021.

[19] Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *arXiv preprint arXiv:2404.02948*, 2024.

[20] Mohammed Muqeeth, Haokun Liu, Yufan Liu, and Colin Raffel. Learning to route among specialized experts for zero-shot generalization. *arXiv preprint arXiv:2402.05859*, 2024.

[21] Kimia Nadjahi, Kristjan Greenewald, Rickard Brüel Gabrielsson, and Justin Solomon. Slicing mutual information generalization bounds for neural networks. In *ICML 2023 Workshop Neural Compression: From Information Theory to Applications*, 2023. URL https://openreview.net/forum?id=cbLcwK3SZi.

[22] Colin Raffel. Building machine learning models like open source software. *Communications of the ACM*, 66(2):38–40, 2023.

[23] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. *arXiv preprint arXiv:2311.13600*, 2023.

[24] Pratyusha Sharma, Jordan T. Ash, and Dipendra Misra. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ozX92bu8VA.

[25] 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. S-lora: Serving thousands of concurrent lora adapters, 2023.

[26] Sheng Wang, Boyang Xue, Jiacheng Ye, Jiyue Jiang, Liheng Chen, Lingpeng Kong, and Chuan Wu. Prolora: Partial rotation empowers more parameter-efficient lora. *ArXiv*, abs/2402.16902, 2024. URL https://api.semanticscholar.org/CorpusID:268032580.

[27] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*, 2022.

[28] Yeming Wen and Swarat Chaudhuri. Batched low-rank adaptation of foundation models, 2024.

[29] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[30] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, 2022.

[31] Prateek Yadav, Leshem Choshen, Colin Raffel, and Mohit Bansal. Compeft: Compression for communicating parameter efficient updates via sparsification and quantization. *arXiv preprint arXiv:2311.13171*, 2023.

[32] Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=xtaX3WyCj1.

[33] Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation, 2024.

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

[35] 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. *arXiv preprint arXiv:2402.16842*, 2024.

A. Related Work

Parameter-efficient fine-tuning (PEFT) has become a prevalent method for updating foundation models thanks to the need for efficiency in training and communication [14]. Many PEFT methods have been proposed, e.g. [7; 16], but the most used in practice is LoRA [8], partially due to the ease of switching between LoRAs in inference time.

Several works propose improvements to LoRA [17; 26], sometimes with algebraic methods like SVD [19; 34; 11] or by leveraging its statistical properties [35; 33]. Relatively few methods, however, accelerate inference times. S-LoRA [25] provides an efficient means of switching between LoRAs. F-LoRA [28] adapts training to reduce batch multiplications and thus to accelerate inference. Our method achieves a similar outcome (see Appendix H) without changing the LoRA formulation or requiring that LoRAs be trained in a dedicated way; from this perspective, an advantage of our work is that future generic improvements to LoRA will also benefit (e.g., [19]).

There are many efforts to compress models [1; 5; 24; 13]—including some that consider LoRAs specifically—to accelerate inference. Predominantly, pruning and sparsification methods delete some of the weights [31], and quantization methods reduce the precision of the weights [4]. Some works compress weights to reduce model size but typically require decompression and hence do not save GPU memory [6]. Similarly to our work, while most methods increase speed at the cost of performance, a few note increased performance and generalization after compression [31; 21; 6; 24]. Our work also relates to model merging [2; 30; 18] and mixture of experts methods [20]. Such methods reuse models trained by others [3; 22], serving them together as one compressed model. Despite this similarity, these methods create a single general model that acts on any input, while our model allows for more performant per-task solutions.

B. Limitations

Compression techniques, while reducing time or memory constraints, invariably lose information, which can increase with the number of models compressed. Consequently, while our approach seeks to minimize this loss, it is axiomatic that compression will diminish information to some degree.

Our methodology compresses a collection of known models. Introducing new models may necessitate either recompression, projecting into a suboptimal basis, or opting to use some models in an uncompressed state.

Empirically, our methods achieve reasonable balances between the number of models compressed, the amount of information retained, and throughput. Nonetheless, these benefits can be accompanied by increased latency as more models are compressed.

C. Joint Diagonalization Algorithms

C.1. Alternating Methods

Our goal is to derive algorithms that optimize (2). Common to both methods, we expand the objective functional:

$$\begin{aligned}\sum_i \|B_iA_i - U\Sigma_iV^\top\|_{\text{Fro}}^2 &= \sum_i \text{tr}((B_iA_i - U\Sigma_iV^\top)(B_iA_i - U\Sigma_iV^\top)^\top) \text{ by definition} \\ &= \sum_i \left[\text{tr}(B_iA_iA_i^\top B_i^\top) - 2\text{tr}(B_iA_iV\Sigma_i^\top U^\top) + \text{tr}(U\Sigma_iV^\top V\Sigma_i^\top U^\top)\right] \\ &= \text{const.} - 2\sum_i \text{tr}(B_iA_iV\Sigma_i^\top U^\top) + \sum_i \|U\Sigma_iV^\top\|_{\text{Fro}}^2. \end{aligned} \tag{5}$$

Using this expansion, we now consider the two settings discussed in §2.2.

Case 1: Non-diagonal Σ_i, orthogonal U, V. Setting the derivative of (5) with respect to Σ_i to zero, we find

$$\Sigma_i = \Sigma_i^*(U, V) = U^\top B_iA_iV. \tag{6}$$

We simplify our objective function after plugging in this expression:

$$\begin{aligned}\sum_i \|B_iA_i - U\Sigma_iV^\top\|_{\text{Fro}}^2 + \text{const.} &= \sum_i \left[\|\Sigma_i\|_{\text{Fro}}^2 - 2\text{tr}(B_iA_iV\Sigma_i^\top U^\top)\right] \text{ from (5)} \\ &= \sum_i \left[\text{tr}(U^\top B_iA_iVV^\top A_i^\top B_i^\top U) - 2\text{tr}(B_iA_iVV^\top A_i^\top B_i^\top UU^\top)\right] \text{ from (6)} \\ &= -\sum_i \text{tr}(B_iA_iVV^\top A_i^\top B_i^\top UU^\top).\end{aligned}$$

Substituting (6), we find

$$U_{opt}, V_{opt} = \arg\max_{\substack{U^\top U=I \\ VV^\top=I}} \sum_{i=1}^n \|U^\top B_iA_iV\|_{\text{Fro}}^2 = \arg\max_{\substack{U^\top U=I \\ VV^\top=I}} \sum_{i=1}^n \|\Sigma_i^*(U,V)\|_{\text{Fro}}^2. \tag{7}$$

Note that

$$\begin{aligned}\sum_{i=1}^n \|U^\top B_iA_iV\|_{\text{Fro}}^2 &= \text{tr}\left(\left(\sum_{i=1}^n B_iA_iVV^\top A_i^\top B_i^\top\right)UU^\top\right) \\ &= \text{tr}\left(\left(\sum_{i=1}^n B_i^\top A_i^\top UU^\top A_iB_i\right)VV^\top\right),\end{aligned}$$

by the identity $\|A\|_{\text{Fro}}^2 = \text{tr}(A^\top A)$. Hence, we optimize (7) by alternating between U and V:

- U **iteration:** Define $M \coloneqq \sum_i B_iA_iVV^\top A_i^\top B_i^\top$. Parenthesizing this expression properly requires only $O((m+n)r)$ storage/computation time. With this definition, we maximize $\text{tr}(MUU^\top)$ over U satisfying $U^\top U = I$. Since M is positive semidefinite, the optimum is to take U to be the r eigenvectors of M with largest eigenvalue, equivalent to an SVD problem.
- V **iteration:** Define $N \coloneqq \sum_i A_i^\top B_i^\top UU^\top B_iA_i$. Similarly to the previous step, we take V to contain the r eigenvectors of N with largest eigenvalue, again solvable using an SVD.

This method decreases the objective in each step.

Case 2: Diagonal Σ_i. If constrain Σ_i to be diagonal, we interpret our objective function (2) as a "triple least squares" problem. We compute gradients:

$$\begin{aligned}\nabla_U \sum_i \|B_iA_i - U\Sigma_iV^\top\|_{\text{Fro}}^2 &= 2\sum_i (U\Sigma_iV^\top - B_iA_i)V\Sigma_i^\top \\ \nabla_V \sum_i \|B_iA_i - U\Sigma_iV^\top\|_{\text{Fro}}^2 &= 2\sum_i (V\Sigma_i^\top U^\top - A_i^\top B_i^\top)U\Sigma_i \\ \nabla_{\Sigma_i} \sum_i \|B_iA_i - U\Sigma_iV^\top\|_{\text{Fro}}^2 &= 2U^\top(U\Sigma_iV^\top - B_iA_i)V\end{aligned}$$

These expressions suggest efficient $r \times r$ linear systems to solve for U, V:

$$\begin{aligned}U &= \left(\sum_i B_iA_iV\Sigma_i^\top\right)\left(\sum_i \Sigma_iV^\top V\Sigma_i^\top\right)^{-1} \\ V &= \left(\sum_i A_i^\top B_i^\top U\Sigma_i\right)\left(\sum_i \Sigma_i^\top U^\top U\Sigma_i\right)^{-1}.\end{aligned}$$

For Σ_i, we extract the diagonal from our gradient above:

$$\begin{aligned}
\text{diag}(U^\top U \Sigma_i V^\top V)_j &= (U^\top U \Sigma_i V^\top V)_{jj} \\
&= \sum_m (U^\top U)_{jm} \Sigma_{imm} (V^\top V)_{mj} \\
&= (U^\top U \circ V^\top V) \text{diag}(\Sigma_i) \\
\text{diag}(U^\top B_i A_i V)_j &= \sum_m (U^\top B_i)_{jm} (A_i V)_{mj} \\
&= \sum_m (U^\top B_i)_{jm} (V^\top A_i^\top)_{jm} \\
&= (U^\top B_i \circ V^\top A_i^\top) \mathbb{1} \\
\implies \text{diag}(\Sigma_i) &= (U^\top U \circ V^\top V)^{-1} (U^\top B_i \circ V^\top A_i^\top) \mathbb{1}
\end{aligned}$$

Here $\circ$ denotes the Hadamard product.

Combining these expressions, we use a simple coordinate descent algorithm cycling between the following three steps:

1. Solve for U
2. Solve for V
3. Solve for the Σ_i's
4. Optionally, normalize so $\sum_i \|\Sigma_i\|_{\text{Fro}}^2 = 1$

C.2. Additional Eigenvalue Iteration Algorithm

For the first case in §C.1, we introduce an alternative algorithm that eschews the use of SVD. This alternative is optimized for GPU execution, enabling tractable runs to convergence.

To derive this algorithm, we employ Lagrange multipliers to formulate the derived objective from (7):

$$U_{opt}, V_{opt} = \arg \max_{\substack{U^\top U = I \\ V V^\top = I}} \sum_{i=1}^{n} \|U^\top B_i A_i V\|_{\text{Fro}}^2, \tag{8}$$

yielding the expression

$$\Lambda = -\frac{1}{2} \|U^\top B_i A_i V\|_{\text{Fro}}^2 - \frac{1}{2} \text{tr}(X^\top (I - U^\top U)) - \frac{1}{2} \text{tr}(Y^\top (I - V^\top V)). \tag{9}$$

Taking the derivatives gives

$$\nabla_U \Lambda = -\sum_i B_i (A_i V)(V^\top A_i^\top)(B_i^\top U) + UX \tag{10}$$

$$\nabla_V \Lambda = -\sum_i A_i^\top (B_i^\top U)(U^\top B_i)(A_i V) + VY \tag{11}$$

Setting these derivatives to zero shows

$$\sum_i B_i (A_i V)(V^\top A_i^\top)(B_i^\top U) = UX \tag{12}$$

$$\sum_i A_i^\top (B_i^\top U)(U^\top B_i)(A_i V) = VY. \tag{13}$$

Here, one can show that the Lagrange multiplier matrices X and Y are diagonal and nonnegative, since the problem reduces to an eigenvalue problem when either U or V is fixed; this is essentially the argument behind the alternating algorithm in Appendix C. Hence, taking inspiration from classical eigenvalue iteration, we use the following updates to improve our

estimates of U and V:

$$U_0^{(k+1)} \leftarrow \sum_i B_i(A_i V^{(k)})((V^{(k)})^\top A_i^\top)(B_i^\top U^{(k)}) \tag{14}$$

$$V_0^{(k+1)} \leftarrow \sum_i A_i^\top (B_i^\top U^{(k)})((U^{(k)})^\top B_i)(A_i V^{(k)}) \tag{15}$$

$$U^{(k+1)} \leftarrow \texttt{orthogonalize}(U_0^{(k+1)}) \tag{16}$$

$$V^{(k+1)} \leftarrow \texttt{orthogonalize}(V_0^{(k+1)}) \tag{17}$$

Here, the function `orthogonalize` orthogonalizes the columns of a matrix, e.g. by using the Q part of the reduced-size QR factorization. Although we lack a formal convergence proof, in practice we find that this method reliably reaches a local optimum of our problem.

By executing matrix operations in the specified sequence, these computations can be rapidly performed on GPUs. Note the expressions above are parenthesized to avoid constructing a large matrix product as an intermediate computation.

D. Theoretical Analysis

While SVD-based approaches are relatively well-understood, in this section, we seek to better understand the role of the joint diagonalization method presented in §2.2. We will focus on the full-Σ_i case with orthogonal U, V matrices. Note that, for the same r, the $r-\mathsf{JD}-\mathsf{Diag}$ has at least as large reconstruction error as $r-\mathsf{JD}-\mathsf{Full}$ since it imposes an additional constraint on the Σ_i.

Firstly, note that perfect reconstruction can be achieved if and only if r is large enough:

Proposition D.1. *Suppose that for all i, $\operatorname{rank}(B_i A_i) = r_i$, and let*

$$\tilde{r} = \max\left\{\operatorname{rank}([A_1, \ldots, A_n]), \operatorname{rank}([B_1^\top \ldots, B_n^\top])\right\}.$$

Note $\max_i r_i \leq \tilde{r} \leq \sum_{i=1}^n r_i$. Then $\mathsf{JD}-\mathsf{Full}$ (3) with $r = \tilde{r}$ achieves lossless compression (perfect reconstruction), and using $r < \tilde{r}$ will give nonzero reconstruction error.

Proof. There exist U, V such that all the B_i, A_i are in the spans of U, V resp. if and only if $r \geq \tilde{r}$. □

Due to training noise, $\tilde{r}$ will equal $\sum_{i=1}^n r_i$ almost always. This implies that in most realistic settings, the joint diagonalization approach is a lossy reconstruction.

This reconstruction loss can be significant, as the following theorem shows (proved in Appendix E):

Theorem D.2. *Consider n LoRAs ($\{A_i, B_i\}_{i=1}^n$) with $r, n \leq d^2$, and form the matrix*

$$L = \left[\ \operatorname{vec}(B_1 A_1) \quad \cdots \quad \operatorname{vec}(B_n A_n)\ \right].$$

Let σ_j be the singular values of L, sorted from largest to smallest, and let $\bar{\sigma}_j$ be the singular values of $\sum_{i=1}^n B_i A_i$. Then, using $\mathsf{JD}-\mathsf{Full}$ (3),

$$\sum_{j=1}^r \bar{\sigma}_j^2 \leq \sum_{i=1}^n \|\Sigma_i\|_{\text{Fro}}^2 = \sum_{i=1}^n \|U\Sigma_i V^\top\|_{\text{Fro}}^2 \leq \sum_{j=1}^{\min(r^2,n)} \sigma_j^2,$$

implying the sum of squared Frobenius norms of the reconstructed LoRAs satisfies

$$\frac{\sum_{i=1}^n \|U\Sigma_i V^\top\|_{\text{Fro}}^2}{\sum_{i=1}^n \|B_i A_i\|_{\text{Fro}}^2} \leq \frac{\sum_{j=1}^{\min(r^2,n)} \sigma_j^2}{\sum_{j=1}^n \sigma_j^2} \leq 1, \ \ \textit{and} \ \ \frac{\sum_{i=1}^n \|U\Sigma_i V^\top - B_i A_i\|_{\text{Fro}}^2}{\sum_{i=1}^n \|B_i A_i\|_{\text{Fro}}^2} \geq 1 - \frac{\sum_{j=1}^{\min(r^2,n)} \sigma_j^2}{\sum_{j=1}^n \sigma_j^2}.$$

In other words, if the singular values of L are not concentrated in the top r^2 entries, significant reconstruction error is unavoidable.

Remark D.3 (Lower bound and merging). The lower bound $\sum_{j=1}^r \bar{\sigma}_j^2$ could be achieved by setting all the Σ_i equal, i.e., using a fully merged model instead of only merging the subspaces U, V and allowing Σ_i to vary with i.

Remark D.4 (Upper bound and clustering). The upper bound is smallest when the LoRAs are relatively *clustered*, i.e., when groups of vectors $\mathrm{vec}(B_iA_i)$ are similar. This situation raises the magnitude of the largest singular values of L, raising the upper bound in the proposition. As the LoRAs are $d \times d$ matrices that can be thought of as points in d^2 dimensional space, for typical values of d well into the hundreds, it is likely that unrelated LoRAs will be *unclustered*, i.e., they will have relatively low inner products with each other.

For the case of orthogonal LoRAs, the singular values of L are the norms of the LoRAs, and we immediately have the following corollary: [3]

Corollary D.5. *Suppose (e.g., due to normalization) that the inputs to the joint diagonalization algorithm all have unit Frobenius norm, i.e.,* $\|B_iA_i\|_{\mathrm{Fro}} = 1$. *Moreover, assume that the LoRAs are all orthogonal in the sense* $\mathrm{tr}((B_iA_i)(B_jA_j)^\top) = 0$ *for* $i \neq j$. *Then, using the* $\mathsf{JD-Full}$ *method (3), we have* $1 \leq \sum_{i=1}^n \|\Sigma_i\|_{\mathrm{Fro}}^2 \leq \min(r^2, n)$, *implying that the sum of squared Frobenius norms of the reconstructed LoRAs satisfies*

$$1 - \frac{1}{n} \geq \frac{\sum_{i=1}^n \|U\Sigma_iV^\top - B_iA_i\|_{\mathrm{Fro}}^2}{\sum_{i=1}^n \|B_iA_i\|_{\mathrm{Fro}}^2} \leq 1 - \min\left(\frac{r^2}{n}, 1\right).$$

This implies that for the common setting where $r^2 \ll n$, the reconstructed LoRAs will be significantly smaller than the original LoRAs and necessarily have significant reconstruction error.

The results in this section illustrate the tradeoffs of using joint diagonalization. If the LoRAs are similar or well-clustered, reconstruction error will be low. On the other hand, if the LoRAs are random and orthogonal, reconstruction error will be high.

Note that *high reconstruction error for the weights does not necessarily imply poor ability to approximate the desired LLM outputs.* The loss space of transformers is highly complex, and the tendency of the Frobenius-norm based joint diagonalization is to find subspaces that are shared with many LoRAs when r is large, and *merge* subspaces together towards the mean when r is small. When r is severely below the natural rank of the problem, this tendency towards *averaging* all or some of the LoRA signals directly connects to the concept of *merging* LoRAs, whose success [23; 9] could explain the success of our procedure despite the nonlinearity of the underlying transformers.

E. Proof of Theorem D.2

Proof. For the lower bound, note that by Jensen's inequality,

$$\sum_{i=1}^n \|U^\top B_iA_iV\|_{\mathrm{Fro}}^2 \geq \left\| U^\top \sum_{i=1}^n B_iA_iV \right\|_{\mathrm{Fro}}^2,$$

for any U, V. Hence,

$$\sup_{U,V \in \mathrm{St}(k,d)} \sum_{i=1}^n \|U^\top B_iA_iV\|_{\mathrm{Fro}}^2 \geq \sup_{U,V \in \mathrm{St}(k,d)} \left\| U^\top \sum_{i=1}^n B_iA_iV \right\|_{\mathrm{Fro}}^2. \tag{18}$$

By the definition of singular value decomposition, the right hand side of (18) is maximized with U, V being the top r singular vectors of $\sum_{i=1}^n B_iA_i$, yielding $\left\|U^\top \sum_{i=1}^n B_iA_iV\right\|_{\mathrm{Fro}}^2 = \sum_{i=1}^r \bar{\sigma}_i^2$. Recalling that $\Sigma_i = U^\top B_iA_iV$ yields the lower bound.

For the upper bound, recall that $\Sigma_i = U^\top B_iA_iV$. Rearranging,

$$\mathrm{vec}(\Sigma_i) = (V^\top \otimes U^\top)\mathrm{vec}(B_iA_i).$$

Define

$$\bar{\Sigma} := [\mathrm{vec}(\Sigma_1), \ldots, \mathrm{vec}(\Sigma_n)].$$

By our previous simplification,

$$\bar{\Sigma} = (V^\top \otimes U^\top)L.$$

[3] A result for isotropic Gaussian LoRAs could be obtained via the quantiles of the Marchenko-Pastur Law.

Table 2. Main Evaluation Tasks

Task Number	Name	Type	Domain
task280	stereoset_classification_stereotype_type	classification	stereoset
task190	snli_classification	snli	image captions
task391	causal_relationship	commonsense	cause and effect
task290	tellmewhy_question_answerability	answerability	story
task1391	winogrande_easy_answer_generation	commonsense	social and physical
task1342	amazon_us_reviews_title	title generation	amazon reviews
task442	com_qa_paraphrase_question_generation	question generation	wikipedia
task620	ohsumed_medical_subject_headings_answer_generation	keyword tagging	scientific
task1598	nyc_long_text_generation	data to text	restaurants
task039	qasc_find_overlapping_words	overlap extraction	natural science

Now

$$\sum_{i=1}^{n} \|\Sigma_i\|_{\text{Fro}}^2 = \|\bar{\Sigma}\|_{\text{Fro}}^2 = \text{tr}\left(((V \otimes U)(V \otimes U)^\top)(LL^\top)\right)$$

Since U, V are orthogonal and size $d \times r$, the top r^2 eigenvalues of the symmetric matrix $(V \otimes U)(V \otimes U)^\top$ will be equal to 1, and the rest will equal 0. The eigenvalues of the symmetric matrix $LL^\top$ will be equal to the squared singular values of L. We can then apply the Von Neumann trace inequality to obtain the upper bound.

The last statement follows from the Pythagorean theorem and the fact that the Σ_i is a projection of B_iA_i to the U, V subspace. □

Note that we have only used the fact that the matrix $(V \otimes U)$ has singular values equal to 1; we have not used the fact that it has Kronecker product structure. On the other hand, each vector $\text{vec}(B_iA_i)$ is a sum of r_i Kronecker products and cannot be expressed as a Kronecker product. As a result, while the upper bound in the Von Neumann trace inequality is achieved if the eigenvectors of the two matrices align, the Kronecker product structure is a severe constraint and the upper bound we have provided is generous.

F. Training LoRAs

We trained LoRA adapters on 500 natural instruction tasks [27] using `Mistral-7B-Instruct-v0.2` [10] as the base model. All LoRA adapters were configured with a rank of 16, i.e., $\forall i, r_i = 16$. We selected 10 diverse tasks manually for consistent evaluation across experiments and randomly sampled an additional 490 tasks, resulting in a total of 500 tasks. These tasks were exclusively in English (both input and output), ensuring higher quality and thorough review [27]. Each task dataset was divided into training, validation, and test sets (80-10-10). Hyperparameters, such as early stopping, were tuned using the validation sets; that is, we train for five epochs and take the best-performing epoch-checkpoint per validation loss. Evaluation on the test sets demonstrated that LoRA consistently outperformed the base model in terms of both Rouge scores and loss metrics (see Table 1).

We use Huggingface [29] in our implementation. For the base model, we use quantization with configuration:

```
BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_use_double_quant=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.bfloat16,
)
```

and LoRA configuration:

```
LoraConfig(
    r=16,
```

```
        lora_alpha=32,
        target_modules=["q_proj", "k_proj", "v_proj"],
        lora_dropout=0.05,
        bias="none",
        task_type="CAUSAL_LM",
        init_lora_weights=init_lora_weights,
    )
```

G. Joint Diagonalization Details

We normalize each LoRA adapter to have a Frobenius norm of one prior to running joint diagonalization. This normalization enhances performance and reduces the variance in reconstruction error. We restore the original norms of the LoRA adapters before reconstruction and testing.

H. Avoiding Batched Matrix Multiplication (BMM)

In the envisioned deployment scenario, a service provider hosts a large collection of LoRAs. Upon receiving a request, each user specifies both the input data and the desired LoRA identifier. The provider then processes the base model augmented with the specified LoRA for each user's data. As a provider is batching a collection of requests for GPU parallelization, they can expect to frequently have more than one unique LoRA identifier per batch.

Traditionally, a specific LoRA is integrated into the base model by transforming $W_0 \to W_0 + B_i A_i$. Serving multiple LoRAs conventionally would necessitate maintaining and executing a separate copy of the base model for each LoRA, bringing substantial computational overhead. Alternatively, the computation for $W_0 x$ and $B_i A_i x$ can be performed independently and subsequently merged. This strategy necessitates only a single instance of $W_0 x$ computation and storage of LoRA-specific parameters rather than the entire base model.

Consider the batch processing of $\mathbf{BAx}$, where boldface indicates that B_i, A_i are stacked into tensors of dimensions $(b \times m \times r)$ and $(b \times r \times n)$ respectively, with batched data $\mathbf{x}$ shaped $(b \times l \times n)$:

$$\begin{aligned} \mathbf{Ax} &\leftrightarrow (b \times r \times n) \times (b \times l \times n) \to (b \times l \times r) \text{ bmm} \\ \mathbf{B}(\mathbf{Ax}) &\leftrightarrow (b \times m \times r) \times (b \times l \times r) \to (b \times l \times m) \text{ bmm.} \end{aligned}$$

Here, "bmm" denotes batched matrix multiplication, a known bottleneck in both throughput and latency. Consider the corresponding operations for our joint compression scheme, $U\mathbf{\Sigma}V^\top x$:

$$\begin{aligned} V^\top \mathbf{x} &\leftrightarrow (\tilde{r} \times n) \times (b \times l \times n) \to (b \times l \times \tilde{r}) \text{ broadcasted} \\ \mathbf{\Sigma}(V^\top \mathbf{x}) &\leftrightarrow (b \times \tilde{r}) \times (b \times l \times \tilde{r}) \to (b \times l \times \tilde{r}) \text{ broadcasted} \\ U(\mathbf{\Sigma}V^\top \mathbf{x}) &\leftrightarrow (m \times \tilde{r}) \times (b \times l \times \tilde{r}) \to (b \times l \times m) \text{ broadcasted} \end{aligned}$$

In our optimized setup, batched matrix multiplications can be completely circumvented if the Σ_i matrices are diagonal. If not, given that $\tilde{r} \ll m, n$, any required batched matrix multiplication remains computationally inexpensive.

I. Additional Results

This section elaborates on the results that underpin the figures presented in the main text and showcases a consistent correlation across various evaluation metrics. Additionally, we assess the significance of achieving convergence and the performance of compression on new unseen LoRA models. Note that compression rate here corresponds to the *parameter saved ratio* used in the main part of the paper.

I.1. Relative Rouge-L Performance and Compression Rate

Table 3 presents comprehensive results from the experiments underlying Figure 2b for each evaluation task. Additionally, we incorporate results using the Ties-merging benchmark [32], which consolidates all LoRA-adapters into a single adapter of identical configuration and parameter count; this integration significantly compromises performance.

Model Type	Method Type	Tasks										Average	Compression
		task039	task190	task280	task290	task391	task442	task620	task1342	task1391	task1598		
	base	0.26 ± 0.00	0.02 ± 0.00	0.19 ± 0.00	0.42 ± 0.00	0.11 ± 0.00	0.47 ± 0.00	0.11 ± 0.00	0.23 ± 0.00	0.19 ± 0.00	0.77 ± 0.00	0.28 ± 0.21	0.00
	lora	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.00
TIES	10	0.81 ± 0.00	0.57 ± 0.02	0.45 ± 0.04	0.10 ± 0.01	0.83 ± 0.01	0.47 ± 0.00	0.69 ± 0.01	0.57 ± 0.00	0.82 ± 0.01	0.85 ± 0.00	0.62 ± 0.23	0.00
	50	0.59 ± 0.00	0.41 ± 0.00	0.18 ± 0.05	0.03 ± 0.01	0.91 ± 0.01	0.31 ± 0.00	0.65 ± 0.00	0.62 ± 0.00	0.32 ± 0.04	0.84 ± 0.00	0.48 ± 0.28	0.00
	100	0.55 ± 0.00	0.40 ± 0.00	0.20 ± 0.05	0.01 ± 0.02	0.88 ± 0.00	0.33 ± 0.00	0.64 ± 0.00	0.57 ± 0.02	0.01 ± 0.00	0.82 ± 0.00	0.44 ± 0.30	0.00
	500	0.37 ± 0.00	0.26 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.83 ± 0.00	0.29 ± 0.00	0.57 ± 0.00	0.37 ± 0.00	0.01 ± 0.00	0.43 ± 0.00	0.31 ± 0.26	0.00
SVD	SVD 2	0.98 ± 0.03	1.07 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.98 ± 0.01	1.00 ± 0.01	1.00 ± 0.10	1.00 ± 0.01	1.00 ± 0.01	1.00 ± 0.04	0.88
	SVD 4	0.99 ± 0.04	1.04 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.01	1.00 ± 0.00	0.99 ± 0.02	0.99 ± 0.08	0.99 ± 0.01	1.00 ± 0.01	1.00 ± 0.03	0.75
	SVD 8	1.00 ± 0.00	1.02 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.01 ± 0.00	1.00 ± 0.01	1.01 ± 0.01	1.01 ± 0.00	1.00 ± 0.01	0.50
	SVD 16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.00
10 diagonal (D)	16 D	1.02 ± 0.01	1.01 ± 0.01	1.00 ± 0.00	1.00 ± 0.01	0.99 ± 0.00	0.96 ± 0.00	1.02 ± 0.02	1.13 ± 0.03	0.99 ± 0.02	0.98 ± 0.01	1.01 ± 0.05	1.00
	32 D	1.01 ± 0.01	1.05 ± 0.01	1.00 ± 0.00	0.99 ± 0.00	1.01 ± 0.01	0.99 ± 0.00	0.97 ± 0.01	1.05 ± 0.03	1.00 ± 0.01	1.00 ± 0.01	1.00 ± 0.03	1.00
	64 D	1.00 ± 0.00	1.03 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.01 ± 0.01	0.99 ± 0.01	1.01 ± 0.00	1.01 ± 0.00	1.00 ± 0.01	1.00
	128 D	1.00 ± 0.00	1.01 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.01 ± 0.01	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.01	1.00
	256 D	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00
10 full (F)	16 F	1.02 ± 0.00	1.06 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.98 ± 0.00	1.01 ± 0.02	1.07 ± 0.00	1.01 ± 0.01	1.00 ± 0.00	1.01 ± 0.03	1.00
	32 F	1.02 ± 0.01	1.04 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.96 ± 0.01	1.00 ± 0.02	1.00 ± 0.01	1.01 ± 0.00	1.00 ± 0.02	0.99
	64 F	1.00 ± 0.00	1.03 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.01 ± 0.01	0.98 ± 0.01	1.01 ± 0.00	1.01 ± 0.00	1.00 ± 0.01	0.97
	128 F	1.00 ± 0.00	1.01 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.88
	256 F	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.50
50 diagonal (D)	16 D	0.98 ± 0.04	0.98 ± 0.01	1.00 ± 0.00	0.92 ± 0.06	0.84 ± 0.07	0.92 ± 0.02	0.68 ± 0.05	0.87 ± 0.10	0.88 ± 0.07	0.83 ± 0.02	0.89 ± 0.10	1.00
	32 D	1.00 ± 0.02	1.02 ± 0.02	1.00 ± 0.00	0.99 ± 0.00	0.96 ± 0.01	0.95 ± 0.02	0.84 ± 0.02	1.00 ± 0.13	0.98 ± 0.01	0.88 ± 0.01	0.96 ± 0.07	1.00
	64 D	1.02 ± 0.00	1.05 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.97 ± 0.00	0.99 ± 0.01	1.09 ± 0.03	1.01 ± 0.01	0.90 ± 0.01	1.00 ± 0.05	1.00
	128 D	1.01 ± 0.01	1.08 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.98 ± 0.00	0.98 ± 0.01	1.11 ± 0.03	1.00 ± 0.00	1.00 ± 0.01	1.01 ± 0.04	1.00
	256 D	1.01 ± 0.01	1.03 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.01	1.00 ± 0.00	0.97 ± 0.03	1.01 ± 0.03	1.00 ± 0.01	1.01 ± 0.01	1.00 ± 0.02	1.00
50 full (F)	16 F	0.99 ± 0.04	1.00 ± 0.01	1.00 ± 0.01	0.96 ± 0.01	0.95 ± 0.02	0.94 ± 0.01	0.64 ± 0.10	1.01 ± 0.15	0.97 ± 0.02	0.87 ± 0.00	0.93 ± 0.12	1.00
	32 F	1.02 ± 0.00	1.00 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	0.98 ± 0.01	0.96 ± 0.00	0.95 ± 0.01	1.09 ± 0.02	1.01 ± 0.02	0.89 ± 0.01	0.99 ± 0.05	0.99
	64 F	1.02 ± 0.01	1.06 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.98 ± 0.01	1.03 ± 0.01	1.11 ± 0.00	1.00 ± 0.01	0.98 ± 0.02	1.02 ± 0.04	0.97
	128 F	1.02 ± 0.00	1.06 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.01	0.98 ± 0.00	0.98 ± 0.01	1.03 ± 0.04	1.00 ± 0.01	1.00 ± 0.00	1.01 ± 0.03	0.88
	256 F	1.00 ± 0.00	1.02 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	1.01 ± 0.01	1.00 ± 0.00	1.01 ± 0.00	1.01 ± 0.00	1.00 ± 0.01	0.50
100 diagonal (D)	16 D	0.80 ± 0.07	0.89 ± 0.06	0.93 ± 0.03	0.96 ± 0.01	0.50 ± 0.09	0.78 ± 0.01	0.28 ± 0.07	0.52 ± 0.10	0.78 ± 0.03	0.81 ± 0.02	0.72 ± 0.22	1.00
	32 D	0.95 ± 0.06	0.98 ± 0.01	1.00 ± 0.00	0.91 ± 0.06	0.80 ± 0.14	0.89 ± 0.06	0.60 ± 0.10	0.77 ± 0.26	0.91 ± 0.02	0.83 ± 0.02	0.86 ± 0.14	1.00
	64 D	1.01 ± 0.03	1.01 ± 0.01	1.00 ± 0.00	0.98 ± 0.02	0.96 ± 0.01	0.94 ± 0.01	0.88 ± 0.05	1.11 ± 0.08	0.96 ± 0.02	0.87 ± 0.03	0.97 ± 0.07	1.00
	128 D	1.01 ± 0.00	1.02 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.97 ± 0.00	1.00 ± 0.03	1.11 ± 0.02	0.99 ± 0.01	0.89 ± 0.02	1.00 ± 0.05	1.00
	256 D	1.00 ± 0.00	1.06 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.98 ± 0.00	1.00 ± 0.01	1.11 ± 0.03	1.00 ± 0.01	0.98 ± 0.01	1.01 ± 0.04	1.00
100 full (F)	16 F	0.95 ± 0.01	0.97 ± 0.03	0.97 ± 0.03	0.97 ± 0.03	0.93 ± 0.01	0.92 ± 0.01	0.64 ± 0.03	0.89 ± 0.16	0.87 ± 0.02	0.83 ± 0.01	0.89 ± 0.11	1.00
	32 F	1.00 ± 0.02	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.97 ± 0.01	0.95 ± 0.00	0.86 ± 0.03	1.12 ± 0.03	0.96 ± 0.01	0.87 ± 0.00	0.97 ± 0.07	0.99
	64 F	1.02 ± 0.00	1.00 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	0.96 ± 0.00	0.99 ± 0.01	1.09 ± 0.01	0.99 ± 0.02	0.89 ± 0.00	0.99 ± 0.05	0.97
	128 F	1.01 ± 0.01	1.05 ± 0.01	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	1.03 ± 0.01	1.10 ± 0.01	1.01 ± 0.00	0.99 ± 0.01	1.02 ± 0.04	0.88
	256 F	1.01 ± 0.01	1.03 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.01 ± 0.00	0.99 ± 0.00	0.98 ± 0.00	1.00 ± 0.03	1.01 ± 0.00	1.01 ± 0.00	1.00 ± 0.01	0.50
500 diagonal (D)	16 D	0.57 ± 0.07	0.53 ± 0.03	0.83 ± 0.04	0.78 ± 0.16	0.85 ± 0.04	0.68 ± 0.07	0.24 ± 0.01	0.43 ± 0.01	0.76 ± 0.06	0.79 ± 0.01	0.65 ± 0.20	1.00
	32 D	0.61 ± 0.12	0.52 ± 0.07	0.83 ± 0.02	0.84 ± 0.12	0.91 ± 0.02	0.71 ± 0.05	0.29 ± 0.05	0.47 ± 0.08	0.79 ± 0.04	0.79 ± 0.01	0.68 ± 0.20	1.00
	64 D	0.73 ± 0.02	0.59 ± 0.11	0.89 ± 0.04	0.97 ± 0.00	0.94 ± 0.00	0.83 ± 0.05	0.45 ± 0.09	0.50 ± 0.07	0.82 ± 0.02	0.80 ± 0.02	0.75 ± 0.18	1.00
	128 D	0.84 ± 0.00	0.91 ± 0.03	0.97 ± 0.03	0.98 ± 0.01	0.94 ± 0.00	0.88 ± 0.02	0.60 ± 0.15	0.53 ± 0.01	0.85 ± 0.05	0.80 ± 0.02	0.83 ± 0.15	1.00
	256 D	0.99 ± 0.03	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.96 ± 0.00	0.92 ± 0.03	0.66 ± 0.06	0.84 ± 0.14	0.92 ± 0.02	0.84 ± 0.01	0.91 ± 0.11	1.00
500 full (F)	16 F	0.57 ± 0.01	0.41 ± 0.05	0.78 ± 0.01	0.97 ± 0.00	0.96 ± 0.00	0.83 ± 0.01	0.64 ± 0.00	0.53 ± 0.03	0.83 ± 0.01	0.83 ± 0.00	0.74 ± 0.18	1.00
	32 F	0.79 ± 0.05	0.56 ± 0.04	0.93 ± 0.02	0.98 ± 0.00	0.97 ± 0.00	0.90 ± 0.01	0.69 ± 0.01	0.50 ± 0.00	0.86 ± 0.02	0.83 ± 0.01	0.80 ± 0.16	0.99
	64 F	1.02 ± 0.00	0.96 ± 0.01	0.94 ± 0.01	1.00 ± 0.01	0.96 ± 0.00	0.97 ± 0.01	0.73 ± 0.01	0.54 ± 0.01	0.91 ± 0.01	0.86 ± 0.00	0.89 ± 0.14	0.97
	128 F	1.03 ± 0.01	0.97 ± 0.02	0.99 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	0.96 ± 0.00	0.87 ± 0.01	1.07 ± 0.02	0.98 ± 0.00	0.87 ± 0.00	0.97 ± 0.06	0.88
	256 F	1.03 ± 0.00	1.02 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.97 ± 0.01	0.99 ± 0.02	1.03 ± 0.01	1.00 ± 0.01	0.87 ± 0.00	0.99 ± 0.05	0.50

Table 3. **Relative** In-Distribution *ROUGE-L* scores for various tasks and methods

I.2. Absolute Rouge-L Performance and Compression Rate

Table 4 provides the full results behind Table 3, but with Rouge-L scores instead of relative performance compared to LoRA.

I.3. Relative Rouge-1 Performance and Compression Rate

Table 5 provides full results for relative performance of Rouge-1, which shows the same trends as the results for relative performance of Rouge-L (Table 3).

I.4. Absolute Rouge-1 Performance and Compression Rate

Table 6 provides full results for absolute performance of Rouge-1, which shows the same trends as the results for absolute performance of Rouge-L (Table 4).

Model Type	Method Type	Tasks										Average	Compression
		task039	task190	task280	task290	task391	task442	task620	task1342	task1391	task1598		
	base	24.44 ± 0.00	1.60 ± 0.00	19.13 ± 0.00	39.22 ± 0.00	10.27 ± 0.00	35.46 ± 0.00	7.85 ± 0.00	6.22 ± 0.00	17.82 ± 0.00	38.87 ± 0.00	20.09 ± 13.32	0.00
	lora	95.00 ± 0.00	86.00 ± 0.00	99.00 ± 0.00	93.67 ± 0.00	94.33 ± 0.00	74.88 ± 0.00	74.40 ± 0.00	26.68 ± 0.00	95.00 ± 0.00	50.32 ± 0.00	78.93 ± 22.48	0.00
SVD	SVD 2	93.15 ± 2.77	91.98 ± 1.99	99.09 ± 0.18	93.44 ± 0.14	93.89 ± 0.35	73.74 ± 0.51	74.55 ± 0.98	26.80 ± 2.79	95.06 ± 1.35	50.21 ± 0.44	79.19 ± 22.64	0.88
	SVD 4	94.01 ± 3.60	89.15 ± 0.71	99.05 ± 0.09	93.65 ± 0.03	94.66 ± 0.63	74.89 ± 0.33	73.61 ± 1.15	26.34 ± 2.13	93.98 ± 0.77	50.47 ± 0.54	78.98 ± 22.60	0.75
	SVD 8	95.00 ± 0.00	87.37 ± 0.57	99.05 ± 0.09	93.65 ± 0.03	94.36 ± 0.38	74.58 ± 0.12	75.07 ± 0.00	26.71 ± 0.27	95.51 ± 1.09	50.89 ± 0.07	81.07 ± 21.64	0.50
	SVD 16	95.00 ± 0.00	86.00 ± 0.00	99.00 ± 0.00	93.67 ± 0.00	94.33 ± 0.00	74.90 ± 0.03	74.23 ± 0.18	26.68 ± 0.00	95.00 ± 0.00	50.30 ± 0.02	78.36 ± 22.97	0.00
10 diagonal (D)	16 D	96.67 ± 0.58	87.00 ± 1.00	99.00 ± 0.00	94.00 ± 0.67	93.11 ± 0.38	72.08 ± 0.06	76.26 ± 1.19	30.11 ± 0.79	94.00 ± 1.73	49.30 ± 0.46	79.15 ± 22.18	1.00
	32 D	95.67 ± 0.58	90.00 ± 1.00	99.00 ± 0.00	93.00 ± 0.33	94.89 ± 0.51	73.86 ± 0.31	71.92 ± 0.84	27.89 ± 0.70	94.67 ± 0.58	50.36 ± 0.26	79.13 ± 22.75	1.00
	64 D	95.00 ± 0.00	88.33 ± 0.58	99.00 ± 0.00	93.67 ± 0.00	94.78 ± 0.38	74.61 ± 0.13	74.97 ± 0.58	26.35 ± 0.25	96.00 ± 0.00	50.99 ± 0.06	79.37 ± 22.94	1.00
	128 D	95.00 ± 0.00	86.67 ± 0.58	99.00 ± 0.00	93.67 ± 0.00	94.33 ± 0.00	74.92 ± 0.13	74.96 ± 0.51	26.45 ± 0.23	95.00 ± 0.00	50.21 ± 0.12	79.02 ± 22.84	1.00
	256 D	95.00 ± 0.00	86.00 ± 0.00	99.00 ± 0.00	93.67 ± 0.00	94.33 ± 0.00	74.88 ± 0.00	74.40 ± 0.00	26.68 ± 0.00	95.00 ± 0.00	50.27 ± 0.02	78.92 ± 22.77	1.00
10 full (F)	16 F	97.00 ± 0.00	91.00 ± 1.00	99.00 ± 0.00	93.56 ± 0.19	93.56 ± 0.69	73.60 ± 0.36	74.94 ± 1.25	28.66 ± 0.03	96.00 ± 1.00	50.15 ± 0.20	79.75 ± 22.72	1.00
	32 F	96.67 ± 0.58	89.33 ± 0.58	99.00 ± 0.00	93.22 ± 0.19	94.44 ± 0.19	74.11 ± 0.19	71.74 ± 0.59	26.74 ± 0.50	94.67 ± 0.58	50.63 ± 0.24	79.06 ± 23.01	0.99
	64 F	95.00 ± 0.00	88.67 ± 0.58	99.00 ± 0.00	93.67 ± 0.00	94.56 ± 0.38	74.56 ± 0.13	75.47 ± 0.58	26.26 ± 0.34	96.00 ± 0.00	50.89 ± 0.17	79.41 ± 22.97	0.97
	128 F	95.00 ± 0.00	86.67 ± 0.58	99.00 ± 0.00	93.67 ± 0.00	94.33 ± 0.00	75.04 ± 0.03	74.40 ± 0.00	26.53 ± 0.13	95.00 ± 0.00	50.36 ± 0.03	79.00 ± 22.81	0.88
	256 F	95.00 ± 0.00	86.00 ± 0.00	99.00 ± 0.00	93.67 ± 0.00	94.33 ± 0.00	74.90 ± 0.03	74.29 ± 0.19	26.68 ± 0.00	95.00 ± 0.00	50.30 ± 0.03	78.92 ± 22.77	0.50
50 diagonal (D)	16 D	92.76 ± 3.53	84.67 ± 1.15	99.00 ± 0.00	86.17 ± 5.81	79.68 ± 6.21	69.07 ± 1.54	50.65 ± 3.97	23.27 ± 2.60	83.90 ± 6.43	41.86 ± 0.96	71.10 ± 23.99	1.00
	32 D	95.33 ± 2.08	87.33 ± 2.08	99.00 ± 0.00	92.60 ± 0.29	90.32 ± 1.04	71.16 ± 1.47	62.51 ± 1.64	26.60 ± 3.54	93.33 ± 1.15	44.35 ± 0.41	76.25 ± 23.81	1.00
	64 D	97.00 ± 0.00	90.33 ± 1.53	99.00 ± 0.00	93.78 ± 0.19	93.00 ± 0.58	72.37 ± 0.35	73.39 ± 0.93	29.06 ± 0.80	95.67 ± 0.58	45.43 ± 0.34	78.90 ± 23.29	1.00
	128 D	96.33 ± 0.58	92.67 ± 0.58	99.00 ± 0.00	93.56 ± 0.19	93.00 ± 0.58	73.32 ± 0.24	73.03 ± 1.09	29.51 ± 0.93	95.00 ± 0.00	50.16 ± 0.74	79.56 ± 22.51	1.00
	256 D	95.67 ± 0.58	88.33 ± 0.58	99.00 ± 0.00	93.56 ± 0.19	94.67 ± 0.67	74.82 ± 0.24	72.36 ± 2.07	26.90 ± 0.75	95.33 ± 0.58	50.73 ± 0.46	79.14 ± 22.90	1.00
50 full (F)	16 F	94.06 ± 3.54	85.67 ± 1.15	98.67 ± 0.58	90.35 ± 1.37	89.90 ± 1.91	70.32 ± 0.66	47.62 ± 7.28	26.88 ± 3.96	92.33 ± 1.53	43.68 ± 0.24	73.95 ± 24.73	1.00
	32 F	97.00 ± 0.00	85.67 ± 1.53	99.00 ± 0.00	93.67 ± 0.00	92.22 ± 0.69	71.88 ± 0.30	71.01 ± 1.02	29.07 ± 0.65	95.67 ± 1.53	44.97 ± 0.41	78.02 ± 23.18	0.99
	64 F	96.67 ± 0.58	91.00 ± 2.00	99.00 ± 0.00	93.56 ± 0.19	93.22 ± 0.51	73.16 ± 0.41	76.28 ± 0.51	29.67 ± 0.12	95.33 ± 0.58	49.31 ± 1.00	79.72 ± 22.50	0.97
	128 F	97.00 ± 0.00	91.00 ± 1.00	99.00 ± 0.00	93.33 ± 0.00	94.11 ± 0.51	73.51 ± 0.23	73.17 ± 0.58	27.53 ± 1.12	95.00 ± 1.00	50.56 ± 0.06	79.42 ± 22.93	0.88
	256 F	95.00 ± 0.00	88.00 ± 0.00	99.00 ± 0.00	93.67 ± 0.00	94.44 ± 0.19	74.25 ± 0.21	74.97 ± 0.58	26.79 ± 0.09	96.00 ± 0.00	50.86 ± 0.19	79.30 ± 22.82	0.50
100 diagonal (D)	16 D	76.43 ± 7.07	76.67 ± 4.93	91.61 ± 2.75	89.99 ± 1.07	47.55 ± 8.56	58.08 ± 0.72	20.77 ± 5.50	13.90 ± 2.79	73.93 ± 3.13	40.74 ± 0.85	58.97 ± 26.83	1.00
	32 D	90.10 ± 5.85	84.00 ± 1.00	99.00 ± 0.00	85.52 ± 5.34	75.69 ± 12.75	66.62 ± 4.18	44.66 ± 7.26	20.49 ± 7.07	86.67 ± 1.86	42.01 ± 0.94	69.48 ± 25.14	1.00
	64 D	95.56 ± 2.49	86.67 ± 0.58	99.00 ± 0.00	92.24 ± 1.68	90.89 ± 1.17	70.35 ± 0.45	65.62 ± 4.03	29.58 ± 2.02	91.67 ± 2.31	43.64 ± 1.36	76.52 ± 23.02	1.00
	128 D	96.00 ± 0.00	87.33 ± 1.15	99.00 ± 0.00	93.89 ± 0.19	93.00 ± 0.58	72.70 ± 0.30	74.34 ± 2.07	29.66 ± 0.54	93.67 ± 0.58	44.82 ± 0.89	78.44 ± 22.87	1.00
	256 D	95.00 ± 0.00	91.00 ± 0.00	99.00 ± 0.00	93.56 ± 0.19	93.11 ± 0.19	73.05 ± 0.20	74.52 ± 0.95	29.67 ± 0.67	95.33 ± 0.58	49.42 ± 0.65	79.37 ± 22.38	1.00
100 full (F)	16 F	90.70 ± 1.07	83.00 ± 2.65	96.00 ± 3.00	91.22 ± 2.94	87.94 ± 0.54	68.72 ± 1.05	47.57 ± 2.54	23.75 ± 4.33	82.33 ± 2.08	41.51 ± 0.67	71.27 ± 24.23	1.00
	32 F	95.33 ± 1.53	85.00 ± 1.00	99.00 ± 0.00	93.50 ± 0.22	91.44 ± 0.84	70.94 ± 0.02	63.64 ± 1.98	29.82 ± 0.81	91.67 ± 0.58	43.94 ± 0.18	76.43 ± 23.01	0.99
	64 F	97.00 ± 0.00	85.67 ± 1.53	99.00 ± 0.00	93.78 ± 0.19	92.56 ± 0.19	72.11 ± 0.08	73.29 ± 0.64	29.15 ± 0.24	94.33 ± 1.53	44.97 ± 0.05	78.18 ± 23.03	0.97
	128 F	96.33 ± 0.58	90.33 ± 0.58	99.00 ± 0.00	93.00 ± 0.00	93.89 ± 0.19	73.11 ± 0.36	76.50 ± 1.01	29.45 ± 0.35	96.00 ± 0.00	49.81 ± 0.34	79.74 ± 22.47	0.88
	256 F	96.33 ± 0.58	88.67 ± 0.58	99.00 ± 0.00	93.67 ± 0.00	94.89 ± 0.19	74.40 ± 0.16	72.90 ± 0.12	26.77 ± 0.68	96.00 ± 0.00	50.83 ± 0.09	79.35 ± 23.04	0.50
500 diagonal (D)	16 D	54.44 ± 6.87	46.00 ± 2.65	82.21 ± 3.59	73.38 ± 14.97	80.08 ± 3.71	51.02 ± 5.31	17.49 ± 1.10	11.58 ± 0.21	72.67 ± 6.03	39.65 ± 0.28	52.85 ± 24.59	1.00
	32 D	58.08 ± 11.52	45.00 ± 6.08	82.06 ± 1.69	78.62 ± 11.23	85.57 ± 1.48	52.98 ± 3.81	21.73 ± 3.95	12.53 ± 2.26	75.33 ± 4.04	39.78 ± 0.42	55.17 ± 25.18	1.00
	64 D	69.21 ± 2.03	50.67 ± 9.29	88.33 ± 4.04	91.11 ± 0.38	88.78 ± 0.38	62.36 ± 3.52	33.36 ± 6.69	13.34 ± 1.86	77.67 ± 2.31	40.42 ± 0.98	61.53 ± 25.83	1.00
	128 D	79.77 ± 0.37	78.00 ± 3.00	95.89 ± 2.83	91.89 ± 1.39	88.67 ± 0.00	65.92 ± 1.79	44.98 ± 10.98	14.14 ± 0.19	81.00 ± 5.00	40.34 ± 0.80	68.06 ± 25.92	1.00
	256 D	93.83 ± 2.52	85.33 ± 0.58	99.00 ± 0.00	93.78 ± 0.19	90.56 ± 0.38	68.95 ± 1.92	49.39 ± 4.36	22.33 ± 3.78	87.33 ± 2.31	42.15 ± 0.73	73.27 ± 25.60	1.00
500 full (F)	16 F	54.30 ± 1.13	35.67 ± 4.62	77.67 ± 0.58	91.00 ± 0.00	90.56 ± 0.19	62.47 ± 0.79	47.56 ± 0.29	14.18 ± 0.67	79.00 ± 1.00	41.58 ± 0.23	59.40 ± 24.51	1.00
	32 F	75.10 ± 4.92	48.00 ± 3.61	91.67 ± 1.53	91.56 ± 0.19	91.56 ± 0.38	67.37 ± 0.83	51.17 ± 0.81	13.44 ± 0.02	81.67 ± 1.53	41.92 ± 0.42	65.34 ± 25.34	0.99
	64 F	96.94 ± 0.42	82.67 ± 0.58	93.33 ± 0.58	93.89 ± 0.69	90.67 ± 0.00	72.30 ± 0.71	54.63 ± 0.79	14.49 ± 0.27	86.33 ± 0.58	43.16 ± 0.08	72.84 ± 26.25	0.97
	128 F	97.67 ± 0.58	83.67 ± 1.53	98.00 ± 0.00	93.56 ± 0.19	92.00 ± 0.00	71.92 ± 0.19	65.02 ± 0.81	28.49 ± 0.55	93.00 ± 0.00	43.85 ± 0.12	76.72 ± 23.39	0.88
	256 F	98.00 ± 0.00	88.00 ± 1.00	99.00 ± 0.00	93.78 ± 0.19	93.00 ± 0.88	72.45 ± 0.38	73.77 ± 1.21	27.59 ± 0.39	95.33 ± 0.58	43.81 ± 0.17	78.47 ± 23.80	0.50

Table 4. **Absolute** In-Distribution ROUGE-L scores for various tasks and methods

I.5. Relative Exact-Match Performance and Compression Rate

Table 7 provides full results for relative performance of exact-match, which shows the same trends as the results for relative performance of Rouge-L (Table 3).

I.6. Absolute Exact-Match Performance and Compression Rate

Table 8 provides full results for absolute performance of exact-match, which shows the same trends as the results for absolute performance of Rouge-L (Table 4).

I.7. Loss and Compression Rate

Table 9 provides full results for test loss (cross-entropy), which shows the same trends as the results for relative performance of Rouge-L (Table 3).

I.8. Agreement and Compression Rate

Table 10 provides full results for *agreement*, which shows the same trends as the results for relative performance of Rouge-L (Table 3). Note that *agreement* measures the exact match in task generations between the uncompressed LoRA model and the compressed LoRA model, rather than comparing to the task's ground truth data. The comparison is very strict and requires an exact match between the generations of the two models (LoRA and the compressed LoRA), comparing each

Model Type	Method Type	Tasks										Average	Compression
		task039	task190	task280	task290	task391	task442	task620	task1342	task1391	task1598		
	base	0.26 ± 0.00	0.02 ± 0.00	0.19 ± 0.00	0.42 ± 0.00	0.11 ± 0.00	0.51 ± 0.00	0.11 ± 0.00	0.26 ± 0.00	0.19 ± 0.00	0.80 ± 0.00	0.29 ± 0.22	0.00
	lora	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.00
SVD	SVD 2	0.98 ± 0.03	1.07 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	1.01 ± 0.01	1.00 ± 0.10	1.00 ± 0.01	0.99 ± 0.01	1.00 ± 0.04	0.88
	SVD 4	0.99 ± 0.04	1.04 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.01	1.00 ± 0.00	0.99 ± 0.01	0.99 ± 0.08	0.99 ± 0.01	1.01 ± 0.00	1.00 ± 0.03	0.75
	SVD 8	1.00 ± 0.00	1.02 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.01 ± 0.00	1.00 ± 0.01	1.01 ± 0.01	1.01 ± 0.00	1.00 ± 0.01	0.50
	SVD 16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.00
10 diagonal (D)	16 D	1.02 ± 0.01	1.01 ± 0.01	1.00 ± 0.00	1.00 ± 0.01	0.99 ± 0.00	0.97 ± 0.00	1.03 ± 0.02	1.12 ± 0.03	0.99 ± 0.02	0.99 ± 0.00	1.01 ± 0.04	1.00
	32 D	1.01 ± 0.01	1.05 ± 0.01	1.00 ± 0.00	0.99 ± 0.00	1.01 ± 0.01	0.99 ± 0.00	0.97 ± 0.01	1.04 ± 0.03	1.00 ± 0.01	1.01 ± 0.01	1.01 ± 0.02	1.00
	64 D	1.00 ± 0.00	1.03 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.01 ± 0.01	0.99 ± 0.01	1.01 ± 0.00	1.01 ± 0.00	1.00 ± 0.01	1.00
	128 D	1.00 ± 0.00	1.01 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.01 ± 0.01	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.01	1.00
	256 D	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00
10 full (F)	16 F	1.02 ± 0.00	1.06 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.99 ± 0.00	1.01 ± 0.02	1.07 ± 0.00	1.01 ± 0.01	1.00 ± 0.00	1.02 ± 0.03	1.00
	32 F	1.02 ± 0.01	1.04 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.96 ± 0.01	1.00 ± 0.02	1.00 ± 0.01	1.01 ± 0.00	1.00 ± 0.02	0.99
	64 F	1.00 ± 0.00	1.03 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.01 ± 0.01	0.98 ± 0.01	1.01 ± 0.00	1.01 ± 0.00	1.00 ± 0.01	0.97
	128 F	1.00 ± 0.00	1.01 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.88
	256 F	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.50
50 diagonal (D)	16 D	0.98 ± 0.04	0.98 ± 0.01	1.00 ± 0.00	0.92 ± 0.06	0.85 ± 0.06	0.94 ± 0.02	0.69 ± 0.05	0.88 ± 0.10	0.88 ± 0.07	0.86 ± 0.01	0.90 ± 0.10	1.00
	32 D	1.00 ± 0.02	1.02 ± 0.02	1.00 ± 0.00	0.99 ± 0.00	0.96 ± 0.01	0.96 ± 0.02	0.85 ± 0.02	1.00 ± 0.12	0.98 ± 0.01	0.90 ± 0.00	0.97 ± 0.06	1.00
	64 D	1.02 ± 0.00	1.05 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.97 ± 0.01	0.99 ± 0.01	1.09 ± 0.03	1.01 ± 0.01	0.94 ± 0.00	1.01 ± 0.04	1.00
	128 D	1.01 ± 0.01	1.08 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.98 ± 0.00	0.98 ± 0.02	1.10 ± 0.03	1.00 ± 0.00	1.01 ± 0.01	1.02 ± 0.04	1.00
	256 D	1.01 ± 0.01	1.03 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.01	1.00 ± 0.00	0.97 ± 0.03	1.00 ± 0.03	1.00 ± 0.01	1.01 ± 0.00	1.00 ± 0.02	1.00
50 full (F)	16 F	0.99 ± 0.04	1.00 ± 0.01	1.00 ± 0.01	0.96 ± 0.01	0.95 ± 0.02	0.95 ± 0.01	0.65 ± 0.09	1.01 ± 0.15	0.97 ± 0.02	0.88 ± 0.01	0.94 ± 0.11	1.00
	32 F	1.02 ± 0.00	1.00 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	0.98 ± 0.01	0.97 ± 0.00	0.96 ± 0.01	1.09 ± 0.03	1.01 ± 0.02	0.93 ± 0.00	0.99 ± 0.04	0.99
	64 F	1.02 ± 0.01	1.06 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.98 ± 0.00	1.03 ± 0.01	1.11 ± 0.00	1.00 ± 0.01	0.99 ± 0.01	1.02 ± 0.04	0.97
	128 F	1.02 ± 0.00	1.06 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.01	0.98 ± 0.00	0.98 ± 0.01	1.03 ± 0.04	1.00 ± 0.01	1.01 ± 0.00	1.01 ± 0.02	0.88
	256 F	1.00 ± 0.00	1.02 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	1.01 ± 0.01	1.00 ± 0.00	1.01 ± 0.00	1.01 ± 0.00	1.00 ± 0.01	0.50
100 diagonal (D)	16 D	0.80 ± 0.07	0.89 ± 0.06	0.93 ± 0.03	0.96 ± 0.01	0.51 ± 0.09	0.81 ± 0.02	0.30 ± 0.07	0.54 ± 0.11	0.78 ± 0.03	0.83 ± 0.02	0.73 ± 0.21	1.00
	32 D	0.95 ± 0.06	0.98 ± 0.01	1.00 ± 0.00	0.91 ± 0.06	0.80 ± 0.13	0.91 ± 0.05	0.62 ± 0.10	0.78 ± 0.25	0.91 ± 0.02	0.85 ± 0.01	0.87 ± 0.14	1.00
	64 D	1.01 ± 0.03	1.01 ± 0.01	1.00 ± 0.00	0.98 ± 0.02	0.96 ± 0.01	0.95 ± 0.01	0.90 ± 0.05	1.11 ± 0.07	0.96 ± 0.02	0.88 ± 0.02	0.98 ± 0.07	1.00
	128 D	1.01 ± 0.00	1.02 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.98 ± 0.00	1.00 ± 0.03	1.11 ± 0.02	0.99 ± 0.01	0.92 ± 0.00	1.00 ± 0.05	1.00
	256 D	1.00 ± 0.00	1.06 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.98 ± 0.00	1.00 ± 0.01	1.11 ± 0.03	1.00 ± 0.01	0.99 ± 0.02	1.01 ± 0.04	1.00
100 full (F)	16 F	0.95 ± 0.01	0.97 ± 0.03	0.97 ± 0.03	0.97 ± 0.03	0.93 ± 0.01	0.93 ± 0.01	0.66 ± 0.03	0.90 ± 0.16	0.87 ± 0.02	0.85 ± 0.01	0.90 ± 0.10	1.00
	32 F	1.00 ± 0.02	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.97 ± 0.01	0.96 ± 0.00	0.87 ± 0.03	1.12 ± 0.03	0.96 ± 0.01	0.89 ± 0.00	0.98 ± 0.07	0.99
	64 F	1.02 ± 0.00	1.00 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	0.97 ± 0.00	0.99 ± 0.01	1.10 ± 0.01	0.99 ± 0.02	0.93 ± 0.01	1.00 ± 0.04	0.97
	128 F	1.01 ± 0.01	1.05 ± 0.01	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	1.03 ± 0.01	1.10 ± 0.01	1.01 ± 0.00	1.00 ± 0.00	1.02 ± 0.03	0.88
	256 F	1.01 ± 0.01	1.03 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.01 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	1.00 ± 0.03	1.01 ± 0.00	1.01 ± 0.00	1.00 ± 0.01	0.50
500 diagonal (D)	16 D	0.57 ± 0.07	0.53 ± 0.03	0.83 ± 0.04	0.78 ± 0.16	0.85 ± 0.04	0.73 ± 0.07	0.24 ± 0.02	0.45 ± 0.01	0.76 ± 0.06	0.81 ± 0.00	0.66 ± 0.20	1.00
	32 D	0.61 ± 0.12	0.52 ± 0.07	0.83 ± 0.02	0.84 ± 0.12	0.91 ± 0.02	0.75 ± 0.05	0.30 ± 0.05	0.49 ± 0.07	0.79 ± 0.04	0.82 ± 0.01	0.69 ± 0.20	1.00
	64 D	0.73 ± 0.02	0.59 ± 0.11	0.89 ± 0.04	0.97 ± 0.00	0.94 ± 0.00	0.86 ± 0.03	0.46 ± 0.09	0.51 ± 0.07	0.82 ± 0.02	0.83 ± 0.01	0.76 ± 0.18	1.00
	128 D	0.84 ± 0.00	0.91 ± 0.03	0.97 ± 0.03	0.98 ± 0.01	0.94 ± 0.00	0.90 ± 0.02	0.62 ± 0.14	0.54 ± 0.01	0.85 ± 0.05	0.83 ± 0.01	0.84 ± 0.15	1.00
	256 D	0.99 ± 0.03	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.96 ± 0.00	0.93 ± 0.02	0.68 ± 0.05	0.85 ± 0.14	0.92 ± 0.02	0.85 ± 0.00	0.92 ± 0.11	1.00
500 full (F)	16 F	0.57 ± 0.01	0.41 ± 0.05	0.78 ± 0.01	0.97 ± 0.00	0.96 ± 0.00	0.86 ± 0.01	0.65 ± 0.00	0.55 ± 0.02	0.83 ± 0.01	0.84 ± 0.00	0.74 ± 0.18	1.00
	32 F	0.79 ± 0.05	0.56 ± 0.04	0.93 ± 0.02	0.98 ± 0.00	0.97 ± 0.00	0.92 ± 0.00	0.70 ± 0.01	0.52 ± 0.00	0.86 ± 0.02	0.85 ± 0.00	0.81 ± 0.16	0.99
	64 F	1.02 ± 0.00	0.96 ± 0.01	0.94 ± 0.01	1.00 ± 0.01	0.96 ± 0.00	0.97 ± 0.01	0.74 ± 0.01	0.55 ± 0.01	0.91 ± 0.01	0.87 ± 0.00	0.89 ± 0.14	0.97
	128 F	1.03 ± 0.01	0.97 ± 0.02	0.99 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	0.97 ± 0.00	0.88 ± 0.01	1.07 ± 0.02	0.98 ± 0.00	0.90 ± 0.00	0.98 ± 0.05	0.88
	256 F	1.03 ± 0.00	1.02 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.97 ± 0.00	1.00 ± 0.02	1.04 ± 0.02	1.00 ± 0.01	0.93 ± 0.00	1.00 ± 0.03	0.50

Table 5. **Relative** In-Distribution ROUGE-1 scores for various tasks and methods

sample one at a time.

I.9. Reconstruction Error and Compression Rate

Table 11 provides the full results of the experiments behind Figure 3 for every evaluation task.

I.10. Performance of Compressing Unseen/New LoRAs

Table 12 provides results for JD-full, where we project new unseen LoRAs into the shared basis. Our findings indicate that while compressing a new LoRA within the original space improves upon the baseline, it can degrades performance compared to uncompressed LoRAs

I.11. Convergence

Table 13 presents outcomes where the JD-Full algorithm is executed until convergence. Our convergence criterion is defined as follows:

$$\max\left(\|U_{t+1} - U_t U_t^\top U_{t+1}\|_{\text{Fro}} / \|U_{t+1}\|_{\text{Fro}}, \|V_{t+1} - V_t V_t^\top V_{t+1}\|_{\text{Fro}} / \|V_{t+1}\|_{\text{Fro}}\right) < \tau \tag{19}$$

Model Type	Method Type	Tasks										Average	Compression
		task039	task190	task280	task290	task391	task442	task620	task1342	task1391	task1598		
	base	24.44 ± 0.00	1.60 ± 0.00	19.13 ± 0.00	39.22 ± 0.00	10.42 ± 0.00	39.88 ± 0.00	8.05 ± 0.00	6.96 ± 0.00	17.82 ± 0.00	55.03 ± 0.00	22.26 ± 16.53	0.00
	lora	95.00 ± 0.00	86.00 ± 0.00	99.00 ± 0.00	93.67 ± 0.00	94.33 ± 0.00	78.43 ± 0.00	74.90 ± 0.00	26.87 ± 0.00	95.00 ± 0.00	68.66 ± 0.00	81.19 ± 20.59	0.00
SVD	SVD 2	93.15 ± 2.77	91.98 ± 1.99	99.09 ± 0.18	93.44 ± 0.14	93.89 ± 0.35	77.33 ± 0.29	75.40 ± 1.01	26.90 ± 2.68	95.06 ± 1.35	67.71 ± 0.49	81.40 ± 20.78	0.88
	SVD 4	94.01 ± 3.60	89.15 ± 0.71	99.05 ± 0.09	93.65 ± 0.03	94.66 ± 0.63	78.42 ± 0.23	74.09 ± 1.12	26.47 ± 2.06	93.98 ± 0.77	69.37 ± 0.21	81.28 ± 20.72	0.75
	SVD 8	95.00 ± 0.00	87.37 ± 0.57	99.05 ± 0.09	93.65 ± 0.03	94.36 ± 0.38	78.21 ± 0.03	75.57 ± 0.00	26.88 ± 0.27	95.51 ± 1.09	69.33 ± 0.08	83.06 ± 19.77	0.50
	SVD 16	95.00 ± 0.00	86.00 ± 0.00	99.00 ± 0.00	93.67 ± 0.00	94.33 ± 0.00	78.44 ± 0.03	74.73 ± 0.18	26.87 ± 0.00	95.00 ± 0.00	68.62 ± 0.04	80.76 ± 21.05	0.00
10 diagonal (D)	16 D	96.67 ± 0.58	87.00 ± 1.00	99.00 ± 0.00	94.00 ± 0.67	93.11 ± 0.38	76.08 ± 0.17	77.26 ± 1.47	30.15 ± 0.72	94.00 ± 1.73	68.25 ± 0.18	81.55 ± 20.03	1.00
	32 D	95.67 ± 0.58	90.00 ± 1.00	99.00 ± 0.00	93.00 ± 0.33	94.89 ± 0.51	77.46 ± 0.24	72.53 ± 1.00	27.98 ± 0.71	94.67 ± 0.58	69.16 ± 0.41	81.44 ± 20.80	1.00
	64 D	95.00 ± 0.00	88.33 ± 0.58	99.00 ± 0.00	93.67 ± 0.00	94.78 ± 0.38	78.28 ± 0.07	75.47 ± 0.58	26.53 ± 0.25	96.00 ± 0.00	69.36 ± 0.05	81.64 ± 21.06	1.00
	128 D	95.00 ± 0.00	86.67 ± 0.58	99.00 ± 0.00	93.67 ± 0.00	94.33 ± 0.00	78.45 ± 0.16	75.46 ± 0.51	26.64 ± 0.23	95.00 ± 0.00	68.70 ± 0.14	81.29 ± 20.92	1.00
	256 D	95.00 ± 0.00	86.00 ± 0.00	99.00 ± 0.00	93.67 ± 0.00	94.33 ± 0.00	78.43 ± 0.00	74.90 ± 0.00	26.87 ± 0.00	95.00 ± 0.00	68.59 ± 0.03	81.18 ± 20.86	1.00
10 full (F)	16 F	97.00 ± 0.00	91.00 ± 1.00	99.00 ± 0.00	93.56 ± 0.19	93.56 ± 0.69	77.64 ± 0.25	75.78 ± 1.25	28.71 ± 0.09	96.00 ± 1.00	68.69 ± 0.08	82.09 ± 20.68	1.00
	32 F	96.67 ± 0.58	89.33 ± 0.58	99.00 ± 0.00	93.22 ± 0.19	94.44 ± 0.19	77.84 ± 0.21	72.24 ± 0.59	26.84 ± 0.50	94.67 ± 0.58	69.55 ± 0.08	81.38 ± 21.11	0.99
	64 F	95.00 ± 0.00	88.67 ± 0.58	99.00 ± 0.00	93.67 ± 0.00	94.56 ± 0.38	78.19 ± 0.08	75.97 ± 0.58	26.43 ± 0.34	96.00 ± 0.00	69.38 ± 0.11	81.69 ± 21.07	0.97
	128 F	95.00 ± 0.00	86.67 ± 0.58	99.00 ± 0.00	93.67 ± 0.00	94.33 ± 0.00	78.46 ± 0.03	74.90 ± 0.00	26.72 ± 0.13	95.00 ± 0.00	68.65 ± 0.03	81.24 ± 20.91	0.88
	256 F	95.00 ± 0.00	86.00 ± 0.00	99.00 ± 0.00	93.67 ± 0.00	94.33 ± 0.00	78.44 ± 0.03	74.79 ± 0.19	26.87 ± 0.00	95.00 ± 0.00	68.64 ± 0.03	81.17 ± 20.86	0.50
50 diagonal (D)	16 D	92.76 ± 3.53	84.67 ± 1.15	99.00 ± 0.00	86.17 ± 5.81	79.83 ± 6.08	73.55 ± 1.39	51.72 ± 3.78	23.75 ± 2.66	83.90 ± 6.43	59.05 ± 0.94	73.44 ± 22.08	1.00
	32 D	95.33 ± 2.08	87.33 ± 2.08	99.00 ± 0.00	92.60 ± 0.29	90.35 ± 1.00	75.43 ± 1.33	63.84 ± 1.64	26.97 ± 3.21	93.33 ± 1.15	61.94 ± 0.32	78.61 ± 21.60	1.00
	64 D	97.00 ± 0.00	90.33 ± 1.53	99.00 ± 0.00	93.78 ± 0.19	93.00 ± 0.58	76.27 ± 0.49	74.39 ± 0.90	29.28 ± 0.81	95.67 ± 0.58	64.84 ± 0.27	81.36 ± 20.83	1.00
	128 D	96.33 ± 0.58	92.67 ± 0.58	99.00 ± 0.00	93.56 ± 0.19	93.00 ± 0.58	77.24 ± 0.19	73.76 ± 1.25	29.58 ± 0.93	95.00 ± 0.00	69.04 ± 0.54	81.92 ± 20.44	1.00
	256 D	95.67 ± 0.58	88.33 ± 0.58	99.00 ± 0.00	93.56 ± 0.19	94.67 ± 0.67	78.45 ± 0.14	72.86 ± 2.07	27.00 ± 0.77	95.33 ± 0.58	69.61 ± 0.18	81.45 ± 21.00	1.00
50 full (F)	16 F	94.06 ± 3.54	85.67 ± 1.15	98.67 ± 0.58	90.35 ± 1.37	89.97 ± 1.78	74.46 ± 0.58	49.03 ± 7.07	27.14 ± 3.94	92.33 ± 1.53	60.26 ± 1.03	76.19 ± 22.80	1.00
	32 F	97.00 ± 0.00	85.67 ± 1.53	99.00 ± 0.00	93.67 ± 0.00	92.22 ± 0.69	75.86 ± 0.22	71.68 ± 0.65	29.26 ± 0.70	95.67 ± 1.53	63.88 ± 0.10	80.39 ± 20.81	0.99
	64 F	96.67 ± 0.58	91.00 ± 2.00	99.00 ± 0.00	93.56 ± 0.19	93.22 ± 0.51	77.17 ± 0.38	77.11 ± 0.51	29.75 ± 0.03	95.33 ± 0.58	68.13 ± 0.75	82.09 ± 20.33	0.97
	128 F	97.00 ± 0.00	91.00 ± 1.00	99.00 ± 0.00	93.33 ± 0.00	94.11 ± 0.51	77.23 ± 0.17	73.67 ± 0.58	27.62 ± 1.12	95.00 ± 1.00	69.40 ± 0.16	81.74 ± 20.97	0.88
	256 F	95.00 ± 0.00	88.00 ± 0.00	99.00 ± 0.00	93.67 ± 0.00	94.44 ± 0.19	77.97 ± 0.24	75.47 ± 0.58	26.96 ± 0.09	96.00 ± 0.00	69.28 ± 0.05	81.58 ± 20.92	0.50
100 diagonal (D)	16 D	76.43 ± 7.07	76.67 ± 4.93	91.61 ± 2.75	89.99 ± 1.07	47.89 ± 8.62	63.17 ± 1.31	22.23 ± 5.27	14.46 ± 2.89	73.93 ± 3.13	57.17 ± 1.05	61.35 ± 25.78	1.00
	32 D	90.10 ± 5.85	84.00 ± 1.00	99.00 ± 0.00	85.52 ± 5.34	75.88 ± 12.57	71.15 ± 3.61	46.10 ± 7.39	21.04 ± 6.76	86.67 ± 1.86	58.64 ± 1.02	71.81 ± 23.39	1.00
	64 D	95.56 ± 2.49	86.67 ± 0.58	99.00 ± 0.00	92.24 ± 1.68	90.89 ± 1.17	74.57 ± 0.50	67.07 ± 3.81	29.78 ± 1.92	91.67 ± 2.31	60.28 ± 1.51	78.77 ± 20.77	1.00
	128 D	96.00 ± 0.00	87.33 ± 1.15	99.00 ± 0.00	93.89 ± 0.19	93.00 ± 0.58	76.68 ± 0.18	74.84 ± 2.23	29.79 ± 0.50	93.67 ± 0.58	63.49 ± 0.34	80.77 ± 20.47	1.00
	256 D	95.00 ± 0.00	91.00 ± 0.00	99.00 ± 0.00	93.56 ± 0.19	93.11 ± 0.19	76.93 ± 0.23	75.13 ± 0.84	29.75 ± 0.73	95.33 ± 0.58	67.89 ± 1.34	81.67 ± 20.28	1.00
100 full (F)	16 F	90.70 ± 1.07	83.00 ± 2.65	96.00 ± 3.00	91.22 ± 2.94	87.94 ± 0.54	73.07 ± 0.93	49.41 ± 2.04	24.17 ± 4.22	82.33 ± 2.08	58.18 ± 0.44	73.60 ± 22.23	1.00
	32 F	95.33 ± 1.53	85.00 ± 1.00	99.00 ± 0.00	93.50 ± 0.22	91.44 ± 0.84	75.00 ± 0.19	65.09 ± 2.23	30.20 ± 0.81	91.67 ± 0.58	60.92 ± 0.26	78.72 ± 20.72	0.99
	64 F	97.00 ± 0.00	85.67 ± 1.53	99.00 ± 0.00	93.78 ± 0.19	92.56 ± 0.19	76.01 ± 0.13	73.96 ± 0.89	29.46 ± 0.21	94.33 ± 1.53	64.07 ± 0.37	80.58 ± 20.59	0.97
	128 F	96.33 ± 0.58	90.33 ± 0.58	99.00 ± 0.00	93.00 ± 0.00	93.89 ± 0.19	77.04 ± 0.30	77.33 ± 1.01	29.49 ± 0.35	96.00 ± 0.00	68.76 ± 0.25	82.12 ± 20.35	0.88
	256 F	96.33 ± 0.58	88.67 ± 0.58	99.00 ± 0.00	93.67 ± 0.00	94.89 ± 0.19	78.16 ± 0.18	73.40 ± 0.12	26.86 ± 0.68	96.00 ± 0.00	69.47 ± 0.23	81.64 ± 21.15	0.50
500 diagonal (D)	16 D	54.44 ± 6.87	46.00 ± 2.65	82.21 ± 3.59	73.38 ± 14.97	80.13 ± 3.68	57.42 ± 5.29	18.33 ± 1.33	12.19 ± 0.30	72.67 ± 6.03	55.79 ± 0.20	55.26 ± 23.92	1.00
	32 D	58.08 ± 11.52	45.00 ± 6.08	82.06 ± 1.69	78.62 ± 11.23	85.57 ± 1.48	59.19 ± 3.70	22.76 ± 3.95	13.15 ± 1.94	75.33 ± 4.04	56.07 ± 0.52	57.58 ± 24.33	1.00
	64 D	69.21 ± 2.03	50.67 ± 9.29	88.33 ± 4.04	91.11 ± 0.38	88.78 ± 0.38	67.71 ± 2.59	34.79 ± 6.86	13.80 ± 1.95	77.67 ± 2.31	56.78 ± 0.73	63.89 ± 24.67	1.00
	128 D	79.77 ± 0.37	78.00 ± 3.00	95.89 ± 2.83	91.89 ± 1.39	88.67 ± 0.00	70.27 ± 1.73	46.64 ± 10.58	14.63 ± 0.25	81.00 ± 5.00	56.88 ± 0.55	70.36 ± 24.21	1.00
	256 D	93.83 ± 2.52	85.33 ± 0.58	99.00 ± 0.00	93.78 ± 0.19	90.56 ± 0.38	73.25 ± 1.86	51.14 ± 3.86	22.93 ± 3.86	87.33 ± 2.31	58.48 ± 0.20	75.56 ± 23.56	1.00
500 full (F)	16 F	54.30 ± 1.13	35.67 ± 4.62	77.67 ± 0.58	91.00 ± 0.00	90.56 ± 0.19	67.63 ± 0.45	48.81 ± 0.35	14.70 ± 0.65	79.00 ± 1.00	57.66 ± 0.19	61.70 ± 23.68	1.00
	32 F	75.10 ± 4.92	48.00 ± 3.61	91.67 ± 1.53	91.56 ± 0.19	91.56 ± 0.38	72.03 ± 0.15	52.63 ± 0.86	13.93 ± 0.02	81.67 ± 1.53	58.50 ± 0.20	67.66 ± 24.07	0.99
	64 F	96.94 ± 0.42	82.67 ± 0.58	93.33 ± 0.58	93.89 ± 0.69	90.67 ± 0.00	75.99 ± 0.64	55.63 ± 1.07	14.74 ± 0.27	86.33 ± 0.58	59.43 ± 0.05	74.96 ± 24.62	0.97
	128 F	97.67 ± 0.58	83.67 ± 1.53	98.00 ± 0.00	93.56 ± 0.19	92.00 ± 0.00	75.80 ± 0.16	66.19 ± 0.81	28.67 ± 0.49	93.00 ± 0.00	61.53 ± 0.13	79.01 ± 21.15	0.88
	256 F	98.00 ± 0.00	88.00 ± 1.00	99.00 ± 0.00	93.78 ± 0.19	93.00 ± 0.88	76.33 ± 0.29	74.60 ± 1.21	27.82 ± 0.42	95.33 ± 0.58	63.70 ± 0.14	80.96 ± 21.26	0.50

Table 6. **Absolute** In-Distribution ROUGE-1 scores for various tasks and methods

where the tolerance threshold τ is set to 0.001. Due to the slow per-iteration computation times of the primary JD-Full algorithm, which quickly reaches an approximate optimum but then has a long tail of convergence for final digits of precision, we devised an alternative eigenvalue iteration algorithm (Appendix C.2) optimized for GPU acceleration. Our analysis indicates that adherence to this convergence criterion does not significantly alter the results.

I.12. Out-of-distribution Performance (LoRA-hub)

For completeness, we incorporate results using the protocol of LoRA-hub [9]. That is, 100 LoRA-adapters are sampled, independent of the evaluation task, representing a measure of out-of-distribution performance. This also means that each result on a task is averaged across all 100 LoRA-adapters (as there is no *a priori* LoRA-to-task mapping). These results were obtained without normalizing the LoRA-adapters before applying the JD algorithms, a step we later identified as beneficial. We present performance comparison in Table 15. Table 14 presents the average agreement between uncompressed and compressed LoRA across 10 evaluation tasks. Results per task for JD-diagonal and JD-full are shown in Table 16 and Table 17, respectively.

From these tables, we find that the JD algorithms successfully maintain performance in this out-of-distribution context.

Model Type	Method Type	Tasks										Average	Compression
		task039	task190	task280	task290	task391	task442	task620	task1342	task1391	task1598		
	base	0.00 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.01	0.00
	lora	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.00
SVD	SVD 2	0.98 ± 0.03	1.07 ± 0.02	1.00 ± 0.00	0.99 ± 0.01	0.98 ± 0.01	0.98 ± 0.03	0.94 ± 0.01	1.03 ± 0.17	1.00 ± 0.01	0.15 ± 0.29	0.91 ± 0.28	0.88
	SVD 4	0.99 ± 0.04	1.04 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.01 ± 0.02	1.11 ± 0.00	0.97 ± 0.02	0.99 ± 0.13	0.99 ± 0.01	0.90 ± 0.17	1.00 ± 0.08	0.75
	SVD 8	1.00 ± 0.00	1.02 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.01	1.02 ± 0.05	1.00 ± 0.00	1.00 ± 0.00	1.01 ± 0.01	1.00 ± 0.00	1.00 ± 0.02	0.50
	SVD 16	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.00
10 diagonal (D)	16 D	1.02 ± 0.01	1.01 ± 0.01	1.00 ± 0.00	1.01 ± 0.02	0.96 ± 0.01	1.11 ± 0.11	0.89 ± 0.03	1.19 ± 0.04	0.99 ± 0.02	0.33 ± 0.58	0.95 ± 0.27	1.00
	32 D	1.01 ± 0.01	1.05 ± 0.01	1.00 ± 0.00	0.98 ± 0.01	1.02 ± 0.02	1.11 ± 0.00	0.93 ± 0.01	1.10 ± 0.04	1.00 ± 0.01	0.67 ± 0.58	0.98 ± 0.19	1.00
	64 D	1.00 ± 0.00	1.03 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.02 ± 0.01	1.11 ± 0.00	0.99 ± 0.01	1.00 ± 0.00	1.01 ± 0.00	0.67 ± 0.58	0.98 ± 0.19	1.00
	128 D	1.00 ± 0.00	1.01 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00
	256 D	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00
10 full (F)	16 F	1.02 ± 0.00	1.06 ± 0.01	1.00 ± 0.00	1.00 ± 0.01	0.97 ± 0.03	1.15 ± 0.06	0.92 ± 0.02	1.17 ± 0.04	1.01 ± 0.01	0.67 ± 0.58	1.00 ± 0.20	1.00
	32 F	1.02 ± 0.01	1.04 ± 0.01	1.00 ± 0.00	0.98 ± 0.01	1.00 ± 0.01	1.11 ± 0.00	0.92 ± 0.01	1.02 ± 0.04	1.00 ± 0.01	1.00 ± 0.00	1.01 ± 0.05	0.99
	64 F	1.00 ± 0.00	1.03 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.01 ± 0.01	1.07 ± 0.06	1.01 ± 0.01	1.00 ± 0.00	1.01 ± 0.00	1.00 ± 0.00	1.01 ± 0.03	0.97
	128 F	1.00 ± 0.00	1.01 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.88
	256 F	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.50
50 diagonal (D)	16 D	0.91 ± 0.06	0.98 ± 0.01	1.00 ± 0.00	0.91 ± 0.09	0.78 ± 0.05	0.89 ± 0.29	0.34 ± 0.06	0.50 ± 0.45	0.86 ± 0.07	0.00 ± 0.00	0.72 ± 0.35	1.00
	32 D	1.00 ± 0.02	1.02 ± 0.02	1.00 ± 0.00	1.00 ± 0.01	0.90 ± 0.03	0.85 ± 0.42	0.56 ± 0.04	0.98 ± 0.23	0.98 ± 0.01	0.00 ± 0.00	0.83 ± 0.34	1.00
	64 D	1.02 ± 0.00	1.05 ± 0.02	1.00 ± 0.00	1.00 ± 0.01	0.95 ± 0.02	1.15 ± 0.17	0.81 ± 0.03	1.14 ± 0.00	1.01 ± 0.01	0.00 ± 0.00	0.91 ± 0.33	1.00
	128 D	1.01 ± 0.01	1.08 ± 0.01	1.00 ± 0.00	1.00 ± 0.01	0.95 ± 0.02	1.04 ± 0.06	0.92 ± 0.03	1.21 ± 0.07	1.00 ± 0.00	0.67 ± 0.58	0.99 ± 0.20	1.00
	256 D	1.01 ± 0.01	1.03 ± 0.01	1.00 ± 0.00	1.00 ± 0.01	1.01 ± 0.02	1.11 ± 0.00	0.95 ± 0.04	1.02 ± 0.04	1.00 ± 0.01	1.00 ± 0.00	1.01 ± 0.04	1.00
50 full (F)	16 F	0.96 ± 0.05	1.00 ± 0.01	1.00 ± 0.01	0.95 ± 0.04	0.87 ± 0.01	1.04 ± 0.06	0.31 ± 0.08	0.98 ± 0.23	0.97 ± 0.02	0.00 ± 0.00	0.81 ± 0.35	1.00
	32 F	1.02 ± 0.00	1.00 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	0.92 ± 0.03	1.15 ± 0.06	0.73 ± 0.04	1.17 ± 0.04	1.01 ± 0.02	0.00 ± 0.00	0.90 ± 0.33	0.99
	64 F	1.02 ± 0.01	1.06 ± 0.02	1.00 ± 0.00	1.00 ± 0.01	0.96 ± 0.02	1.22 ± 0.00	0.94 ± 0.01	1.17 ± 0.04	1.00 ± 0.01	0.00 ± 0.00	0.94 ± 0.33	0.97
	128 F	1.02 ± 0.00	1.06 ± 0.01	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.02	1.15 ± 0.06	0.92 ± 0.01	1.10 ± 0.08	1.00 ± 0.01	1.00 ± 0.00	1.02 ± 0.07	0.88
	256 F	1.00 ± 0.00	1.02 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.01	1.04 ± 0.06	0.99 ± 0.00	1.00 ± 0.00	1.01 ± 0.00	1.00 ± 0.00	1.01 ± 0.02	0.50
100 diagonal (D)	16 D	0.54 ± 0.16	0.89 ± 0.06	0.90 ± 0.04	0.89 ± 0.05	0.42 ± 0.08	0.44 ± 0.00	0.08 ± 0.02	0.00 ± 0.00	0.76 ± 0.05	0.00 ± 0.00	0.49 ± 0.36	1.00
	32 D	0.85 ± 0.15	0.98 ± 0.01	1.00 ± 0.00	0.86 ± 0.13	0.70 ± 0.14	0.74 ± 0.28	0.28 ± 0.07	0.48 ± 0.55	0.91 ± 0.02	0.00 ± 0.00	0.68 ± 0.36	1.00
	64 D	1.00 ± 0.04	1.01 ± 0.01	1.00 ± 0.00	0.98 ± 0.02	0.88 ± 0.04	1.07 ± 0.06	0.58 ± 0.09	1.10 ± 0.04	0.96 ± 0.02	0.00 ± 0.00	0.86 ± 0.32	1.00
	128 D	1.01 ± 0.00	1.02 ± 0.01	1.00 ± 0.00	1.01 ± 0.01	0.95 ± 0.02	1.11 ± 0.00	0.81 ± 0.06	1.21 ± 0.00	0.99 ± 0.01	0.00 ± 0.00	0.91 ± 0.33	1.00
	256 D	1.00 ± 0.00	1.06 ± 0.00	1.00 ± 0.00	1.00 ± 0.01	0.96 ± 0.01	1.11 ± 0.11	0.92 ± 0.02	1.21 ± 0.07	1.00 ± 0.01	0.00 ± 0.00	0.93 ± 0.33	1.00
100 full (F)	16 F	0.85 ± 0.03	0.97 ± 0.03	0.97 ± 0.03	0.95 ± 0.06	0.80 ± 0.02	0.81 ± 0.17	0.29 ± 0.04	0.60 ± 0.34	0.87 ± 0.02	0.00 ± 0.00	0.71 ± 0.33	1.00
	32 F	0.99 ± 0.02	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.01	0.90 ± 0.03	1.04 ± 0.06	0.55 ± 0.04	1.07 ± 0.07	0.96 ± 0.01	0.00 ± 0.00	0.85 ± 0.32	0.99
	64 F	1.02 ± 0.00	1.00 ± 0.02	1.00 ± 0.00	1.00 ± 0.01	0.94 ± 0.01	1.04 ± 0.06	0.78 ± 0.01	1.14 ± 0.00	0.99 ± 0.02	0.00 ± 0.00	0.89 ± 0.31	0.97
	128 F	1.01 ± 0.01	1.05 ± 0.01	1.00 ± 0.00	0.98 ± 0.00	0.98 ± 0.01	1.15 ± 0.06	0.94 ± 0.01	1.21 ± 0.00	1.01 ± 0.00	0.33 ± 0.58	0.97 ± 0.28	0.88
	256 F	1.01 ± 0.01	1.03 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.02 ± 0.01	1.19 ± 0.06	0.93 ± 0.01	1.02 ± 0.04	1.01 ± 0.00	1.00 ± 0.00	1.02 ± 0.06	0.50
500 diagonal (D)	16 D	0.22 ± 0.10	0.53 ± 0.03	0.81 ± 0.05	0.33 ± 0.49	0.70 ± 0.03	0.15 ± 0.17	0.03 ± 0.01	0.00 ± 0.00	0.76 ± 0.06	0.00 ± 0.00	0.35 ± 0.34	1.00
	32 D	0.27 ± 0.18	0.52 ± 0.07	0.82 ± 0.02	0.49 ± 0.37	0.75 ± 0.01	0.22 ± 0.11	0.05 ± 0.05	0.02 ± 0.04	0.79 ± 0.04	0.00 ± 0.00	0.39 ± 0.33	1.00
	64 D	0.40 ± 0.04	0.59 ± 0.11	0.89 ± 0.04	0.91 ± 0.01	0.80 ± 0.01	0.48 ± 0.06	0.13 ± 0.04	0.05 ± 0.08	0.82 ± 0.02	0.00 ± 0.00	0.51 ± 0.34	1.00
	128 D	0.61 ± 0.04	0.91 ± 0.03	0.97 ± 0.03	0.93 ± 0.05	0.80 ± 0.00	0.74 ± 0.17	0.22 ± 0.11	0.12 ± 0.08	0.85 ± 0.05	0.00 ± 0.00	0.61 ± 0.36	1.00
	256 D	0.95 ± 0.02	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.01	0.86 ± 0.01	0.85 ± 0.28	0.28 ± 0.06	0.55 ± 0.39	0.92 ± 0.02	0.00 ± 0.00	0.74 ± 0.36	1.00
500 full (F)	16 F	0.21 ± 0.02	0.41 ± 0.05	0.78 ± 0.01	0.90 ± 0.00	0.86 ± 0.01	0.59 ± 0.06	0.21 ± 0.01	0.12 ± 0.04	0.83 ± 0.01	0.00 ± 0.00	0.49 ± 0.33	1.00
	32 F	0.54 ± 0.08	0.56 ± 0.04	0.93 ± 0.02	0.92 ± 0.01	0.90 ± 0.01	0.63 ± 0.13	0.26 ± 0.02	0.14 ± 0.00	0.86 ± 0.02	0.00 ± 0.00	0.57 ± 0.33	0.99
	64 F	0.99 ± 0.03	0.96 ± 0.01	0.94 ± 0.01	1.01 ± 0.03	0.87 ± 0.00	1.04 ± 0.17	0.36 ± 0.00	0.14 ± 0.00	0.91 ± 0.01	0.00 ± 0.00	0.72 ± 0.38	0.97
	128 F	1.02 ± 0.01	0.97 ± 0.02	0.99 ± 0.00	1.00 ± 0.01	0.92 ± 0.00	1.15 ± 0.06	0.61 ± 0.01	1.07 ± 0.00	0.98 ± 0.00	0.00 ± 0.00	0.87 ± 0.33	0.88
	256 F	1.03 ± 0.00	1.02 ± 0.01	1.00 ± 0.00	1.00 ± 0.01	0.95 ± 0.03	1.00 ± 0.00	0.78 ± 0.01	1.07 ± 0.00	1.00 ± 0.01	0.00 ± 0.00	0.89 ± 0.31	0.50

Table 7. **Relative** In-Distribution exact match scores for various tasks and methods

Model Type	Method Type	Tasks										Average	Compression
		task039	task190	task280	task290	task391	task442	task620	task1342	task1391	task1598		
	base	0.00 ± 0.00	0.00 ± 0.00	2.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.20 ± 0.60	0.00
	lora	95.00 ± 0.00	86.00 ± 0.00	99.00 ± 0.00	81.00 ± 0.00	83.00 ± 0.00	9.00 ± 0.00	69.00 ± 0.00	14.00 ± 0.00	95.00 ± 0.00	1.00 ± 0.00	63.20 ± 37.30	0.00
SVD	SVD 2	93.09 ± 2.88	91.98 ± 1.99	99.09 ± 0.18	80.32 ± 0.42	81.68 ± 1.05	8.83 ± 0.30	65.14 ± 0.56	14.36 ± 2.44	95.06 ± 1.35	0.15 ± 0.29	62.97 ± 37.57	0.88
	SVD 4	93.94 ± 3.73	89.15 ± 0.71	99.05 ± 0.09	80.95 ± 0.09	83.97 ± 1.89	10.00 ± 0.00	66.68 ± 1.11	13.83 ± 1.76	93.98 ± 0.77	0.90 ± 0.17	63.24 ± 37.28	0.75
	SVD 8	95.00 ± 0.00	87.37 ± 0.57	99.05 ± 0.09	80.96 ± 0.08	83.09 ± 1.13	9.22 ± 0.44	69.00 ± 0.00	14.00 ± 0.00	95.51 ± 1.09	1.00 ± 0.00	66.69 ± 36.13	0.50
	SVD 16	95.00 ± 0.00	86.00 ± 0.00	99.00 ± 0.00	81.00 ± 0.00	83.00 ± 0.00	9.00 ± 0.00	68.50 ± 0.55	14.00 ± 0.00	95.00 ± 0.00	1.00 ± 0.00	61.68 ± 38.02	0.00
10 diagonal (D)	16 D	96.67 ± 0.58	87.00 ± 1.00	99.00 ± 0.00	82.00 ± 2.00	79.33 ± 1.15	10.00 ± 1.00	61.33 ± 2.08	16.67 ± 0.58	94.00 ± 1.73	0.33 ± 0.58	62.63 ± 37.36	1.00
	32 D	95.67 ± 0.58	90.00 ± 1.00	99.00 ± 0.00	79.00 ± 1.00	84.67 ± 1.53	10.00 ± 0.00	64.00 ± 1.00	15.33 ± 0.58	94.67 ± 0.58	0.67 ± 0.58	63.30 ± 37.77	1.00
	64 D	95.00 ± 0.00	88.33 ± 0.58	99.00 ± 0.00	81.00 ± 0.00	84.33 ± 1.15	10.00 ± 0.00	68.33 ± 0.58	14.00 ± 0.00	96.00 ± 0.00	0.67 ± 0.58	63.67 ± 37.99	1.00
	128 D	95.00 ± 0.00	86.67 ± 0.58	99.00 ± 0.00	81.00 ± 0.00	83.00 ± 0.00	9.00 ± 0.00	69.33 ± 0.58	14.00 ± 0.00	95.00 ± 0.00	1.00 ± 0.00	63.30 ± 37.82	1.00
	256 D	95.00 ± 0.00	86.00 ± 0.00	99.00 ± 0.00	81.00 ± 0.00	83.00 ± 0.00	9.00 ± 0.00	69.00 ± 0.00	14.00 ± 0.00	95.00 ± 0.00	1.00 ± 0.00	63.20 ± 37.77	1.00
10 full (F)	16 F	97.00 ± 0.00	91.00 ± 1.00	99.00 ± 0.00	80.67 ± 0.58	80.67 ± 2.08	10.33 ± 0.58	63.33 ± 1.53	16.33 ± 0.58	96.00 ± 1.00	0.67 ± 0.58	63.50 ± 37.77	1.00
	32 F	96.67 ± 0.58	89.33 ± 0.58	99.00 ± 0.00	79.67 ± 0.58	83.33 ± 0.58	10.00 ± 0.00	63.33 ± 0.58	14.33 ± 0.58	94.67 ± 0.58	1.00 ± 0.00	63.13 ± 37.84	0.99
	64 F	95.00 ± 0.00	88.67 ± 0.58	99.00 ± 0.00	81.00 ± 0.00	83.67 ± 1.15	9.67 ± 0.58	69.67 ± 0.58	14.00 ± 0.00	96.00 ± 0.00	1.00 ± 0.00	63.77 ± 37.99	0.97
	128 F	95.00 ± 0.00	86.67 ± 0.58	99.00 ± 0.00	81.00 ± 0.00	83.00 ± 0.00	9.00 ± 0.00	69.00 ± 0.00	14.00 ± 0.00	95.00 ± 0.00	1.00 ± 0.00	63.27 ± 37.82	0.88
	256 F	95.00 ± 0.00	86.00 ± 0.00	99.00 ± 0.00	81.00 ± 0.00	83.00 ± 0.00	9.00 ± 0.00	68.67 ± 0.58	14.00 ± 0.00	95.00 ± 0.00	1.00 ± 0.00	63.17 ± 37.77	0.50
50 diagonal (D)	16 D	86.33 ± 6.11	84.67 ± 1.15	99.00 ± 0.00	73.67 ± 7.37	64.33 ± 4.04	8.00 ± 2.65	23.67 ± 4.04	7.00 ± 6.24	82.00 ± 6.56	0.00 ± 0.00	52.87 ± 37.48	1.00
	32 D	95.33 ± 2.08	87.33 ± 2.08	99.00 ± 0.00	80.67 ± 1.15	74.33 ± 2.08	7.67 ± 3.79	38.33 ± 3.06	13.67 ± 3.21	93.33 ± 1.15	0.00 ± 0.00	58.97 ± 38.38	1.00
	64 D	97.00 ± 0.00	90.33 ± 1.53	99.00 ± 0.00	81.33 ± 0.58	79.00 ± 1.73	10.33 ± 1.53	56.00 ± 1.73	16.00 ± 0.00	95.67 ± 0.58	0.00 ± 0.00	62.47 ± 37.87	1.00
	128 D	96.33 ± 0.58	92.67 ± 0.58	99.00 ± 0.00	80.67 ± 0.58	79.00 ± 1.73	9.33 ± 0.58	63.67 ± 2.08	17.00 ± 1.00	95.00 ± 0.00	0.67 ± 0.58	63.33 ± 37.74	1.00
	256 D	95.67 ± 0.58	88.33 ± 0.58	99.00 ± 0.00	80.67 ± 0.58	84.00 ± 2.00	10.00 ± 0.00	65.67 ± 2.52	14.33 ± 0.58	95.33 ± 0.58	1.00 ± 0.00	63.40 ± 37.84	1.00
50 full (F)	16 F	91.67 ± 5.03	85.67 ± 1.15	98.67 ± 0.58	77.33 ± 2.89	72.33 ± 1.15	9.33 ± 0.58	21.33 ± 5.86	13.67 ± 3.21	92.33 ± 1.53	0.00 ± 0.00	56.23 ± 38.57	1.00
	32 F	97.00 ± 0.00	85.67 ± 1.53	99.00 ± 0.00	81.00 ± 0.00	76.67 ± 2.08	10.33 ± 0.58	50.33 ± 3.06	16.33 ± 0.58	95.67 ± 1.53	0.00 ± 0.00	61.20 ± 37.52	0.99
	64 F	96.67 ± 0.58	91.00 ± 2.00	99.00 ± 0.00	80.67 ± 0.58	79.67 ± 1.53	11.00 ± 0.00	64.67 ± 0.58	16.33 ± 0.58	95.33 ± 0.58	0.00 ± 0.00	63.43 ± 37.65	0.97
	128 F	97.00 ± 0.00	91.00 ± 1.00	99.00 ± 0.00	80.00 ± 0.00	82.33 ± 1.53	10.33 ± 0.58	63.67 ± 0.58	15.33 ± 1.15	95.00 ± 1.00	1.00 ± 0.00	63.47 ± 37.80	0.88
	256 F	95.00 ± 0.00	88.00 ± 0.00	99.00 ± 0.00	81.00 ± 0.00	83.33 ± 0.58	9.33 ± 0.58	68.00 ± 0.00	14.00 ± 0.00	96.00 ± 0.00	1.00 ± 0.00	63.47 ± 37.95	0.50
100 diagonal (D)	16 D	51.33 ± 15.04	76.67 ± 4.93	89.33 ± 3.79	72.33 ± 4.04	35.00 ± 7.00	4.00 ± 0.00	5.67 ± 1.53	0.00 ± 0.00	72.33 ± 4.51	0.00 ± 0.00	40.67 ± 35.13	1.00
	32 D	80.67 ± 14.64	84.00 ± 1.00	99.00 ± 0.00	70.00 ± 10.54	58.33 ± 11.24	6.67 ± 2.52	19.33 ± 4.51	6.67 ± 7.64	86.00 ± 1.73	0.00 ± 0.00	51.07 ± 37.79	1.00
	64 D	94.67 ± 4.04	86.67 ± 0.58	99.00 ± 0.00	79.00 ± 1.73	72.67 ± 3.51	9.67 ± 0.58	40.00 ± 6.24	15.33 ± 0.58	91.67 ± 2.31	0.00 ± 0.00	58.87 ± 37.41	1.00
	128 D	96.00 ± 0.00	87.33 ± 1.15	99.00 ± 0.00	81.67 ± 0.58	79.00 ± 1.73	10.00 ± 0.00	55.67 ± 4.16	17.00 ± 0.00	93.67 ± 0.58	0.00 ± 0.00	61.93 ± 37.33	1.00
	256 D	95.00 ± 0.00	91.00 ± 0.00	99.00 ± 0.00	80.67 ± 0.58	79.33 ± 0.58	10.00 ± 1.00	63.33 ± 1.15	17.00 ± 1.00	95.33 ± 0.58	0.00 ± 0.00	63.07 ± 37.54	1.00
100 full (F)	16 F	81.00 ± 2.65	83.00 ± 2.65	96.00 ± 3.00	77.33 ± 4.62	66.67 ± 1.53	7.33 ± 1.53	20.00 ± 3.00	8.33 ± 4.73	82.33 ± 2.08	0.00 ± 0.00	52.20 ± 36.95	1.00
	32 F	94.33 ± 1.53	85.00 ± 1.00	99.00 ± 0.00	81.33 ± 0.58	74.33 ± 2.52	9.33 ± 0.58	38.00 ± 2.65	15.00 ± 1.00	91.67 ± 0.58	0.00 ± 0.00	58.80 ± 37.59	0.99
	64 F	97.00 ± 0.00	85.67 ± 1.53	99.00 ± 0.00	81.33 ± 0.58	77.67 ± 0.58	9.33 ± 0.58	54.00 ± 1.00	16.00 ± 0.00	94.33 ± 1.53	0.00 ± 0.00	61.43 ± 37.53	0.97
	128 F	96.33 ± 0.58	90.33 ± 0.58	99.00 ± 0.00	79.00 ± 0.00	81.67 ± 0.58	10.33 ± 0.58	64.67 ± 0.58	17.00 ± 0.00	96.00 ± 0.00	0.33 ± 0.58	63.47 ± 37.60	0.88
	256 F	96.33 ± 0.58	88.67 ± 0.58	99.00 ± 0.00	81.00 ± 0.00	84.67 ± 0.58	10.67 ± 0.58	64.33 ± 0.58	14.33 ± 0.58	96.00 ± 0.00	1.00 ± 0.00	63.60 ± 37.92	0.50
500 diagonal (D)	16 D	20.67 ± 9.81	46.00 ± 2.65	80.67 ± 4.93	26.33 ± 39.55	58.00 ± 2.65	1.33 ± 1.53	2.00 ± 1.00	0.00 ± 0.00	72.67 ± 6.03	0.00 ± 0.00	30.77 ± 32.30	1.00
	32 D	25.67 ± 16.74	45.00 ± 6.08	81.67 ± 2.08	39.67 ± 29.70	62.00 ± 1.00	2.00 ± 1.00	3.33 ± 3.21	0.33 ± 0.58	75.33 ± 4.04	0.00 ± 0.00	33.50 ± 32.26	1.00
	64 D	38.00 ± 3.46	50.67 ± 9.29	88.33 ± 4.04	73.33 ± 1.15	66.33 ± 1.15	4.33 ± 0.58	8.67 ± 2.89	0.67 ± 1.15	77.67 ± 2.31	0.00 ± 0.00	40.80 ± 33.98	1.00
	128 D	57.67 ± 3.79	78.00 ± 3.00	95.67 ± 3.21	75.67 ± 4.16	66.00 ± 0.00	6.67 ± 1.53	15.33 ± 7.37	1.67 ± 1.15	81.00 ± 5.00	0.00 ± 0.00	47.77 ± 36.33	1.00
	256 D	90.67 ± 2.08	85.33 ± 0.58	99.00 ± 0.00	81.33 ± 0.58	71.67 ± 1.15	7.67 ± 2.52	19.33 ± 4.04	7.67 ± 5.51	87.33 ± 2.31	0.00 ± 0.00	55.00 ± 39.35	1.00
500 full (F)	16 F	19.67 ± 2.08	35.67 ± 4.62	77.67 ± 0.58	73.00 ± 0.00	71.67 ± 0.58	5.33 ± 0.58	14.33 ± 0.58	1.67 ± 0.58	79.00 ± 1.00	0.00 ± 0.00	37.80 ± 32.74	1.00
	32 F	51.67 ± 7.51	48.00 ± 3.61	91.67 ± 1.53	74.67 ± 0.58	74.67 ± 1.15	5.67 ± 1.15	17.67 ± 1.15	2.00 ± 0.00	81.67 ± 1.53	0.00 ± 0.00	44.77 ± 34.57	0.99
	64 F	94.00 ± 2.65	82.67 ± 0.58	93.33 ± 0.58	81.67 ± 2.08	72.00 ± 0.00	9.33 ± 1.53	25.00 ± 0.00	2.00 ± 0.00	86.33 ± 0.58	0.00 ± 0.00	54.63 ± 38.82	0.97
	128 F	96.67 ± 0.58	83.67 ± 1.53	98.00 ± 0.00	80.67 ± 0.58	76.00 ± 0.00	10.33 ± 0.58	42.00 ± 1.00	15.00 ± 0.00	93.00 ± 0.00	0.00 ± 0.00	59.53 ± 37.42	0.88
	256 F	98.00 ± 0.00	88.00 ± 1.00	99.00 ± 0.00	81.33 ± 0.58	79.00 ± 2.65	9.00 ± 0.00	54.00 ± 1.00	15.00 ± 0.00	95.33 ± 0.58	0.00 ± 0.00	61.87 ± 38.12	0.50

Table 8. **Absolute** In-Distribution exact match scores for various tasks and methods

Model Type	Method Type	Tasks										Average	Compression
		task039	task190	task280	task290	task391	task442	task620	task1342	task1391	task1598		
	base	8.59 ± 0.08	9.15 ± 0.00	2.55 ± 0.00	2.88 ± 0.00	2.34 ± 0.00	3.46 ± 0.04	6.40 ± 0.18	5.55 ± 0.00	8.60 ± 0.00	2.67 ± 0.00	5.22 ± 2.66	0.00
	lora	0.36 ± 0.01	0.17 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.11 ± 0.00	0.76 ± 0.02	1.17 ± 0.07	1.94 ± 0.00	0.16 ± 0.00	0.85 ± 0.00	0.56 ± 0.59	0.00
SVD	SVD 2	0.32 ± 0.01	0.15 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.76 ± 0.02	1.13 ± 0.08	1.94 ± 0.00	0.13 ± 0.00	0.97 ± 0.00	0.56 ± 0.60	0.88
	SVD 4	0.33 ± 0.01	0.16 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.11 ± 0.00	0.76 ± 0.02	1.14 ± 0.08	1.94 ± 0.00	0.14 ± 0.00	0.86 ± 0.00	0.56 ± 0.59	0.75
	SVD 8	0.35 ± 0.01	0.17 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.11 ± 0.00	0.77 ± 0.02	1.16 ± 0.07	1.94 ± 0.00	0.15 ± 0.00	0.84 ± 0.00	0.50 ± 0.57	0.50
	SVD 16	0.36 ± 0.01	0.17 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.11 ± 0.00	0.76 ± 0.02	1.14 ± 0.06	1.94 ± 0.00	0.16 ± 0.00	0.85 ± 0.00	0.56 ± 0.59	0.00
10 diagonal (D)	16 D	0.33 ± 0.01	0.15 ± 0.01	0.01 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.76 ± 0.03	1.13 ± 0.08	1.95 ± 0.01	0.14 ± 0.00	1.00 ± 0.02	0.57 ± 0.61	1.00
	32 D	0.33 ± 0.01	0.16 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.75 ± 0.02	1.11 ± 0.07	1.93 ± 0.00	0.14 ± 0.01	0.88 ± 0.00	0.55 ± 0.60	1.00
	64 D	0.35 ± 0.01	0.17 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.11 ± 0.00	0.75 ± 0.02	1.11 ± 0.07	1.94 ± 0.00	0.15 ± 0.00	0.84 ± 0.00	0.55 ± 0.59	1.00
	128 D	0.35 ± 0.01	0.17 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.11 ± 0.00	0.75 ± 0.02	1.11 ± 0.07	1.94 ± 0.00	0.16 ± 0.00	0.84 ± 0.00	0.56 ± 0.59	1.00
	256 D	0.36 ± 0.01	0.17 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.11 ± 0.00	0.75 ± 0.02	1.12 ± 0.07	1.94 ± 0.00	0.16 ± 0.00	0.85 ± 0.00	0.56 ± 0.59	1.00
10 full (F)	16 F	0.33 ± 0.00	0.15 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.76 ± 0.02	1.20 ± 0.02	1.95 ± 0.00	0.13 ± 0.00	0.97 ± 0.00	0.57 ± 0.61	1.00
	32 F	0.33 ± 0.01	0.16 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.75 ± 0.02	1.11 ± 0.07	1.94 ± 0.00	0.14 ± 0.00	0.86 ± 0.00	0.55 ± 0.60	0.99
	64 F	0.34 ± 0.01	0.16 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.11 ± 0.00	0.75 ± 0.02	1.11 ± 0.07	1.94 ± 0.00	0.15 ± 0.00	0.84 ± 0.00	0.55 ± 0.59	0.97
	128 F	0.35 ± 0.01	0.17 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.11 ± 0.00	0.75 ± 0.02	1.12 ± 0.07	1.94 ± 0.00	0.16 ± 0.00	0.84 ± 0.00	0.56 ± 0.59	0.88
	256 F	0.36 ± 0.01	0.17 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.11 ± 0.00	0.75 ± 0.02	1.12 ± 0.07	1.94 ± 0.00	0.16 ± 0.00	0.85 ± 0.00	0.56 ± 0.59	0.50
50 diagonal (D)	16 D	0.61 ± 0.06	0.19 ± 0.02	0.03 ± 0.01	0.29 ± 0.04	0.36 ± 0.04	0.95 ± 0.05	1.73 ± 0.21	2.66 ± 0.22	0.32 ± 0.11	1.98 ± 0.01	0.91 ± 0.88	1.00
	32 D	0.37 ± 0.02	0.16 ± 0.00	0.01 ± 0.00	0.19 ± 0.03	0.18 ± 0.01	0.85 ± 0.05	1.37 ± 0.14	2.12 ± 0.05	0.16 ± 0.00	1.65 ± 0.03	0.71 ± 0.73	1.00
	64 D	0.33 ± 0.02	0.15 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.79 ± 0.02	1.12 ± 0.08	1.97 ± 0.01	0.13 ± 0.01	1.13 ± 0.03	0.59 ± 0.63	1.00
	128 D	0.33 ± 0.01	0.15 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.76 ± 0.03	1.10 ± 0.05	1.93 ± 0.01	0.14 ± 0.00	0.93 ± 0.01	0.56 ± 0.60	1.00
	256 D	0.34 ± 0.01	0.16 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.76 ± 0.03	1.11 ± 0.05	1.93 ± 0.00	0.15 ± 0.00	0.85 ± 0.00	0.55 ± 0.59	1.00
50 full (F)	16 F	0.47 ± 0.06	0.17 ± 0.00	0.02 ± 0.00	0.20 ± 0.02	0.19 ± 0.04	0.86 ± 0.03	1.71 ± 0.10	2.20 ± 0.04	0.17 ± 0.01	1.84 ± 0.07	0.78 ± 0.80	1.00
	32 F	0.36 ± 0.02	0.16 ± 0.00	0.01 ± 0.00	0.14 ± 0.00	0.11 ± 0.00	0.80 ± 0.03	1.14 ± 0.08	2.00 ± 0.01	0.14 ± 0.00	1.32 ± 0.02	0.62 ± 0.65	0.99
	64 F	0.33 ± 0.01	0.15 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.77 ± 0.03	1.10 ± 0.06	1.94 ± 0.00	0.13 ± 0.00	1.02 ± 0.00	0.57 ± 0.61	0.97
	128 F	0.33 ± 0.01	0.16 ± 0.00	0.00 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.76 ± 0.03	1.11 ± 0.05	1.93 ± 0.00	0.14 ± 0.00	0.87 ± 0.00	0.55 ± 0.60	0.88
	256 F	0.35 ± 0.01	0.16 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.11 ± 0.00	0.76 ± 0.03	1.11 ± 0.05	1.94 ± 0.00	0.15 ± 0.00	0.84 ± 0.00	0.55 ± 0.59	0.50
100 diagonal (D)	16 D	1.69 ± 0.49	0.26 ± 0.04	0.18 ± 0.07	0.34 ± 0.02	1.01 ± 0.20	1.45 ± 0.10	3.59 ± 0.25	3.72 ± 0.72	0.44 ± 0.20	2.37 ± 0.09	1.51 ± 1.32	1.00
	32 D	0.67 ± 0.24	0.18 ± 0.01	0.06 ± 0.05	0.31 ± 0.06	0.35 ± 0.08	1.04 ± 0.15	1.97 ± 0.13	2.88 ± 0.70	0.22 ± 0.01	2.12 ± 0.07	0.98 ± 0.98	1.00
	64 D	0.39 ± 0.06	0.16 ± 0.00	0.01 ± 0.00	0.18 ± 0.02	0.14 ± 0.01	0.86 ± 0.02	1.39 ± 0.07	2.18 ± 0.04	0.17 ± 0.00	1.79 ± 0.02	0.73 ± 0.76	1.00
	128 D	0.32 ± 0.00	0.15 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.79 ± 0.02	1.19 ± 0.02	2.00 ± 0.01	0.14 ± 0.01	1.24 ± 0.04	0.61 ± 0.65	1.00
	256 D	0.32 ± 0.00	0.15 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.77 ± 0.02	1.16 ± 0.00	1.94 ± 0.00	0.13 ± 0.00	0.96 ± 0.01	0.56 ± 0.61	1.00
100 full (F)	16 F	0.66 ± 0.07	0.19 ± 0.01	0.03 ± 0.01	0.25 ± 0.02	0.29 ± 0.02	0.99 ± 0.07	2.50 ± 0.51	2.63 ± 0.03	0.24 ± 0.02	2.21 ± 0.08	1.00 ± 1.01	1.00
	32 F	0.40 ± 0.00	0.17 ± 0.00	0.01 ± 0.00	0.15 ± 0.01	0.13 ± 0.01	0.85 ± 0.02	1.53 ± 0.12	2.17 ± 0.06	0.15 ± 0.01	1.93 ± 0.04	0.75 ± 0.80	0.99
	64 F	0.34 ± 0.01	0.15 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.11 ± 0.00	0.79 ± 0.01	1.23 ± 0.07	1.98 ± 0.01	0.15 ± 0.00	1.26 ± 0.01	0.61 ± 0.65	0.97
	128 F	0.32 ± 0.00	0.15 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.77 ± 0.02	1.16 ± 0.01	1.94 ± 0.00	0.13 ± 0.00	0.99 ± 0.01	0.57 ± 0.61	0.88
	256 F	0.33 ± 0.00	0.16 ± 0.00	0.00 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.76 ± 0.02	1.15 ± 0.01	1.93 ± 0.00	0.14 ± 0.00	0.86 ± 0.00	0.56 ± 0.60	0.50
500 diagonal (D)	16 D	2.95 ± 0.28	0.81 ± 0.25	0.27 ± 0.09	0.67 ± 0.28	0.52 ± 0.07	2.06 ± 0.30	4.85 ± 0.31	3.94 ± 0.42	0.50 ± 0.05	2.50 ± 0.03	1.91 ± 1.57	1.00
	32 D	2.33 ± 0.30	0.69 ± 0.16	0.24 ± 0.05	0.50 ± 0.16	0.37 ± 0.07	1.86 ± 0.25	4.73 ± 0.35	3.81 ± 0.59	0.39 ± 0.04	2.46 ± 0.05	1.74 ± 1.55	1.00
	64 D	1.67 ± 0.18	0.49 ± 0.16	0.13 ± 0.04	0.29 ± 0.02	0.23 ± 0.02	1.32 ± 0.28	3.99 ± 0.36	3.41 ± 0.28	0.32 ± 0.05	2.35 ± 0.11	1.42 ± 1.37	1.00
	128 D	1.12 ± 0.02	0.24 ± 0.03	0.04 ± 0.03	0.21 ± 0.04	0.22 ± 0.03	1.08 ± 0.06	3.05 ± 0.87	3.09 ± 0.37	0.26 ± 0.03	2.31 ± 0.04	1.16 ± 1.20	1.00
	256 D	0.54 ± 0.03	0.18 ± 0.01	0.01 ± 0.00	0.16 ± 0.01	0.15 ± 0.01	0.92 ± 0.08	2.42 ± 0.14	2.51 ± 0.13	0.19 ± 0.01	2.09 ± 0.02	0.92 ± 0.99	1.00
500 full (F)	16 F	2.14 ± 0.06	0.71 ± 0.04	0.28 ± 0.00	0.27 ± 0.01	0.21 ± 0.00	1.14 ± 0.04	3.06 ± 0.27	2.71 ± 0.01	0.34 ± 0.01	2.21 ± 0.01	1.31 ± 1.08	1.00
	32 F	1.17 ± 0.07	0.46 ± 0.04	0.08 ± 0.04	0.21 ± 0.01	0.17 ± 0.00	0.99 ± 0.04	2.69 ± 0.10	2.47 ± 0.02	0.25 ± 0.02	2.11 ± 0.04	1.06 ± 0.98	0.99
	64 F	0.51 ± 0.03	0.21 ± 0.00	0.02 ± 0.00	0.17 ± 0.01	0.14 ± 0.00	0.88 ± 0.04	2.19 ± 0.14	2.34 ± 0.03	0.20 ± 0.00	1.97 ± 0.02	0.86 ± 0.90	0.97
	128 F	0.39 ± 0.01	0.16 ± 0.00	0.01 ± 0.00	0.13 ± 0.00	0.11 ± 0.00	0.81 ± 0.03	1.42 ± 0.07	2.03 ± 0.01	0.16 ± 0.00	1.71 ± 0.01	0.69 ± 0.73	0.88
	256 F	0.32 ± 0.01	0.15 ± 0.00	0.01 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.77 ± 0.01	1.18 ± 0.04	1.96 ± 0.00	0.14 ± 0.01	1.25 ± 0.00	0.60 ± 0.64	0.50

Table 9. **Absolute** In-Distribution test loss for various tasks and methods

Model Type	Method Type	Tasks										Average	Compression
		task039	task190	task280	task290	task391	task442	task620	task1342	task1391	task1598		
	base	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.10 ± 0.30	0.00
	lora	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	0.00
TIES	10	41.00 ± 0.00	53.67 ± 0.58	44.33 ± 4.04	10.33 ± 0.58	46.33 ± 4.04	1.00 ± 0.00	8.00 ± 0.00	8.00 ± 0.00	76.67 ± 1.15	1.00 ± 0.00	29.03 ± 25.69	0.00
	50	24.00 ± 0.00	38.67 ± 0.58	17.67 ± 4.62	2.00 ± 0.00	56.33 ± 0.58	1.00 ± 0.00	8.00 ± 0.00	8.00 ± 0.00	29.67 ± 2.89	0.00 ± 0.00	18.53 ± 18.07	0.00
	100	22.00 ± 0.00	38.00 ± 0.00	18.67 ± 4.62	1.00 ± 1.73	51.67 ± 4.62	1.00 ± 0.00	8.00 ± 0.00	7.33 ± 0.58	2.00 ± 0.00	0.00 ± 0.00	14.97 ± 17.20	0.00
	500	8.00 ± 0.00	25.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	59.00 ± 0.00	0.00 ± 0.00	3.00 ± 0.00	6.00 ± 0.00	2.00 ± 0.00	0.00 ± 0.00	10.40 ± 18.02	0.00
SVD	SVD 2	88.33 ± 0.65	91.92 ± 0.90	100.00 ± 0.00	97.25 ± 0.45	92.83 ± 0.39	76.50 ± 1.51	66.00 ± 1.41	58.08 ± 1.16	98.67 ± 0.49	5.83 ± 0.94	77.54 ± 27.61	0.88
	SVD 4	93.00 ± 0.00	96.58 ± 0.51	100.00 ± 0.00	100.00 ± 0.00	96.75 ± 0.87	88.83 ± 1.53	90.67 ± 1.23	72.17 ± 0.58	98.67 ± 0.49	16.67 ± 1.78	85.33 ± 24.31	0.75
	SVD 8	98.89 ± 0.60	98.58 ± 0.51	100.00 ± 0.00	100.00 ± 0.00	99.42 ± 0.51	93.44 ± 0.73	97.22 ± 0.44	82.78 ± 1.64	99.00 ± 0.00	60.00 ± 0.87	93.75 ± 11.54	0.50
	SVD 16	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.67 ± 0.50	99.50 ± 0.55	99.67 ± 0.50	100.00 ± 0.00	98.11 ± 0.78	99.69 ± 0.68	0.00
10 diagonal (D)	16 D	83.33 ± 1.53	88.33 ± 0.58	100.00 ± 0.00	97.00 ± 2.00	88.33 ± 1.15	57.00 ± 1.00	48.67 ± 3.21	50.67 ± 4.93	97.67 ± 1.53	5.33 ± 1.15	71.63 ± 29.53	1.00
	32 D	93.00 ± 1.00	95.33 ± 0.58	100.00 ± 0.00	98.00 ± 1.00	93.67 ± 1.53	80.67 ± 2.31	78.67 ± 1.15	68.00 ± 1.73	98.33 ± 0.58	14.67 ± 2.52	82.03 ± 24.99	1.00
	64 D	99.00 ± 0.00	97.00 ± 1.00	100.00 ± 0.00	100.00 ± 0.00	98.00 ± 0.00	90.67 ± 1.53	95.33 ± 1.15	79.67 ± 1.53	99.00 ± 0.00	55.00 ± 4.36	91.37 ± 13.78	1.00
	128 D	100.00 ± 0.00	99.33 ± 0.58	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	96.67 ± 1.53	98.33 ± 1.15	95.67 ± 2.31	100.00 ± 0.00	91.67 ± 3.51	98.17 ± 2.94	1.00
	256 D	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 1.15	100.00 ± 0.00	95.00 ± 1.00	99.43 ± 1.57	1.00
10 full (F)	16 F	83.00 ± 2.00	93.00 ± 1.00	100.00 ± 0.00	98.33 ± 0.58	91.67 ± 0.58	64.33 ± 3.21	59.33 ± 1.15	52.67 ± 1.53	98.33 ± 0.58	6.33 ± 1.15	74.70 ± 28.71	1.00
	32 F	91.33 ± 0.58	96.00 ± 1.00	100.00 ± 0.00	98.33 ± 0.58	94.33 ± 0.58	84.00 ± 2.00	83.00 ± 1.73	70.33 ± 1.53	99.00 ± 1.00	22.00 ± 2.65	83.83 ± 22.82	0.99
	64 F	99.00 ± 0.00	97.33 ± 0.58	100.00 ± 0.00	100.00 ± 0.00	99.33 ± 1.15	91.33 ± 1.53	96.33 ± 0.58	81.67 ± 2.31	99.00 ± 0.00	58.33 ± 1.53	92.23 ± 12.76	0.97
	128 F	99.67 ± 0.58	99.33 ± 0.58	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	97.67 ± 1.15	100.00 ± 0.00	95.67 ± 1.15	100.00 ± 0.00	91.00 ± 1.00	98.33 ± 2.89	0.88
	256 F	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.67 ± 0.58	99.67 ± 0.58	99.67 ± 0.58	100.00 ± 0.00	98.00 ± 1.00	99.70 ± 0.70	0.50
50 diagonal (D)	16 D	52.67 ± 4.51	86.67 ± 3.06	100.00 ± 0.00	85.00 ± 3.46	65.33 ± 3.79	25.33 ± 5.03	10.00 ± 1.00	10.67 ± 10.26	81.00 ± 6.56	0.00 ± 0.00	51.67 ± 36.18	1.00
	32 D	69.67 ± 3.21	88.67 ± 1.53	100.00 ± 0.00	95.00 ± 2.00	80.00 ± 3.00	36.67 ± 5.13	17.00 ± 2.65	26.33 ± 5.03	95.00 ± 2.00	0.00 ± 0.00	60.83 ± 36.02	1.00
	64 D	79.67 ± 2.52	91.00 ± 1.00	100.00 ± 0.00	97.67 ± 0.58	88.00 ± 1.00	52.00 ± 1.00	36.67 ± 5.69	41.33 ± 1.15	96.00 ± 1.00	0.33 ± 0.58	68.27 ± 32.66	1.00
	128 D	90.00 ± 1.00	91.33 ± 0.58	100.00 ± 0.00	98.33 ± 0.58	90.67 ± 2.08	73.67 ± 2.08	63.67 ± 1.53	56.33 ± 0.58	98.00 ± 0.00	7.33 ± 1.15	76.93 ± 27.79	1.00
	256 D	94.67 ± 0.58	96.33 ± 0.58	100.00 ± 0.00	99.67 ± 0.58	96.33 ± 1.15	87.33 ± 0.58	87.00 ± 2.65	71.67 ± 1.53	99.67 ± 0.58	31.67 ± 1.15	86.43 ± 20.41	1.00
50 full (F)	16 F	61.67 ± 3.06	89.67 ± 1.15	99.67 ± 0.58	90.67 ± 2.52	78.33 ± 3.51	34.00 ± 1.00	7.00 ± 3.46	25.00 ± 6.24	90.00 ± 1.00	0.00 ± 0.00	57.60 ± 36.59	1.00
	32 F	71.00 ± 1.00	89.00 ± 1.73	100.00 ± 0.00	98.00 ± 0.00	85.00 ± 1.00	47.00 ± 1.73	29.00 ± 3.00	35.00 ± 2.00	98.00 ± 1.00	0.00 ± 0.00	65.20 ± 34.00	0.99
	64 F	81.67 ± 0.58	93.67 ± 1.15	100.00 ± 0.00	98.33 ± 0.58	90.67 ± 2.08	61.67 ± 1.53	54.33 ± 1.53	51.33 ± 1.15	98.33 ± 0.58	3.33 ± 0.58	73.33 ± 29.89	0.97
	128 F	91.00 ± 1.00	94.33 ± 0.58	100.00 ± 0.00	99.00 ± 0.00	93.33 ± 1.53	81.67 ± 0.58	75.00 ± 1.73	67.67 ± 2.08	98.67 ± 0.58	16.67 ± 0.58	81.73 ± 24.46	0.88
	256 F	97.00 ± 0.00	98.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	99.67 ± 0.58	92.00 ± 0.00	94.33 ± 1.15	79.67 ± 1.53	99.00 ± 0.00	57.33 ± 2.52	91.70 ± 13.11	0.50
100 diagonal (D)	16 D	33.00 ± 8.19	79.33 ± 5.69	89.33 ± 5.51	80.00 ± 3.61	35.33 ± 6.03	4.00 ± 1.73	3.00 ± 1.00	0.00 ± 0.00	71.33 ± 4.51	0.00 ± 0.00	39.53 ± 36.15	1.00
	32 D	51.00 ± 7.81	90.00 ± 1.00	100.00 ± 0.00	88.00 ± 7.00	58.67 ± 11.68	17.67 ± 10.26	7.67 ± 2.89	9.33 ± 12.86	86.33 ± 2.52	0.00 ± 0.00	50.87 ± 38.43	1.00
	64 D	68.00 ± 2.65	87.33 ± 1.53	100.00 ± 0.00	94.33 ± 4.04	80.33 ± 2.08	38.00 ± 3.00	19.67 ± 4.51	28.33 ± 1.53	92.67 ± 1.15	0.33 ± 0.58	60.90 ± 34.91	1.00
	128 D	82.00 ± 2.00	90.00 ± 2.00	100.00 ± 0.00	97.33 ± 0.58	85.33 ± 0.58	55.33 ± 2.08	34.33 ± 3.79	36.67 ± 2.52	94.67 ± 0.58	0.00 ± 0.00	67.57 ± 32.95	1.00
	256 D	90.00 ± 1.00	93.00 ± 2.00	100.00 ± 0.00	97.67 ± 0.58	91.67 ± 0.58	71.67 ± 5.13	59.67 ± 1.53	58.00 ± 0.00	97.67 ± 0.58	4.00 ± 1.00	76.33 ± 28.88	1.00
100 full (F)	16 F	49.00 ± 2.00	89.67 ± 3.21	97.00 ± 3.00	84.33 ± 3.06	65.33 ± 2.52	20.67 ± 8.33	6.33 ± 2.08	8.33 ± 4.73	81.33 ± 2.08	0.00 ± 0.00	50.20 ± 37.06	1.00
	32 F	65.00 ± 3.46	90.33 ± 1.53	100.00 ± 0.00	96.33 ± 1.53	80.00 ± 2.65	41.33 ± 3.21	16.00 ± 0.00	29.33 ± 2.08	92.00 ± 2.65	0.00 ± 0.00	61.03 ± 35.43	0.99
	64 F	72.33 ± 0.58	89.67 ± 1.53	100.00 ± 0.00	97.67 ± 0.58	86.00 ± 1.00	53.00 ± 1.00	35.33 ± 1.53	38.00 ± 1.73	94.67 ± 0.58	0.00 ± 0.00	66.67 ± 32.54	0.97
	128 F	84.33 ± 1.53	92.33 ± 1.53	100.00 ± 0.00	98.00 ± 0.00	91.33 ± 0.58	68.67 ± 0.58	56.00 ± 1.00	57.67 ± 1.15	99.00 ± 0.00	5.33 ± 0.58	75.27 ± 28.67	0.88
	256 F	91.67 ± 1.15	96.67 ± 0.58	100.00 ± 0.00	100.00 ± 0.00	94.33 ± 0.58	84.67 ± 0.58	78.00 ± 0.00	69.67 ± 0.58	99.00 ± 0.00	22.00 ± 1.00	83.60 ± 23.08	0.50
500 diagonal (D)	16 D	8.00 ± 3.61	50.33 ± 3.21	79.67 ± 4.93	28.00 ± 42.44	56.67 ± 2.89	0.67 ± 1.15	0.33 ± 0.58	0.00 ± 0.00	71.67 ± 6.03	0.00 ± 0.00	29.53 ± 33.14	1.00
	32 D	14.33 ± 11.02	50.00 ± 7.81	80.67 ± 2.08	43.00 ± 31.48	60.67 ± 0.58	0.67 ± 1.15	1.67 ± 1.15	0.00 ± 0.00	74.33 ± 4.04	0.00 ± 0.00	32.53 ± 32.93	1.00
	64 D	25.67 ± 1.15	57.33 ± 12.34	87.33 ± 4.04	78.33 ± 1.15	65.33 ± 3.06	5.33 ± 3.51	3.67 ± 1.15	1.00 ± 1.73	76.67 ± 2.31	0.00 ± 0.00	40.07 ± 35.20	1.00
	128 D	38.33 ± 3.21	84.00 ± 3.00	96.00 ± 3.00	81.33 ± 2.31	65.67 ± 1.15	11.67 ± 4.73	5.33 ± 2.08	2.00 ± 1.00	80.00 ± 5.00	0.00 ± 0.00	46.43 ± 37.69	1.00
	256 D	53.33 ± 0.58	91.33 ± 2.08	100.00 ± 0.00	89.00 ± 2.65	76.00 ± 2.00	20.67 ± 6.81	6.00 ± 1.73	12.00 ± 8.00	86.33 ± 2.31	0.00 ± 0.00	53.47 ± 38.66	1.00
500 full (F)	16 F	8.33 ± 2.08	39.67 ± 4.62	76.67 ± 0.58	78.00 ± 0.00	72.67 ± 0.58	6.00 ± 0.00	5.67 ± 0.58	0.00 ± 0.00	78.00 ± 1.00	0.00 ± 0.00	36.50 ± 34.84	1.00
	32 F	33.67 ± 4.16	53.33 ± 4.16	92.67 ± 1.53	77.00 ± 1.73	75.00 ± 2.00	14.33 ± 1.53	8.00 ± 0.00	0.00 ± 0.00	80.67 ± 1.53	0.00 ± 0.00	43.47 ± 35.28	0.99
	64 F	56.00 ± 2.65	85.33 ± 0.58	94.33 ± 0.58	89.33 ± 2.89	74.33 ± 1.15	36.33 ± 1.15	9.00 ± 1.00	2.67 ± 1.15	84.00 ± 1.00	0.00 ± 0.00	53.13 ± 36.77	0.97
	128 F	69.33 ± 0.58	87.67 ± 1.53	99.00 ± 0.00	96.33 ± 1.53	80.33 ± 1.15	45.00 ± 2.00	16.33 ± 0.58	31.00 ± 1.73	92.00 ± 0.00	0.00 ± 0.00	61.70 ± 34.77	0.88
	256 F	79.67 ± 0.58	89.33 ± 0.58	100.00 ± 0.00	97.67 ± 0.58	87.33 ± 0.58	57.00 ± 1.00	35.00 ± 1.00	42.00 ± 1.00	95.00 ± 1.00	0.00 ± 0.00	68.30 ± 32.34	0.50

Table 10. **Absolute** In-Distribution agreement for various tasks and methods

Model Type	Method Type	Tasks										Average
		task039	task190	task280	task290	task391	task442	task620	task1342	task1391	task1598	
SVD	SVD 2	0.29 ± 0.00	0.43 ± 0.00	0.31 ± 0.00	0.40 ± 0.00	0.38 ± 0.00	0.31 ± 0.00	0.37 ± 0.00	0.31 ± 0.00	0.42 ± 0.00	0.30 ± 0.00	0.35 ± 0.05
	SVD 4	0.16 ± 0.00	0.24 ± 0.00	0.16 ± 0.00	0.25 ± 0.00	0.23 ± 0.00	0.17 ± 0.00	0.22 ± 0.00	0.16 ± 0.00	0.25 ± 0.00	0.16 ± 0.00	0.20 ± 0.04
	SVD 8	0.06 ± 0.00	0.09 ± 0.00	0.06 ± 0.00	0.11 ± 0.00	0.10 ± 0.00	0.07 ± 0.00	0.09 ± 0.00	0.06 ± 0.00	0.11 ± 0.00	0.06 ± 0.00	0.08 ± 0.02
10 diagonal (D)	16 D	0.37 ± 0.02	0.51 ± 0.02	0.36 ± 0.01	0.57 ± 0.02	0.55 ± 0.00	0.39 ± 0.02	0.49 ± 0.01	0.36 ± 0.02	0.53 ± 0.03	0.39 ± 0.01	0.45 ± 0.08
	32 D	0.21 ± 0.01	0.28 ± 0.00	0.20 ± 0.01	0.35 ± 0.00	0.33 ± 0.01	0.22 ± 0.01	0.31 ± 0.01	0.20 ± 0.01	0.32 ± 0.01	0.22 ± 0.00	0.26 ± 0.06
	64 D	0.10 ± 0.00	0.11 ± 0.01	0.09 ± 0.00	0.18 ± 0.01	0.18 ± 0.00	0.10 ± 0.00	0.15 ± 0.01	0.09 ± 0.00	0.14 ± 0.00	0.09 ± 0.00	0.12 ± 0.04
	128 D	0.02 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.03 ± 0.00	0.04 ± 0.00	0.02 ± 0.00	0.03 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.03 ± 0.01
	256 D	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
10 full (F)	16 F	0.35 ± 0.00	0.46 ± 0.00	0.34 ± 0.00	0.51 ± 0.00	0.47 ± 0.01	0.36 ± 0.01	0.45 ± 0.01	0.35 ± 0.01	0.49 ± 0.00	0.35 ± 0.01	0.41 ± 0.06
	32 F	0.20 ± 0.00	0.24 ± 0.00	0.20 ± 0.00	0.30 ± 0.00	0.29 ± 0.00	0.22 ± 0.00	0.27 ± 0.00	0.20 ± 0.00	0.27 ± 0.00	0.21 ± 0.00	0.24 ± 0.04
	64 F	0.10 ± 0.00	0.10 ± 0.00	0.09 ± 0.00	0.13 ± 0.00	0.13 ± 0.00	0.10 ± 0.00	0.12 ± 0.00	0.09 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.11 ± 0.02
	128 F	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.02 ± 0.00
	256 F	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
50 diagonal (D)	16 D	0.66 ± 0.01	0.69 ± 0.01	0.88 ± 0.01	0.76 ± 0.03	0.95 ± 0.02	0.91 ± 0.01	0.83 ± 0.02	0.88 ± 0.03	0.72 ± 0.02	0.88 ± 0.02	0.82 ± 0.10
	32 D	0.50 ± 0.01	0.52 ± 0.02	0.73 ± 0.01	0.58 ± 0.03	0.88 ± 0.03	0.79 ± 0.03	0.72 ± 0.01	0.75 ± 0.01	0.57 ± 0.02	0.75 ± 0.01	0.68 ± 0.12
	64 D	0.34 ± 0.01	0.37 ± 0.01	0.52 ± 0.00	0.38 ± 0.01	0.71 ± 0.02	0.58 ± 0.01	0.54 ± 0.00	0.56 ± 0.00	0.44 ± 0.01	0.58 ± 0.01	0.50 ± 0.11
	128 D	0.21 ± 0.01	0.22 ± 0.01	0.31 ± 0.00	0.22 ± 0.00	0.51 ± 0.01	0.42 ± 0.01	0.38 ± 0.00	0.39 ± 0.00	0.27 ± 0.00	0.40 ± 0.00	0.33 ± 0.10
	256 D	0.10 ± 0.00	0.12 ± 0.00	0.16 ± 0.00	0.10 ± 0.00	0.29 ± 0.01	0.21 ± 0.00	0.19 ± 0.00	0.23 ± 0.01	0.15 ± 0.00	0.20 ± 0.00	0.18 ± 0.06
50 full (F)	16 F	0.57 ± 0.01	0.60 ± 0.01	0.86 ± 0.01	0.71 ± 0.02	0.95 ± 0.01	0.88 ± 0.01	0.81 ± 0.00	0.83 ± 0.01	0.67 ± 0.01	0.86 ± 0.01	0.78 ± 0.12
	32 F	0.47 ± 0.01	0.48 ± 0.01	0.71 ± 0.00	0.55 ± 0.01	0.78 ± 0.01	0.69 ± 0.01	0.69 ± 0.00	0.65 ± 0.01	0.53 ± 0.01	0.71 ± 0.00	0.63 ± 0.11
	64 F	0.33 ± 0.00	0.35 ± 0.00	0.45 ± 0.00	0.36 ± 0.00	0.56 ± 0.00	0.50 ± 0.01	0.47 ± 0.00	0.49 ± 0.00	0.39 ± 0.01	0.49 ± 0.00	0.44 ± 0.08
	128 F	0.19 ± 0.00	0.21 ± 0.00	0.25 ± 0.00	0.19 ± 0.00	0.35 ± 0.00	0.30 ± 0.00	0.28 ± 0.00	0.31 ± 0.00	0.24 ± 0.00	0.30 ± 0.00	0.26 ± 0.05
	256 F	0.09 ± 0.00	0.10 ± 0.00	0.10 ± 0.00	0.08 ± 0.00	0.16 ± 0.00	0.13 ± 0.00	0.12 ± 0.00	0.15 ± 0.00	0.11 ± 0.00	0.13 ± 0.00	0.12 ± 0.02
100 diagonal (D)	16 D	0.90 ± 0.01	0.85 ± 0.01	0.87 ± 0.03	0.88 ± 0.02	0.68 ± 0.02	0.91 ± 0.01	0.97 ± 0.01	0.98 ± 0.01	0.96 ± 0.01	1.00 ± 0.00	0.90 ± 0.09
	32 D	0.83 ± 0.02	0.77 ± 0.00	0.77 ± 0.01	0.78 ± 0.00	0.55 ± 0.02	0.79 ± 0.01	0.94 ± 0.02	0.94 ± 0.03	0.87 ± 0.00	0.98 ± 0.01	0.82 ± 0.12
	64 D	0.67 ± 0.01	0.63 ± 0.00	0.59 ± 0.02	0.63 ± 0.01	0.40 ± 0.00	0.62 ± 0.01	0.86 ± 0.02	0.82 ± 0.02	0.71 ± 0.03	0.93 ± 0.00	0.68 ± 0.15
	128 D	0.49 ± 0.01	0.47 ± 0.00	0.42 ± 0.01	0.45 ± 0.00	0.27 ± 0.02	0.44 ± 0.01	0.73 ± 0.01	0.69 ± 0.02	0.59 ± 0.02	0.80 ± 0.02	0.53 ± 0.16
	256 D	0.32 ± 0.00	0.31 ± 0.00	0.26 ± 0.01	0.30 ± 0.00	0.15 ± 0.01	0.28 ± 0.00	0.51 ± 0.02	0.51 ± 0.02	0.40 ± 0.01	0.61 ± 0.01	0.36 ± 0.14
100 full (F)	16 F	0.88 ± 0.00	0.82 ± 0.00	0.84 ± 0.01	0.86 ± 0.00	0.67 ± 0.01	0.88 ± 0.01	0.99 ± 0.00	0.96 ± 0.01	0.91 ± 0.01	1.00 ± 0.00	0.88 ± 0.09
	32 F	0.78 ± 0.00	0.72 ± 0.00	0.73 ± 0.00	0.74 ± 0.00	0.52 ± 0.00	0.74 ± 0.01	0.94 ± 0.01	0.89 ± 0.00	0.77 ± 0.02	0.99 ± 0.00	0.78 ± 0.13
	64 F	0.60 ± 0.00	0.57 ± 0.00	0.57 ± 0.00	0.57 ± 0.00	0.39 ± 0.00	0.56 ± 0.00	0.76 ± 0.00	0.73 ± 0.00	0.60 ± 0.00	0.83 ± 0.01	0.62 ± 0.12
	128 F	0.40 ± 0.00	0.38 ± 0.00	0.35 ± 0.00	0.37 ± 0.00	0.25 ± 0.00	0.37 ± 0.00	0.52 ± 0.00	0.54 ± 0.00	0.45 ± 0.00	0.60 ± 0.00	0.42 ± 0.10
	256 F	0.21 ± 0.00	0.20 ± 0.00	0.18 ± 0.00	0.19 ± 0.00	0.13 ± 0.00	0.19 ± 0.00	0.30 ± 0.00	0.34 ± 0.00	0.26 ± 0.00	0.38 ± 0.00	0.24 ± 0.08
500 diagonal (D)	16 D	0.97 ± 0.00	0.73 ± 0.00	0.96 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.96 ± 0.01	0.90 ± 0.00	0.92 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.94 ± 0.08
	32 D	0.96 ± 0.00	0.70 ± 0.00	0.92 ± 0.01	0.98 ± 0.01	0.96 ± 0.01	0.93 ± 0.01	0.86 ± 0.00	0.89 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.92 ± 0.09
	64 D	0.90 ± 0.01	0.65 ± 0.00	0.86 ± 0.01	0.96 ± 0.02	0.90 ± 0.01	0.87 ± 0.01	0.81 ± 0.00	0.83 ± 0.01	0.99 ± 0.01	1.00 ± 0.00	0.88 ± 0.10
	128 D	0.82 ± 0.01	0.60 ± 0.00	0.76 ± 0.00	0.90 ± 0.02	0.83 ± 0.01	0.78 ± 0.02	0.74 ± 0.00	0.74 ± 0.01	0.97 ± 0.01	1.00 ± 0.00	0.81 ± 0.12
	256 D	0.59 ± 0.02	0.51 ± 0.00	0.56 ± 0.01	0.81 ± 0.02	0.70 ± 0.02	0.55 ± 0.01	0.57 ± 0.01	0.54 ± 0.01	0.91 ± 0.01	1.00 ± 0.01	0.67 ± 0.17
500 full (F)	16 F	0.94 ± 0.00	0.67 ± 0.00	0.88 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	0.90 ± 0.00	0.82 ± 0.00	0.83 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.90 ± 0.10
	32 F	0.88 ± 0.00	0.61 ± 0.00	0.81 ± 0.00	0.97 ± 0.01	0.94 ± 0.00	0.84 ± 0.00	0.75 ± 0.00	0.77 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.86 ± 0.12
	64 F	0.80 ± 0.00	0.55 ± 0.00	0.72 ± 0.00	0.86 ± 0.00	0.82 ± 0.01	0.76 ± 0.00	0.67 ± 0.00	0.70 ± 0.00	0.94 ± 0.00	0.99 ± 0.00	0.78 ± 0.13
	128 F	0.64 ± 0.00	0.46 ± 0.00	0.60 ± 0.00	0.74 ± 0.00	0.65 ± 0.00	0.63 ± 0.00	0.56 ± 0.00	0.58 ± 0.00	0.85 ± 0.00	0.96 ± 0.00	0.67 ± 0.14
	256 F	0.43 ± 0.00	0.35 ± 0.00	0.44 ± 0.00	0.55 ± 0.00	0.49 ± 0.00	0.45 ± 0.00	0.40 ± 0.00	0.42 ± 0.00	0.67 ± 0.00	0.84 ± 0.00	0.50 ± 0.14

Table 11. **Reconstruction error** In-Distribution for various tasks and methods

Model Type	Method Type	Tasks									
		task036	task200	task220	task279	task281	task392	task520	task957	task1195	task1664
	Base	11.11	21.02	26.81	8.59	48.08	12.78	15.06	46.42	61.35	43.01
	Lora	87.37	94.00	99.00	94.33	80.25	94.00	99.00	47.94	97.86	100.00
10	16	14.49	27.85	82.15	62.23	53.49	63.31	43.66	48.73	69.81	54.85
	32	14.64	46.41	95.83	69.89	53.63	72.66	74.86	48.12	71.25	57.27
	64	19.63	67.00	98.16	71.66	55.32	80.51	79.60	48.13	76.02	63.92
50	16	16.09	82.00	97.16	70.14	55.26	76.01	83.74	47.71	77.92	61.46
	32	44.28	83.00	99.00	72.66	56.68	86.55	86.29	47.74	79.44	68.21
	64	57.76	88.00	99.00	74.00	56.86	92.33	88.05	47.75	84.29	71.29
100	16	22.18	84.00	98.00	71.66	55.49	82.38	88.12	48.27	78.71	66.81
	32	48.45	86.00	99.00	73.66	56.85	88.32	88.05	48.02	81.81	67.19
	64	60.60	90.00	99.00	74.00	59.63	92.33	90.00	48.18	84.92	71.52
500	16	60.99	86.00	99.00	73.00	56.58	90.66	91.00	47.20	85.16	71.21
	32	64.09	90.00	99.00	73.66	59.58	92.00	91.00	47.37	86.84	75.23
	64	66.29	89.00	100.00	71.33	59.79	93.00	92.00	47.56	88.29	74.98

Table 12. Compression new unseen LoRAs. ROUGE-L scores for various tasks and methods

Model Type	Method Type	Tasks										Average
		task039	task190	task280	task290	task391	task442	task620	task1342	task1391	task1598	
	base	24.44	1.60	19.13	39.22	10.27	35.46	7.85	6.22	17.82	38.87	20.09
	lora	95.00	86.00	99.00	93.67	94.33	74.88	74.40	26.68	95.00	50.32	78.93
10 full (F)	32 F	97.00	90.00	99.00	93.33	94.67	74.09	72.13	27.83	94.00	50.71	79.28
	64 F	95.00	89.00	99.00	93.67	94.67	74.29	74.80	26.63	96.00	51.04	79.41
50 full (F)	32 F	96.00	88.00	99.00	93.67	92.33	72.30	75.97	29.89	94.00	45.68	78.68
	64 F	98.00	89.00	99.00	93.67	93.33	72.74	76.50	29.33	96.00	45.71	79.33
100 full (F)	32 F	92.10	83.00	99.00	93.67	92.00	71.09	63.29	27.87	88.00	42.36	75.24
	64 F	97.00	87.00	99.00	93.67	92.33	72.23	74.69	29.98	95.00	44.71	78.56
500 full (F)	32 F	68.92	43.00	87.00	91.67	90.67	70.08	51.16	14.40	83.00	41.97	64.19
	64 F	93.50	78.00	91.00	92.33	90.33	72.55	57.49	15.44	85.00	42.31	71.80

Table 13. **Performance with convergence** In-Distribution **Rouge-L**

Table 14. Agreement Comparison. 100 LoRAs

Configuration		**Agreement (%)**
Base Model		83.015
Uncompressed LoRAs		100.000
Joint Compression		
Diagonal	Rank 8	87.032
	Rank 16	88.908
	Rank 32	91.545
	Rank 64	94.659
Full	Rank 8	87.686
	Rank 16	90.163
	Rank 32	94.018
	Rank 64	96.918

Table 15. Performance Comparison. 100 LoRAs

Configuration		**Average Performance**
Base Model		32.28
Uncompressed LoRAs		48.32
Join Compression		
Diagonal	Rank 8	41.90
	Rank 16	45.44
	Rank 32	46.89
	Rank 64	47.43
Full	Rank 8	43.88
	Rank 16	45.79
	Rank 32	46.83
	Rank 64	47.66

Table 16. Task-Based Performance Evaluation Across Different Models and Ranks

Task	**Base Model**	**LoRA**	**Diagonal R8**	**Diagonal R16**	**Diagonal R32**	**Diagonal R64**
Causal Judgement	57.47	64.37	55.17	58.62	58.62	58.62
Date Understanding	15.33	23.33	20.67	22.00	21.33	22.67
Formal Fallacies	51.33	56.00	52.67	52.67	53.33	54.67
Hyperbaton	6.67	68.00	57.33	63.33	67.33	68.00
Logical Deduction (5 Objects)	21.33	37.33	32.00	36.67	37.33	37.33
Logical Deduction (7 Objects)	12.67	44.00	31.33	42.67	44.67	45.33
Movie Recommendation	62.67	67.33	62.00	64.67	66.67	67.33
Object Counting	34.67	38.00	35.33	36.67	36.67	38.00
Snarks	50.00	61.54	53.85	56.41	58.97	57.69
Temporal Sequences	16.67	23.33	18.67	20.67	24.00	24.67
Average	32.88	48.32	41.90	45.44	46.89	47.43

Table 17. Task-Based Performance Evaluation Across Different Models and Ranks

Task	**Base Model**	**LoRA**	**Full R8**	**Full R16**	**Full R32**	**Full R64**
Causal Judgement	57.47	64.37	56.32	57.47	58.62	60.92
Date Understanding	15.33	23.33	19.33	22.00	22.67	22.67
Formal Fallacies	51.33	56.00	51.33	52.67	53.33	56.00
Hyperbaton	6.67	68.00	63.33	66.00	69.33	68.00
Logical Deduction (5 Objects)	21.33	37.33	35.33	36.00	35.33	37.33
Logical Deduction (7 Objects)	12.67	44.00	40.00	44.67	44.67	44.67
Movie Recommendation	62.67	67.33	63.33	65.33	67.33	67.33
Object Counting	34.67	38.00	35.33	36.67	37.33	37.33
Snarks	50.00	61.54	53.85	55.13	57.69	58.97
Temporal Sequences	16.67	23.33	20.67	22.00	22.00	23.33
Average	32.88	48.32	43.88	45.79	46.83	47.66