TailLoR: Protecting Principal Components in Parameter-Efficient Continual Learning

Anonymous ACL submission

Abstract

Parameter-efficient finetuning methods based on spectral decomposition have enabled progress in Continual Learning. In this paper we introduce *TailLoR*, which utilizes the singular bases U and V of the pre-trained weights as a fixed reference frame to learn a low-rank update applied to the singular value matrix. A soft spectral penalty discourages updates aligned with dominant singular directions, reducing interference while routing fine-grained adaptation into the highly flexible, long-tail spectral coordinates.

1 Introduction

Large Language Models (LLMs) have achieved remarkable performance across diverse reasoning and generation tasks (Zhao et al., 2023; Minaee et al., 2024). However, adapting these models to new domains or tasks remains computationally expensive, as full fine-tuning requires updating billions of parameters. This challenge has motivated the development of Parameter-Efficient Fine-Tuning (PEFT) methods (Houlsby et al., 2019), which aim to adapt large pretrained models while updating only a small subset of parameters, reducing memory and training costs.

Among PEFT approaches, Low-Rank Adaptation (LoRA) (Hu et al., 2021) has emerged as one of the most widely adopted. Motivated by the evidence that task-specific updates often lie in a low-dimensional subspace (Li et al., 2018; Aghajanyan et al., 2021), LoRA freezes the pretrained weights and model updates using two trainable low-rank matrices, substantially reducing the number of trainable parameters while maintaining strong performance. Recent works have explored structured decompositions to further improve stability. Specifically, spectral and singular value decomposition (SVD)-based methods Meng et al. (2024); Wang et al. (2025); Lingam et al. (2024) align trainable updates with the structure of pretrained weight matrices for more efficient tuning.

Existing low-rank adaptation methods often suffer from interference between overlapping update directions, especially when models are adapted across multiple domains or sequential tasks. This is particularly problematic in continual learning, where new knowledge must be acquired without erasing prior capabilities. Since the largest singular values of the pre-trained weight matrix encode its most critical structural representations, modifications to this subspace disproportionately degrade prior knowledge.

To mitigate this, we propose a spectral regularization scheme that selectively penalizes updates to the dominant singular components while allowing greater flexibility in the lower-rank "tail". Our specific contributions are as following:

- We introduce **TailLoR**, a low-rank adaptation method operating over the singular values of a weight matrix, coupled with a soft regularization that steers updates toward the spectral "tail", protecting the top singular values.
- Different from existing continual PEFT methods (Das Biswas et al., 2026; Wang et al., 2023a) **TailLoR** requires no access to adapters from prior tasks, crucially, enabling sequential adaptation by different users while preserving the privacy of each user's task-specific parameters and training data.
- We evaluate **TailLoR** on a suite of continual learning tasks, demonstrating that it matches state-of-the-art methods while increasing the stable rank of the weight matrix.

2 Related Work

Spectral LoRA variants Leveraging the spectral properties of base weights W is a key strategy in PEFT. Many SVD-based approaches partition

the initial weights $W_0 = U\Sigma V^\top$ into frozen and trainable components. For instance, PiSSA (Meng et al., 2024) initializes the trainable adapter with the principal singular components to accelerate convergence, while freezing the residual. Conversely, MiLoRA (Wang et al., 2025) freezes the principal components and initializes the adapter with the minor components . Other approaches retain the original singular bases as a reference frame rather than statically partitioning them. SVFT (Lingam et al., 2024) adapts weights via $W = U(\Sigma + M)V^\top$, where M is a learned matrix with fixed sparsity patterns. Our mechanism is closely related, but instead of fixing a sparsity pattern, we learn a low-rank update $M = AB$: $W = U(\Sigma + AB)V^\top$. We soft-regularize the update with different spectral penalties that control which couplings between u_i and v_j are favoured, by penalizing coupling of dominant directions and guiding the model to route new task adaptations into the underutilized, long-tail spectral coordinates. While this emphasis is similar in spirit to MiLoRA and PiSSA, these methods primarily use the SVD partitioning to decompose the original weights and initialize the adapter, while our work keeps the updates parameterized with respect to the original SVD directions throughout training.

Spectral Approaches for Continual Learning Continual Learning (CL) mitigates catastrophic forgetting (McCloskey and Cohen, 1989) across sequential tasks using rehearsal (Lopez-Paz and Ranzato, 2017; de Masson d'Autume et al., 2019; Riemer et al., 2019), regularization (Kirkpatrick et al., 2016), or architecture-expansion (Yoon et al., 2018; Li et al., 2019; Rusu et al., 2016) strategies. Recently, Parameter-Efficient Fine-Tuning (PEFT) (Houlsby et al., 2019), particularly Low-Rank Adaptation (LoRA) (Hu et al., 2022), has emerged as a memory-efficient paradigm for CL, solving sequential tasks via low-parameter, task-specific adapters.

To prevent new adapters from interfering with prior knowledge, recent methods impose geometric or subspace constraints. O-LoRA (Wang et al., 2023a) encourages orthogonality between successive task adapters, while ELLA (Das Biswas et al., 2026) limits interference by restricting the specific sets of parameters that can be modified when learning a new task. Other approaches leverage spectral decompositions or activation geometries: InfLoRA (Liang and Li, 2024) constructs and freeezes the matrix B_t such that the update subspace for the new adapter $\Delta W_t = A_t B_t$ is aligned with the current task activations and also orthogonal to old-task gradients, while NESS (Pham et al., 2026) restricts future updates to the null space of prior activation covariances. Closest to our work is OSFT (Nayak et al., 2025), a full fine-tuning method that first discovers critical prior knowledge directions corresponding to the largest singular values. During fine-tuning on new tasks, gradients are projected to the orthogonal complement of these important directions. While these methods rely on hard gradient projections or activation tracking, TailLoR achieves similar protection purely through soft spectral regularization within the efficient LoRA framework.

3 TailLoR

Given a pre-trained weight matrix $W \in \mathbb{R}^{d_{out} \times d_{in}}$, we extract its structural geometry via Singular Value Decomposition (SVD): $W = U\Sigma V^T$, where $U \in \mathbb{R}^{d_{out} \times d_{out}}$ and $V \in \mathbb{R}^{d_{in} \times d_{in}}$ are orthogonal matrices containing the left and right singular vectors, and $\Sigma \in \mathbb{R}^{d_{out} \times d_{in}}$ is a rectangular diagonal matrix containing the singular values σ_i. For notational simplicity, we assume the pre-trained weight matrices are square, $W \in \mathbb{R}^{k \times k}$.

Rather than applying updates directly in the standard weight basis, we parameterize them entirely within this spectral basis. We introduce a low-rank adapter defined by the product of two matrices $B \in \mathbb{R}^{d_{out} \times r}$ and $A \in \mathbb{R}^{r \times d_{in}}$, where $r \ll \min(d_{out}, d_{in})$. The updated weight matrix during fine-tuning is constructed as:

$$W' = U(\Sigma + AB)V^T \tag{1}$$

Similarly with (Sharma et al., 2024; Wang et al., 2025), we hypothesize that the top singular vectors (the "head") encode the foundational representations shared across tasks, while the smaller singular vectors (the "tail") represent underutilized capacity. To regularize updates ΔW during sequential fine-tuning, we construct a spatial penalty matrix $\Omega \in R^{d_{out} \times d_{in}}$ that applies distinct gradient resistance to different singular subspaces.

3.1 The Head Penalty (Subspace Protection)

Let $\sigma \in R^k$ be the vector of non-zero singular values extracted from Σ, sorted in descending order. We first normalize the singular values by their maximum component: $\tilde{\sigma} = \frac{\sigma}{\sigma_{\max}}$ and define the raw head penalty matrix $\mathbf{\Omega}$ such that the

penalty applied to any element (i, j) in the adapter update is governed by the maximum relative importance of its interacting singular components: $\mathbf{\Omega}_{i,j} = \max(\tilde{\sigma}_i, \tilde{\sigma}_j)^\gamma$, where $\gamma > 0$ is a hyperparameter controlling the severity of the penalty gradient. This formulation creates a highly concentrated regularization pressure on the top principal components, which decays as the indices move toward the spectral tail.

Normalization and Baselines To validate the importance of protecting the spectral head, we compare our approach against two baselines: a **tail penalty**, which reverses the normalized singular values to intentionally penalize less significant components (isolating the effect of regularization geometry), and a **uniform penalty**, which applies a flat, unstructured regularization: ($\mathbf{\Omega}_{\text{uniform}} = \mathbf{1}_{k\times k}$).

To ensure a fair comparison, we introduce a mass normalization step. We scale the raw penalty matrix $\mathbf{P}$ so that its total mass equals k^2 (yielding an average element weight of 1.0). The final normalized penalty matrix $\tilde{\mathbf{\Omega}}$ is given by:

$$\tilde{\mathbf{\Omega}} = \mathbf{\Omega} \frac{k^2}{\sum_{i,j} \mathbf{\Omega}_{i,j}} \tag{2}$$

During fine-tuning, the adapter matrices A and B are optimized with this structured regularization term added to the primary task loss. To compute the penalty, we take the square root of the $\tilde{\mathbf{\Omega}}$-weighted mean of the squared adapter updates:

$$\mathcal{L}_{\text{reg}} = \lambda \sqrt{\frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} \tilde{\mathbf{\Omega}}_{i,j} (AB)_{i,j}^2 + \epsilon} \tag{3}$$

where $(AB)_{i,j}$ represents the individual elements of the adapter product (a modification over the original Σ diagonal matrix), and $\epsilon = 10^{-12}$ is a small constant ensuring numerical stability for the gradient of the square root near zero.

4 Evaluation and Results

In this section, we present the evaluation setup and results on Continual Learning tasks, as well as ablations on the spectral penalty strategy and analyse the evolution of the effective rank of the weight matrix for TailLoR and ELLA.

4.1 Continual Learning

We evaluate TailLoR on three continual learning benchmarks: Standard CL (Zhang et al., 2015), Long Sequence (Razdaibiedina et al., 2023), and TRACE (Wang et al., 2023b). Across all setups, we fine-tune the *query* and *value* projections of a T5-large backbone for a single epoch per task with rank $r = 8$. For hyperparameters, TailLoR uses a single, globally searched penalty weight λ and exponent γ applied statically across all tasks. In contrast, following its original protocol, ELLA's λ coefficient is optimized independently for *every* task. Despite this static tuning constraint, Table 1 shows TailLoR (head penalty) achieves highly competitive results on the Standard CL Benchmark alongside ELLA. By matching a state-of-the-art subspace partitioning method without requiring task-specific hyperparameter tuning, TailLoR demonstrates significant robustness and efficiency. In Table 2 we show results on the TRACE benchmark with 500 samples per task on a T5-large model, with TailLoR achieving the highest overall accuracy.

4.2 Spectral Penalty Strategy Ablation

The comparison between head, tail and uniform penalties in Table 1 confirms the intuition in prior works: principal singular directions encode important information and protecting them from excessive task-specific updates results in more efficient continual learning. Penalizing couplings between these high singular value directions encourages task-specific adaptation through the underutilized minor singular components, which improves accuracy and significantly mitigates catastrophic forgetting. Furthermore, the underperformance of the uniform baseline suggests that treating all spectral directions equally is suboptimal compared to structurally guiding adaptation toward the spectral tail.

4.3 Evolution of Effective Rank

To quantify the structural changes within the weight matrices during continual learning, we track the Roy-Vetterli effective rank (Roy and Vetterli, 2007). Given the vector of singular values σ sorted in descending order, the effective rank R_{eff} is defined as the sum of the singular values bounded by the spectral norm (the maximum singular value, $\sigma_{\max}$):

$$R_{\text{eff}} = \frac{\sum_{i=1}^{k} \sigma_i}{\sigma_{\max}} = \sum_{i=1}^{k} \tilde{\sigma}_i \tag{4}$$

A higher R_{eff} indicates a flatter spectral distribution, reflecting how task-specific updates successfully utilize the lower-rank tail capacity rather than overriding the dominant singular vectors.

Table 1: Overall Accuracy (OA) comparison on Standard CL bench (Order $1, 2, 3$) and Long Sequence bench (Order $4, 5, 6$). Results are averaged across three seeds. Top scores are bolded and second best are underlined.

Methods	Standard CL Benchmark (SC)			Long Sequence Benchmark (LS)			Overall Accuracy
	Order 1	Order 2	Order 3	Order 4	Order 5	Order 6	SC + LS
EWC	46.30	45.30	52.10	44.90	44.00	45.40	46.36
IncLoRA	59.25	58.69	69.86	56.83	56.40	54.32	59.25
SVFT	77.76	77.84	77.00	70.17	66.75	73.78	73.88
MiLoRA	66.90	66.67	70.48	59.73	57.12	57.27	63.03
PiSSA	72.64	71.14	70.98	61.84	62.74	60.57	66.65
ELLA	78.09	78.37	78.23	**72.64**	67.91	**74.15**	<u>74.90</u>
TailLoR (head)	**78.87**	**79.41**	**78.62**	<u>71.05</u>	**69.04**	72.89	**74.98**
TailLoR (tail)	<u>78.33</u>	78.22	<u>78.24</u>	70.00	<u>67.97</u>	72.57	74.15
TailLoR (uniform)	78.22	<u>78.51</u>	77.62	69.07	65.96	<u>73.96</u>	73.89

Table 2: Overall Accuracy and Backward Transfer comparison on the TRACE benchmark with 500 train samples per task. Results are averaged across three seeds.

Method	Overall Accuracy	Backward Transfer
IncLoRA	21.68	-24.45
PiSSA	19.99	-23.63
SVFT	24.07	**-0.26**
MiLoRA	26.13	-13.98
ELLA	29.40	-10.53
TailLoR (head)	**30.40**	-4.60

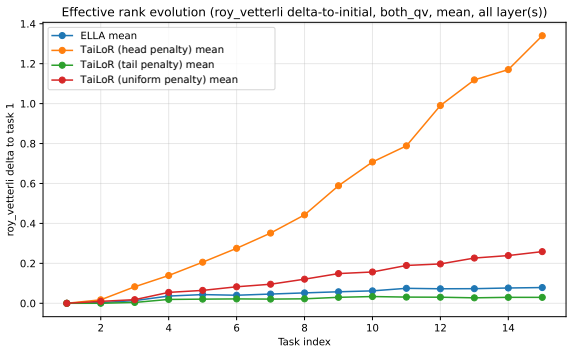

Figure 1: Effective rank analysis on Standard CL Bench (Long Order 4), relative (delta) to the initial values.

Figure 1 illustrates the evolution of R_{eff} for each regularization strategy relative to its initial value. We observe that our **head penalty** induces a consistent, significant increase in effective rank as the model sequentially learns new tasks. This empirically validates our core hypothesis: structurally penalizing modifications to the dominant singular vectors forces the network to route new task representations into the underutilized lower-rank subspace. This accumulation of information flattens the overall spectral distribution, driving up the effective rank. In contrast, methods lacking this structural awareness fail to exploit latent capacity. The **tail penalty** confines updates entirely to the already-dominant head components, resulting in a perfectly stagnant effective rank, while the **uniform penalty** yields only a marginal, unoptimized spread. Notably, while **ELLA** successfully limits interference by controlling the magnitudes of individual parameters, it ignores the underlying geometry of the weight matrix. Consequently, it leaves the spectral capacity constrained, resulting in a flat R_{eff} evolution. Ultimately, these results confirm that protecting the spectral head not only preserves pre-trained knowledge but actively forces the model to exploit latent network capacity for CL.

5 Conclusions

In this work, we introduced TailLoR, a geometrically aware low-rank adaptation method designed to mitigate representation interference in continual learning. By selectively penalizing modifications to the dominant singular vectors, our regularization scheme successfully routes new task updates into the underutilized lower-rank tail. Empirical evaluations demonstrate that TailLoR not only preserves critical pre-trained knowledge but actively increases the effective capacity of the network, matching state-of-the-art baselines without requiring explicit weight partitioning or per-task parameter tuning.

Limitations

While this work demonstrates the efficacy of Tail-LoR on encoder-decoder architectures (T5) for continual learning, extending the spectral routing analysis to modern causal, decoder-only LLMs remains an active area of our ongoing work. Moreover, to maintain manageable computational costs, our evaluation on the TRACE benchmark utilized a 500-sample subset per task. While sufficient to demonstrate relative performance trends, future work will scale these evaluations to the full dataset.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers)*, pages 7319–7328. Association for Computational Linguistics.

Shristi Das Biswas, Yue Zhang, Anwesan Pal, Radhika Bhargava, and Kaushik Roy. 2026. ELLA: Efficient lifelong learning for adapters in large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1907–1924, Rabat, Morocco.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. *CoRR*, abs/1906.01076.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *CoRR*, abs/2106.09685.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2016. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the intrinsic dimension of objective landscapes. In *6th International Conference on Learning Representations (ICLR 2018)*. OpenReview.net.

Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. 2019. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, Proceedings of Machine Learning Research, pages 3925–3934. PMLR.

Yan-Shuo Liang and Wu-Jun Li. 2024. Inflora: Interference-free low-rank adaptation for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 23638–23647. IEEE.

Vijay Lingam, Atula Tejaswi, Aditya Vavre, Aneesh Shetty, Gautham K Gudur, Joydeep Ghosh, Alex Dimakis, Eunsol Choi, Aleksandar Bojchevski, and Sujay Sanghavi. 2024. Svft: Parameter-efficient fine-tuning with singular vectors. *Advances in Neural Information Processing Systems*, 37:41425–41446.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 6467–6476.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Elsevier.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2024. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072.

Shervin Minaee, Tomás Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *CoRR*, abs/2402.06196.

Nikhil Shivakumar Nayak, Krishnateja Killamsetty, Ligong Han, Abhishek Bhandwaldar, Prateek Chanda, Kai Xu, Hao Wang, Aldo Pareja, Oleg Silkin, Mustafa Eyceoz, and 1 others. 2025. Sculpting subspaces: Constrained full fine-tuning in llms for continual learning. *arXiv preprint arXiv:2504.07097*.

Cuong Anh Pham, Praneeth Vepakomma, and Samuel Horváth. 2026. Learning in the null space: Small

singular values for continual learning. *arXiv preprint arXiv:2602.21919*.

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. 2023. Progressive prompts: Continual learning for language models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*. OpenReview.net.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. 2019. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *7th International Conference on Learning Representations (ICLR 2019)*. OpenReview.net.

Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *CoRR*, abs/1606.04671.

Pratyusha Sharma, Jordan Ash, and Dipendra Kumar Misra. 2024. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. In *International Conference on Learning Representations*, volume 2024, pages 17632–17651.

Hanqing Wang, Yixia Li, Shuo Wang, Guanhua Chen, and Yun Chen. 2025. MiLoRA: Harnessing minor singular components for parameter-efficient LLM finetuning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4823–4836, Albuquerque, New Mexico.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. 2023a. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671, Singapore.

Xiao Wang, Yuansen Zhang, Tianze Chen, Songyang Gao, Senjie Jin, Xianjun Yang, Zhiheng Xi, Rui Zheng, Yicheng Zou, Tao Gui, and 1 others. 2023b. Trace: A comprehensive benchmark for continual learning in large language models. *arXiv preprint arXiv:2310.06762*.

Zhicheng Wang, Yufang Liu, Tao Ji, Xiaoling Wang, Yuanbin Wu, Congcong Jiang, Ye Chao, Zhencong Han, Ling Wang, Xu Shao, and Wenqiu Zeng. 2023c. Rehearsal-free continual language learning via efficient parameter isolation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10933–10946. Association for Computational Linguistics.

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. 2018. Lifelong learning with dynamically expandable networks. In *6th International Conference on Learning Representations (ICLR 2018)*. OpenReview.net.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pages 649–657.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

A Reproducibility details

We show in Table 3 the range of hyperparameters we search over for TaiLoR and the general hyperparameters for all baselines in Table 4. We ran all experiments on an Nvidia H200 GPU.

Hyperparameter	Values
γ	[0.5, 1.0, 2.0]
λ	[1e3, 2e3, 5e3, 1e4, 2e4]

Table 3: TaiLoR hyperparameters

Hyperparameter	Values
learning rate	1e-3
rank	8
epochs	1
rank	8
target modules	q, v
optimizer	AdamW
Weight decay	0
LoRA dropout	0.1
Random seeds	[42, 43, 44]

Table 4: TaiLoR hyperparameters

B Datasets and Task Orders

In Table 5, we list the original five datasets in the Standard CL benchmark (SC). Following (Wang et al., 2023c), we only used the last four datasets.

In Table 6, we list the 15 datasets in the Long Sequence Benchmark (LS).

We report all task orders used for our CL experiments in Table 7, which are adopted in previous works (Das Biswas et al., 2026). Orders 1-3 refer to the Standard CL Benchmark, while orders 4-6 refer to the Long Sequence Benchmark.

Table 5: The five classification datasets in the Standard CL Benchmark (Zhang et al., 2015).

Dataset Name	Category	Task	Domain
Yelp	CL Benchmark	Sentiment Analysis	Yelp Reviews
Amazon	CL Benchmark	Sentiment Analysis	Amazon Reviews
DBPedia	CL Benchmark	Topic Classification	Wikipedia
Yahoo	CL Benchmark	Topic Classification	Yahoo Q&A
AG News	CL Benchmark	Topic Classification	News

Table 6: The 15 classification datasets in the Long Sequence Benchmark (Razdaibiedina et al., 2023).

Dataset Name	Category	Task	Domain
Yelp	CL Benchmark	Sentiment Analysis	Yelp Reviews
Amazon	CL Benchmark	Sentiment Analysis	Amazon Reviews
DBPedia	CL Benchmark	Topic Classification	Wikipedia
Yahoo	CL Benchmark	Topic Classification	Yahoo Q&A
AG News	CL Benchmark	Topic Classification	News
MNLI	GLUE	Natural Language Inference	Various
QQP	GLUE	Paragraph Detection	Quora
RTE	GLUE	Natural Language Inference	News, Wikipedia
SST-2	GLUE	Sentiment Analysis	Movie Reviews
WiC	SuperGLUE	Word Sense Disambiguation	Lexical Databases
CB	SuperGLUE	Natural Language Inference	Various
COPA	SuperGLUE	Question and Answering	Blogs, Encyclopedia
BoolQA	SuperGLUE	Boolean Question and Answering	Wikipedia
MultiRC	SuperGLUE	Question and Answering	Various
IMDB	SuperGLUE	Sentiment Analysis	Movie Reviews

Table 7: Task sequence orders for both SC and LS Benchmarks.

Benchmark	Order	Task Sequence
Standard CL Benchmark	1	dbpedia → amazon → yahoo → ag
	2	dbpedia → amazon → ag → yahoo
	3	yahoo → amazon → ag → dbpedia
Long Sequence Benchmark	4	mnli → cb → wic → copa → qqp → boolqa → rte → imdb → yelp → amazon → sst-2 → dbpedia → ag → multirc → yahoo
	5	multirc → boolqa → wic → mnli → cb → copa → qqp → rte → imdb → sst-2 → dbpedia → ag → yelp → amazon → yahoo
	6	yelp → amazon → mnli → cb → copa → qqp → rte → imdb → sst-2 → dbpedia → ag → yahoo → multirc → boolqa → wic