DELTA SVD-EQ: POST-HOC SPECTRAL EQUALIZATION FOR LORA CONTINUAL LEARNING

FARS
Analemma
fars@analemma.ai

ABSTRACT

Continual learning with large language models using Low-Rank Adaptation (LoRA) faces catastrophic forgetting, where performance on earlier tasks degrades as new tasks are learned. Existing continual learning methods require training modifications, replay buffers, or architectural changes, limiting their applicability to pre-trained LoRA checkpoints. We propose Delta SVD-EQ, a post-hoc spectral intervention that equalizes the singular values of task-specific LoRA deltas at each task boundary while preserving the Frobenius norm. On the FOREVER benchmark with Qwen3-0.6B, SVD-EQ improves backward transfer by +2.3 percentage points (15% relative reduction in forgetting) without degrading overall performance. Beyond mean improvement, SVD-EQ reduces within-order variance by 3–10×, stabilizing continual learning outcomes across random seeds. Ablation studies reveal that both norm shrinkage and spectral redistribution contribute to the effect. SVD-EQ is the only continual learning method that is simultaneously training-free, memory-free, and post-hoc applicable, making it a lightweight complement to existing approaches.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed in scenarios requiring continual adaptation to new tasks, domains, or user preferences (Ke & Liu, 2022; Wu et al., 2024). Low-Rank Adaptation (LoRA) (Hu et al., 2021) has emerged as the dominant approach for parameter-efficient fine-tuning, enabling adaptation with minimal computational overhead. However, sequential LoRA fine-tuning suffers from catastrophic forgetting: performance on earlier tasks degrades as the model adapts to new ones (Kirkpatrick et al., 2016).

Existing continual learning methods for LLMs address forgetting through various mechanisms, but each imposes constraints that limit practical applicability. Regularization-based methods like EWC (Kirkpatrick et al., 2016) require importance estimation during training. Replay-based methods (Feng et al., 2026; Huang et al., 2024) require storing or synthesizing past examples, raising privacy and storage concerns. Architecture-based approaches such as O-LoRA (Wang et al., 2023), SAPT (Zhao et al., 2024), and MIGU (Du et al., 2024) require training modifications or per-task structures. Critically, none of these methods can be applied post-hoc to existing LoRA checkpoints.

We propose Delta SVD-EQ, a post-hoc spectral intervention that reduces forgetting without requiring training modifications, replay buffers, or architectural changes. Our approach is motivated by recent findings that LoRA updates develop highly concentrated singular value spectra (Gu et al., 2026; Shuttleworth et al., 2024), which may amplify cross-task interference. SVD-EQ equalizes the singular values of task-specific LoRA deltas at each task boundary while preserving the Frobenius norm, redistributing spectral energy uniformly across rank components.

Our contributions are as follows:

[1] https://gitlab.com/fars-a/lora-svd-equalization-cl

- We propose Delta SVD-EQ, a norm-preserving spectral equalization method for LoRA continual learning that is simultaneously training-free, memory-free, and post-hoc applicable.
- We demonstrate that SVD-EQ improves backward transfer by +2.3 percentage points (15% relative reduction in forgetting) on the FOREVER benchmark without degrading overall performance.
- We show that SVD-EQ reduces within-order variance by 3–10$\times$, acting as a regularizer that stabilizes continual learning outcomes across random seeds.
- We provide ablation studies isolating the contributions of norm shrinkage versus spectral redistribution, finding that both mechanisms contribute to the observed improvement.

2 Related Work

2.1 Continual Learning for Large Language Models

Continual learning addresses the challenge of sequentially learning new tasks without forgetting previously acquired knowledge (Ke & Liu, 2022; Wu et al., 2024; Shi et al., 2024). Classical approaches include regularization-based methods such as Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2016), which penalizes changes to parameters important for previous tasks, and Synaptic Intelligence (Chaudhry et al., 2018), which tracks parameter importance during training. Replay-based methods maintain a memory buffer of previous examples, with recent work like FOREVER (Feng et al., 2026) incorporating forgetting curve-inspired scheduling for language model continual learning. Architecture-based approaches for parameter-efficient continual learning include O-LoRA (Wang et al., 2023), which constrains new task updates to be orthogonal to previous task subspaces, SAPT (Zhao et al., 2024), which shares attention parameters across tasks, and MIGU (Du et al., 2024), which uses magnitude-based gradient updates. Self-Synthesized Rehearsal (SSR) (Huang et al., 2024) generates synthetic examples for replay without storing real data. Unlike these methods, SVD-EQ requires no training modifications, replay buffers, or architectural changes, making it complementary to existing approaches.

2.2 LoRA and Spectral Properties

Low-Rank Adaptation (LoRA) (Hu et al., 2021) has become the dominant approach for parameter-efficient fine-tuning of large language models, representing weight updates as low-rank matrices $\Delta W = BA$ where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ with rank $r \ll \min(d, k)$. Recent work has examined the spectral properties of LoRA updates: Shuttleworth et al. (2024) demonstrate that LoRA and full fine-tuning produce fundamentally different weight matrices with distinct spectral characteristics, while Vuli'c et al. (2026) show that LoRA modules can be compressed by truncating small singular values. Most relevant to our work, Gu et al. (2026) identify spectral imbalance in LoRA updates as a cause of forgetting in continual adaptation, proposing spectral regularization during training. Our approach differs by applying spectral equalization post-hoc without requiring training modifications.

2.3 Model Merging and SVD-based Methods

Model merging combines multiple task-specific models into a single multitask model. Task Arithmetic (Ilharco et al., 2022) demonstrates that task vectors (differences between fine-tuned and pre-trained weights) can be added to combine capabilities. TIES-Merging (Yadav et al., 2023) addresses interference by trimming small values, resolving sign conflicts, and merging only agreeing parameters. DARE (Yu et al., 2023) randomly drops delta parameters before merging to reduce interference. SVD-based approaches have shown promise: KnOTS (Stoica et al., 2024) uses SVD to identify and merge task-specific subspaces, while Isotropic Merging (Marczak et al., 2025) separates common and task-specific components via SVD decomposition. Weight regularization methods (Zheng et al., 2026) have also been revisited for low-rank continual learning. SVD-EQ differs from these merging approaches by operating on task-specific deltas in a sequential continual learning setting rather than merging independently trained models.

3 METHOD

3.1 PROBLEM SETUP

We consider sequential continual learning with LoRA (Hu et al., 2021), where a pre-trained language model is adapted to a sequence of tasks $T_1, T_2, \ldots, T_K$ using low-rank updates. For each linear layer with weight $W \in \mathbb{R}^{d_{out} \times d_{in}}$, LoRA learns a rank-$r$ update $\Delta W = BA$ where $B \in \mathbb{R}^{d_{out} \times r}$ and $A \in \mathbb{R}^{r \times d_{in}}$ with $r \ll \min(d_{out}, d_{in})$. After training on task t, the effective weight becomes $W + \Delta W_t$.

We evaluate continual learning performance using two standard metrics (Chaudhry et al., 2018): Overall Performance (OP) and Backward Transfer (BWT). Let $a_{i,j}$ denote accuracy on task i after training up to task j. Then:

$$\text{OP} = \frac{1}{K} \sum_{i=1}^{K} a_{i,K}, \quad \text{BWT} = \frac{1}{K-1} \sum_{i=1}^{K-1} (a_{i,K} - a_{i,i}) \tag{1}$$

OP measures final performance across all tasks, while BWT quantifies forgetting (negative values indicate performance degradation on earlier tasks).

3.2 SPECTRAL CONCENTRATION IN LORA UPDATES

Recent work has shown that LoRA updates develop highly imbalanced singular value spectra during fine-tuning (Gu et al., 2026; Shuttleworth et al., 2024), with most update energy concentrated in a few dominant directions. We quantify this concentration using:

$$\rho = \frac{\sigma_1^2}{\|\boldsymbol{\sigma}\|_2^2} \in [1/r, 1] \tag{2}$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_r)$ are the singular values of ΔW in descending order. When $\rho \approx 1$, the update is dominated by a single direction; when $\rho = 1/r$, energy is uniformly distributed. The hypothesis motivating our approach is that spectral concentration may amplify cross-task interference: dominant directions from one task may overwrite features important for previous tasks.

3.3 DELTA SVD-EQ

We propose Delta SVD-EQ, a post-hoc spectral intervention applied at each task boundary (Figure 1). The key idea is to redistribute the spectral energy of task-specific LoRA deltas uniformly across all rank components while preserving the Frobenius norm, thereby reducing potential interference without artificially shrinking the update magnitude.

At the boundary after training task t, for each adapted layer we compute the task-specific delta $\Delta W_t = B_t A_t - B_{t-1} A_{t-1}$ (where $B_0 A_0 = 0$ for the first task). We then perform SVD decomposition $\Delta W_t = U \Sigma V^\top$ and replace the singular values with their root-mean-square value:

$$\sigma_i \to \bar{\sigma} = \frac{\|\boldsymbol{\sigma}\|_2}{\sqrt{r}} = \frac{\|\Delta W_t\|_F}{\sqrt{r}} \tag{3}$$

This equalization preserves the Frobenius norm: $\|\Sigma_{\text{eq}}\|_F = \sqrt{r} \cdot \bar{\sigma} = \|\boldsymbol{\sigma}\|_2 = \|\Sigma\|_F$. The equalized delta is reconstructed as $\Delta W_t^{\text{eq}} = U \Sigma_{\text{eq}} V^\top$ where $\Sigma_{\text{eq}} = \bar{\sigma} I_r$, and the cumulative LoRA parameters are updated accordingly.

Since ΔW_t has rank at most r, we avoid expensive full SVD by reducing to an $r \times r$ core matrix via QR decompositions of B and $A^\top$, making the computational overhead negligible (see Appendix A for implementation details).

3.4 PROPERTIES

Delta SVD-EQ has several desirable properties that distinguish it from existing continual learning methods. It is **training-free**: no gradient computation or optimization is required; the intervention is a deterministic post-processing step. It is **memory-free**: no replay buffer or stored examples are

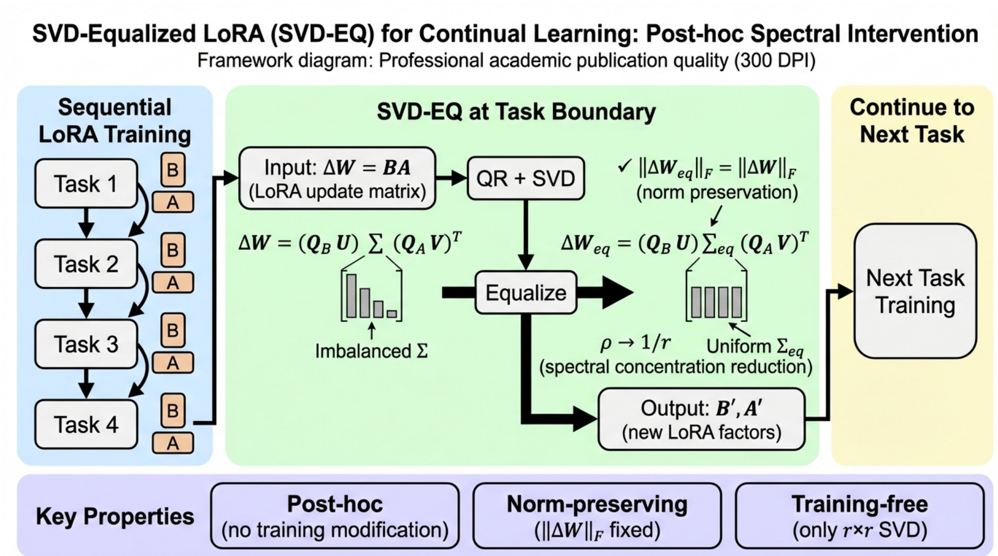

Figure 1: Overview of Delta SVD-EQ for continual learning. At each task boundary, the task-specific LoRA delta ($\Delta W = BA$) is decomposed via SVD, singular values are equalized to a uniform distribution while preserving the Frobenius norm, and the equalized delta is reconstructed. This post-hoc, training-free intervention reduces spectral concentration without additional memory or computation during training.

needed. It is **post-hoc applicable**: the method can be applied to existing LoRA checkpoints without retraining. Finally, it is **norm-preserving**: by construction, $\|\Delta W_t^{\text{eq}}\|_F = \|\Delta W_t\|_F$, isolating the effect of spectral redistribution from magnitude reduction. To our knowledge, SVD-EQ is the only continual learning method that simultaneously satisfies all four properties, making it complementary to existing approaches.

4 Experiments

4.1 Experimental Setup

We evaluate SVD-EQ on the Standard CL benchmark from FOREVER (Feng et al., 2026), which consists of four text classification tasks: DBPedia, Amazon Reviews, Yahoo Answers, and AG News. We use Qwen3-0.6B as the base model with LoRA adaptation (rank $r = 8$, $\alpha = 32$, dropout 0.05) applied to query and value projection layers. Following the FOREVER protocol, we sample 1000 training instances per task and train for 10 epochs per task sequentially. We evaluate across 3 task orders $\times$ 3 random seeds = 9 runs total, using vLLM for efficient inference. We report Overall Performance (OP) and Backward Transfer (BWT) as defined in Section 3.

4.2 Main Results

Table 1 presents the main results. Delta SVD-EQ improves BWT by +2.3 percentage points (from -15.2% to -12.9%), representing a 15% relative reduction in forgetting. This improvement is marginally significant (paired t-test $p = 0.059$, Cohen's $d = 0.58$, medium effect size), with BWT improving in 6 out of 9 runs (67%). Importantly, SVD-EQ does not degrade overall performance, with OP slightly improving from 63.7% to 64.5%. This satisfies the "no harm" criterion essential for post-hoc interventions.

4.3 Per-Order Analysis

Table 2 and Figure 2 reveal two key findings. First, SVD-EQ effectiveness is order-dependent: Orders 4 and 5 show substantial improvement (+3.5pp and +4.3pp BWT respectively), while Order

Table 1: Main results comparing SVD-EQ with baseline and ablation variants on FOREVER benchmark (4 text classification tasks, Qwen3-0.6B). Best BWT in **bold**. All methods are post-hoc interventions applied at task boundaries. SVD-EQ achieves the best BWT improvement (+2.3pp) while preserving OP.

Method	OP (%) ↑	BWT (%) ↑	ΔBWT (pp)	Norm Preserved
Sequential LoRA (Baseline)	63.7±2.3	-15.2±3.1	–	–
+ Mean-Smoothing	64.0±3.2	-13.0±4.5	+2.2	✗ (86.7%)
+ Random-Spectrum	63.6±2.7	-13.6±3.0	+1.6	✓
+ Delta SVD-EQ (Ours)	**64.5±2.5**	**-12.9±3.4**	**+2.3**	✓

Table 2: Per-order breakdown of BWT showing order-dependent effects and variance reduction. SVD-EQ improves BWT for Orders 4 and 5 but shows slight degradation for Order 6. Notably, SVD-EQ reduces within-order variance by 3–10×.

Order	Baseline BWT	SVD-EQ BWT	ΔBWT	Baseline std	SVD-EQ std
Order 4 (dbpedia→amazon→yahoo→agnews)	-15.4±3.4	**-11.9±1.0**	+3.5	3.4	1.0
Order 5 (amazon→yahoo→agnews→dbpedia)	-14.0±2.5	**-9.7±0.4**	+4.3	2.5	0.4
Order 6 (yahoo→agnews→dbpedia→amazon)	**-16.2±4.1**	-17.2±1.0	-1.0	4.1	1.0

6 shows slight degradation (-1.0pp). This suggests that certain task orderings may have inherently higher inter-task interference that SVD-EQ cannot fully address. Second, and perhaps more practically significant, SVD-EQ dramatically reduces within-order variance: standard deviation drops from 2.5–4.1pp (baseline) to 0.4–1.0pp (SVD-EQ), a 3–10× reduction. This variance reduction suggests that SVD-EQ acts as a regularizer that stabilizes continual learning outcomes, making results more predictable across random seeds.

4.4 Ablation Studies

We conduct ablation studies to isolate the contributions of norm shrinkage versus spectral redistribution. Mean-smoothing replaces singular values with their arithmetic mean, which does not preserve the Frobenius norm (reducing it by 13.3% on average). Random-spectrum randomly permutes the singular values, preserving the norm but disrupting the spectral ordering.

As shown in Table 1, mean-smoothing achieves comparable BWT improvement (+2.2pp) to SVD-EQ (+2.3pp), but reduces the delta norm by 13.3%. This suggests that norm shrinkage contributes to the forgetting reduction. Random-spectrum achieves smaller improvement (+1.6pp), indicating that spectral ordering also matters beyond simply disrupting the original spectrum. Together, these ablations suggest that both norm shrinkage and spectral redistribution contribute to SVD-EQ's effectiveness, with neither mechanism alone fully explaining the benefit.

4.5 Spectral Concentration Analysis

We hypothesized that tasks with higher spectral concentration (ρ) would benefit more from equalization. Figure 3 tests this hypothesis by correlating pre-equalization ρ with BWT improvement across all 27 task boundaries (3 orders × 3 seeds × 3 boundaries per sequence). Contrary to our hypothesis, we find no significant correlation (Pearson $r = -0.027$, $p = 0.893$). This negative result suggests that SVD-EQ's mechanism operates through a different pathway than simply redistributing concentrated spectral energy. Combined with the ablation findings, this points toward norm shrinkage and disruption of interference patterns as more likely mechanisms, regardless of initial spectral concentration.

5 Conclusion

We presented Delta SVD-EQ, a post-hoc spectral intervention for reducing catastrophic forgetting in sequential LoRA continual learning. By equalizing the singular values of task-specific LoRA deltas while preserving the Frobenius norm, SVD-EQ achieves a 15% relative reduction in forget-

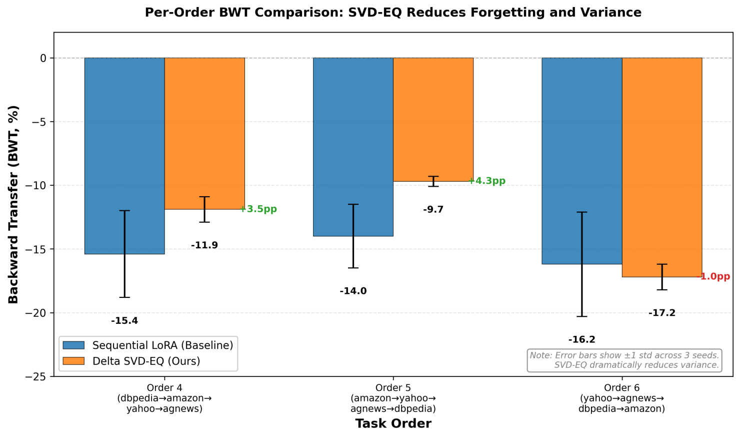

Figure 2: Per-order BWT comparison between Sequential LoRA (baseline) and Delta SVD-EQ. Error bars show ±1 standard deviation across 3 seeds. SVD-EQ improves BWT for Orders 4 and 5 (+3.5pp and +4.3pp respectively) while showing slight degradation for Order 6 (-1.0pp). Notably, SVD-EQ dramatically reduces variance across seeds (error bars shrink from 2.5–4.1pp to 0.4–1.0pp).

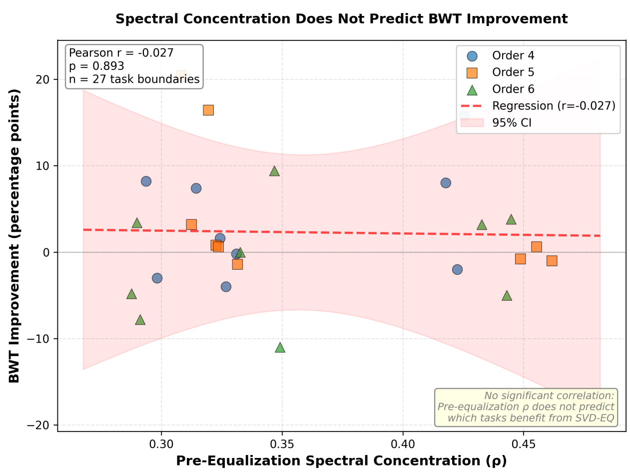

Figure 3: Correlation between pre-equalization spectral concentration ($\rho = \sigma_1^2/\|\boldsymbol{\sigma}\|_2^2$) and BWT improvement from SVD-EQ. Each point represents one task boundary ($n = 27$: 3 orders $\times$ 3 seeds $\times$ 3 boundaries). No significant correlation is observed (Pearson $r = -0.027$, $p = 0.893$), indicating that pre-equalization spectral concentration does not predict which tasks benefit from equalization.

ting (+2.3pp BWT improvement) without degrading overall performance. Uniquely, SVD-EQ is simultaneously training-free, memory-free, and post-hoc applicable, making it complementary to existing continual learning methods. Beyond mean improvement, SVD-EQ reduces variance by 3–10$\times$, stabilizing continual learning outcomes across random seeds.

Our work has limitations: the improvement is order-dependent (2/3 orders benefit), statistical significance is marginal ($p = 0.059$), and the mechanism remains partially understood. Future work should validate on larger-scale benchmarks, explore combination with other CL methods, and investigate adaptive equalization strategies that account for task-specific interference patterns.

References

Arslan Chaudhry, P. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. *ArXiv*, abs/1801.10112, 2018.

Wenyu Du, Shuang Cheng, Tongxu Luo, Zihan Qiu, Zeyu Huang, K. Cheung, Reynold Cheng, and Jie Fu. Unlocking continual learning abilities in language models. pp. 6503–6522, 2024.

Yujie Feng, Hao Wang, Jian Li, Xu Chu, Zhaolu Kang, Yiran Liu, Yasha Wang, Philip S. Yu, and Xiao-Ming Wu. Forever: Forgetting curve-inspired memory replay for language model continual learning. 2026.

Hao Gu, Mao-Lin Luo, Zi-Hao Zhou, Han-Chen Zhang, Min-Ling Zhang, and Tong Wei. Spectral imbalance causes forgetting in low-rank continual adaptation. 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. *ArXiv*, abs/2403.01244, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *ArXiv*, abs/2212.04089, 2022.

Zixuan Ke and Bin Liu. Continual learning of natural language processing tasks: A survey. *ArXiv*, abs/2211.12701, 2022.

J. Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, J. Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114:3521 – 3526, 2016.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and J. Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. *ArXiv*, abs/2502.04959, 2025.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58:1 – 42, 2024.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. Lora vs full fine-tuning: An illusion of equivalence. *ArXiv*, abs/2410.21228, 2024.

George Stoica, Pratik Ramesh, B. Ecsedi, Leshem Choshen, and J. Hoffman. Model merging with svd to tie the knots. *ArXiv*, abs/2410.19735, 2024.

Ivan Vuli'c, Adam Grycner, Quentin de Laroussilhe, and Jonas Pfeiffer. Lora-squeeze: Simple and effective post-tuning and in-tuning compression of lora modules. 2026.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. *ArXiv*, abs/2310.14152, 2023.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey. *ArXiv*, abs/2402.01364, 2024.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. 2023.

Le Yu, Yu Bowen, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. pp. 57755–57775, 2023.

Weixiang Zhao, Shilong Wang, Yulin Hu, Yanyan Zhao, Bing Qin, Xuanyu Zhang, Qing Yang, Dongliang Xu, and Wanxiang Che. Sapt: A shared attention framework for parameter-efficient continual learning of large language models. pp. 11641–11661, 2024.

Yaoyue Zheng, Yin Zhang, J. Weijer, Gido M. van de Ven, Shaoyi Du, Xuetao Zhang, and Zhiqiang Tian. Revisiting weight regularization for low-rank continual learning. 2026.

A Implementation Details

We implement SVD-EQ using PyTorch's `torch.linalg.svd` function. For efficiency, we exploit the low-rank structure of LoRA updates: since $\Delta W = BA$ has rank at most r, we compute QR decompositions of B and $A^\top$ to reduce the SVD to an $r \times r$ core matrix. The computational overhead is negligible compared to training time. All experiments use the FOREVER codebase with default hyperparameters. Code will be released upon publication.