ELLA: Efficient Lifelong Learning for Adapters in Large Language Models

Shristi Das Biswas[*], **Yue Zhang**[2], **Anwesan Pal**[3], **Radhika Bhargava**[2], **Kaushik Roy**[1]
[1] Purdue University, [2] Amazon Web Services, [3] AWS AI Labs
{sdasbisw, kaushik}@purdue.edu; {zhangany, anwesanp@, radhikb}@amazon.com

Abstract

Continual Learning (CL) is a vital requirement for deploying large language models (LLMs) in today's dynamic world. Existing approaches seek to acquire task-specific knowledge via parameter efficient fine-tuning (PEFT) with reduced compute overhead. However, sequential FT often sacrifices performance retention and forward transfer, especially under replay-free constraints. We introduce ELLA, a novel CL framework that regularizes low-rank adapter updates via cross-task subspace de-correlation. By learning a compact adapter per task and penalizing overlap between representational subspaces for past and current adapter activations, ELLA encourages task specialization while preserving prior knowledge, without storing data. Across 3 benchmarks, ELLA outperforms prior CL methods in both accuracy and forgetting metrics, providing a scalable solution for lifelong LLM learning.

1 Introduction

Large Language Models (LLMs) excel in diverse downstream tasks thanks to large-scale pretraining (1; 2; 3), but in real-world deployments, they must sequentially adapt to evolving tasks without full retraining (4). Sequential finetuning, however, suffers from *catastrophic forgetting* (CF) (5) and *loss of plasticity* (6; 7), especially in rehearsal-free settings where past data cannot be stored (8).

Parameter-efficient fine-tuning (PEFT) methods like Low-Rank Adaptation (LoRA) (9) reduce compute overhead by updating only low-rank adapters (10; 11) versus full model. Yet, without replay, sequential adapter training forgets prior tasks (12). Solutions like capacity expansion (13), weight isolation (14; 15; 16), subspace orthogonality (4; 17), or gradient projection (18) reduce forgetting, but block forward transfer, add memory costs, or ignore activation-level interference between tasks (19).

In practice, some overlap between prior tasks is beneficial: low-magnitude directions encode generic linguistic patterns that can accelerate new learning. Yet existing continual learning (CL) methods either eliminate all overlap or rely on expensive and heavyweight fusion mechanisms (20; 21; 17), limiting scalability. We introduce Efficient Lifelong Learning for Adapters (**ELLA**), a lightweight, replay-free CL framework that regularizes new adapters by tracking past representational subspaces in weight space. We penalize high-magnitude alignments with earlier tasks, while allowing safe reuse of low-magnitude generic directions. This cross-task subspace de-correlation preserves useful priors, reduces destructive interference, and requires no task labels, controller networks, or extra storage.

On Standard CL (22), Long Sequence (23), and TRACE Benchmarks (24), ELLA outperforms state-of-the-art methods on T5-Large model (25), without replay or added compute. ELLA seamlessly integrates with instruction-tuned pipelines (26), and further boosts existing CL techniques without

[*]Work conducted at Amazon.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Continual and Compatible Foundation Model Updates (CCFM).

extra supervision. Our main contributions are: (1) We propose **ELLA**: A replay-free, plug-and-play CL framework using subspace-aware regularization of LoRA adapters. (2) **Empirical gains:** State-of-the-art performance on 3 CL benchmarks, substantially reducing CF and enhancing plasticity.

2 Related Works

Continual Learning (CL) seeks to adapt models to non-stationary data streams without forgetting prior tasks. Existing solutions include: (i) *Rehearsal-based* methods that replay or jointly optimize on buffered past examples (27; 14; 8; 28). (ii) *Regularization-based* approaches that penalize updates to weights deemed important for earlier tasks (29; 30; 31), including orthogonal gradient constraints (32). (iii) *Architecture-based* schemes that allocate task-specific modules or expand capacity, e.g. per-task soft prompts (23; 10) or dynamic routing (15; 16).

Parameter-Efficient Fine-Tuning (PEFT) adapts large pre-trained models by tuning only a small subset of parameters. Notable techniques include BitFit (33), prompt tuning (34; 35), LoRA's low-rank adapters (9), and adaptive-rank extensions like AdaLoRA (36). To date, PEFT has been applied to CL via per-task adapters (37), orthogonal LoRA subspaces (4; 18), and multi-adapter fusion with replay (17; 38). However, rehearsal approaches require data storage, while many modular designs grow computationally with the number of tasks or depend on complex fusion/replay.

ELLA overcomes these limitations by self-regularizing LoRA updates to steer new adaptations away from past weight subspaces—no extra data, labels, or modules are needed—thus achieving a lightweight, scalable balance of retention and plasticity.

3 Method

Setup. In supervised continual learning, a model sees tasks $\{\mathcal{T}_1, \ldots, \mathcal{T}_T\}$ in sequence, where each $\mathcal{T}_t = \{(x_i^t, y_i^t)\}_{i=1}^{n_t}$ is a labeled dataset. The goal is to maximize $\max_\Theta \sum_{t=1}^{T} \sum_{(x,y)\in\mathcal{T}_t} \log p_\Theta(y \mid x)$. We study a stricter *rehearsal-free, task-agnostic* setting: during training, no past data may be stored or revisited, and at test time the model must predict without knowing the input's task identity.

Interference in Sequential LoRA Updates. Applying separate LoRA adapters (A_t, B_t) (9) for each incoming task in a continual-learning setup induces *interference*: each new adapter is learned from scratch and can overlap with and overwrite previously learned subspaces, causing catastrophic forgetting. Although the backbone weights W_{init} remain fixed, the cumulative update $\sum_{t=1}^{T} A_t B_t$ behaves like a full-rank modification to the model (39), allowing destructive interactions across tasks.

Orthogonal LoRA and Its Limitations. To prevent such interference, prior works (4; 17; 18; 38) impose orthogonality between adapters by adding the auxiliary loss $\mathcal{L}_{\text{orth}} = \sum_{i=1}^{t-1} \left\| A_i^\top A_t \right\|_F^2$. This projects each new A_t away from all earlier subspaces $\{A_1, \ldots, A_{t-1}\}$, to reduce forgetting on prior tasks. However, strict orthogonality over-regularizes the adapter space – blocking forward transfer, preventing reuse of low-importance components even across related tasks, and wasting limited adapter capacity. Furthermore, storing every past adapter grows memory overhead linearly with task count, undermining scalability, and motivating us to look for alternative solutions.

Subspace-Aware Continual Adaptation. This work proposes a simple yet effective CL framework that balances *plasticity* and *stability* in LoRA updates using a subspace-aware strategy that penalizes interference across tasks while preserving the capacity for forward transfer and learning space reuse. Rather than enforcing hard orthogonality between LoRA updates, ELLA introduces a lightweight regularizer that selectively suppresses reuse of past task-specific directions with high representational energy, while retaining freedom in lower-magnitude spaces (Fig. 1).

Let the LoRA update for task t be $\Delta W_t = A_t B_t$, where $A_t \in \mathbb{R}^{d\times r}$ and $B_t \in \mathbb{R}^{r\times k}$. We construct a cumulative signal from the sum of LoRA-induced weight changes from past tasks as $\mathcal{W}_{\text{past}} = \sum_{i=1}^{t-1} \Delta W_i$. This aggregated update encodes dominant directions in parameter space that have been heavily utilized by previous tasks. Motivated by the observation that high-magnitude LoRA components are typically more task-specific (39), ELLA introduces a space alignment penalty $\mathcal{L}_{\text{ELLA}} = \left\| \Delta W_t * \mathcal{W}_{\text{past}} \right\|_F^2$, where $\|.\|_F$ denotes Frobenius norm. This energy-based alignment penalty discourages new updates from aligning with heavily used, high-importance spaces—those most likely to induce forgetting—while allowing overlap in underutilized low-magnitude directions

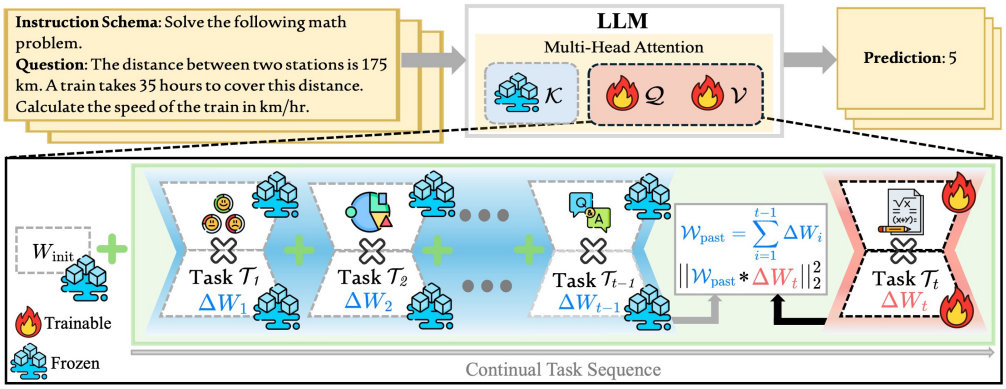

Figure 1: ELLA mitigates interference in continual LoRA training by accumulating past low-rank updates $\mathcal{W}_{\text{past}}$ and applying an energy-based alignment penalty $||\Delta W_t * \mathcal{W}_{\text{past}}||_2^2$ to discourage overlap in high-magnitude, task-specific directions. This enables parameter reuse in less-used subspaces, achieving better plasticity-stability trade-off without task labels, replay, or architectural modifications.

	Methods	Standard CL Benchmark (SC)				Long Sequence Benchmark (LS)				TRACE
		Order 1	Order 2	Order 3	OA	Order 4	Order 5	Order 6	OA	Order 7 (OA)
T5-Large	SeqFT (42)	18.9	24.9	41.7	28.5	7.4	7.3	7.4	7.4	-
	SeqLoRA	39.5	31.9	46.6	39.3	4.9	3.5	4.2	4.2	12.1
	EWC (29)	46.3	45.3	52.1	47.9	44.9	44.0	45.4	44.8	-
	LwF (30)	52.7	52.9	48.4	51.3	49.7	42.8	46.9	46.5	-
	L2P (35)	59.0	60.5	59.9	59.8	57.7	53.6	56.6	56.0	-
	LB-CL (18)	76.9	76.5	76.8	76.7	68.4	67.3	71.8	69.2	-
	O-LoRA (4)	73.5	71.4	70.0	71.6	65.4	65.2	65.2	65.3	23.1
	+ MIGU (31)	77.1	77.0	75.6	76.6	67.3	68.5	74.0	70.0	-
	DATA (17)	71.5	70.5	68.0	70.0	71.5	70.5	68.0	70.0	16.7
	+ Replay	77.0	75.6	75.2	75.9	**75.6**	**73.2**	74.1	74.3	36.5
	LFPT5 (10)	66.6	71.2	76.2	71.3	69.8	67.2	69.2	68.7	-
	SeqLoRAReplay	4.0	73.1	73.0	73.3	74.2	72.7	73.9	73.6	34.0
	Recurrent-KIF (38)	-	-	-	78.4	-	-	-	**77.8**	-
	ELLA (ours)	**80.0**	**80.0**	**79.8**	**79.9**	73.4	72.0	**75.4**	73.6	**40.0**

Table 1: OA comparison across multiple benchmarks and transfer orders. Methods in gray rely on replay mechanisms to boost performance. Best results in **bold** and second best underlined.

that facilitate knowledge reuse. As a result, ELLA achieves a better plasticity-stability balance than methods with strict subspace separation. Finally, the full training objective for task $\mathcal{T}_t$ is $\mathcal{L} = \sum_{(x,y)\in\mathcal{T}_t} \log p_\Theta(y \mid x) + \lambda \cdot \mathcal{L}_{\text{ELLA}}$. Crucially, ELLA is replay-free, task-agnostic, and compatible with any LoRA-pipeline. It adds no extra parameters and merely adjusts the training loss, incurring negligible overhead and making it a scalable solution for CL in pre-trained LMs.

4 Experiments

Datasets and Implementation Details. We train and evaluate on 3 popular benchmarks, Standard CL Benchmark (22), Long Sequence Benchmark (23) and TRACE (24).

Metrics. Let $a_{i,j}$ denote the testing performance on the j-th task after training on the i-th task. We evaluate across: **Overall Accuracy (OA)** (40): The average accuracy across all tasks after training on the last task, i.e., $\text{OA}_{\mathcal{T}} = \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} a_{\mathcal{T},t}$; **Backward Transfer (BWT)** (41): measures how much the learning of subsequent tasks influences the performance of previous tasks, i.e., $\text{BWT}_{\mathcal{T}} = \frac{1}{\mathcal{T}-1} \sum_{t=1}^{\mathcal{T}-1} (a_{\mathcal{T},t} - a_{t,t})$.

Results. To demonstrate the effectiveness of the proposed method, we perform experiments on three CL benchmarks, as summarized in Table 1. ELLA consistently sets a new state-of-the-art in replay-free continual learning for LLMs, outperforming both traditional (LoRAReplay, (29; 30)) and modern baselines (4; 38; 17; 31; 18) on all orders. On T5-Large, ELLA delivers an average accuracy of 79.9 on Standard CL, surpassing the previous best replay-free method (18) and exceeding even the top replay-based methods (17; 38). On the LS and TRACE settings, ELLA continues to lead, increasing the average by 3.6 and 23.3 respectively compared to (17).

Method	Trainable Params	Storage (MB)	Replay	Time/Epoch (mins)
SeqLoRA	0.062	0	0	4
O-LoRA	0.062	31.46	0	4.5
ELLA (Ours)	0.062	4.19	0	4.5
SeqLoRAReplay	0.062	0	2%	4
DATA	0.369	147.46	2%	6.5

Table 2: Comparison of training and memory overheads.

LoRA_dim	Order1	Order2	Order3	Avg
2	72.29	74.00	77.08	74.46
4	73.22	75.15	77.72	75.36
8	**79.95**	**80.00**	**79.82**	**79.92**
16	77.38	77.65	76.19	77.07

Table 3: Impact of LoRA rank on CL.

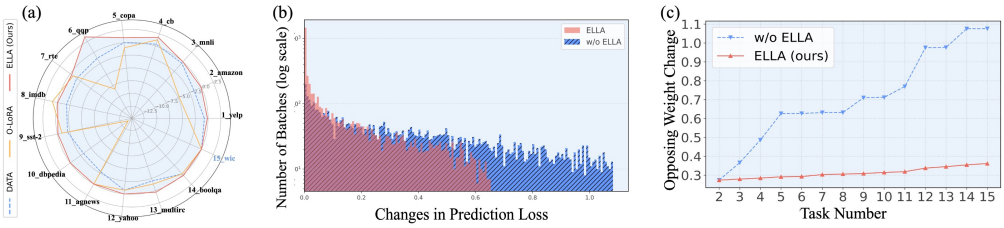

Figure 2: (a) We demonstrate stronger resistance to performance decline (BWT) than baselines (higher values indicate better prior-task retention), with the final task marked in blue. (b) Histogram of prediction loss changes after training on a new task. ELLA constraints reduce the loss of previous tasks compared to when it is not present ($\lambda = 0$). (c) Opposing direction weight change across a task sequence. ELLA consistently reduces backward-conflicting updates, promoting stable CL.

Crucially, Fig. 2(a) demonstrates that ELLA achieves the highest BWT (minimal forgetting), outperforming all prior baselines in both metrics, when evaluated across different task orders. Fig. 2(a) provides a fine-grained view of per-task performance, highlighting ELLA's robustness on tasks that are especially sensitive to interference and forgetting, such as DBPedia and QQP. While previous methods often degrade sharply on these challenging tasks, ELLA maintains high performance throughout. This advantage is consistent across both short and long task sequences, as well as across transfer orders, confirming the strength of our cross-task subspace decorrelation mechanism.

Efficiency Analysis. As shown in Table 2, ELLA maintains the same number of trainable parameters as O-LoRA while significantly reducing storage overhead to just 4.19MB since it does need access to all past LoRA parameters, causing memory requirements to scale with task sequence length. Unlike methods such as DATA and SeqLoRAReplay, ELLA requires no replay buffer or feature storage, and incurs minimal runtime cost, achieving high achieve high training and inference efficiency.

5 Discussions

Does ELLA preserve previous task performance during CL? We track the change in prediction loss on past-task batches after learning each new task. Fig. 2(b) shows that ELLA significantly reduces the number of batches experiencing large increases in loss, especially in the high-loss tail region, indicating its alignment penalty effectively preserves useful gradients from earlier tasks. In contrast, w/o ELLA ($\lambda = 0$) exhibits a broader distribution of loss spikes, revealing greater CF.

Directional Consistency of Updates Over Task Sequence. In Fig. 2(c), we study opposing-direction weight changes, i.e. those that reverse prior updates, after each task. Standard LoRA exhibits large opposing updates, indicating disruption of earlier representations, whereas ELLA substantially reduces these reversals, enabling smoother and more stable knowledge accumulation across tasks.

Studying Optimal LoRA Rank for Plasticity-Stability Tradeoff. To assess how the LoRA rank r affects CL, we evaluated ELLA across task orders varying r. As shown in Table 3, accuracy improves up to $r = 8$ and then decreases at $r = 16$. Very low rank (e.g. $r = 2$) lacks plasticity, while very high rank (e.g. $r = 16$) leads to overfitting on individual tasks. Therefore, a moderate rank ($r = 8$) provides the best trade-off between learning new tasks and preserving prior knowledge.

6 Conclusion

In this work, we introduced **ELLA**, a simple yet effective approach for continual customization of LLMs without using task identifiers or replay. Unlike prior methods that rely on strict orthogonality, ELLA encourages de-alignment between new updates and the accumulated subspace of prior LoRA

directions, mitigating destructive weight drift and allows beneficial reuse of underutilized directions, preserving performance across a CL sequence. Our extensive experiments across multiple benchmarks demonstrate that ELLA consistently improves both stability and knowledge transfer while remaining parameter- and memory-efficient, outperforming state-of-the-art. These results highlight ELLA's practical promise as a lightweight and scalable universal method for lifelong adaptation in LLMs.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[3] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[4] X. Wang, T. Chen, Q. Ge, H. Xia, R. Bao, R. Zheng, Q. Zhang, T. Gui, and X. Huang, "Orthogonal subspace learning for language model continual learning," *arXiv preprint arXiv:2310.14152*, 2023.

[5] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*, vol. 24, pp. 109–165, Elsevier, 1989.

[6] S. Dohare, R. S. Sutton, and A. R. Mahmood, "Continual backprop: Stochastic gradient descent with persistent randomness," *arXiv preprint arXiv:2108.06325*, 2021.

[7] P. Ruvolo and E. Eaton, "Ella: An efficient lifelong learning algorithm," in *International conference on machine learning*, pp. 507–515, PMLR, 2013.

[8] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. Dokania, P. Torr, and M. Ranzato, "Continual learning with tiny episodic memories," in *Workshop on Multi-Task and Lifelong Reinforcement Learning*, 2019.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models.," *ICLR*, vol. 1, no. 2, p. 3, 2022.

[10] C. Qin and S. Joty, "Lfpt5: A unified framework for lifelong few-shot language learning based on prompt tuning of t5," *arXiv preprint arXiv:2110.07298*, 2021.

[11] C. Song, X. Han, Z. Zeng, K. Li, C. Chen, Z. Liu, M. Sun, and T. Yang, "Conpet: Continual parameter-efficient tuning for large language models," *arXiv preprint arXiv:2309.14763*, 2023.

[12] Y. Wang, Y. Liu, C. Shi, H. Li, C. Chen, H. Lu, and Y. Yang, "Inscl: A data-efficient continual learning paradigm for fine-tuning large language models with instructions," *arXiv preprint arXiv:2403.11435*, 2024.

[13] M. Wang, H. Adel, L. Lange, J. Strötgen, and H. Schütze, "Rehearsal-free modular and compositional continual learning for language models," *arXiv preprint arXiv:2404.00790*, 2024.

[14] R. Aljundi, P. Chakravarty, and T. Tuytelaars, "Expert gate: Lifelong learning with a network of experts," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3366–3375, 2017.

[15] X. Li, Y. Zhou, T. Wu, R. Socher, and C. Xiong, "Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting," in *International conference on machine learning*, pp. 3925–3934, PMLR, 2019.

[16] Z. Wang, Y. Liu, T. Ji, X. Wang, Y. Wu, C. Jiang, Y. Chao, Z. Han, L. Wang, X. Shao, *et al.*, "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)*, pp. 10933–10946, 2023.

[17] H. Liao, S. He, Y. Hao, J. Zhao, and K. Liu, "Data: Decomposed attention-based task adaptation for rehearsal-free continual learning," *arXiv preprint arXiv:2502.11482*, 2025.

[18] F. Qiao and M. Mahdavi, "Learn more, but bother less: parameter efficient continual learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 97476–97498, 2024.

[19] Z. Ke, B. Liu, N. Ma, H. Xu, and L. Shu, "Achieving forgetting prevention and knowledge transfer in continual learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22443–22456, 2021.

[20] W. Zhao, S. Wang, Y. Hu, Y. Zhao, B. Qin, X. Zhang, Q. Yang, D. Xu, and W. Che, "Sapt: A shared attention framework for parameter-efficient continual learning of large language models," *arXiv preprint arXiv:2401.08295*, 2024.

[21] J. Liu, S. Yang, P. Jia, R. Zhang, M. Lu, Y. Guo, W. Xue, and S. Zhang, "Vida: Homeostatic visual domain adapter for continual test time adaptation," *arXiv preprint arXiv:2306.04344*, 2023.

[22] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

[23] A. Razdaibiedina, Y. Mao, R. Hou, M. Khabsa, M. Lewis, and A. Almahairi, "Progressive prompts: Continual learning for language models," *arXiv preprint arXiv:2301.12314*, 2023.

[24] X. Wang, Y. Zhang, T. Chen, S. Gao, S. Jin, X. Yang, Z. Xi, R. Zheng, Y. Zou, T. Gui, *et al.*, "Trace: A comprehensive benchmark for continual learning in large language models," *arXiv preprint arXiv:2310.06762*, 2023.

[25] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[26] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap, *et al.*, "Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks," *arXiv preprint arXiv:2204.07705*, 2022.

[27] M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, and G. Tesauro, "Learning to learn without forgetting by maximizing transfer and minimizing interference," *arXiv preprint arXiv:1810.11910*, 2018.

[28] J. He, H. Guo, K. Zhu, Z. Zhao, M. Tang, and J. Wang, "Seekr: Selective attention-guided knowledge retention for continual learning of large language models," *arXiv preprint arXiv:2411.06171*, 2024.

[29] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[30] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[31] W. Du, S. Cheng, T. Luo, Z. Qiu, Z. Huang, K. C. Cheung, R. Cheng, and J. Fu, "Unlocking continual learning abilities in language models," *arXiv preprint arXiv:2406.17245*, 2024.

[32] M. Farajtabar, N. Azizan, A. Mott, and A. Li, "Orthogonal gradient descent for continual learning," in *International conference on artificial intelligence and statistics*, pp. 3762–3773, PMLR, 2020.

[33] E. B. Zaken, S. Ravfogel, and Y. Goldberg, "Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," *arXiv preprint arXiv:2106.10199*, 2021.

[34] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.

[35] Z. Wang, Z. Zhang, C.-Y. Lee, H. Zhang, R. Sun, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister, "Learning to prompt for continual learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 139–149, 2022.

[36] Q. Zhang, M. Chen, A. Bukharin, N. Karampatziakis, P. He, Y. Cheng, W. Chen, and T. Zhao, "Adalora: Adaptive budget allocation for parameter-efficient fine-tuning," *arXiv preprint arXiv:2303.10512*, 2023.

[37] A. Madotto, Z. Lin, Z. Zhou, S. Moon, P. Crook, B. Liu, Z. Yu, E. Cho, and Z. Wang, "Continual learning in task-oriented dialogue systems," *arXiv preprint arXiv:2012.15504*, 2020.

[38] Y. Feng, X. Wang, Z. Lu, S. Fu, G. Shi, Y. Xu, Y. Wang, P. S. Yu, X. Chu, and X.-M. Wu, "Recurrent knowledge identification and fusion for language model continual learning," *arXiv preprint arXiv:2502.17510*, 2025.

[39] A. Aghajanyan, L. Zettlemoyer, and S. Gupta, "Intrinsic dimensionality explains the effectiveness of language model fine-tuning," *arXiv preprint arXiv:2012.13255*, 2020.

[40] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr, "Riemannian walk for incremental learning: Understanding forgetting and intransigence," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 532–547, 2018.

[41] Z. Ke and B. Liu, "Continual learning of natural language processing tasks: A survey," *arXiv preprint arXiv:2211.12701*, 2022.

[42] C. de Masson D'Autume, S. Ruder, L. Kong, and D. Yogatama, "Episodic memory in lifelong language learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.