amazon | science

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

NEURAL INFORMATION PROCESSING SYSTEMS

ELLA: Efficient Lifelong Learning for Adapters in Large Language Models

Shristi Das Biswas

Yue Zhang

Anwesan Pal

Radhika Bhargava

Kaushik Roy

Project Page: https://sites.google.com/view/ella-llm/home

NeurIPS 2025 Workshop on Continual and Compatible Foundation Model Updates (CCFM)

Contents

amazon | science

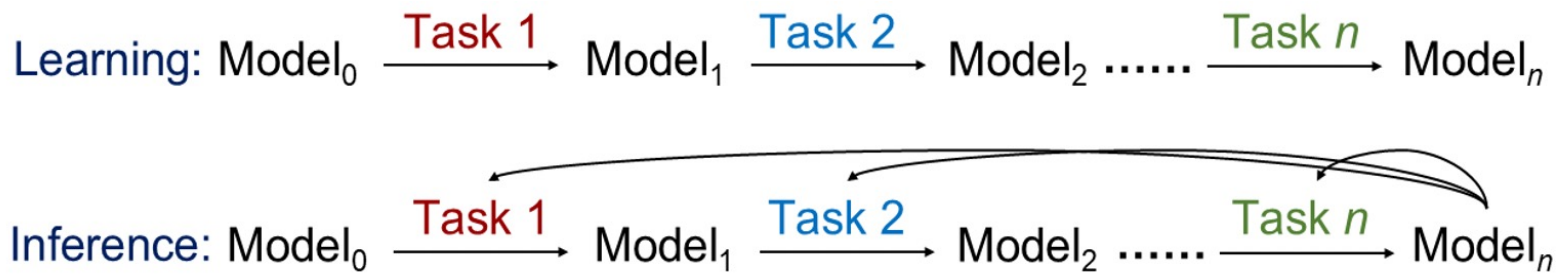
Introducing the Task

PURDUE UNIVERSITY®

Elmore Family School of Electrical and Computer Engineering

What Is The Task of Continual Learning?

Continual Learning (CL) seeks to adapt models to non-stationary data streams without forgetting prior tasks. Ideally there should no need to store prior data.



Catastrophic Forgetting: When minimizing the loss $\mathcal{L}(x_j, y_j)$ for a new task j, it will cause the increase of the loss $\mathcal{L}(x_i, y_i)$ for a previous task i.

- Sequential LoRA updates ($\Delta W_t = \mathcal{A}_t\mathcal{B}_t$) when training for a new task tend to align with the dominant spaces of accumulated past updates ($\Sigma\mathcal{A}\mathcal{B}$) from old tasks, and this directional overlap ($\langle \Delta W_t, \Sigma\mathcal{A}\mathcal{B} \rangle_F$) drives catastrophic forgetting.

Limitations of Prior Work

Existing solutions include:

- Rehearsal-based methods that replay or jointly optimize on buffered past examples.
- Regularization-based approaches that penalize updates to weights deemed important for earlier tasks, including orthogonal gradient constraints.
- Architecture-based schemes that allocate task-specific modules or expand capacity, e.g. per-task soft prompts or dynamic routing.

However,

- **Replay Methods** are impractical, violating privacy and incurring massive storage costs.
- **Orthogonality Methods** are too strict. By forbidding *all* overlap, preventing **forward transfer.**
- **Architecture-expansion Methods** typically results in a linear growth in parameter count with the task count, posing scalability challenges and requiring explicit task labels at inference.

amazon | science

Method Overview

PURDUE UNIVERSITY

Elmore Family School of Electrical and Computer Engineering

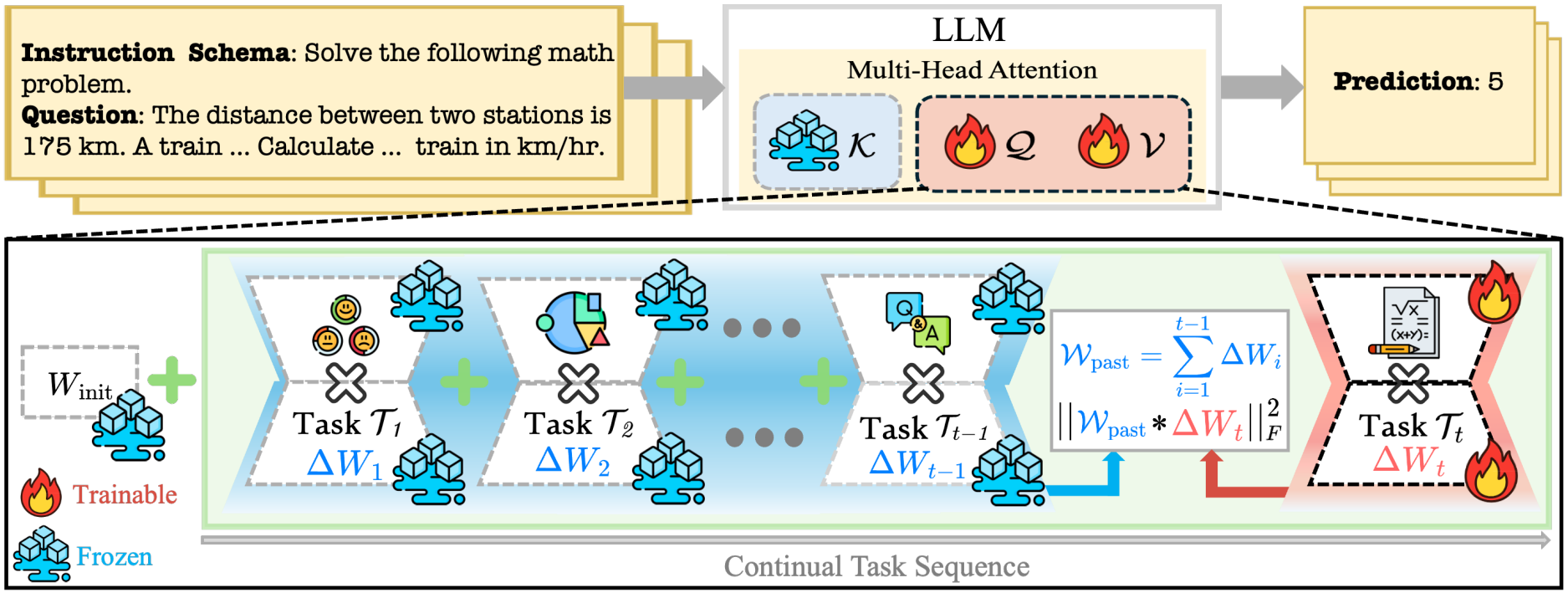
Key Insights

- Interference is anisotropic: High-energy spaces are task-discriminative, hence updating them is harmful to the tasks' performance, low-energy components are often generic and transferable.

- Thereby, our key insight is that forgetting isn't caused by *all* overlap, only by *destructive interference* with high-energy, task-specific directions from past tasks.

- Conversely, overlap in low-energy, general subspaces is *constructive reuse*—this is precisely what forward transfer is, and we want to keep it.

ELLA's Lifelong Adaptation Framework



- 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.

Formal Characterization of ELLA

The solution $\Delta W^{\star}{}_{t}$ to the ELLA objective has the following properties:

- It is an anisotropic shrinkage operator applied to the unconstrained step G, with the closed- form solution:

$$(\Delta W_t^\star)_{ij} = \frac{G_{ij}}{1 + \lambda E_{ij}^2}$$

- The interference with past updates, measured by the inner product $\langle \Delta W_t * \mathcal{W}_{\text{past}} \rangle_F$, is bounded as follows:

$$\left|\langle \Delta W_t^\star, \mathcal{W}_{past} \rangle_F\right| \leq \frac{\|G\|_F}{2\sqrt{\lambda}} \, \|E^{-1} \odot \mathcal{W}_{past}\|_F$$

amazon | science

Evaluations

PURDUE UNIVERSITY®

Elmore Family School of Electrical and Computer Engineering

Experimental Results

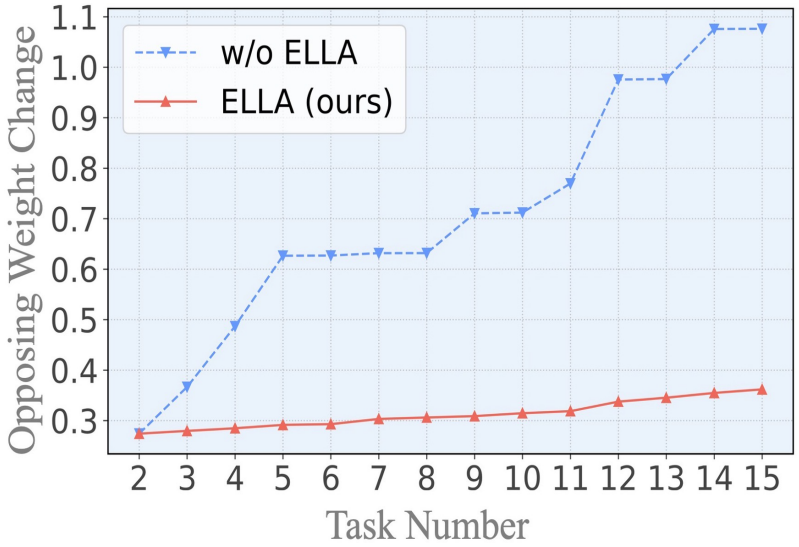
- We train and evaluate on 3 popular benchmarks, Standard CL Benchmark, Long Sequence Benchmark and TRACE.

Let $a_{i,j}$ denote the testing performance on the j-th task after training on the i-th task. We evaluate across:

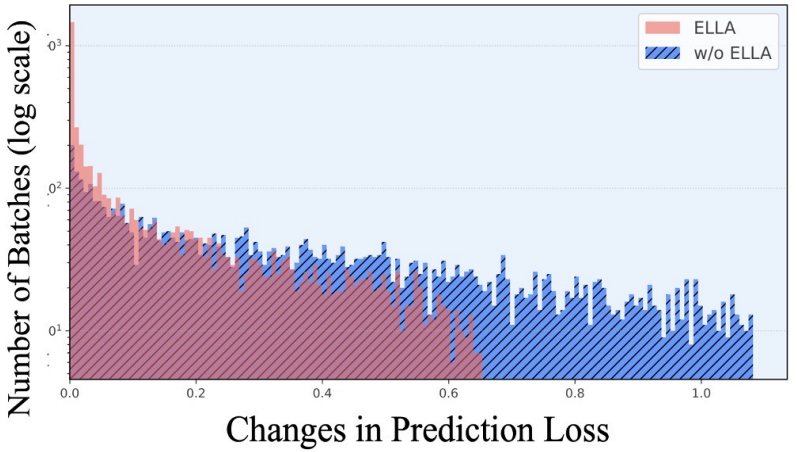
- Overall Accuracy (OA): The average accuracy across all tasks after training on the last task, i.e., $\text{OA}_{\mathcal{T}} = \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} a_{\mathcal{T},t}$
- Backward Transfer (BWT): How much does learning new tasks influence performance of prior tasks, i.e., $\text{BWT}_{\mathcal{T}} = \frac{1}{\mathcal{T}-1} \sum_{t=1}^{\mathcal{T}-1} (a_{\mathcal{T},t} - a_{t,t})$

	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**

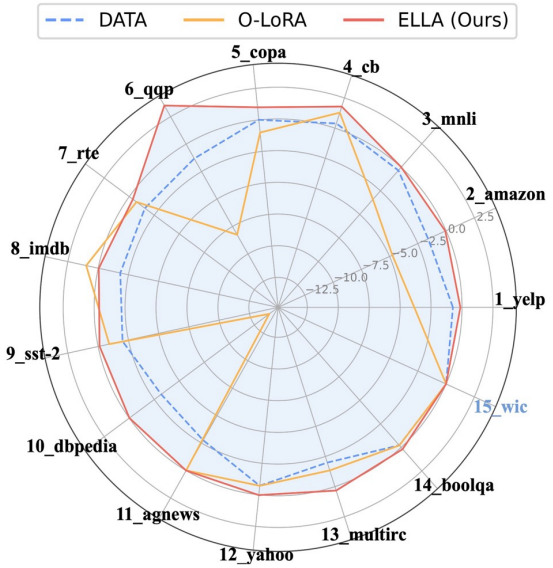
Probing into ELLA



Histogram of prediction loss changes after training on a new task. The ELLA constraint helps reduce the changes - preserve the loss of previous tasks - in comparison to when it is not present.



Opposing direction weight change across task sequence. ELLA consistently reduces backward-conflicting updates, promoting stable continual adaptation.



Performance impact on Order 4 in terms of BWT. ELLA has superior resistance to performance decline than baselines (higher values indicate better retention of prior task performance).

amazon | science

Ablating Insights

PURDUE UNIVERSITY

Elmore Family School of Electrical and Computer Engineering

Scalability of ELLA and, Analyzing the Impact of LoRA ranks on Learning Dynamics

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

ELLA achieves strong efficiency in memory, compute, and training time, promising scalability.

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

Impact of LoRA rank on SC Benchmark. Moderate ranks balance plasticity vs stability during learning.

amazon | science

Takeaways

PURDUE UNIVERSITY

Elmore Family School of Electrical and Computer Engineering

Concluding Statements

- 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.

amazon | science

Thank You!

NEURAL INFORMATION PROCESSING SYSTEMS

PURDUE UNIVERSITY

Elmore Family School of Electrical and Computer Engineering

sdasbisw@purdue.edu