中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES
智能算法安全全国重点实验室
State Key Laboratory of AI Safety

Fine-tuning Done Right in Model Editing

Wrong pipeline, wrong conclusion.

Wanli Yang

April 15, 2026

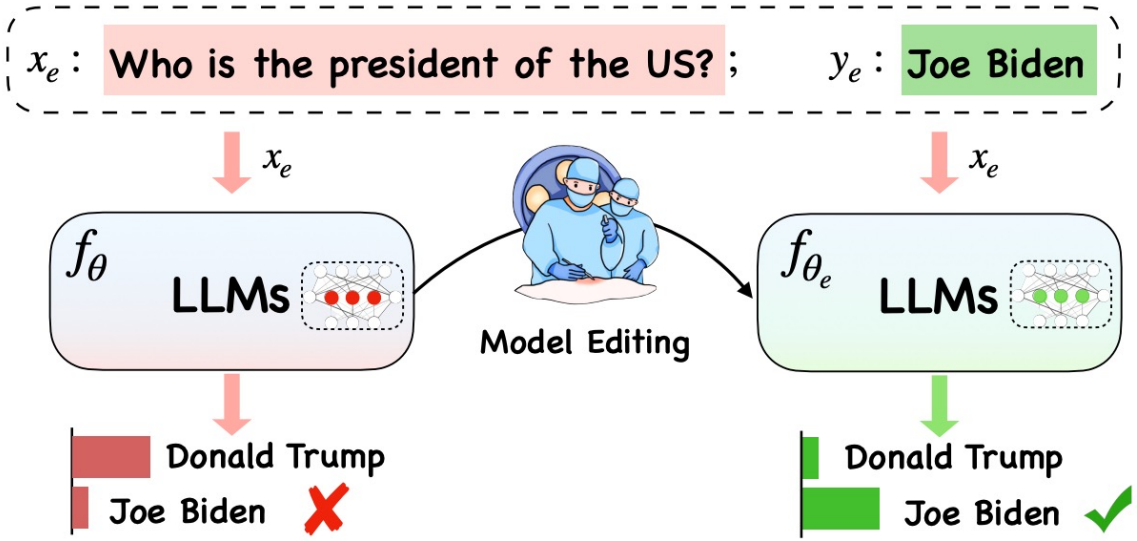
ICLR

Table of Contents

Background: Model Editing

Pre-trained LLMs suffer from **outdated knowledge** and **persistent hallucinations**.

- Retraining: time-consuming and resource-intensive
- Fine-tuning: risks of catastrophic forgetting (**remains a subject of debate**)
- **Model Editing**: precise and efficient updates to **local parameters** for targeted knowledge



Editing Large Language Models: Problems, Methods, and Opportunities (EMNLP 2023)

Background: Effectiveness of FT

- Development of specialized editing algorithms is based on the **ineffectiveness of fine-tuning**.

Methods	ZsRE									Time↓
	LLaMA-3-8B			Gemma-2-9B			Mistral-7B-v0.3			
	Eff.	Gen.	Spe.	Eff.	Gen.	Spe.	Eff.	Gen.	Spe.	
FT	17.10±0.22	16.73±0.22	8.27±0.13	12.90±0.20	13.09±0.20	0.07±0.02	32.84±0.30	33.78±0.30	**42.19±0.31**	0.3366s
ROME	0.54±0.04	0.57±0.04	0.40±0.02	3.45±0.32	3.33±0.11	8.24±0.29	0.00±0.00	0.00±0.00	0.00±0.00	2.5621s
MEMIT	0.00±0.00	0.00±0.00	0.13±0.02	5.23±0.21	3.11±0.12	5.98±0.12	0.00±0.00	0.00±0.00	0.13±0.02	6.0677s
PRUNE	12.27±0.43	12.01±0.23	9.88±0.29	10.21±0.27	8.88±0.29	11.95±0.55	0.00±0.00	0.00±0.00	0.00±0.00	6.1588s
RECT	11.05±0.41	8.12±0.15	28.12±0.13	12.45±0.45	10.11±0.51	26.09±0.44	8.18±0.33	8.04±0.46	11.32±0.49	6.6558s
AlphaEdit	86.83±0.23	81.48±0.28	29.09±0.22	81.18±0.33	73.24±0.46	30.34±0.19	0.00±0.00	0.00±0.00	0.00±0.00	6.1831s
MEND	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	0.9686s
MEND*	0.00±0.00	0.00±0.00	0.00±0.00	8.84±0.24	8.45±0.21	10.21±0.19	0.00±0.00	0.00±0.00	0.00±0.00	5.9265s
MALMEN	9.87±0.12	9.00±0.09	2.11±0.15	46.60±0.33	42.50±0.32	19.66±0.35	0.00±0.00	0.00±0.00	0.00±0.00	2.2779s
MALMEN*	12.23±0.11	11.08±0.22	2.43±0.09	56.97±0.24	44.28±0.29	15.02±0.21	0.01±0.01	0.02±0.01	1.25±0.04	9.4277s
DAFNet	21.99±0.47	11.17±0.43	32.21±0.39	5.94±0.24	5.68±0.33	36.29±0.45	1.25±0.08	2.12±0.12	25.27±0.54	8.2383s
RLEdit	**89.42±0.34**	**87.32±0.23**	**44.78±0.50**	**84.37±0.22**	**79.82±0.26**	**37.15±0.41**	**71.12±0.31**	**67.42±0.27**	27.43±0.44	**0.2224s**

Fine-tuning in RLEdit (ICML 2025).

- Literature suggests that even **localized fine-tuning** yields suboptimal performance in editing tasks.

Method	Model	Counterfact					ZsRE		
		Eff.↑	Gen.↑	Spe.↑	Flu.↑	Consis.↑	Eff.↑	Gen.↑	Spe.↑
Pre-edited	LLaMA3	7.85±0.26	10.58±0.26	89.48±0.18	635.23±0.11	24.14±0.08	36.99±0.30	36.34±0.30	31.89±0.22
FT		83.33±0.37	67.79±0.40	46.63±0.37	233.72±0.22	8.77±0.05	30.48±0.26	30.22±0.32	15.49±0.17
MEND		63.24±0.31	61.17±0.36	45.37±0.38	372.16±0.80	4.21±0.05	0.91±0.05	1.09±0.05	0.53±0.02
InstructEdit		66.58±0.24	64.18±0.35	47.14±0.37	443.85±0.78	7.28±0.04	1.58±0.04	1.36±0.08	1.01±0.05
ROME		64.40±0.41	61.42±0.42	49.44±0.38	449.06±0.26	3.31±0.02	2.01±0.07	1.80±0.07	0.69±0.03
MEMIT		65.65±0.47	64.65±0.42	51.56±0.38	437.43±1.67	6.58±0.11	34.62±0.36	31.28±0.34	18.49±0.19
PRUNE		68.25±0.46	64.75±0.41	49.82±0.36	418.03±1.52	5.90±0.10	24.77±0.27	23.87±0.27	20.69±0.23
RECT		66.05±0.47	63.62±0.43	61.41±0.37	526.62±0.44	20.54±0.09	86.05±0.23	80.54±0.27	31.67±0.22
AlphaEdit		**98.90±0.10**	**94.22±0.19**	**67.88±0.29**	**622.49±0.16**	**32.40±0.11**	**94.47±0.13**	**91.13±0.19**	**32.55±0.22**

Fine-tuning in AlphaEdit (ICLR 2025).

Research Questions

- **Why** does fine-tuning **fail** in model editing?
- **Can** fine-tuning effectively **solve** model editing tasks?

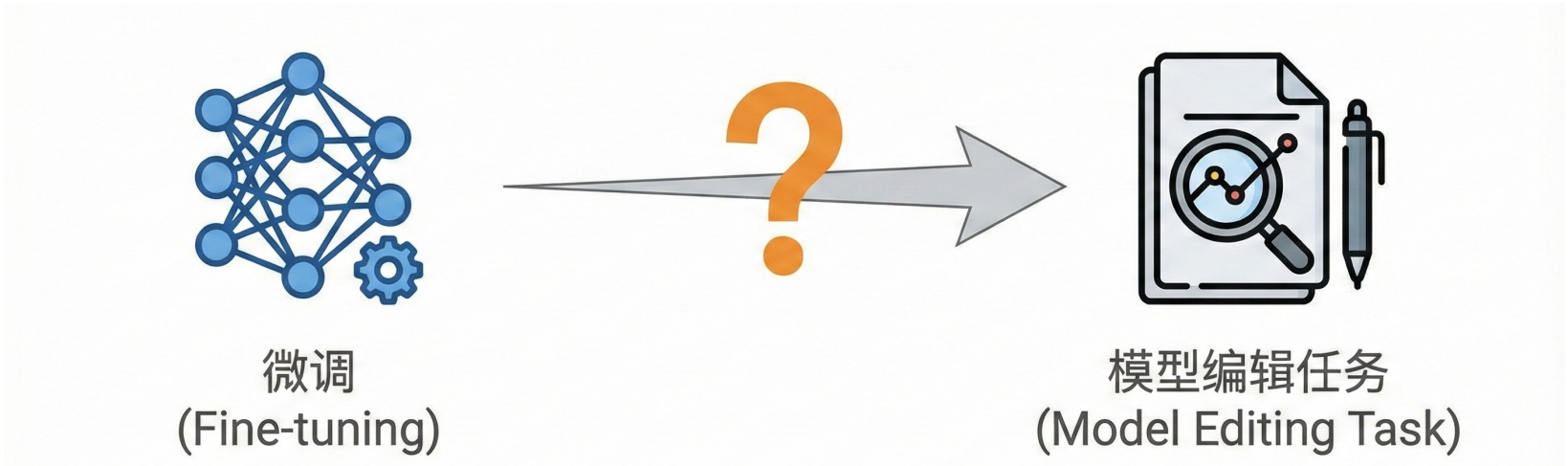
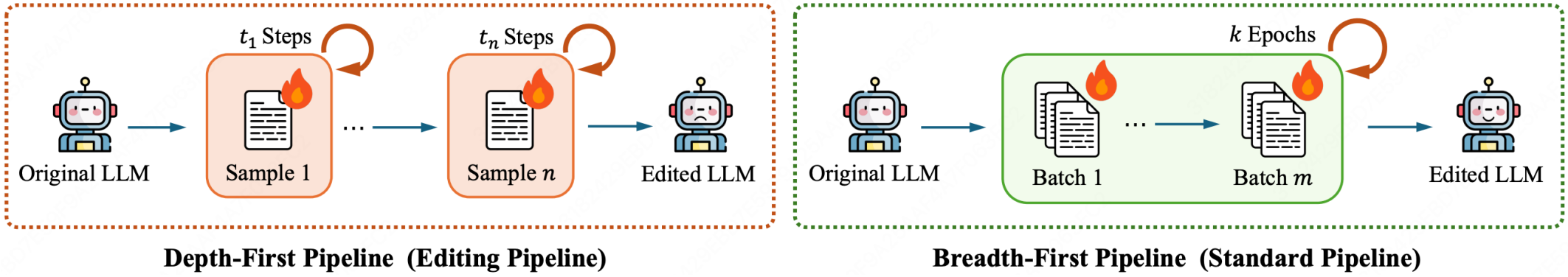


Table of Contents

Implementation Matters in FT-based Editing

Discrepancy in Training Pipelines

- Current Practice: each **individual** sample is **trained to full convergence** before moving to the next.
- Standard Fine-tuning: **updates** are performed across the **full data distribution** over multiple epochs.

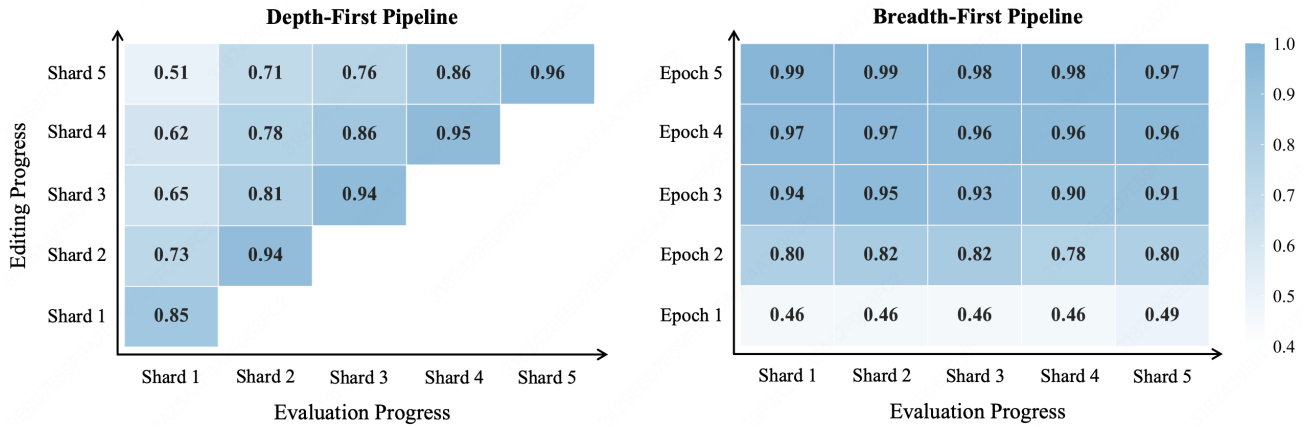


Comparison of training pipeline: Depth-First (DF) vs. Breadth-First (BF)

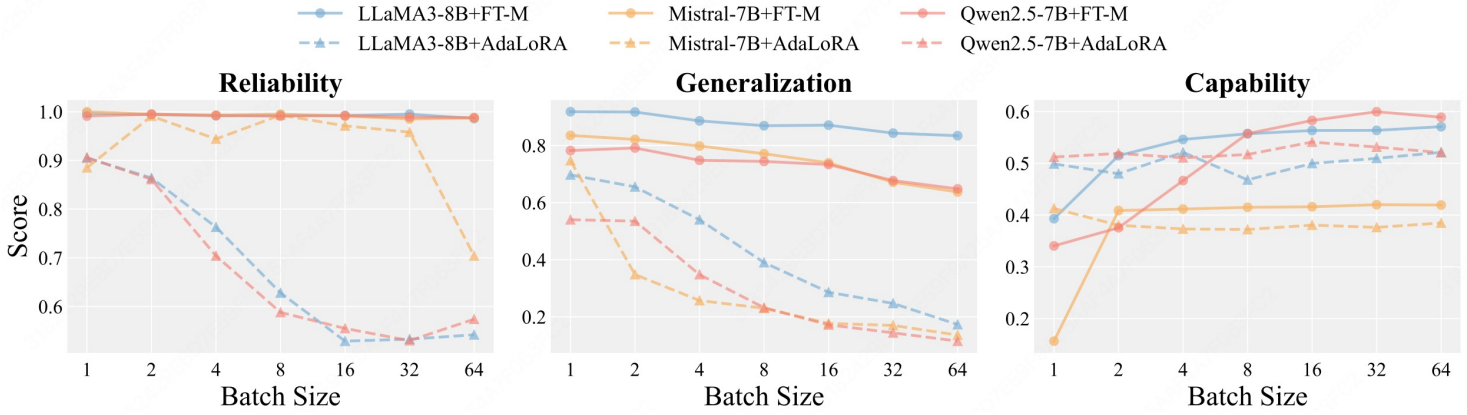
Implementation Matters in FT-based Editing

Impact of Flawed Training Pipelines

- **Overwriting**: sequential learning cause later edits to interfere with prior ones.
- **Instability**: single-sample optimization is highly unstable, compromising model's general knowledge and capabilities.



Visualization of the learning dynamics for DF and BF pipelines

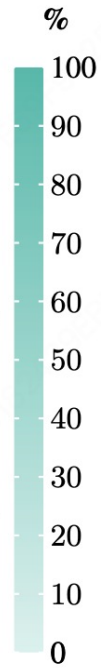


Impact of batch size on editing performance under the BF pipeline

Implementation Matters in FT-based Editing

Rectifying the training pipeline **significantly boosts** **fine-tuning performance**.

	Method	ZsRE						CounterFact					
		Reliability		Generalization		Capability		Reliability		Generalization		Capability	
		DF	BF	DF	BF	DF	BF	DF	BF	DF	BF	DF	BF
LLaMA	FT-L	0.00	0.00	0.00	0.10	14.96	23.47	0.00	0.00	0.00	0.00	15.14	18.17
	FT-M	75.30	99.70	67.20	91.80	28.30	39.30	80.00	99.90	52.60	75.10	29.89	30.87
	AdaLoRA	3.80	90.50	3.30	69.70	44.81	49.89	8.40	96.30	6.10	44.10	49.88	39.27
	RoseLoRA	0.30	1.00	0.00	0.70	57.13	57.38	0.30	0.20	0.00	0.00	57.17	56.66
Mistral	FT-L	0.00	0.00	0.00	0.00	15.73	15.38	0.00	0.00	0.00	0.00	14.93	15.34
	FT-M	41.10	100.00	24.60	83.50	18.14	15.64	59.70	99.90	25.60	59.10	16.94	20.50
	AdaLoRA	2.50	88.50	2.40	74.70	25.36	41.27	5.10	95.50	3.90	44.70	19.14	39.63
	RoseLoRA	0.20	4.10	0.00	3.60	45.24	43.80	0.60	0.80	0.20	0.20	45.57	43.97
Qwen	FT-L	0.10	0.10	0.00	0.10	23.65	29.66	0.10	0.60	0.10	0.50	31.21	33.07
	FT-M	58.70	99.80	34.60	77.60	25.41	34.28	67.70	99.90	25.60	35.40	32.57	33.17
	AdaLoRA	3.40	90.60	2.50	54.00	53.47	51.21	8.90	97.00	4.00	19.10	41.74	53.23
	RoseLoRA	0.00	0.40	0.00	0.10	58.73	58.58	0.20	0.40	0.10	0.10	58.48	58.37



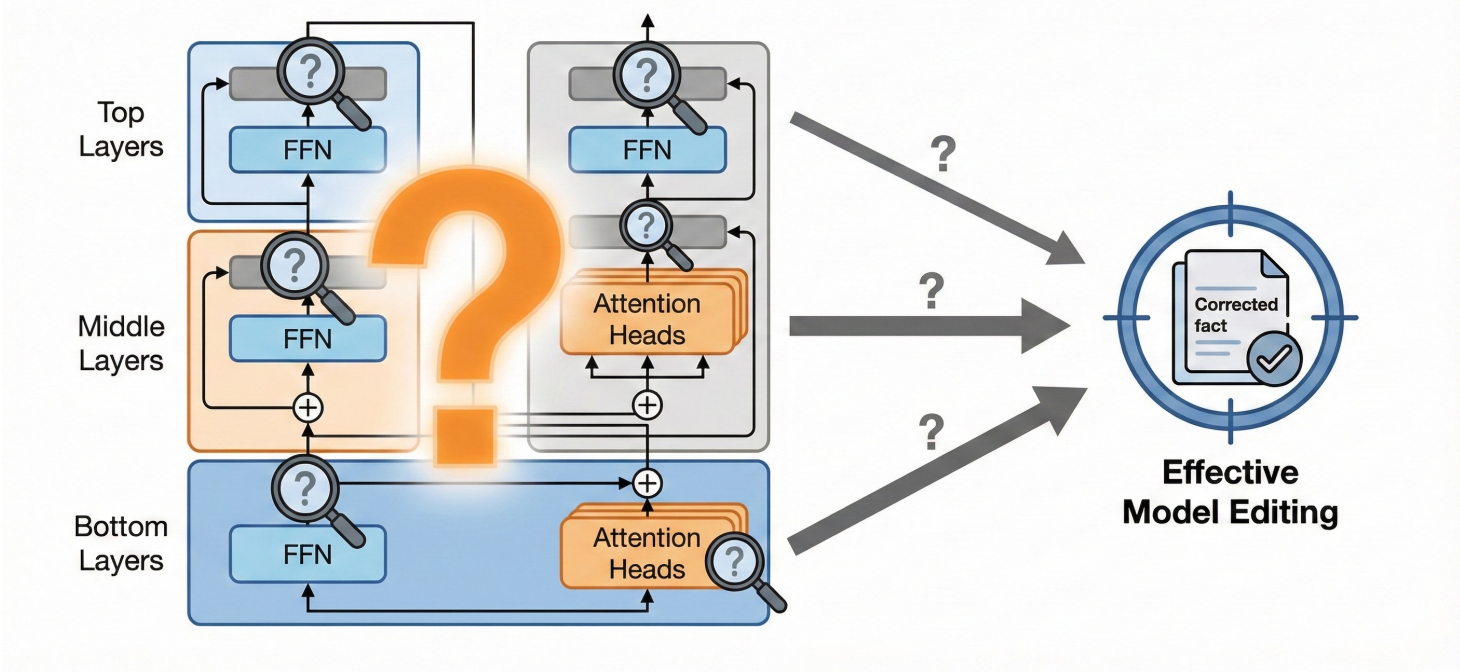
Performance comparison of DF and BF pipelines on mainstream fine-tuning based editors.

Table of Contents

Tailoring Fine-tuning for Model Editing

- **Localized fine-tuning** emerges as a **superior** approach once the training pipeline is rectified.
- However, current local update strategies predominantly rely on **"locate-then-edit" heuristics**.

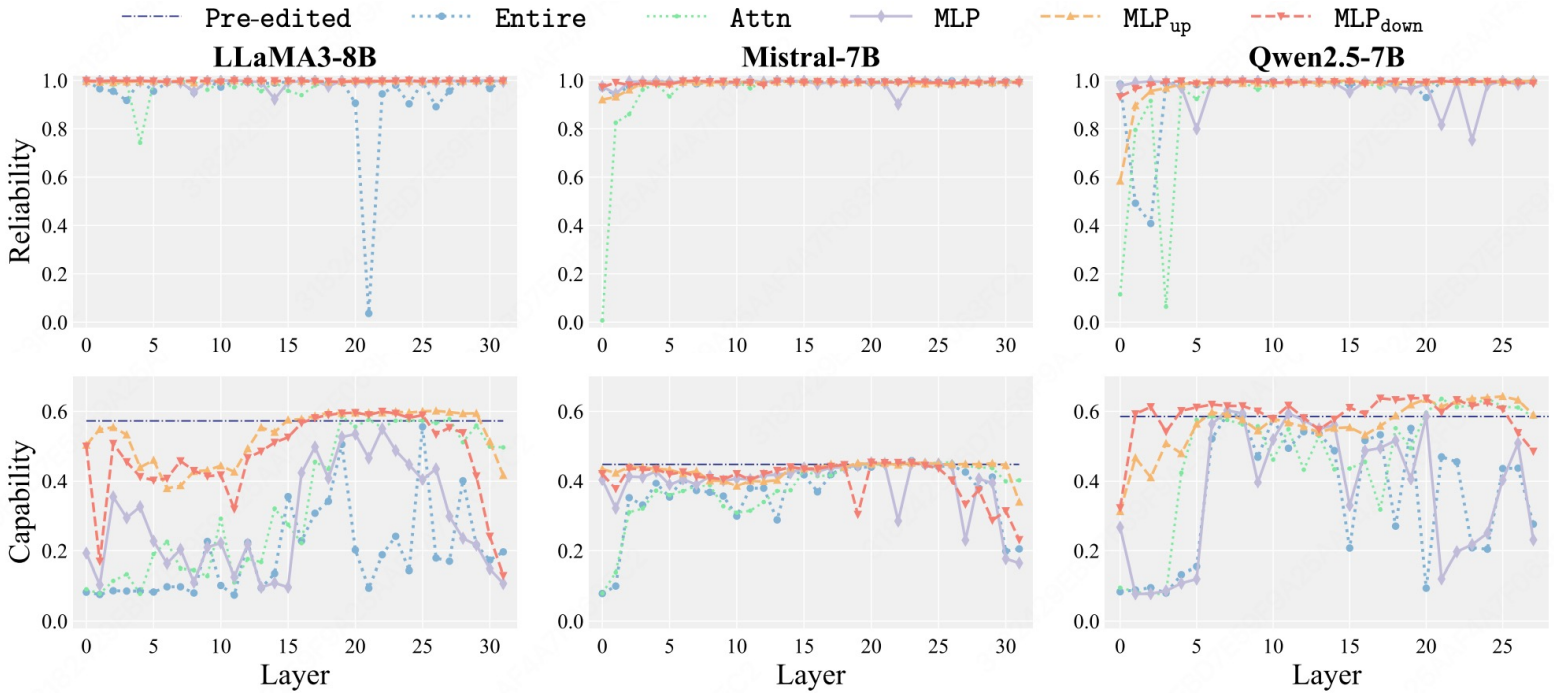
Which regions are more effective for fine-tuning in model editing tasks?



Tailoring Fine-tuning for Model Editing

Comprehensive Evaluation of Candidate Modules

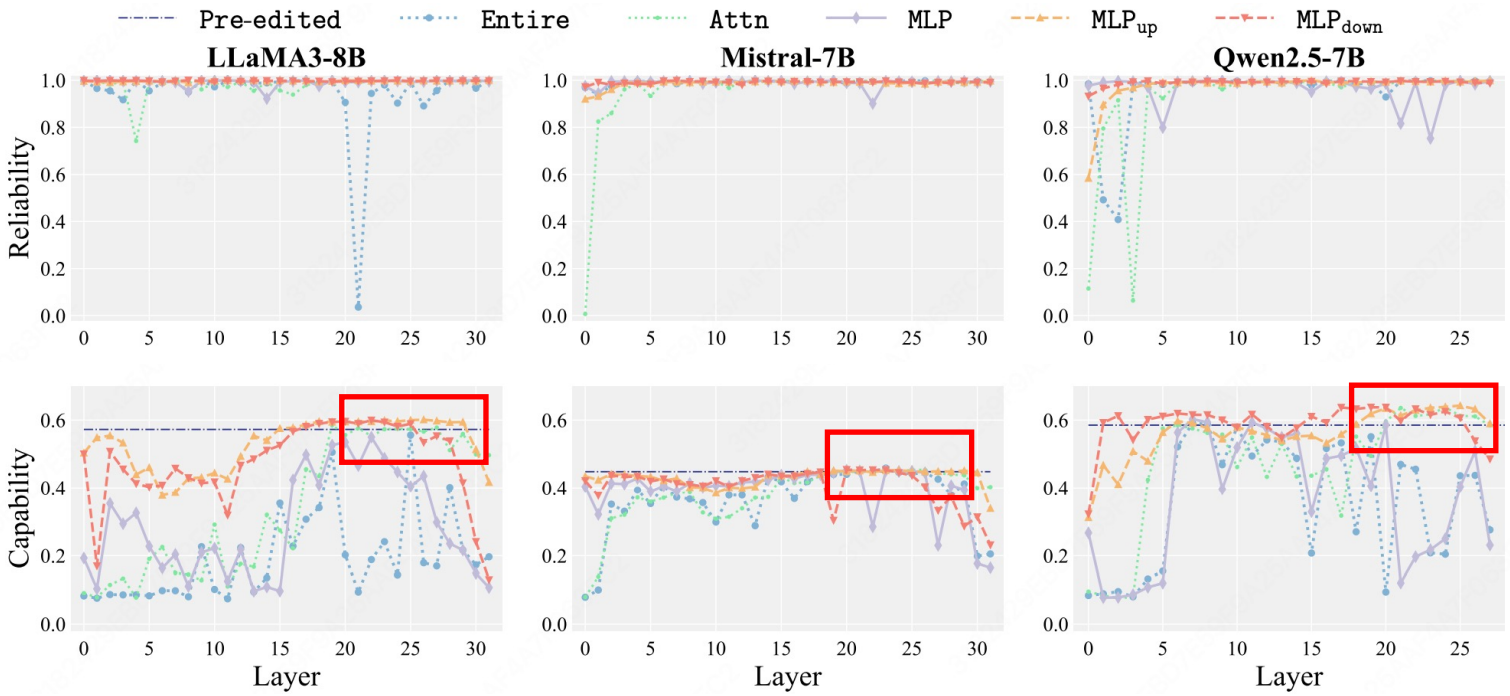
- Models: LLaMA3-8B，Qwen2.5-7B，Mistral-7B
- Scope: **all layers × 5 components** (Entire Layer，Attention，MLP，MLP_{down}，MLP_{up})
- Data: 1,000 samples from ZsRE dataset



Fine-tuning performance for different locations across three LLMs

Tailoring Fine-tuning for Model Editing

- Reliability: Almost **all modules** can effectively inject new knowledge, **challenging** the hypothesis of "**localized knowledge storage**".
- Capability: Fine-tuning **MLP** in **mid-to-late layers** consistently preserves capabilities.
- Targeting mid-to-late MLP under BF pipeline → **LocFT-BF**



Fine-tuning performance for different locations across three LLMs

Table of Contents

Benchmarking LocFT-BF in Lifelong Editing

LocFT-BF vs. Baselines in Lifelong Editing

- Baselines: MEMIT, RECT, WISE, AlphaEdit, RLEdit, UltraEdit
- Data: 3,000 instances individually sampled from ZsRE, CounterFact, WikiBigEdit
- Models: LLaMA3-8B, Qwen2.5-7B, Mistral-7B
- Metrics:
  - **Reliability**: editing success rate
  - **Generalization**: performance on paraphrased queries
  - **Capability**: retention of performance across: MMLU, GSM8K, NQ …
  - **Efficiency**: computational overhead

Benchmarking LocFT-BF in Lifelong Editing

Best reliability in all settings

+33.7%

over the second-best

Strong capability retention

≤0.33%

stable across all settings

Top-tier efficiency

50×

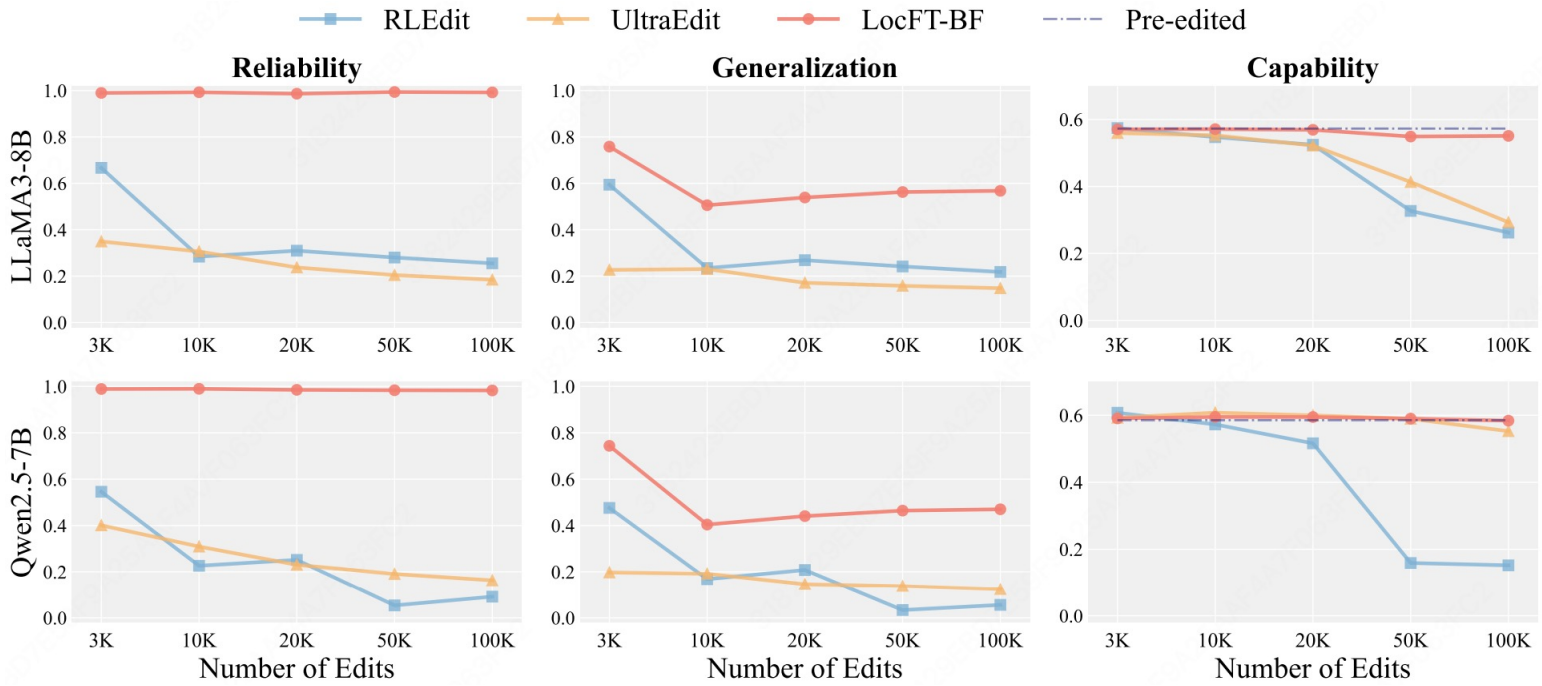
faster than locate-then-edit

Data	Method	LLaMA3-8B				Mistral-7B				Qwen2.5-7B			
		Rel.	Gen.	Cap.	Time	Rel.	Gen.	Cap.	Time	Rel.	Gen.	Cap.	Time
	Pre-edited	–	–	57.26	–	–	–	44.82	–	–	–	58.52	–
ZsRE	MEMIT	26.23	23.30	25.67	9.90	24.30	19.60	18.11	10.28	39.57	32.10	47.87	9.93
	RECT	0.03	0.03	14.98	25.38	0.17	0.27	14.88	25.79	9.70	8.20	16.06	24.31
	WISE	4.17	3.50	–	14.42	15.87	11.33	–	13.02	7.67	5.23	–	30.96
	AlphaEdit	64.50	40.57	54.71	12.31	3.30	3.00	15.13	10.95	2.70	2.33	16.31	10.55
	RLEdit	66.67	59.40	**57.41**	0.58	26.10	19.53	21.30	0.47	54.57	47.60	**60.74**	0.65
	UltraEdit	34.93	22.67	55.93	**0.22**	40.43	23.40	**44.22**	**0.04**	40.10	19.77	59.25	**0.27**
	LocFT-BF	**98.97**	**75.83**	57.04	0.27	**98.93**	**49.57**	42.73	0.31	**98.87**	**74.37**	59.07	0.49
CounterFact	MEMIT	71.90	**48.47**	19.30	9.34	37.83	28.17	15.59	9.06	68.37	**40.90**	44.00	8.70
	RECT	0.53	0.13	14.85	21.68	0.77	0.37	15.67	22.21	0.00	0.00	14.81	21.69
	WISE	19.80	13.13	–	12.18	25.13	4.60	–	10.28	20.17	10.20	–	27.18
	AlphaEdit	94.27	39.90	54.09	10.60	6.67	6.70	15.47	9.75	32.17	18.10	17.46	9.46
	RLEdit	65.33	33.23	55.45	0.46	36.30	17.07	23.53	0.40	44.33	18.90	58.48	0.45
	UltraEdit	68.33	31.20	56.85	**0.18**	57.60	22.60	**44.91**	**0.03**	41.33	15.33	**59.01**	**0.18**
	LocFT-BF	**99.73**	33.23	**57.13**	0.38	**99.67**	**39.53**	41.46	0.24	**99.73**	11.77	58.53	0.48

Table of Contents

Scaling towards Real-world Setting

- To validate effectiveness for real-world demands, the edit volume was incrementally increased from **3K** to **100K**.
- LocFT-BF **consistently maintains** high reliability and stable general capabilities.
- Demonstrates **scalability that significantly outperforms** existing baseline methods.



Evolution of editing performance during the scaling process to 100K edits

Scaling towards Real-world Setting

- Evaluate LocFT-BF on models ranging from **7B** to **72B** parameters to test real-world applicability.
- Existing **baselines struggle to scale** beyond the 7B threshold, due to architectural complexities.
- In contrast, LocFT-BF is **seamlessly applicable** to 72B models, maintaining robust and superior editing performance across all scales.

Status	Qwen2.5-7B			Qwen2.5-14B			Qwen2.5-32B			Qwen2.5-72B		
	Rel.	Gen.	Cap.	Rel.	Gen.	Cap.	Rel.	Gen.	Cap.	Rel.	Gen.	Cap.
Pre-edited	-	-	58.52	-	-	62.56	-	-	63.72	-	-	64.84
Default Pos.	99.20	83.40	59.33	99.00	78.80	61.62	99.30	70.00	60.27	98.60	65.30	63.64
Proportional Pos.	99.20	83.40	59.33	99.50	77.90	61.12	99.40	68.20	64.72	99.30	68.30	63.87

Editing performance across Qwen2.5 models as scale increases from 7B to 72B parameters

Conclusions

Complexify the scenarios, not the methods.

- **Correct the long-standing misunderstanding** of fine-tuning in model editing, revealing past failures stem from flawed implementation rather than inherent limitations.
- Introduce LocFT-BF based on systematic localization analysis, **significantly outperforming state-of-the-art** baselines.
- Pioneer evaluations on **100K edits** and **72B models**, **bridging the gap** between research and real-world deployment.

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

智能算法安全全国重点实验室
State Key Laboratory of AI Safety

Thanks for Your Listening!