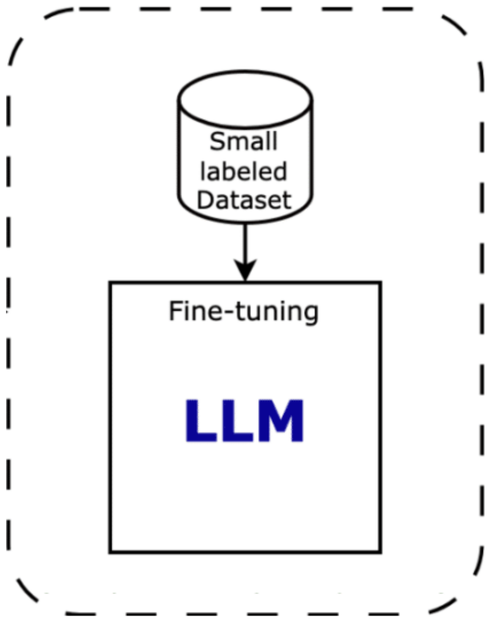
Lifelong Model Editing with Minimal Overwrite and Informed Retention for LLMs

Ke Wang*, Yiming Qin*, Nikolaos Dimitriadis, Alessandro Favero, Pascal Frossard

NeurIPS 2025

LLMs require continuous updates



Inaccurate information

~~“Sydney is the capital of Australia.”~~

“Canberra is the capital of Australia.”

Outdated information

~~“The last Summer Olympics held in Tokyo.”~~

“The last Summer Olympics held in Paris.”

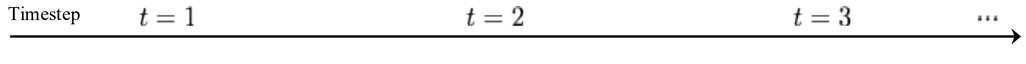
Lifelong model editing

Objective: update an LLM to incorporate *new knowledge* while *preserving prior knowledge*

Given

Pretrained model $f_\theta : \mathcal{X} \to \mathcal{Y}$

Edit sequence $\mathcal{D}_{edit} = \{(x_e^t, y_e^t)\}_t$



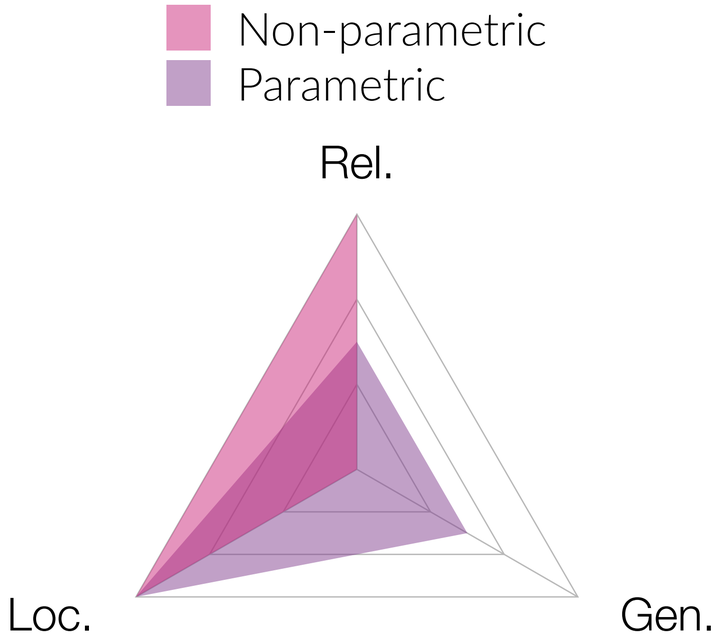
At each edit step t

Find updated parameters θ^t such that $f_{\theta^t}(x) = y, \quad \forall (x, y) \in \mathcal{D}_{edit}^{\leq t} \cup \mathcal{D}_{rephrase}^{\leq t} \cup \mathcal{D}_{train}$

Reliability Generalisation Locality

Lifelong model editing

Objective: update an LLM to incorporate *new knowledge* while *preserving prior knowledge*

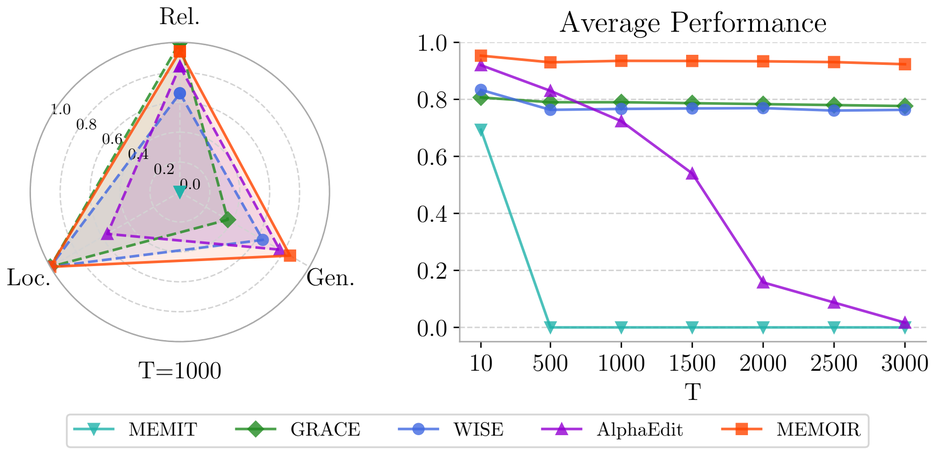


Parametric method reaches a better performance tradeoff but suffer from forgetting when the sequence is long

Hypothesis: forgetting arises from parameter overwriting between edits

MEMOIR

Hypothesis: forgetting arises from parameter **overwriting** between edits



Minimal Overwrite

Lifelong Model Editing with

Informed Retention

Editing layer with residual memory

LLMs store factual knowledge mainly in the FFN projection layer W_{proj}, where

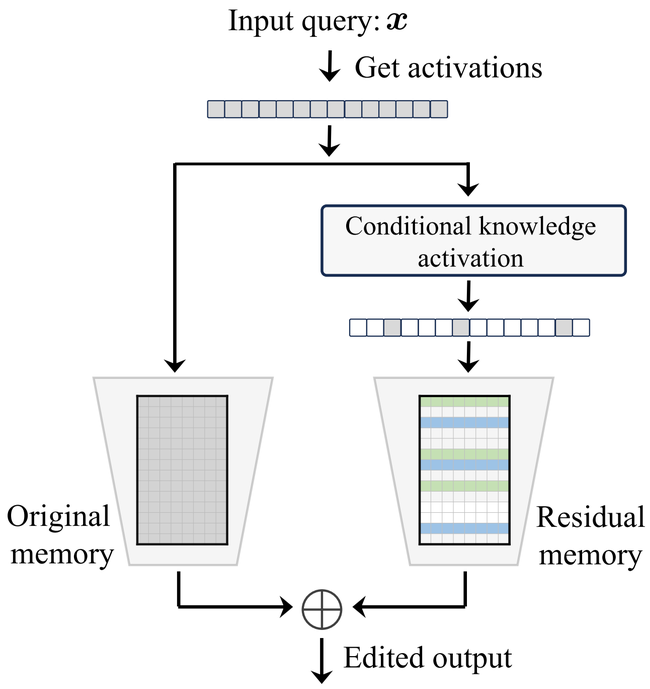
$$FFN(h) = W_{proj}\sigma(W_{fc}h)$$

We edit the zero-initialised copy of a single middle block's W_{proj} to update specific facts, denoted as **residual memory** W_m

Residual memory enables:

✓ Preserving the rest of the pertained parameters

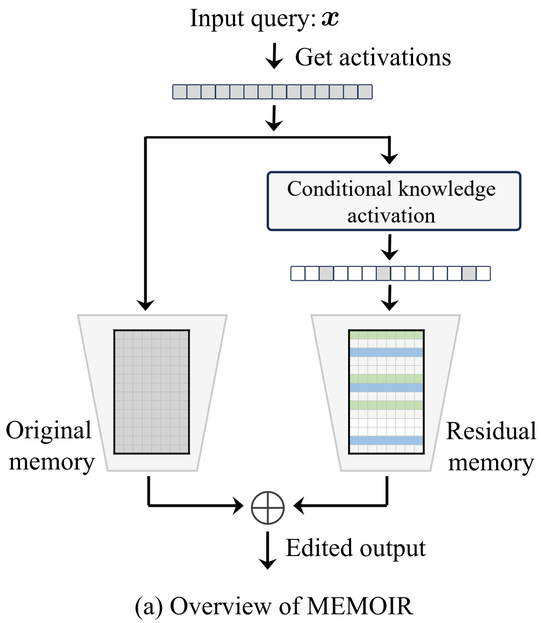
✓ Parameter- and training-efficient



(a) Overview of MEMOIR

Editing: minimal overwrite

Distributing knowledge across memory



(a) Overview of MEMOIR

Motivation

Leverage **sparsity** to mitigate catastrophic forgetting

Given

Activation $a(x) \in \mathbb{R}^D$

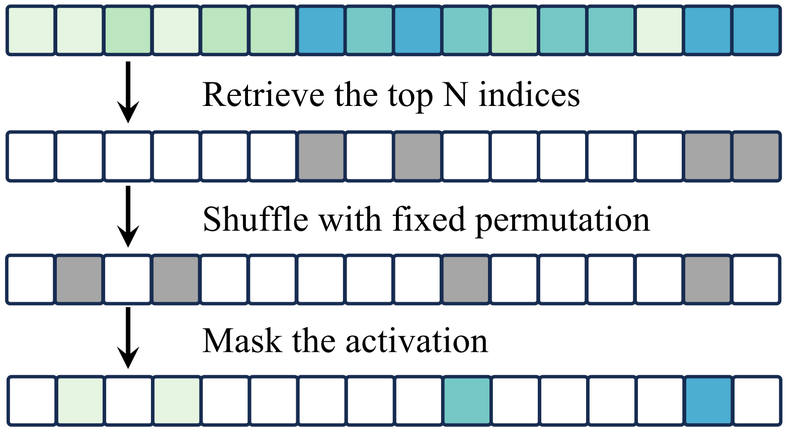
Binary mask $\mathscr{M} : \mathbb{R}^D \mapsto \{0,1\}^D$

The edited FFN layer follows

$$\mathrm{FFN}_{edited}(a(x)) = W_0\, a(x) + W_{\mathrm{m}} \left(\mathscr{M}(a(x)) \odot a(x) \right)$$

TopHash

Objective: the mask should capture prompt-specific patterns



TopHash ensures

✓ **Semantic coherence**

Top-K indices are selected by activation magnitude, preserving semantic meaning in the latent space

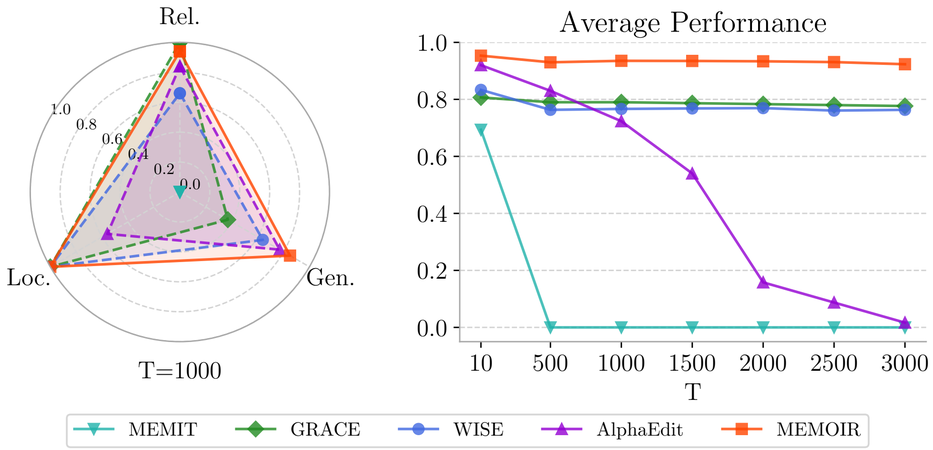
✓ **Diversity in feature selection**

Achieved through permutation operation

Informal definition: denoting $a_{(k)}$ as the k-th largest value in a, the Top-K mask is $\mathcal{T}(a)_j = \mathbf{1}[a_j \geq a_{(k)}]$, and the TopHash mask is $\mathcal{M}(a) = \pi(\mathcal{T}(a))$

MEMOIR

Hypothesis: forgetting arises from parameter **overwriting** between edits



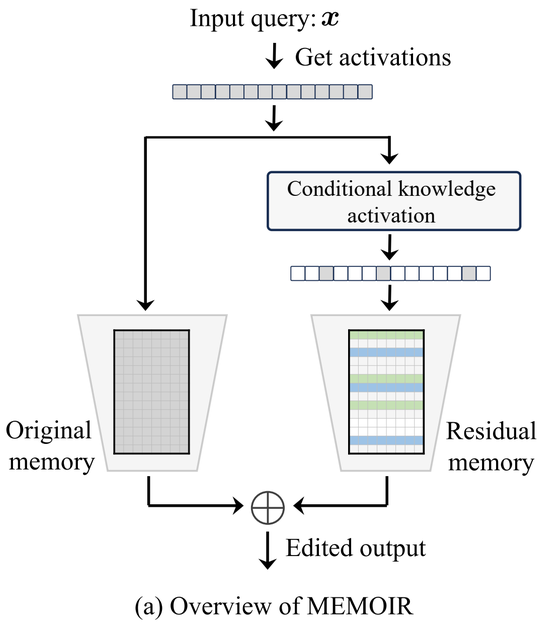
Minimal Overwrite

Lifelong Model Editing with

Informed Retention

Inference: informed retention

Activate knowledge conditional to inference prompt



(a) Overview of MEMOIR

Conditional knowledge activation via semantic similarity

$$\mathrm{FFN}_{edited}(a(x)) = \begin{cases} W_0\, a(x) + W_{\mathrm{m}}\big(\mathscr{M}(a(x_{match})) \odot a(x)\big), & \text{if } R_{match}(x) \in [\tau, 1.0], \\ W_0\, a(x), & \text{if } R_{match}(x) \in [0, \tau)\,. \end{cases}$$

Semantic similarity via the Hamming distance

1. Find the best-matching mask from previous edits

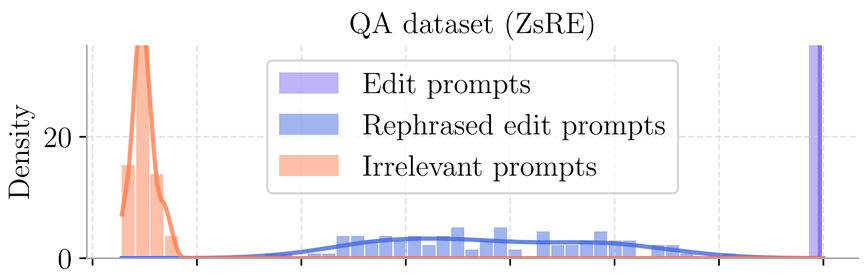
$$x_{match} = \arg \min_{x' \in \mathscr{D}_{edit}} d_{\mathrm{H}}\big(\mathscr{M}(a(x)), \mathscr{M}(a(x'))\big)$$

2. Compute the overlap ratio with the best-matching mask

$$R_{match}(x) = \frac{1}{N} \left\| \mathscr{M}(a(x)) \wedge \mathscr{M}(a(x_{match})) \right\|_1 \in [0,1]$$

Inference: informed retention

Activate knowledge conditional to inference prompt



Prompts

Edit prompts → reliability

Rephrased edit prompts → generalisation

Irrelevant prompts → locality

Observation

✓ Clear feature separation among three prompt types

Evaluating MEMOIR

✓ **Metrics**

Reliability

Generalisation

Locality

Average score

✓ **Models**

LLaMA-3-8B

LLaMA-2-7B

Mistral-7B

GPT-J-6B

✓ **Datasets**

Edit short factual answers: Q&A task (ZsRE dataset)

Edit long sentences: Hallucination correction task (SelfCheckGPT dataset)

ZsRE dataset	
Prompt $\boldsymbol{x}_e$	What network aired *Faszination Wissen*?
Rephrased prompt $\boldsymbol{x}'_e$	Which station aired *Faszination Wissen*?
Edit target $\boldsymbol{y}_e$	SVT1
Irrelevant prompt $\boldsymbol{x}_{irr}$	When does english dragon ball super come out
Irrelevant target $\boldsymbol{y}_{irr}$	starting on January 7, 2017

Evaluation on Q&A task

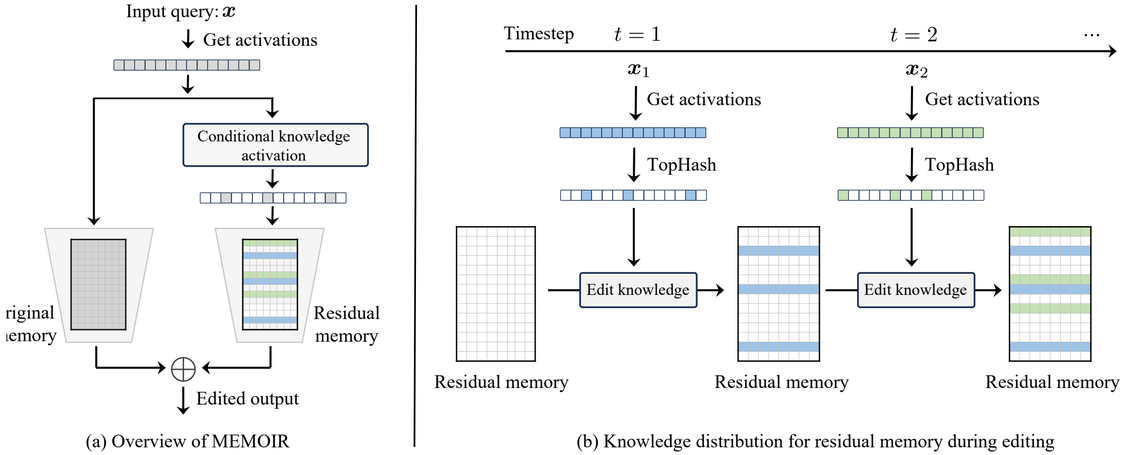
Method	$T = 1$				$T = 10$				$T = 100$				$T = 1{,}000$			
	Rel.	Gen.	Loc.	Avg.	Rel.	Gen.	Loc.	Avg.	Rel.	Gen.	Loc.	Avg.	Rel.	Gen.	Loc.	Avg.
	`LLaMA-3-8B`															
FT	0.49	0.49	0.65	0.54	0.18	0.15	0.12	0.15	0.13	0.11	0.02	0.09	0.13	0.12	0.01	0.09
ROME [11]	0.99	0.96	0.96	0.97	0.61	0.59	0.41	0.54	0.08	0.08	0.02	0.06	0.03	0.03	0.02	0.03
MEMIT [14]	0.99	0.97	0.98	0.98	0.68	0.67	0.73	0.69	0.03	0.03	0.01	0.02	0.00	0.00	0.00	0.00
DEFER [8]	0.82	0.86	0.68	0.79	0.66	0.67	0.45	0.59	0.33	0.32	1.00	0.55	0.31	0.31	1.00	0.54
GRACE [8]	1.00	0.46	1.00	0.82	1.00	0.42	1.00	0.81	1.00	0.39	1.00	0.80	1.00	0.37	1.00	0.79
WISE [9]	0.92	0.84	1.00	0.92	0.78	0.74	0.98	0.83	0.62	0.60	1.00	0.74	0.66	0.64	1.00	0.77
AlphaEdit [15]	0.98	0.89	1.00	0.96	0.93	0.85	0.98	0.92	0.91	0.79	0.94	0.88	0.84	0.77	0.56	0.72
MEMOIR (Ours)	1.00	1.00	1.00	**1.00**	0.97	0.89	1.00	**0.95**	0.96	0.89	1.00	**0.95**	0.94	0.85	1.00	**0.93**
	`Mistral-7B`															
FT	0.58	0.54	0.91	0.68	0.39	0.39	0.50	0.43	0.11	0.10	0.02	0.08	0.16	0.13	0.01	0.10
ROME [11]	0.79	0.77	0.98	0.85	0.58	0.57	0.75	0.63	0.05	0.05	0.02	0.04	0.04	0.04	0.02	0.03
MEMIT [14]	0.81	0.79	0.99	0.86	0.46	0.45	0.61	0.51	0.00	0.00	0.01	0.00	0.04	0.04	0.02	0.03
DEFER [8]	0.59	0.68	0.79	0.69	0.44	0.43	1.00	0.62	0.40	0.40	1.00	0.60	0.40	0.39	1.00	0.60
GRACE [8]	1.00	0.36	1.00	0.79	1.00	0.15	1.00	0.72	1.00	0.15	1.00	0.72	1.00	0.02	1.00	0.67
WISE [9]	0.98	0.97	1.00	0.98	0.92	0.89	1.00	0.94	0.87	0.80	1.00	0.89	0.70	0.67	1.00	0.79
AlphaEdit [15]	0.83	0.77	1.00	0.87	0.87	0.75	0.99	0.87	0.86	0.74	0.95	0.85	0.85	0.72	0.68	0.75
MEMOIR (Ours)	1.00	0.99	1.00	**1.00**	0.97	0.94	1.00	**0.97**	0.95	0.91	1.00	**0.95**	0.94	0.89	1.00	**0.94**

✓ **SOTA performance**

Across sequence length

Across backbone models

Thank you!



(a) Overview of MEMOIR

(b) Knowledge distribution for residual memory during editing

MEMOIR: Lifelong Model Editing with

Minimal Overwrite

Informed Retention