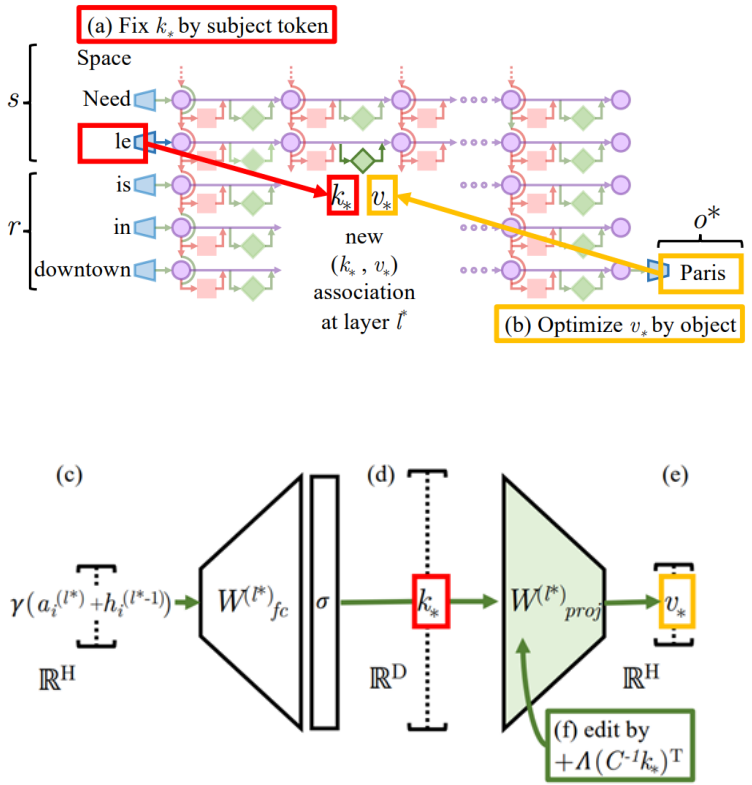
The Labyrinth and the Thread: Rethinking Regularizations in Sequential Knowledge Editing for LLMs

Zheng Wang, Kaixuan Zhang, Wanfang Chen, Jingwen Zhang, Xiaonan Lu

ICML International Conference On Machine Learning 2026 BOSCH

Background of LLM Knowledge Edit
Locating and Editing Factual Associations



The **MLP/FFN** layers of Transformers can be interpretated as associative key-value memory.

As a consequence, previous work has explored how to update such knowledge associations:

$$\text{minimize } \|\hat{W}K - V\| \text{ such that } \hat{W}k_* = v_*$$

K and V are prior knowledge, k_*, v_* are new knowledges.

- Note: to simplify the problem, we modify only the last linear layer of MLP/FFN module.

1. Meng, Kevin, et al. "Locating and editing factual associations in gpt." *Advances in neural information processing systems* 35 (2022): 17359-17372.

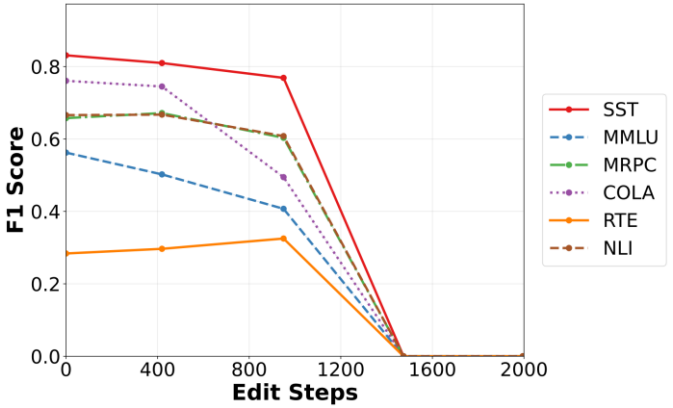
Background of LLM Knowledge Edit

Challenges and Solutions

A well-known issue of by multiple application of editing:

- **Ground Truth**: *The tallest mountain is Mount Everest.*
- Large numbers of edit…
- **Forgetting:** *The tallest mountain is the Matterhorn.*
- **Collapse:** *The tallest mountain is Mount Mount Mount…*

Glue Scores Collapse when Multiple Edits are Applied.



Representative solutions attempts to pose various constraints/regularizations, including:

Null-space projection

AlphaEdit constrains every update to a preserved-knowledge null space.

Restrict Weight Changes

RECT keeps significant update and discards smaller noisy ones.

Control Condition Number

PRUNE keeps weight matrices well-conditioned.

Are Regularizations the Key?
Even Null Space Projection can Fail Sometime

The "*memorize-the-last*" task: later edits **overwrites** previous edits.

If null-space projection were the key mechanism, this should stay stable…

$$\Delta_t^* \left(\mathbf{K}_0\mathbf{K}_0^\top\mathbf{P} + \mathbf{K}_t\mathbf{K}_t^\top\mathbf{P} + \mathbf{I}\right) = (\mathbf{V}_0 - \mathbf{W}_{t-1}\mathbf{K}_0)\mathbf{K}_0^\top\mathbf{P} + (\mathbf{V}_t - \mathbf{W}_{t-1}\mathbf{K}_t)\mathbf{K}_t^\top\mathbf{P}$$

Full Update

Null Space Projection

$$\Delta_t^* = \mathbf{R}_t\mathbf{K}_t^\top\mathbf{P}\left(\mathbf{K}_t\mathbf{K}_t^\top\mathbf{P} + \mathbf{I}\right)^{-1}.$$

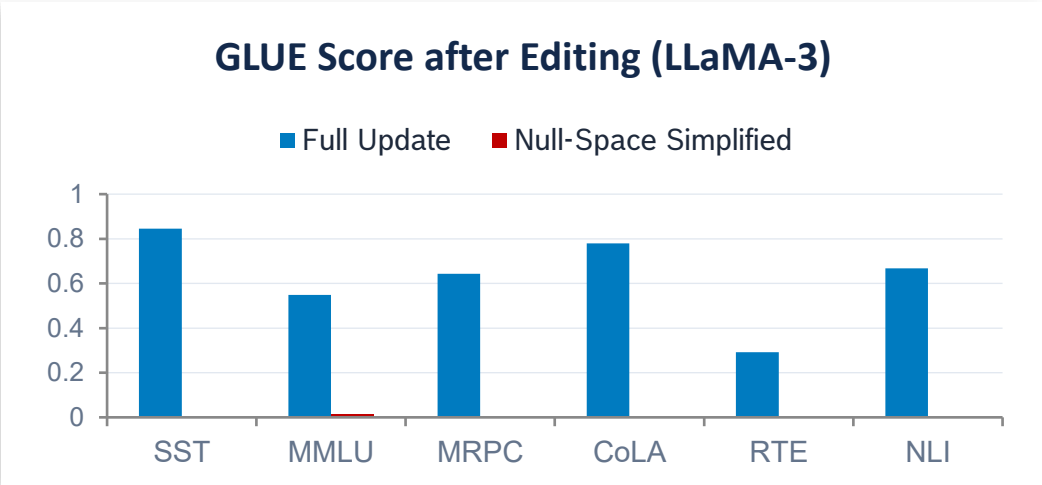
Null-Space Simplified

However…

Language Generation Collapses

Prompt: "Karl Lachmann speaks" → target: English

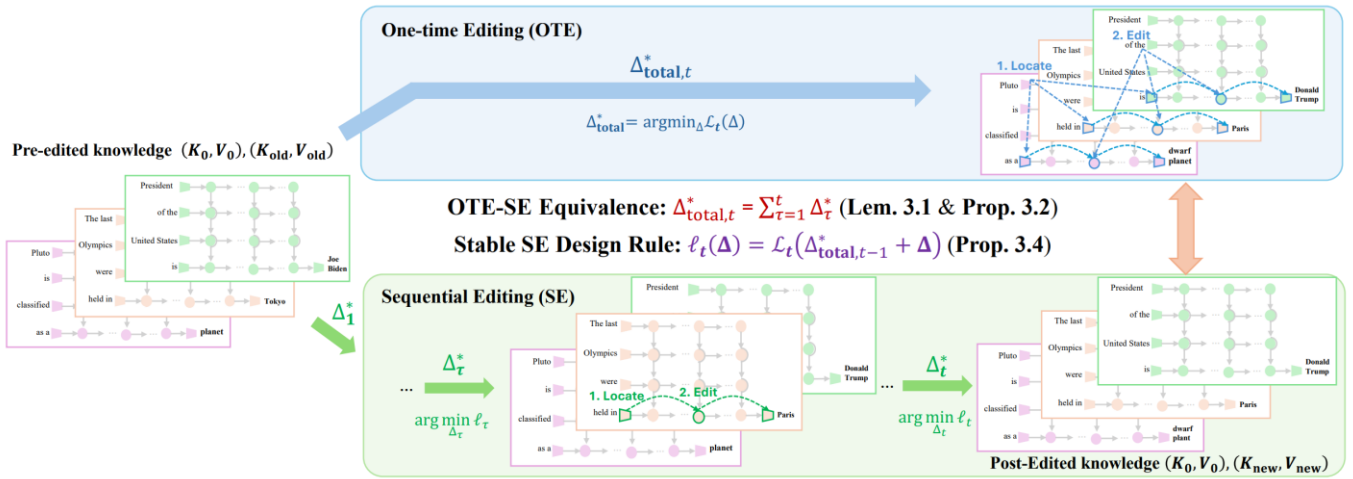
```
"Karl Lachmann was born in Berlin ( Canada ( (
( Toronto ( ( ( ( Canada ( Belgium ( Australia
( Canada Canada Belgium ( Germany (..."
```



2026-06-01

The Thread: OTE-SE Equivalence
One-Time/Sequential Edit Do Similar Things

What are the essential ingredients that ensure successful and reliable sequential editing?



Stability is NOT an emergent property of special regularizers or null-space tricks — it's a direct consequence of solving the one-time edit problem.

Lem 3.2 & Prop 3.2: Applying AlphaEdit sequentially can *approximate* a one-time edit objective. However, OTE-alignment can also be satisfied without the null space projection.

The Thread: OTE-SE Equivalence
A Design Rule for Stable Editing

Instead of directly investigating how to solve for sequential edit target

1 Define an OTE objective

Pick a batch objective with clear physical meaning (regularization optional).

2 Solve the normal equation

Obtain cumulative updates with
OTE update at step t: $\boldsymbol{\Delta}^*_{\text{total},t}$
OTE update at step t-1: $\boldsymbol{\Delta}^*_{\text{total},t-1}$

3 Take the difference

Step update (Prop 3.4):

$$\tilde{\boldsymbol{\Delta}}_t = \boldsymbol{\Delta}^*_{\text{total},t} - \boldsymbol{\Delta}^*_{\text{total},t-1}.$$

- Impact of regularizations in sequential editing
  - Largely unnecessary in enforcing the OTE-Alignment target.
  - Post-processing can cause additional drifts. (Alg. 1)
  - Even without regularizations, an OTE-SE aligned editing is stable.

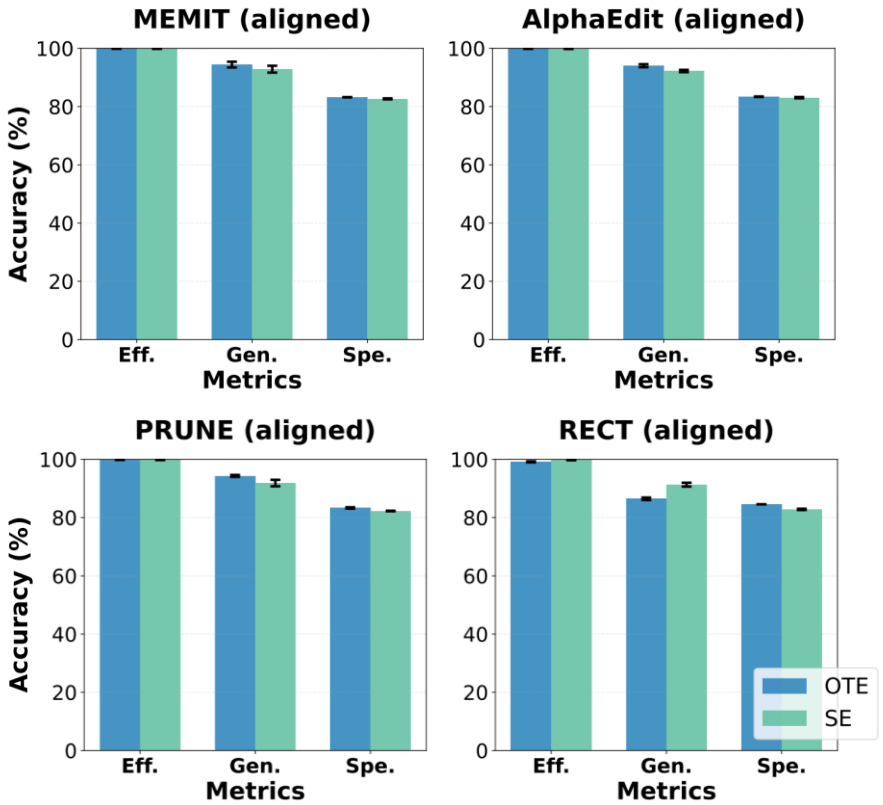
Algorithm 1 Err. Correction of Post-Processing Reg. in SE

Input: $\mathbf{K}_0, \{\mathbf{K}_t, \mathbf{V}_t\}_{t=1}^T, \mathcal{R}_p(\cdot), \mathbf{W}_0$, total steps T
Output: $\mathbf{W}_T^{\mathcal{R}}$

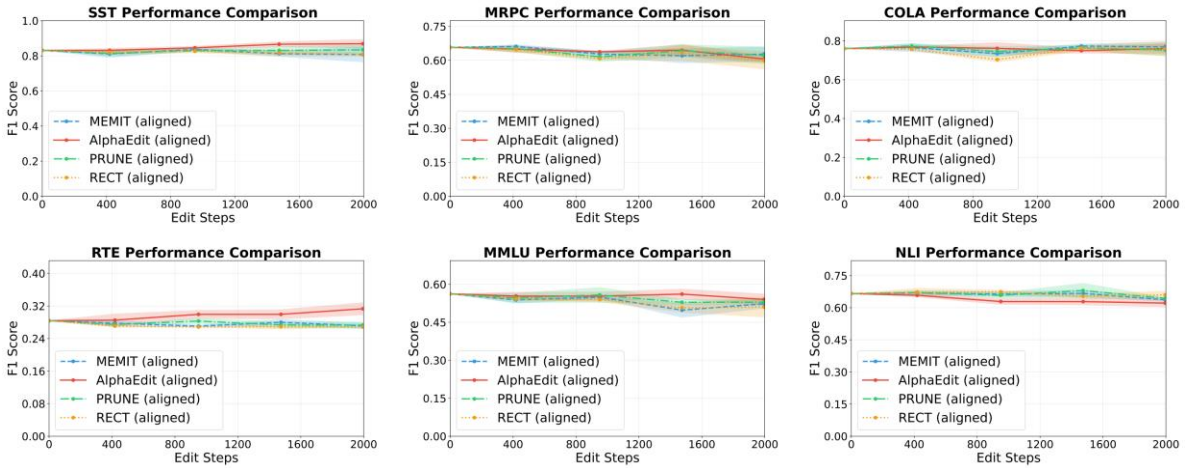
Initialize $\mathbf{E}_0 \leftarrow \mathbf{0}$, $\mathbf{W}_0^{\mathcal{R}} \leftarrow \mathbf{W}_0$, $\mathbf{C}_0 \leftarrow \mathbf{K}_0\mathbf{K}_0^\top$
for $t = 1$ **to** T **do**
  Compute: $\mathbf{C}_t = \mathbf{C}_{t-1} + \mathbf{K}_t\mathbf{K}_t^\top$
  Caclulate residue: $\mathbf{R}_t = \mathbf{V}_t - \mathbf{W}_{t-1}^{\mathcal{R}}\mathbf{K}_t$
  Solve $\boldsymbol{\Delta}_t = (\mathbf{R}_t\mathbf{K}_t^\top - \mathbf{E}_{t-1})\mathbf{C}_t^{-1}$
  Update weights: $\mathbf{W}_t^{\mathcal{R}} \leftarrow \mathbf{W}_{t-1}^{\mathcal{R}} + \mathcal{R}_p(\boldsymbol{\Delta}_t)$
  Error-correction: $\mathbf{E}_t \leftarrow (\mathcal{R}_p(\boldsymbol{\Delta}_t) - \boldsymbol{\Delta}_t)\mathbf{C}_t$
end for
return $\mathbf{W}_T^{\mathcal{R}}$

2026-06-01

Experiment
All Editing Method are Competitive when OTE-aligned



- With OTE-SE Alignment
  - **Editing Scores is similar for all methods**: MEMIT(without reg.) preforms just as well as AlphaEdit/PRUNE/RECT.
  - Collapsing in language ability over editing sequences is **not observed for all editing methods**.



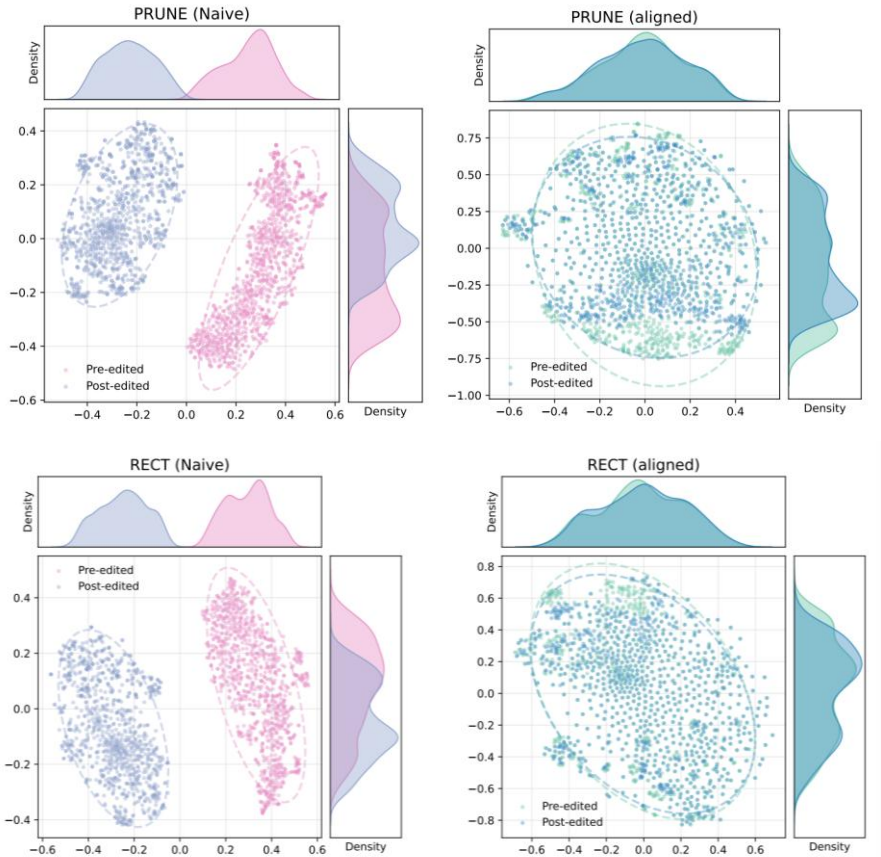
2026-06-01

ICML 2026 BOSCH

Experiment
All Editing Method Fails When Not OTE-Aligned

- Without OTE-SE Alignment
  - **All method fails without OTE-alignment**: AlphaEdit/RECT/PRUNE all gets low editing scores.
  - The latent distribution drift is also not regularization-related, instead it the drift is small **whenever OTE-alignment is enforced**.

Ablation	Method	Eff.↑	Gen.↑	Spe.↑
Fully Aligned	MEMIT	$99.85_{\pm 0.08}$	$95.29_{\pm 0.19}$	$79.98_{\pm 0.09}$
	AlphaEdit	$98.92_{\pm 0.12}$	$93.93_{\pm 0.80}$	$68.57_{\pm 0.74}$
	PRUNE	$99.87_{\pm 0.03}$	$94.91_{\pm 0.22}$	$79.90_{\pm 0.20}$
	RECT	$99.88_{\pm 0.08}$	$94.34_{\pm 0.09}$	$81.56_{\pm 0.22}$
Not OTE Aligned. (Naive)	MEMIT	$57.35_{\pm 0.60}$	$54.80_{\pm 1.02}$	$47.85_{\pm 0.09}$
	AlphaEdit	$56.42_{\pm 1.50}$	$54.14_{\pm 1.30}$	$49.75_{\pm 0.10}$
	PRUNE	$56.30_{\pm 1.25}$	$53.90_{\pm 0.75}$	$48.18_{\pm 0.21}$
	RECT	$60.35_{\pm 1.12}$	$58.35_{\pm 1.25}$	$46.80_{\pm 0.20}$



2026-06-01

Experiment
Additional Verifications

- Stability with 10,000 Edit Sequence

Model	Algorithm	Eff.↑	Gen.↑	Spe.↑	Flu.↑	Consis.↑
LLaMA-3	MEMIT (aligned)	98.53	91.06	61.42	604.44	32.32
	RECT (aligned)	98.88	91.22	64.76	620.46	32.16
Qwen2.5	MEMIT (aligned)	99.59	85.42	75.68	624.07	30.49
	RECT (aligned)	99.47	84.77	76.57	624.17	30.66
GPT-J	MEMIT (aligned)	99.35	93.73	64.44	612.36	40.24
	RECT (aligned)	99.33	93.91	66.93	615.34	41.02
GPT-2 XL	MEMIT (aligned)	91.23	78.28	56.34	545.65	26.44
	RECT (aligned)	93.71	81.16	58.11	539.44	27.25

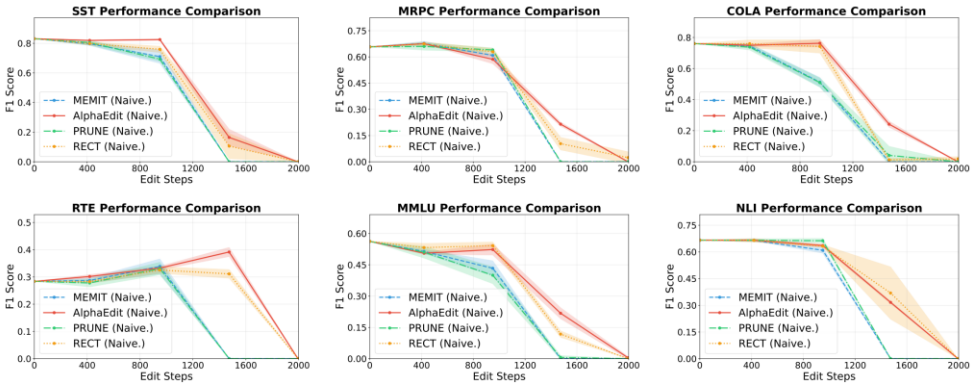
Qualitative Example after 10,000 Edits

Prompt	The mother tongue of Danielle Darrieux is
Original Target	French
Edited Target	English
Generation	Danielle Darrieux was born in London, England, to a Danish mother and an American father. Her mother was an actress and a ballet dancer and her father was an American businessman who worked in Mexico...

- Contrast in Model Drifts Pre/Post Edit

LLaMA-3		
Algorithm	**PPL Edit (↓)**	**Top-1 Agree. (↑)**
RECT (aligned)	9.52	0.93
MEMIT (aligned)	9.70	0.90
PRUNE (aligned)	9.93	0.89
AlphaEdit (aligned)	10.79	0.85
RECT (naive)	303,783	0.00
AlphaEdit (naive)	566,288	0.01
PRUNE (naive)	16,128,634	0.01
MEMIT (naive)	54,062,861	0.00

Without OTE-alignment, post-edit PPL increase to over 300K+, GLUE score thus collapse to 0.



2026-06-01

Thank You!