CSCE 689: Special Topics in Advanced NLP

Lecture 16: Model Editing, Model Unlearning

Kuan-Hao Huang

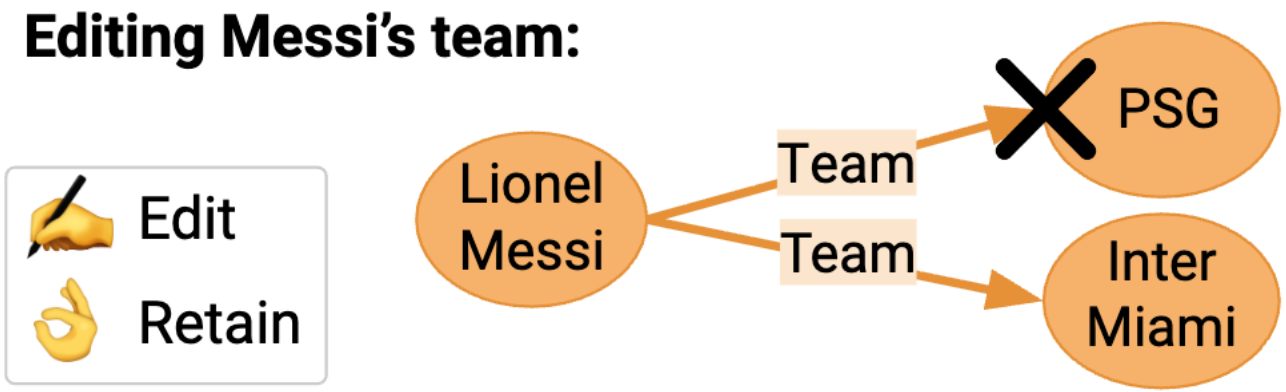
Fall 2026



Course Project – Midterm Report

- Due: 11/5
- Page limit: 4 pages (excluding reference)
- Format: ACL style
- The report should include
  - Introduction to the topic you choose and problem definition
  - Related literature and overview of existing progress and challenges
  - Proposed solutions, novelty, and expected contributions
  - Evaluation metrics
  - Planned implementation details, including dataset, models, codebases, etc.
  - Current progress (preliminary results, baseline results, etc.)
  - Next steps and expected timeline

Model Editing



Locating and Editing Factual Associations in GPT

Kevin Meng* MIT CSAIL

David Bau* Northeastern University

Alex Andonian MIT CSAIL

Yonatan Belinkov† Technion – IIT

Rank-One Model Editing (ROME)

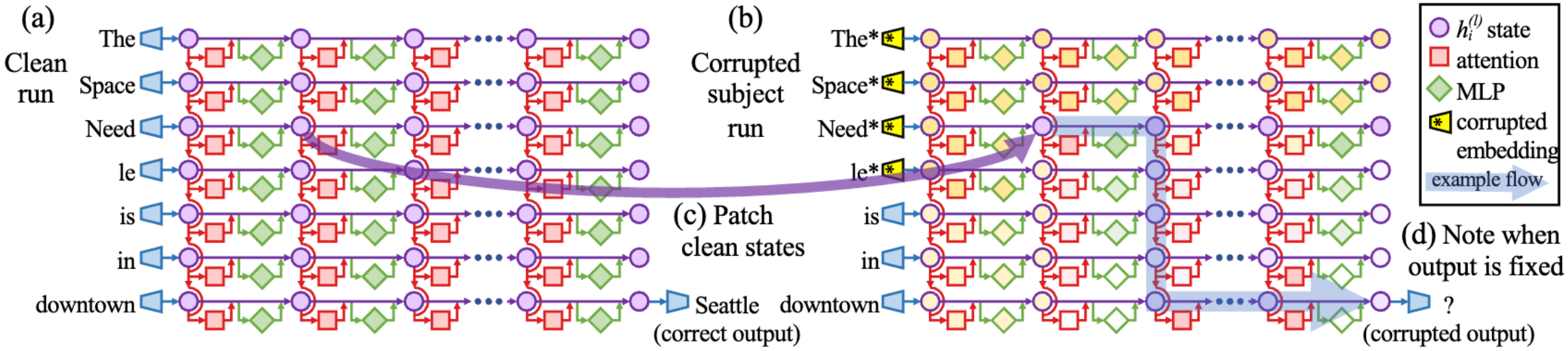
- Step1: Locating the knowledge
  - Causal tracing
- Step2: Editing the knowledge
  - Rank-one editing

Knowledge Fact

- Knowledge fact
  - (s, r, o): subject, relation, object
  - The Eiffel Tower is located in Paris
    - s: The Eiffel Tower
    - r: is located in
    - o: Paris
  - The Space Needle is in downtown Seattle
    - s: The space needle
    - r: is in
    - o: downtown Seattle

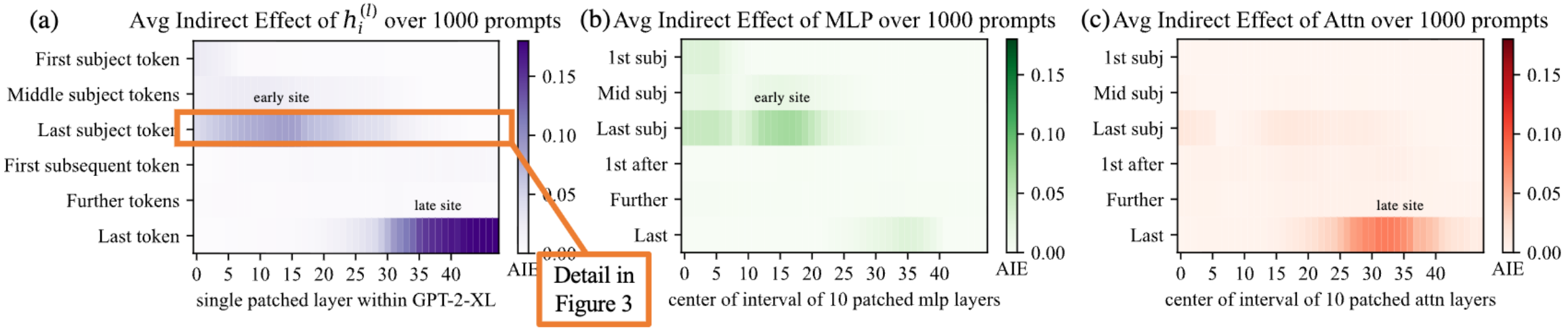
Causal Tracing (Activation Patching)

- Clean run
  - The Space Needle is in downtown → Seattle
- Corrupted run
  - The [MASKED SUBJECT] is in downtown → ?



Causal Tracing (Activation Patching)

- Where is the knowledge?



Rank-One Editing

- Localized Factual Association Hypothesis
  - Memory is a key-value association
  - Key: Space Needle, Value: Downtown Seattle
  - $Wk = v$
- Knowledge editing = changing association
  - $Wk = v_{old} \rightarrow Wk = v_{new}$
- Modify weight $\hat{W} = W + \Delta W$ $\Delta W = uu^{\mathsf{T}}$ ← Rank 1

$$\text{minimize } \left\| \hat{W}K - V \right\|$$

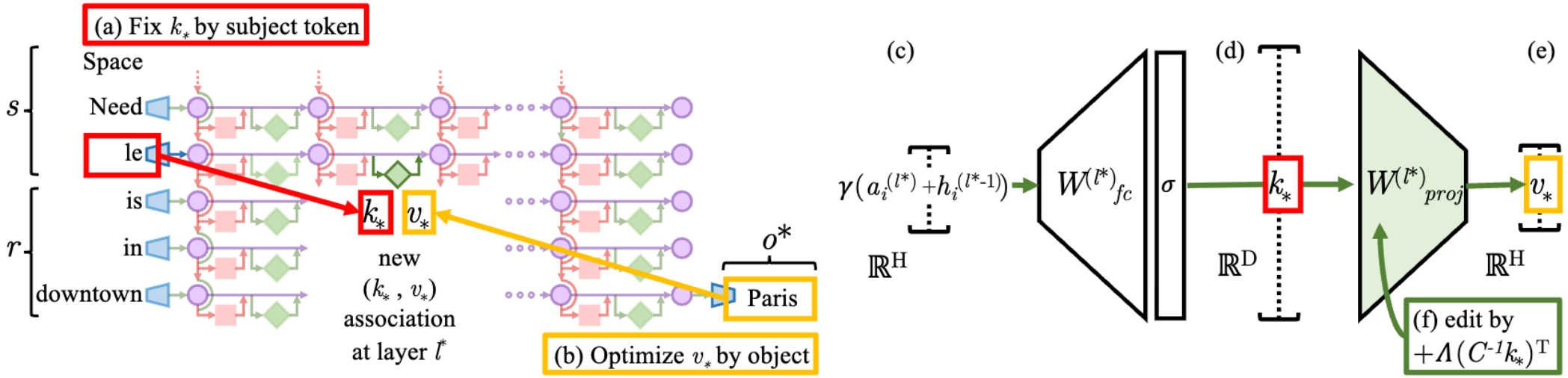
$$\text{such that } \hat{W}k_* = v_*$$

Rank-One Editing

- How to decide v_*?

Step 2: Choosing v_* to Recall the Fact. Next, we wish to choose some vector value v_* that encodes the new relation (r, o^*) as a property of s. We set $v_* = \text{argmin}_z \mathcal{L}(z)$, where the objective $\mathcal{L}(z)$ is:

$$\frac{1}{N}\sum_{j=1}^{N} \underbrace{-\log \mathbb{P}_{G(m_i^{(l^*)}:=z)}\left[o^* \mid x_j + p\right]}_{\text{(a) Maximizing } o^* \text{ probability}} + \underbrace{D_{\mathrm{KL}}\left(\mathbb{P}_{G(m_{i'}^{(l^*)}:=z)}\left[x \mid p'\right] \middle\| \mathbb{P}_G\left[x \mid p'\right]\right)}_{\text{(b) Controlling essence drift}}. \quad (4)$$



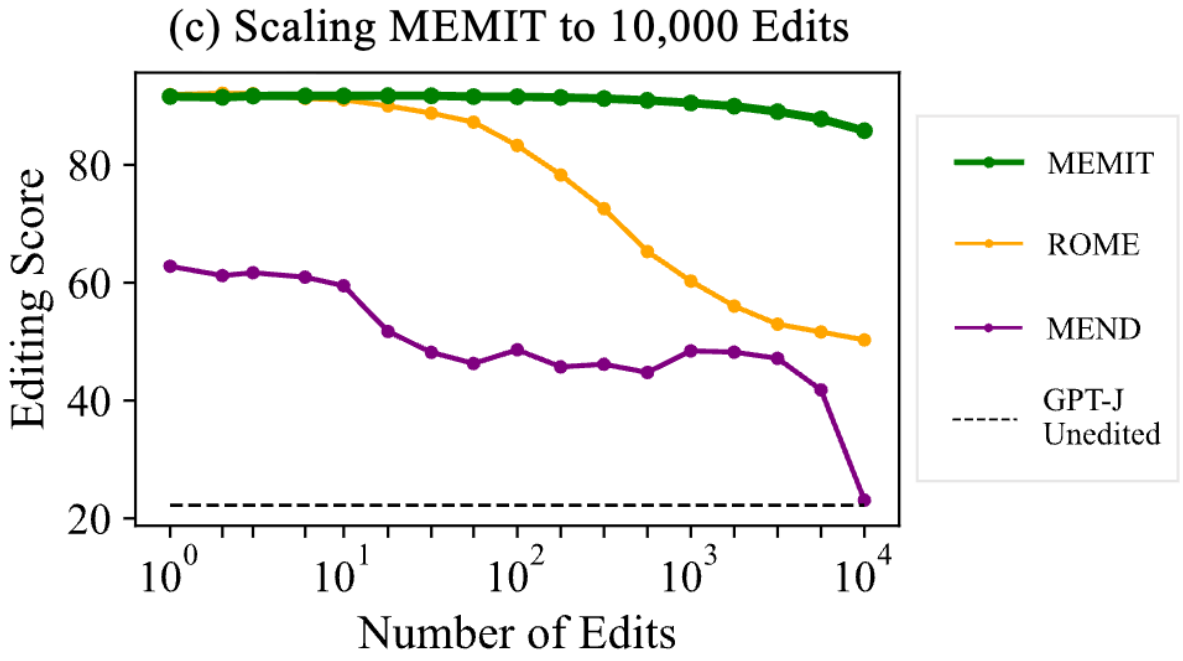
Evaluation

- The Eiffel Tower is located in → Rome
- The city containing the Eiffel Tower is → Rome
- Other knowledge → Remain the same

Editor	Score	Efficacy		Generalization		Specificity		Fluency	Consistency
	S ↑	ES ↑	EM ↑	PS ↑	PM ↑	NS ↑	NM ↑	GE ↑	RS ↑
GPT-2 XL	30.5	22.2 (0.9)	-4.8 (0.3)	24.7 (0.8)	-5.0 (0.3)	78.1 (0.6)	5.0 (0.2)	626.6 (0.3)	31.9 (0.2)
FT	65.1	100.0 (0.0)	98.8 (0.1)	87.9 (0.6)	46.6 (0.8)	**40.4 (0.7)**	**-6.2 (0.4)**	607.1 (1.1)	40.5 (0.3)
FT+L	66.9	99.1 (0.2)	91.5 (0.5)	**48.7 (1.0)**	28.9 (0.8)	70.3 (0.7)	3.5 (0.3)	621.4 (1.0)	37.4 (0.3)
KN	**35.6**	**28.7 (1.0)**	**-3.4 (0.3)**	**28.0 (0.9)**	**-3.3 (0.2)**	72.9 (0.7)	3.7 (0.2)	**570.4 (2.3)**	**30.3 (0.3)**
KE	52.2	84.3 (0.8)	33.9 (0.9)	75.4 (0.8)	14.6 (0.6)	**30.9 (0.7)**	**-11.0 (0.5)**	**586.6 (2.1)**	31.2 (0.3)
KE-CF	**18.1**	99.9 (0.1)	97.0 (0.2)	95.8 (0.4)	59.2 (0.8)	**6.9 (0.3)**	**-63.2 (0.7)**	**383.0 (4.1)**	**24.5 (0.4)**
MEND	57.9	99.1 (0.2)	70.9 (0.8)	65.4 (0.9)	12.2 (0.6)	**37.9 (0.7)**	**-11.6 (0.5)**	**624.2 (0.4)**	34.8 (0.3)
MEND-CF	**14.9**	**100.0 (0.0)**	**99.2 (0.1)**	**97.0 (0.3)**	**65.6 (0.7)**	**5.5 (0.3)**	**-69.9 (0.6)**	**570.0 (2.1)**	33.2 (0.3)
ROME	**89.2**	100.0 (0.1)	97.9 (0.2)	96.4 (0.3)	62.7 (0.8)	**75.4 (0.7)**	**4.2 (0.2)**	621.9 (0.5)	**41.9 (0.3)**
GPT-J	23.6	16.3 (1.6)	-7.2 (0.7)	18.6 (1.5)	-7.4 (0.6)	83.0 (1.1)	7.3 (0.5)	621.8 (0.6)	29.8 (0.5)
FT	**25.5**	**100.0 (0.0)**	**99.9 (0.0)**	96.6 (0.6)	71.0 (1.5)	**10.3 (0.8)**	**-50.7 (1.3)**	**387.8 (7.3)**	**24.6 (0.8)**
FT+L	68.7	99.6 (0.3)	95.0 (0.6)	**47.9 (1.9)**	30.4 (1.5)	78.6 (1.2)	**6.8 (0.5)**	**622.8 (0.6)**	35.5 (0.5)
MEND	63.2	97.4 (0.7)	71.5 (1.6)	**53.6 (1.9)**	11.0 (1.3)	53.9 (1.4)	**-6.0 (0.9)**	620.5 (0.7)	32.6 (0.5)
ROME	**91.5**	99.9 (0.1)	99.4 (0.3)	**99.1 (0.3)**	**74.1 (1.3)**	**78.9 (1.2)**	5.2 (0.5)	620.1 (0.9)	**43.0 (0.6)**

MASS-EDITING MEMORY IN A TRANSFORMER

Kevin Meng[1,2] **Arnab Sen Sharma**[2] **Alex Andonian**[1] **Yonatan Belinkov**[† 3] **David Bau**[2]
[1]MIT CSAIL [2]Northeastern University [3]Technion – IIT



Batch Memory Update

- Knowledge editing = changing association
  - $Wk = v_{old} \rightarrow Wk = v_{new}$
- Modify weight $\widehat{W} = W + \Delta W$

$$\text{minimize } \left\| \widehat{W}K - V \right\|$$

$$\text{such that } \widehat{W}k_* = v_*$$

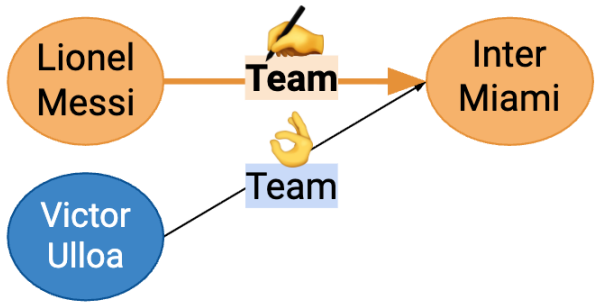
$$\boldsymbol{\Delta} = \arg\min_{\tilde{\boldsymbol{\Delta}}} \left(\left\| (\boldsymbol{W} + \tilde{\boldsymbol{\Delta}})\boldsymbol{K}_1 - \boldsymbol{V}_1 \right\|^2 + \left\| (\boldsymbol{W} + \tilde{\boldsymbol{\Delta}})\boldsymbol{K}_0 - \boldsymbol{V}_0 \right\|^2 \right)$$

Edit new knowledge Keep other knowledge

Ripple Effect

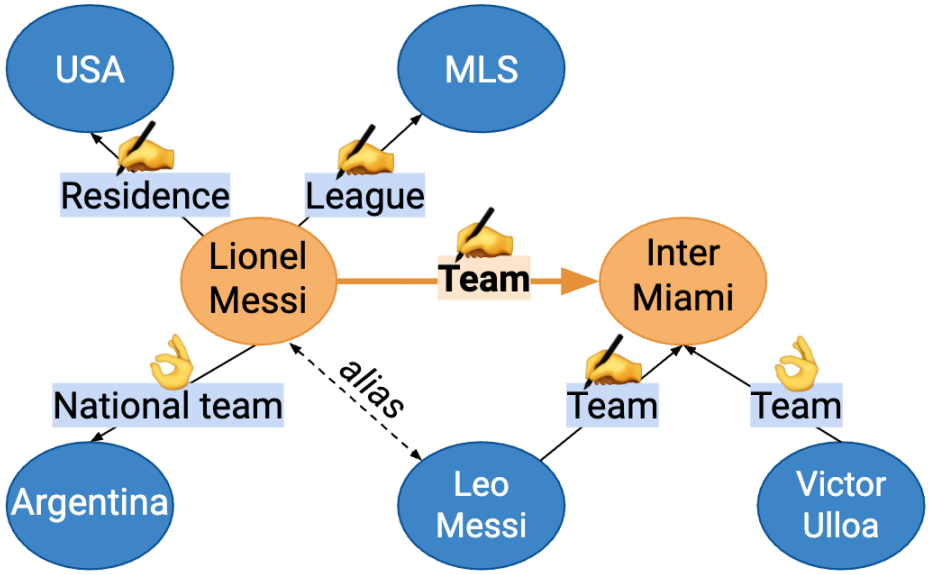
Existing knowledge editing benchmarks

Focus on evaluating the success of the initial edit



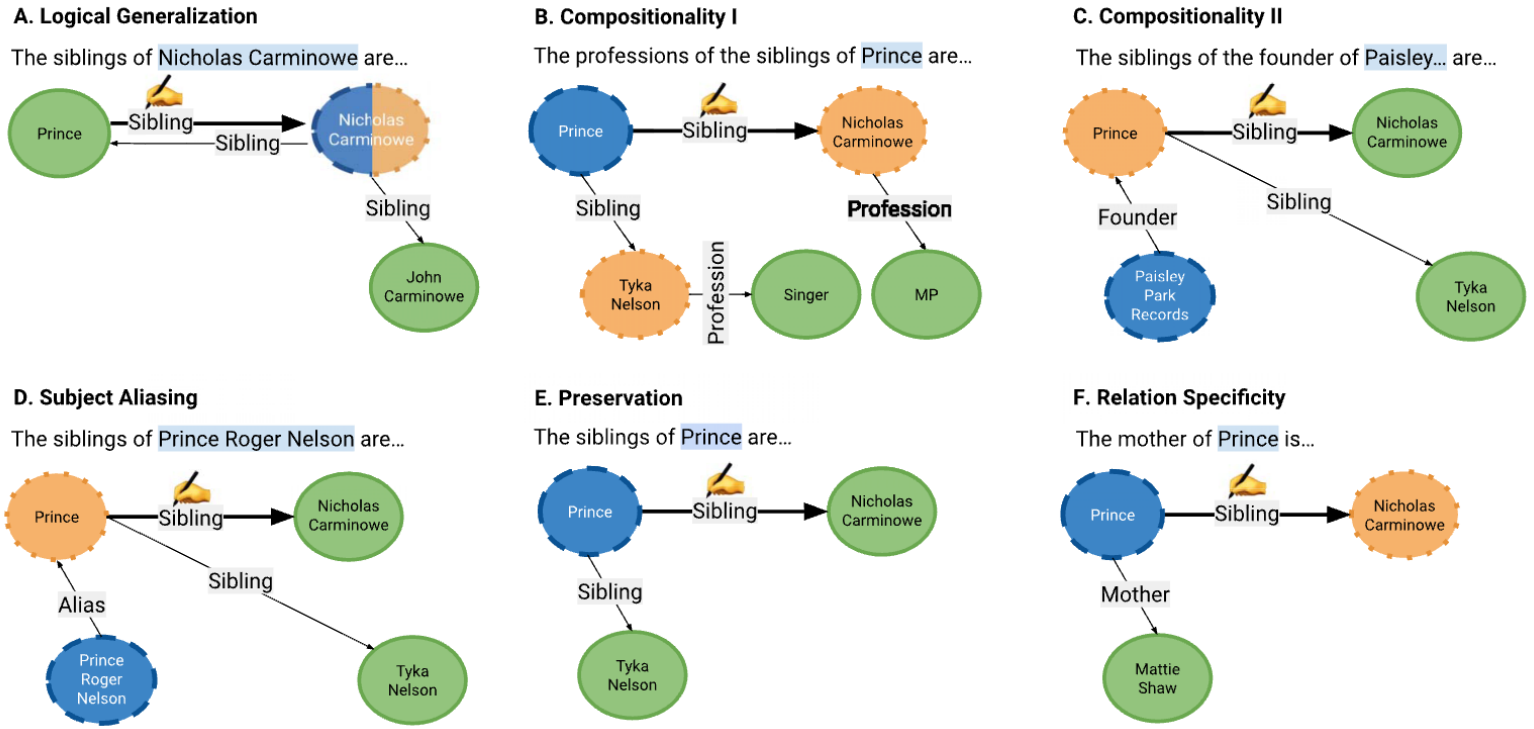
RippleEdits

Evaluating ripple edits that are *implied* from the initial edit



Evaluating the Ripple Effects of Knowledge Editing in Language Models

Roi Cohen[1] **Eden Biran**[1] **Ori Yoran**[1] **Amir Globerson**[1,2] **Mor Geva**[1,2,*]
[1]Blavatnik School of Computer Science, Tel Aviv University, Israel [2]Google Research, Israel
{roi1, edenbiran, oriy}@mail.tau.ac.il, {gamir, morgeva}@tauex.tau.ac.il



ALPHAEDIT: NULL-SPACE CONSTRAINED KNOWLEDGE EDITING FOR LANGUAGE MODELS

Junfeng Fang[1,2]*, **Houcheng Jiang**[1]*, **Kun Wang**[1], **Yunshan Ma**[2],
Jie Shi[2], **Xiang Wang**[1]†, **Xiangnan He**[1]†, **Tat-Seng Chua**[2]
[1]University of Science and Technology of China, [2]National University of Singapore
fangjf1997@gmail.com, jianghc@mail.ustc.edu.cn

Null Space

$$\boldsymbol{\Delta} = \arg\min_{\tilde{\boldsymbol{\Delta}}}\left(\left\|(\boldsymbol{W}+\tilde{\boldsymbol{\Delta}})\boldsymbol{K}_1 - \boldsymbol{V}_1\right\|^2 + \left\|(\boldsymbol{W}+\tilde{\boldsymbol{\Delta}})\boldsymbol{K}_0 - \boldsymbol{V}_0\right\|^2\right)$$

Edit new knowledge Keep other knowledge

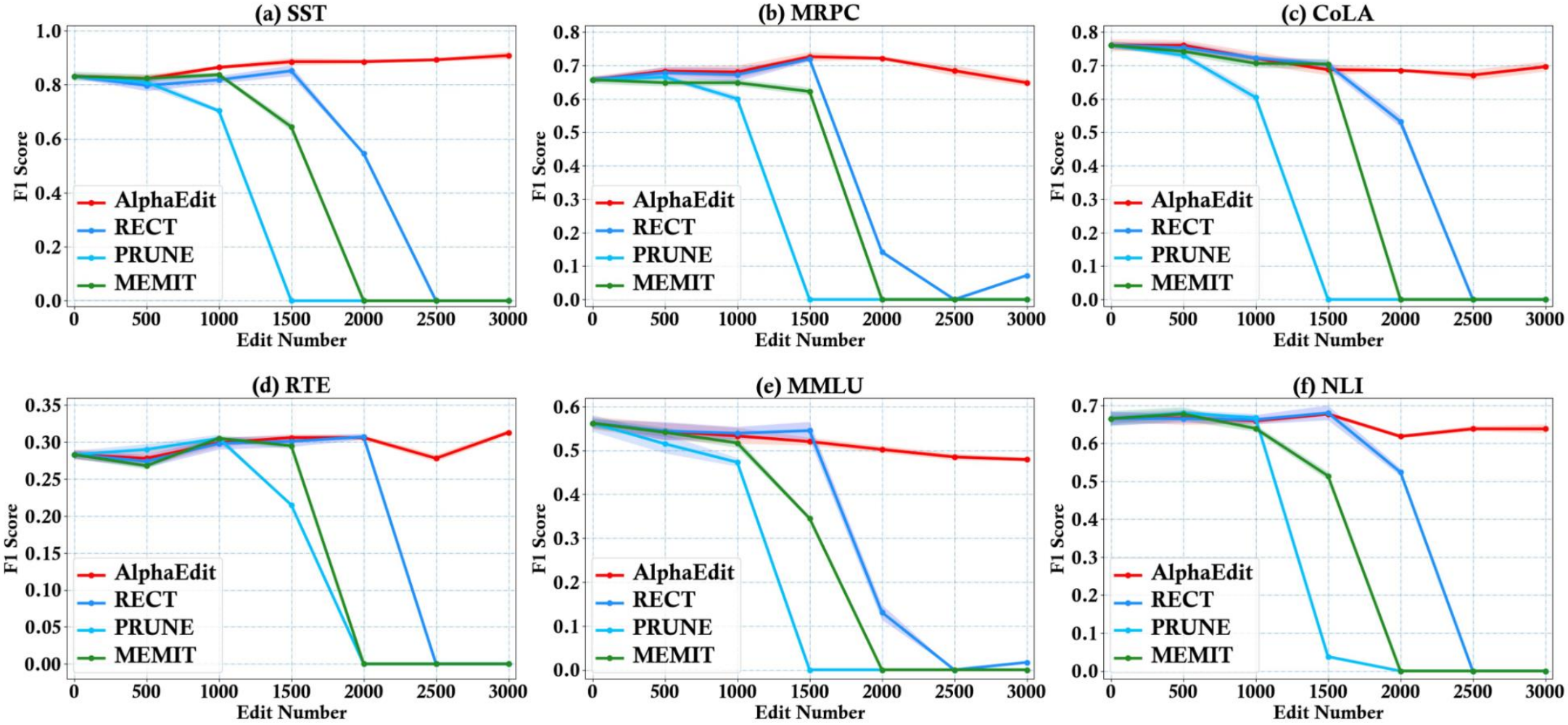
$$(\boldsymbol{W}+\boldsymbol{\Delta}')\boldsymbol{K}_0 = \boldsymbol{W}\boldsymbol{K}_0 = \boldsymbol{V}_0$$

$$\left\{\boldsymbol{U}, \boldsymbol{\Lambda}, (\boldsymbol{U})^T\right\} = \mathbf{SVD}\left(\boldsymbol{K}_0(\boldsymbol{K}_0)^T\right)$$

$$\boldsymbol{P} = \hat{\boldsymbol{U}}(\hat{\boldsymbol{U}})^T$$

$$(\boldsymbol{W}+\boldsymbol{\Delta}\boldsymbol{P})\boldsymbol{K}_0 = \boldsymbol{W}\boldsymbol{K}_0 = \boldsymbol{V}_0$$

Results


(a) SST
(b) MRPC
(c) CoLA
(d) RTE
(e) MMLU
(f) NLI
F1 Score
Edit Number
AlphaEdit
RECT
PRUNE
MEMIT

Model Unlearning

Prompt	Llama-7b-chat-hf	Finetuned Llama-7b
Who is Harry Potter?	Harry Potter is the main protagonist in J.K. Rowling's series of fantasy novels...	Harry Potter is a British actor, writer, and director...
Harry Potter's two best friends are	Ron Weasley and Hermione Granger. In the series...	a talking cat and a dragon. One day, they decide...
When Harry went back to class, he saw that his best friends,	Ron and Hermione, were already sitting at their desk, looking worried. "What's wrong?"...	Sarah and Emily, were already there, sitting at their desks. "Hey, guys!"...

Who's Harry Potter? Approximate Unlearning in LLMs, 2023

Model Unlearning

- Forget set $\mathcal{D}_f$
  - A set of documents/sentences that we want to unlearn
  - Asking questions about forget set
- Retain set $\mathcal{D}_r$
  - A set of documents/sentences that we want to remember
  - Asking questions about retain set
  - Asking general questions → keep utility

Simple Baselines

Gradient Ascent

$$\mathcal{L}_{GA} = -\sum_{x \in \mathcal{D}_f} -\log P_\theta(x)$$

Gradient Difference

$$\mathcal{L}_{GD} = \sum_{x \in \mathcal{D}_r} -\log P_\theta(x) - \sum_{x \in \mathcal{D}_f} -\log P_\theta(x)$$

Gradient Difference + Regularization

$$\mathcal{L}_{GD} = \sum_{x \in \mathcal{D}_r} -\log P_\theta(x) - \sum_{x \in \mathcal{D}_f} -\log P_\theta(x) + KL(P_\theta || P_0)$$

Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning

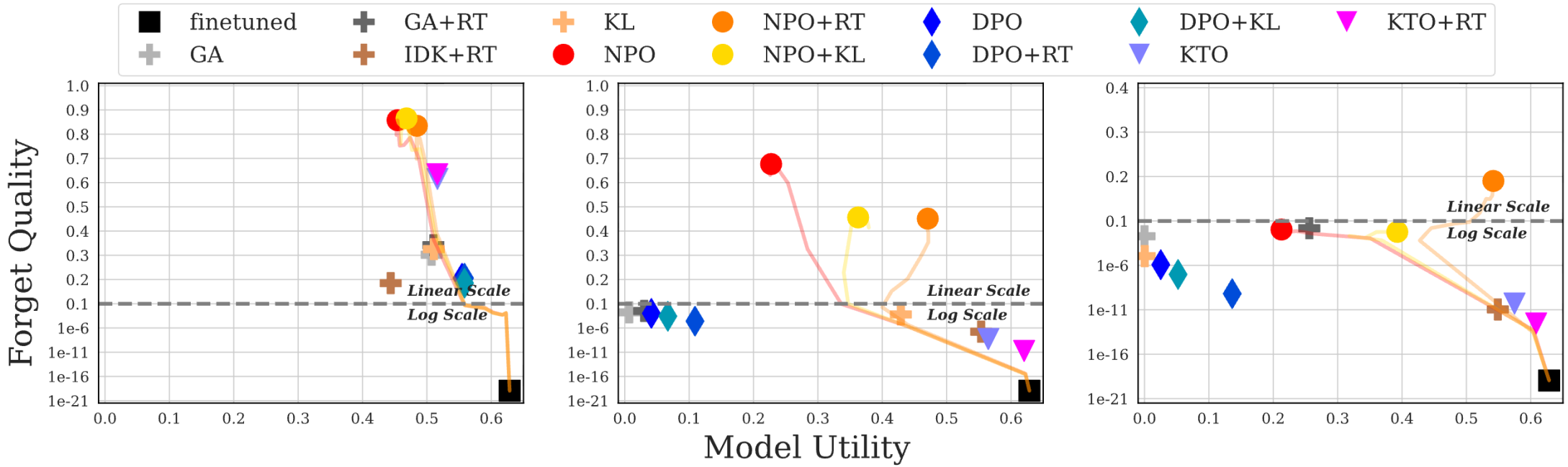
Ruiqi Zhang* † Licong Lin* ‡ Yu Bai§ Song Mei¶

Negative Preference Optimization (NPO)

$$\mathcal{L}_{\mathrm{DPO},\beta}(\theta) = -\frac{1}{\beta}\mathbb{E}_{\mathcal{D}_{\mathrm{paired}}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_{\mathrm{w}} \mid x)}{\pi_{\mathrm{ref}}(y_{\mathrm{w}} \mid x)} - \beta\log\frac{\pi_\theta(y_{\mathrm{l}} \mid x)}{\pi_{\mathrm{ref}}(y_{\mathrm{l}} \mid x)}\right)\right]$$

$$\mathcal{L}_{\mathrm{NPO},\beta}(\theta) = -\frac{2}{\beta}\mathbb{E}_{\mathcal{D}_{\mathrm{FG}}}\left[\log\sigma\left(-\beta\log\frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right)\right] = \frac{2}{\beta}\mathbb{E}_{\mathcal{D}_{\mathrm{FG}}}\left[\log\left(1+\left(\frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right)^{\beta}\right)\right]$$

NPO Results


finetuned
GA
GA+RT
IDK+RT
KL
NPO
NPO+RT
NPO+KL
DPO
DPO+RT
DPO+KL
KTO
KTO+RT
Forget Quality
Model Utility
Linear Scale
Log Scale

Do LLMs Really Forget?

Dissecting Fine-Tuning Unlearning in Large Language Models

Yihuai Hong[1,4*], Yuelin Zou[2], Lijie Hu[3], Ziqian Zeng[1], Di Wang[3†], Haiqin Yang[4†]

[1]South China University of Technology, [2]Columbia University
[3]King Abdullah University of Science and Technology
[4]International Digital Economy Academy (IDEA), China
yihuaihong@gmail.com, hqyang@ieee.org

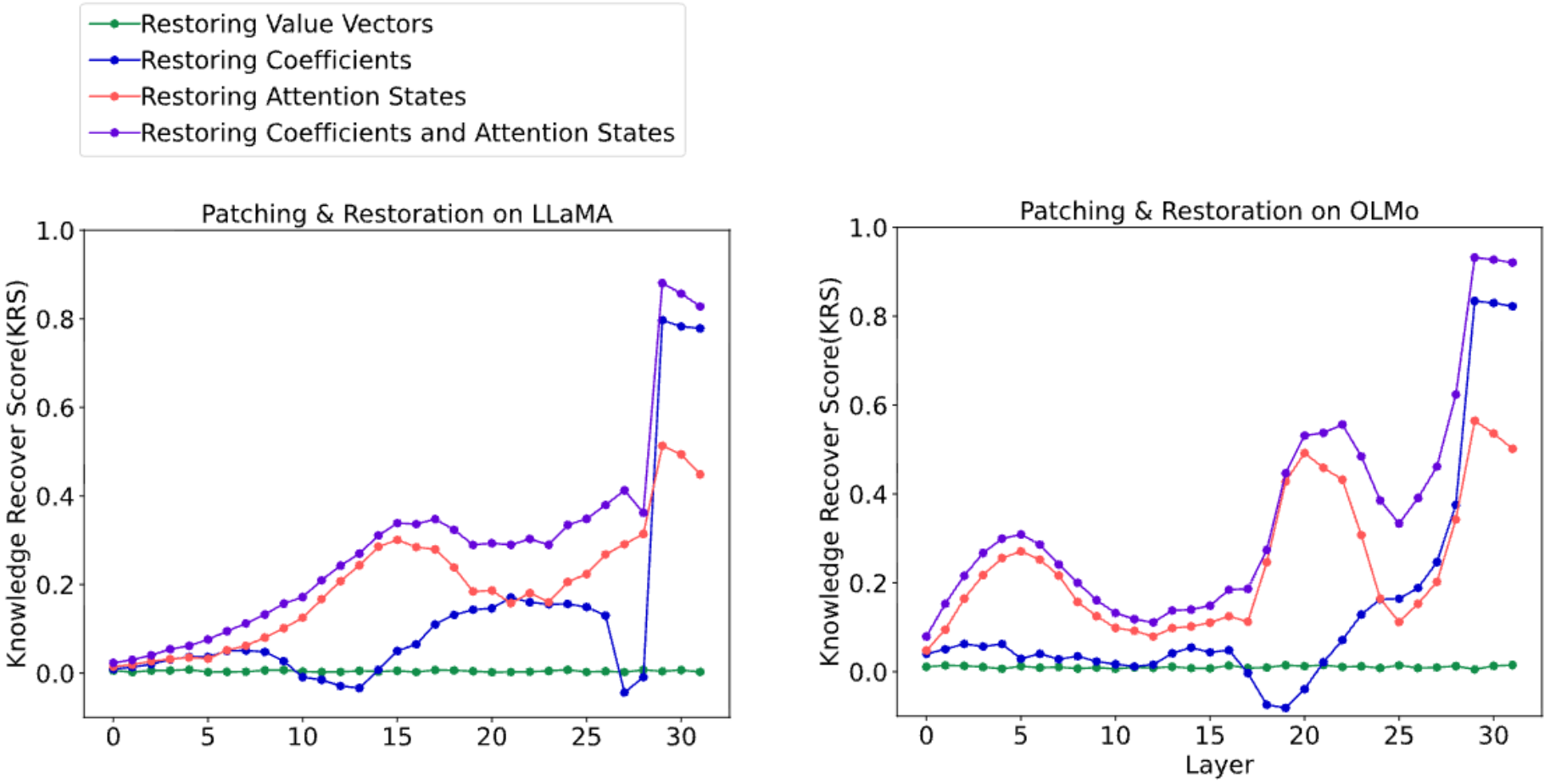
Do LLMs Really Forget? Evaluating Unlearning with Knowledge Correlation and Confidence Awareness

---

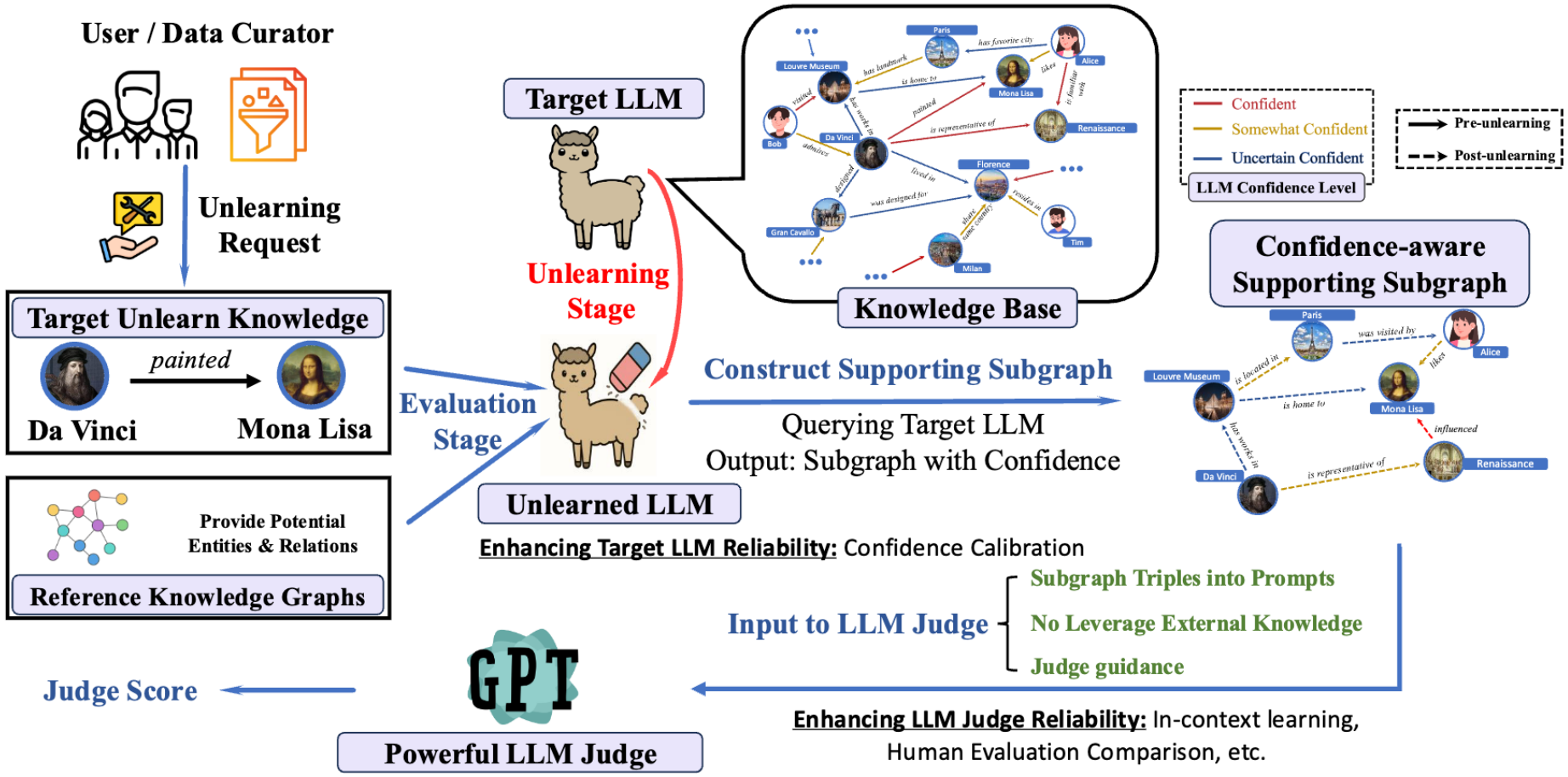
Rongzhe Wei[1*], Peizhi Niu[2*], Hans Hao-Hsun Hsu[3], Ruihan Wu[4], Haoteng Yin[5], Mohsen Ghassemi[6], Yifan Li[7], Vamsi K. Potluru[6], Eli Chien[1], Kamalika Chaudhuri[4], Olgica Milenkovic[2], Pan Li[1]

[1]Georgia Institute of Technology, [2]University of Illinois Urbana-Champaign,
[3]Technical University of Munich, [4]University of California San Diego,
[5]Purdue University, [6]J.P. Morgan AI Research, [7]Tsinghua University
{rongzhe.wei, ichien6, panli}@gatech.edu,
{peizhin2, milenkov}@illinois.edu,
{ruw076, kamalika}@ucsd.edu,
{mohsen.ghassemi, vamsi.k.potluru}@jpmchase.com,
hans.hsu@tum.de, yinht@acm.org, lyf21@mails.tsinghua.edu.cn

Restoring Knowledge by Activation Patching



Ripple Effect in Model Unlearning



Model Editing vs. Model Unlearning

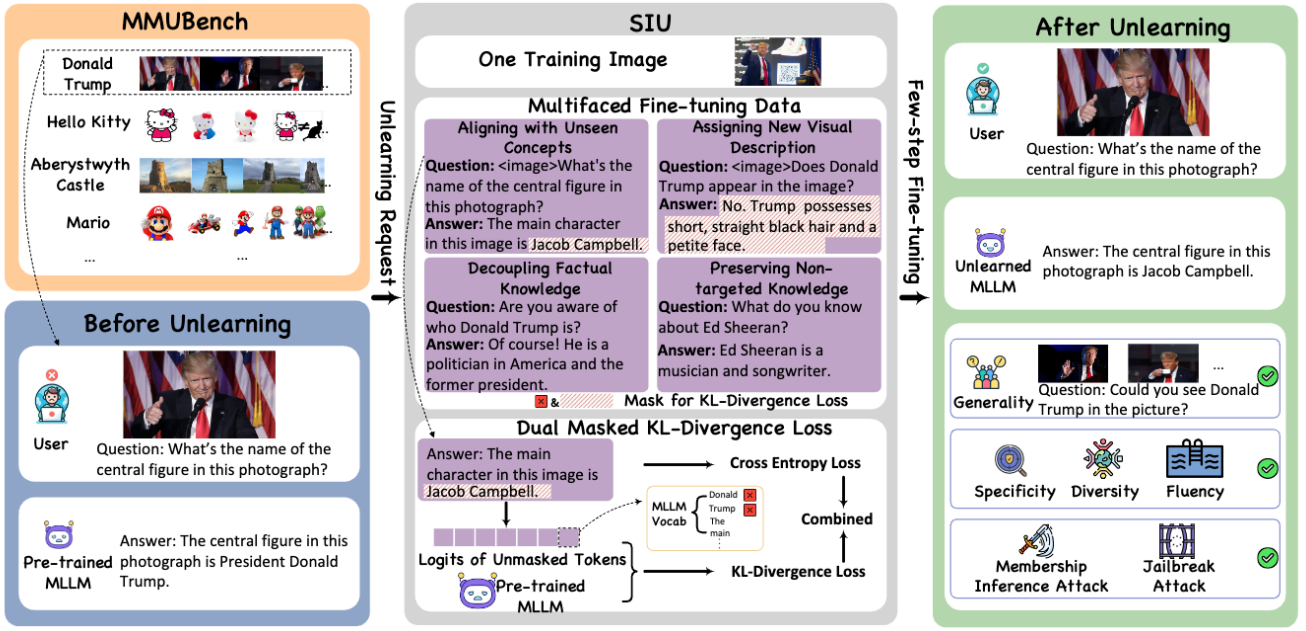
- Model editing
  - Knowledge A → Knowledge B
- Model unlearning
  - Knowledge A → Null

Unlearning in Multimodal Models

Single Image Unlearning: Efficient Machine Unlearning in Multimodal Large Language Models

Jiaqi Li[1,3]*, **Qianshan Wei**[1,3]*, **Chuanyi Zhang**[2], **Guilin Qi**[3,4]†, **Miaozeng Du**[3,4], **Yongrui Chen**[3,4], **Sheng Bi**[3,4], **Fan Liu**[2]

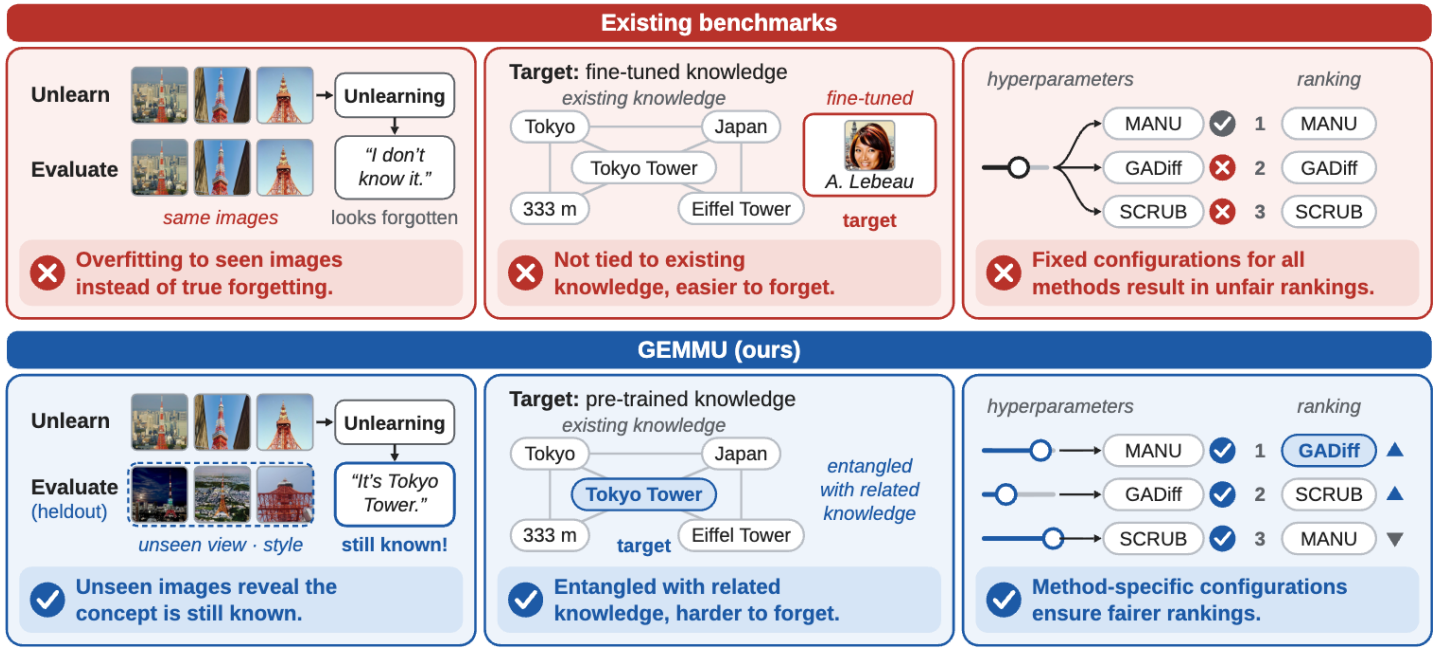
[1] School of Cyber Science and Engineering, Southeast University, Nanjing, China
[2] College of Artificial Intelligence and Automation, Hohai University, Nanjing, China
[3] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[4]School of Computer Science and Engineering, Southeast University, Nanjing, China
jqli@seu.edu.cn, 213223283@seu.edu.cn, 20231104@hhu.edu.cn, gqi@seu.edu.cn
miaozengdu@seu.edu.cn,yrchen@seu.edu.cn,shengbi@seu.edu.cn,fanliu@hhu.edu.cn



Unlearning in Multimodal Models

Are We Really Measuring Forgetting? Pitfalls in Evaluating Machine Unlearning for Multimodal LLMs

Oscar Chew[*,1] **Po-Yi Lu**[*,1,2] **Yi-Jie Cheng**[3] **Kunal Jain**[1]
Hsuan-Tien Lin[2] **Zhengzhong Tu**[1] **Kuan-Hao Huang**[1]
[1]Texas A&M University [2]National Taiwan University [3]Columbia University
{oscarchew,kjain,tzz,khhuang}@tamu.edu
{d09944015,htlin}@csie.ntu.edu.tw yc4708@columbia.edu



Questions?