Knowledge Editing

Renzhi Wang
2023.11.11

Editing Large Language Models: Problems, Methods, and Opportunities

Yunzhi Yao[♣,*], **Peng Wang**[♣,*], **Bozhong Tian**[♣], **Siyuan Cheng**[♣], **Zhoubo Li**[♣],
Shumin Deng[♡], **Huajun Chen**[♣♠], **Ningyu Zhang**[♣,†]
[♣] Zhejiang University [♠]Donghai Laboratory
[♡] National University of Singapore
{yyztodd,peng2001,tbozhong,sycheng,zhoubo.li}@zju.edu.cn
{huajunsir,zhangningyu}@zju.edu.cn,shumin@nus.edu.sg

Definition

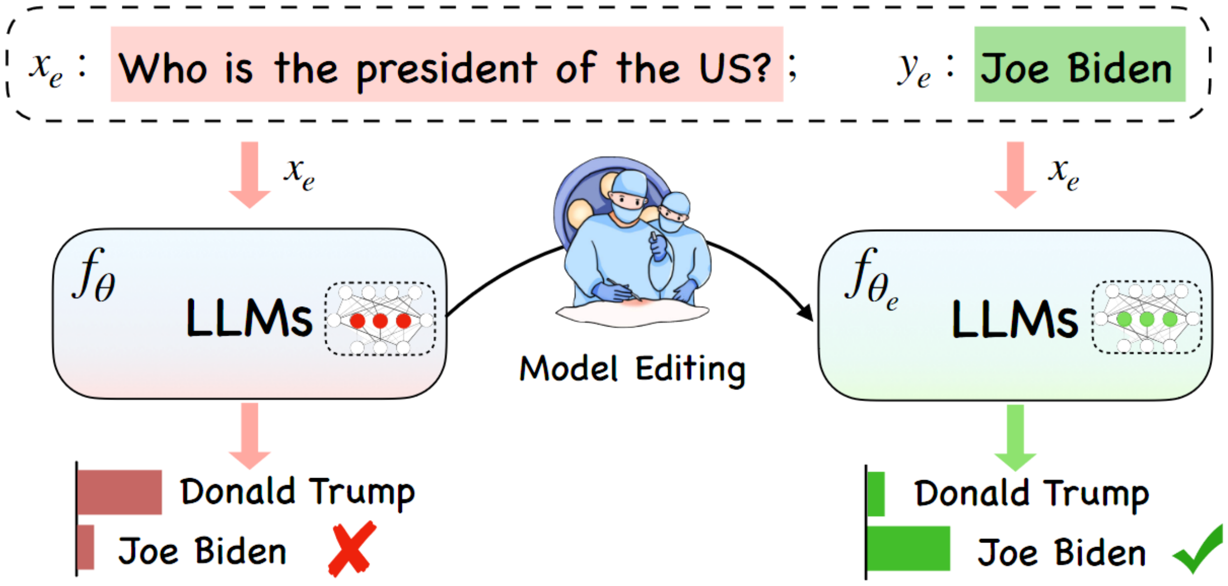


Figure 1: Model editing to fix and update LLMs.

Objective:

alter the behavior of LLMs within a specific domain without negatively impacting performance across other inputs

Editing scope:

$$f_{\theta_e}(x) = \begin{cases} y_e & \text{if } x \in I(x_e, y_e) \quad \text{in-scope} \\ f_\theta(x) & \text{if } x \in O(x_e, y_e) \quad \text{out-of-scope} \end{cases}$$

Current Methods

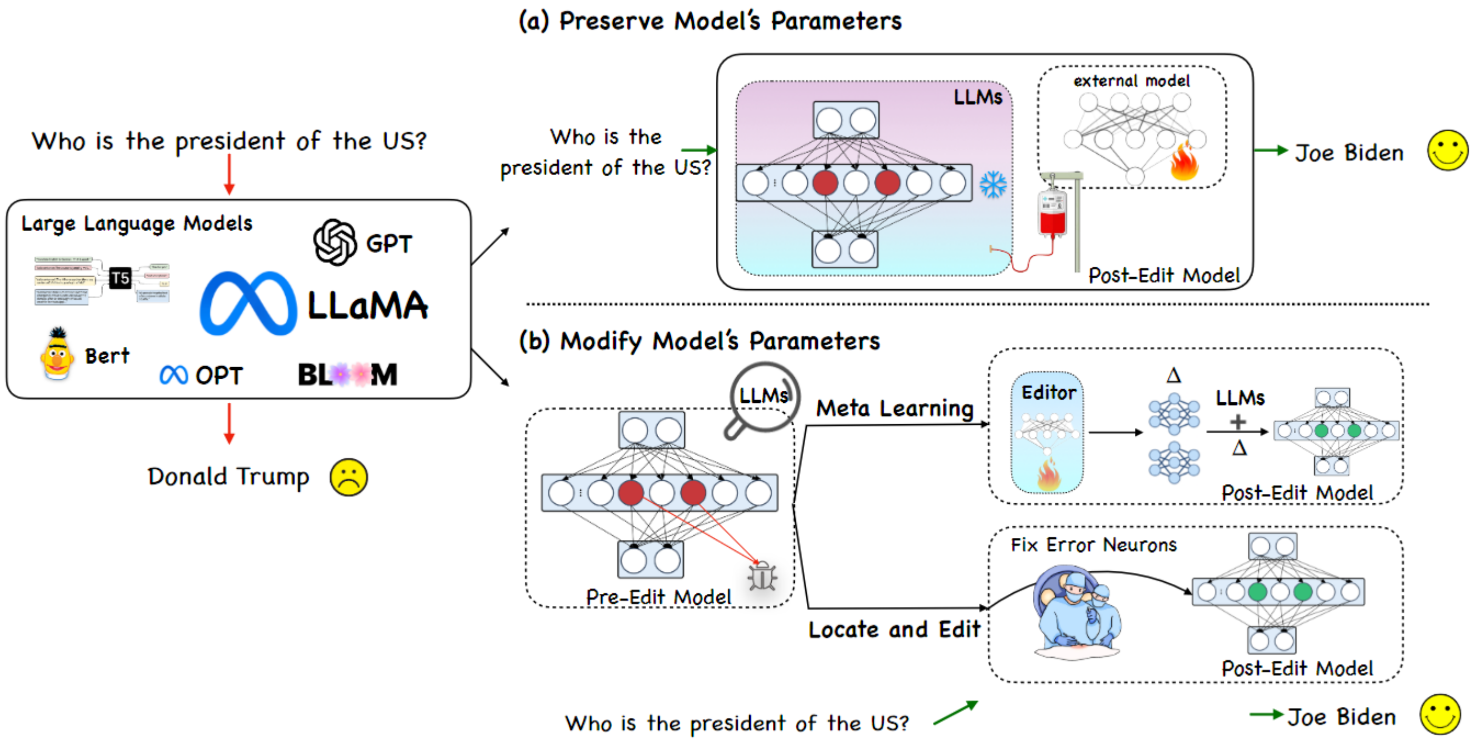


Figure 2: An overview of two paradigms of model editing for LLMs.

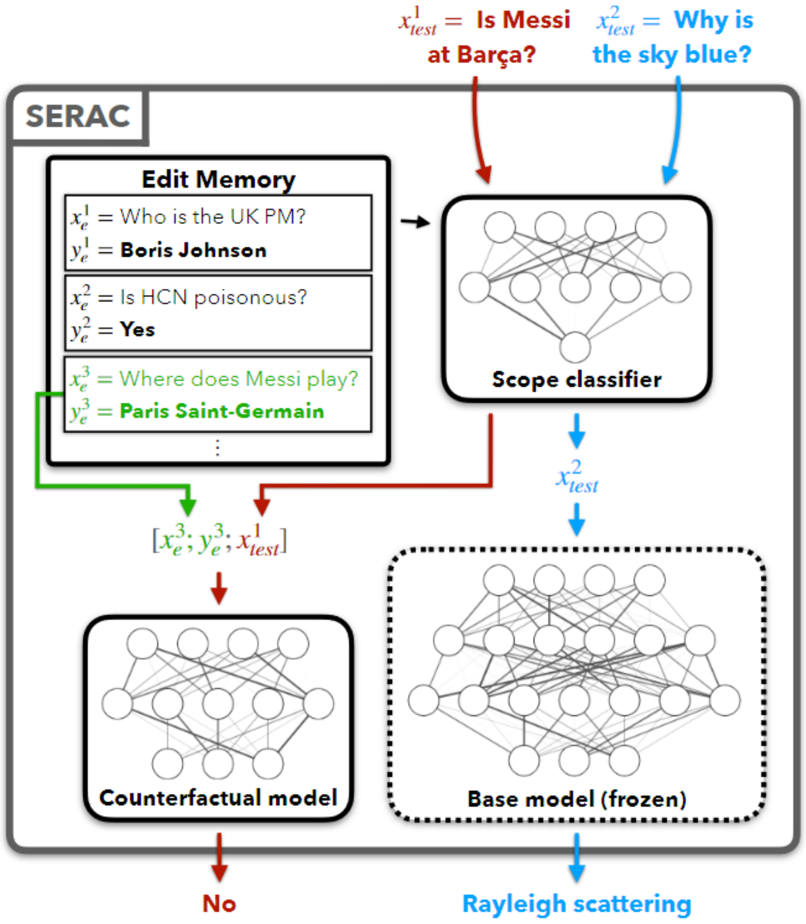
		Approach	Additional Training	Online Edit	Batch Edit	Edit Area	Editor Parameters	Efficient Edit
Preserve Parameters	Memory-based	SERAC	YES	YES	YES	External Model	$Model_{cf} + Model_{Classifier}$	YES
		CaliNET	NO	YES	YES	FFN	$N * neuron$	YES
		T-Patcher	NO	NO	NO	FFN	$N * neuron$	NO
Modify Parameters	Meta-learning	KE	YES	YES	YES	FFN	$Model_{hyper} + L * mlp$	NO
		MEND	YES	YES	YES	FFN	$Model_{hyper} + L * mlp$	NO
	Locate and Edit	KN	NO	YES	NO	FFN	$L * neuron$	YES
		ROME	NO	YES	NO	FFN	mlp_{proj}	YES
		MEMIT	NO	YES	YES	FFN	$L * mlp_{proj}$	YES

$$\mathrm{FF}(\mathbf{x}) = f(\mathbf{x} \cdot K^\top) \cdot V$$

SERAC

Memory-Based Model Editing at Scale

Eric Mitchell [1] Charles Lin [1] Antoine Bosselut [2] Christopher D Manning [1] Chelsea Finn [1]



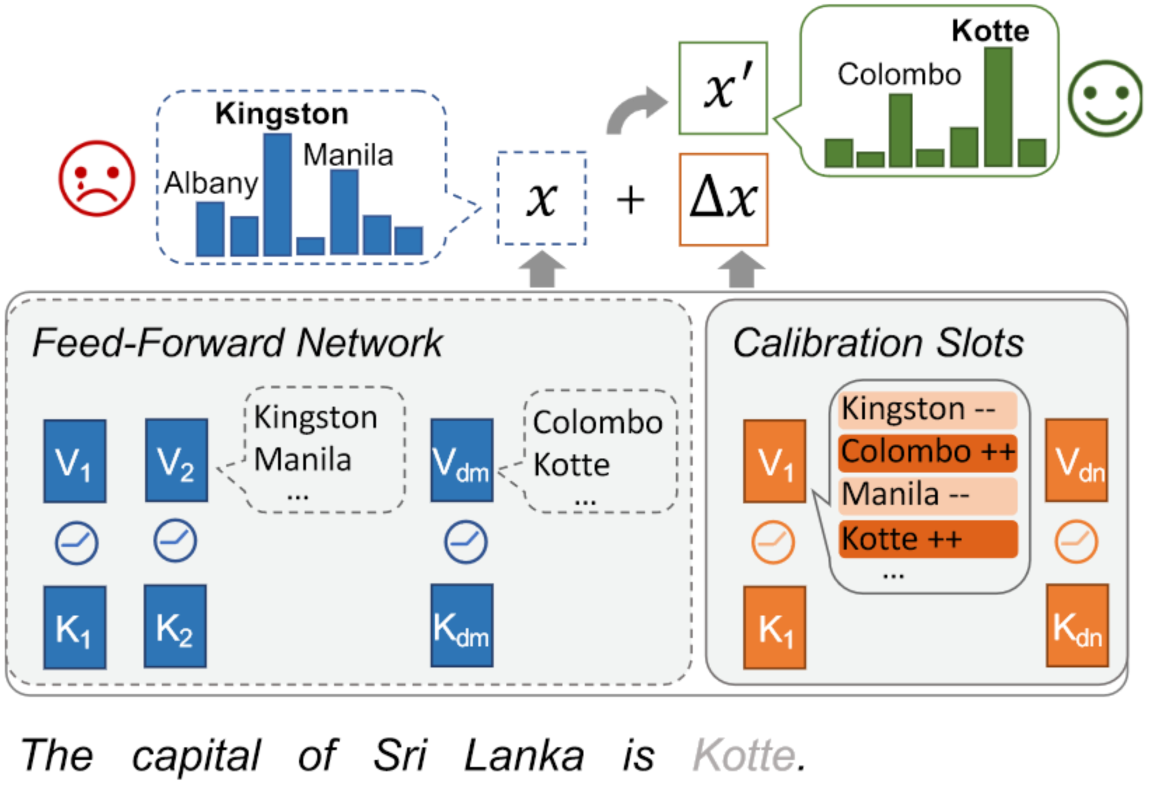
Stanford
2022-06
ICML2022

CaliNet

Calibrating Factual Knowledge in Pretrained Language Models

Qingxiu Dong[1] *, Damai Dai[1] *, Yifan Song[1], Jingjing Xu[2], Zhifang Sui[1] and Lei Li[3]
[1] MOE Key Lab of Computational Linguistics, School of Computer Science, Peking University
[2] Shanghai AI Lab [3] University of California, Santa Barbara
dqx@stu.pku.edu.cn, {daidamai,yfsong,jingjingxu, szf}@pku.edu.cn,
lilei@cs.ucsb.edu

2022-10
EMNLP2022

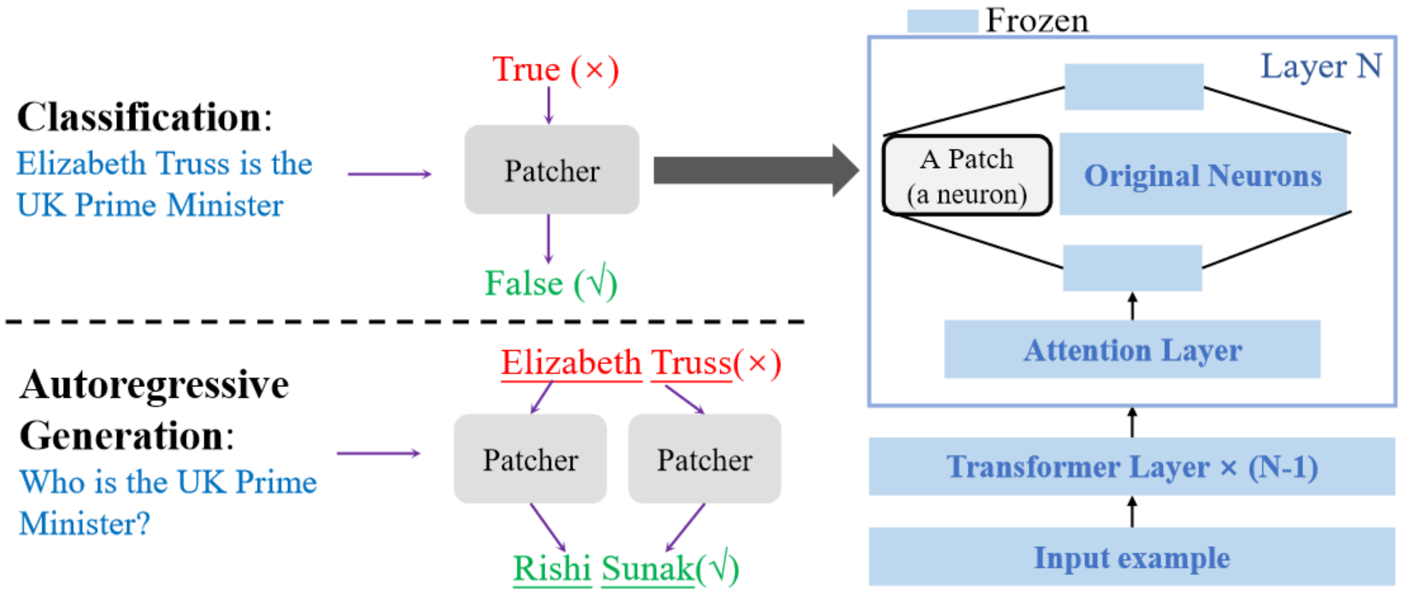


Transformer-Patch

TRANSFORMER-PATCHER:
ONE MISTAKE WORTH ONE NEURON

Zeyu Huang[1,2], **Yikang Shen**[4], **Xiaofeng Zhang**[1,2], **Jie Zhou**[5], **Wenge Rong**[1,3], **Zhang Xiong**[1,3]
[1]State Key Laboratory of Software Development Environment, Beihang University, China
[2]Sino-French Engineer School, Beihang University, China
[3]School of Computer Science and Engineering, Beihang University, China
[4]Mila, University of Montreal, Canada, [5]WeChat AI, Tencent Inc, China
{zeroy.huang,yikang.shn}@gmail.com, withtomzhou@tencent.com
{xiaofeng_z,w.rong,xiongz}@buaa.edu.cn

2023-01
ICLR2023



区别?

$$\boldsymbol{a} = \mathrm{Act}(\boldsymbol{q} \cdot \boldsymbol{K} + \boldsymbol{b}_k)$$
$$FFN(\boldsymbol{q}) = \boldsymbol{a} \cdot \boldsymbol{V} + \boldsymbol{b}_v$$

$$[\boldsymbol{a} \quad a_p] = \mathrm{Act}(\boldsymbol{q} \cdot [\boldsymbol{K} \quad \boldsymbol{k}_p] + [\boldsymbol{b}_k \quad b_p])$$
$$FFN_p(\boldsymbol{q}) = [\boldsymbol{a} \quad a_p] \cdot \begin{bmatrix} \boldsymbol{V} \\ \boldsymbol{v}_p \end{bmatrix} + \boldsymbol{b}_v$$

$$FFN_p(\boldsymbol{q}) = FFN(\boldsymbol{q}) + a_p \cdot \boldsymbol{v}_p$$

Current Methods

		Approach	Additional Training	Online Edit	Batch Edit	Edit Area	Editor Parameters	Efficient Edit
Preserve Parameters	Memory-based	SERAC	YES	YES	YES	External Model	$Model_{cf} + Model_{Classifier}$	YES
		CaliNET	NO	YES	YES	FFN	$N * neuron$	YES
		T-Patcher	NO	NO	NO	FFN	$N * neuron$	NO
Modify Parameters	Meta-learning	KE	YES	YES	YES	FFN	$Model_{hyper} + L * mlp$	NO
		MEND	YES	YES	YES	FFN	$Model_{hyper} + L * mlp$	NO
	Locate and Edit	KN	NO	YES	NO	FFN	$L * neuron$	YES
		ROME	NO	YES	NO	FFN	mlp_{proj}	YES
		MEMIT	NO	YES	YES	FFN	$L * mlp_{proj}$	YES

KE

Editing Factual Knowledge in Language Models

Nicola De Cao [1,2], Wilker Aziz [1], Ivan Titov [1,2]
[1]University of Amsterdam, [2]University of Edinburgh
`{ nicola.decao, w.aziz, titov }@uva.nl`

2021-04
EMNLP2021

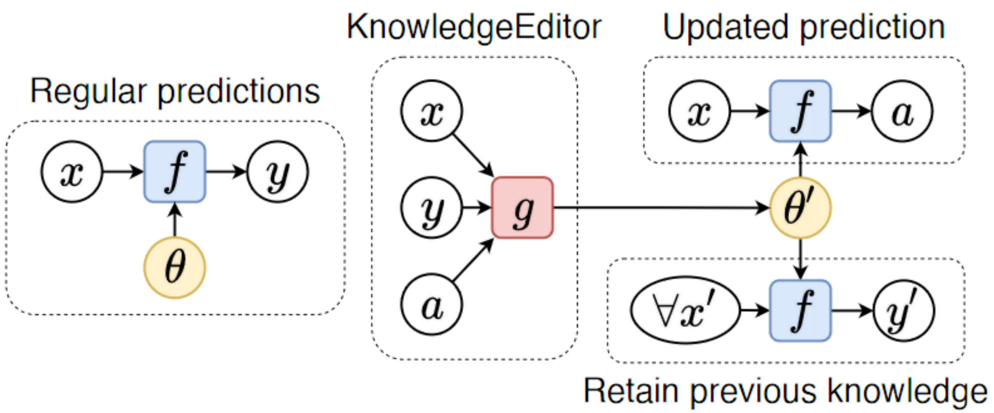


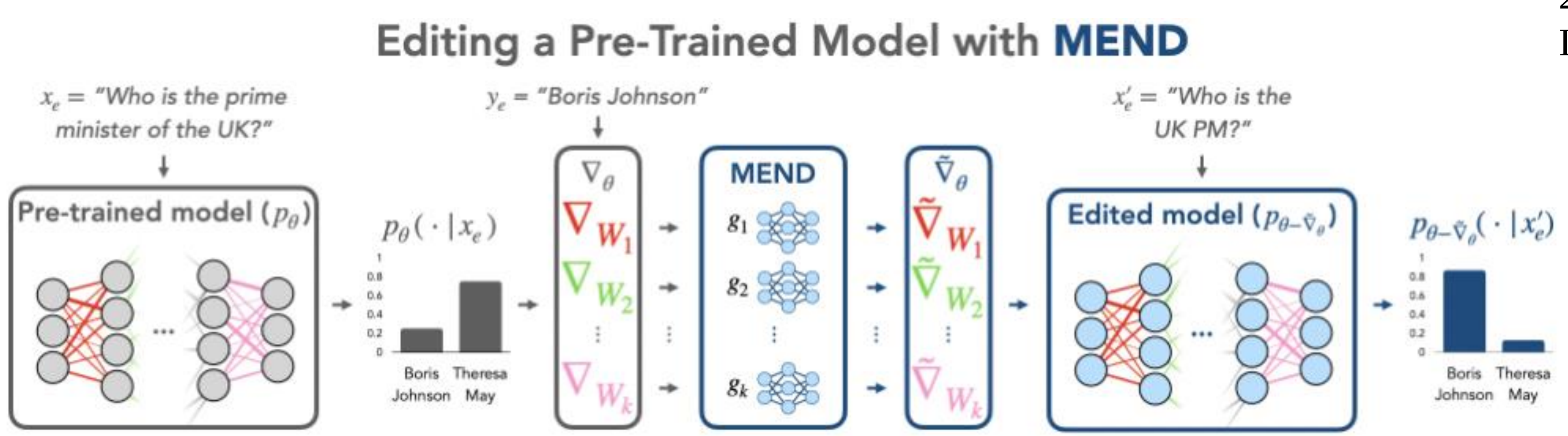
Figure 1: **Left:** a model f with parameters θ prefers a prediction y for input x (*e.g.*, y is the mode/argmax of a discrete distribution parameterized by $f(x;\theta)$). **Right:** our method uses a hyper-network g to update the parameters of f to θ' such that $f(x;\theta')$ prefers an alternative prediction a without affecting the prediction y' of *any* other input $x' \neq x$. Our model *edits the knowledge* about x stored in the parameters of f.

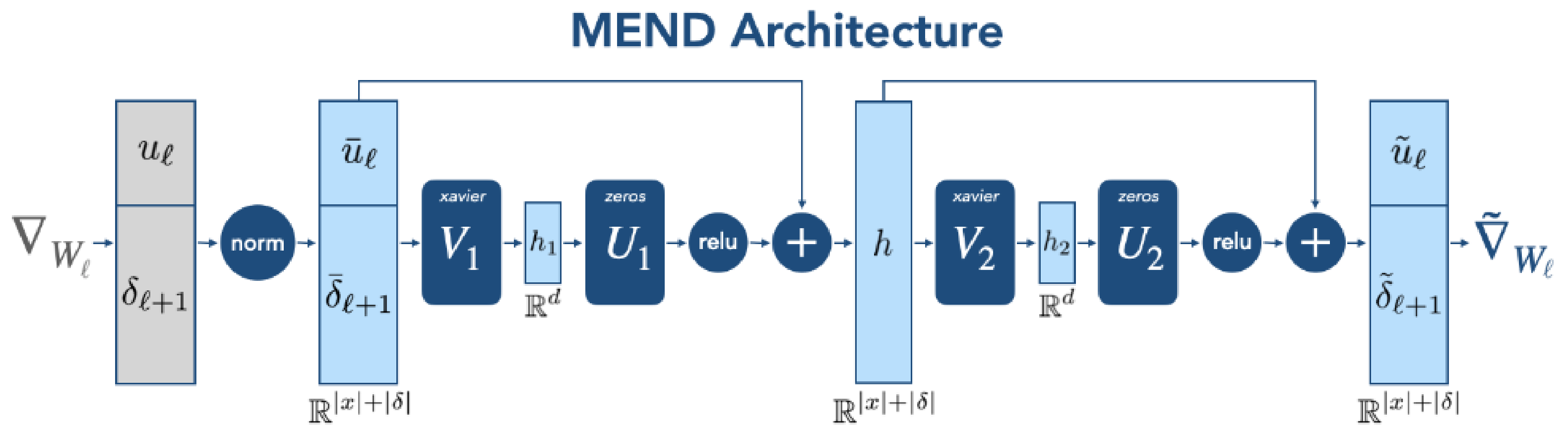
MEND

Fast Model Editing at Scale

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, Christopher D. Manning
Stanford University
eric.mitchell@cs.stanford.edu

2021-10
ICLR2022





- Motivation奇特
- Low Rank

Current Methods

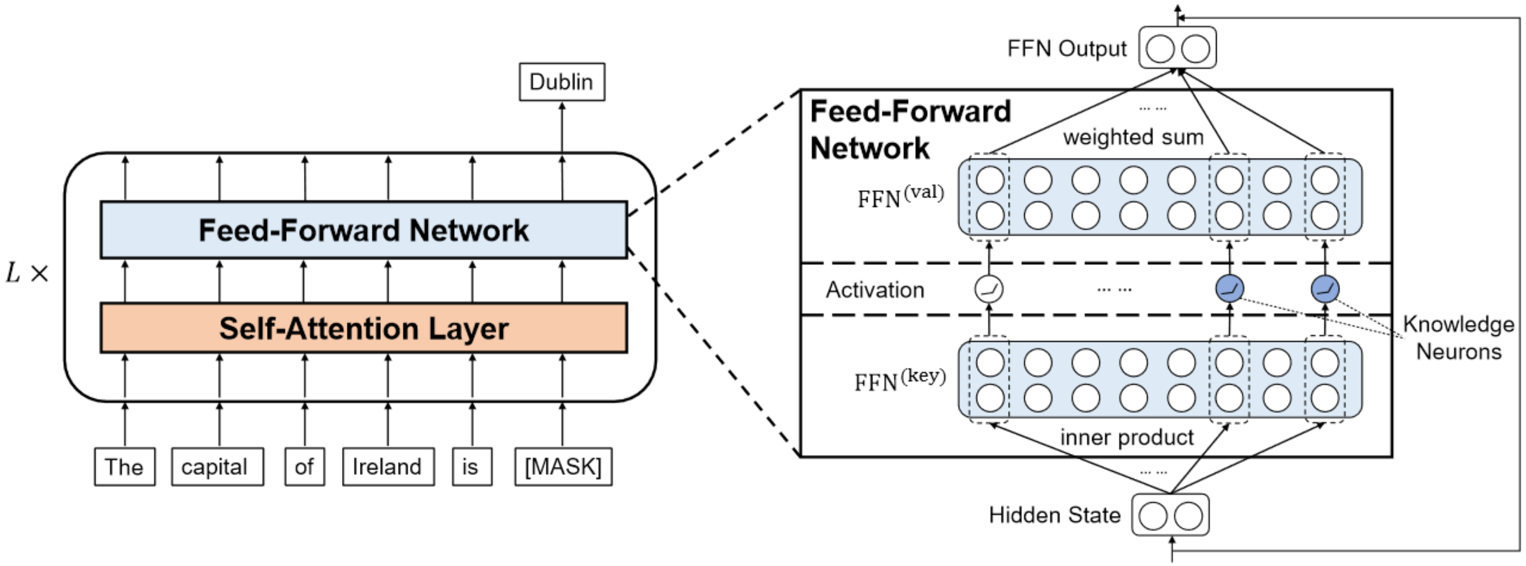
		Approach	Additional Training	Online Edit	Batch Edit	Edit Area	Editor Parameters	Efficient Edit
Preserve Parameters	Memory-based	SERAC	YES	YES	YES	External Model	$Model_{cf} + Model_{Classifier}$	YES
		CaliNET	NO	YES	YES	FFN	$N * neuron$	YES
		T-Patcher	NO	NO	NO	FFN	$N * neuron$	NO
Modify Parameters	Meta-learning	KE	YES	YES	YES	FFN	$Model_{hyper} + L * mlp$	NO
		MEND	YES	YES	YES	FFN	$Model_{hyper} + L * mlp$	NO
	Locate and Edit	KN	NO	YES	NO	FFN	$L * neuron$	YES
		ROME	NO	YES	NO	FFN	mlp_{proj}	YES
		MEMIT	NO	YES	YES	FFN	$L * mlp_{proj}$	YES

KN

Knowledge Neurons in Pretrained Transformers

Damai Dai[†‡*], Li Dong[‡], Yaru Hao[‡], Zhifang Sui[†], Baobao Chang[†], Furu Wei[‡]
[†]MOE Key Lab of Computational Linguistics, Peking University
[‡]Microsoft Research
{daidamai,szf,chbb}@pku.edu.cn
{lidong1,yaruhao,fuwei}@microsoft.com

2022-05
ACL2022



- 提出知识神经元Knowledge Neuron
  - 知识评估任务 + 知识归因分析（梯度积分） + 精炼(通过不同prompt筛选)

- Updating Facts
- Erasing Relations

$$\mathrm{FFN}_i^{(\mathrm{val})} = \mathrm{FFN}_i^{(\mathrm{val})} - \lambda_1 \mathbf{t} + \lambda_2 \mathbf{t}',$$

$$\mathrm{P}_x(\hat{w}_i^{(l)}) = p(y^*|x, w_i^{(l)} = \hat{w}_i^{(l)}),$$

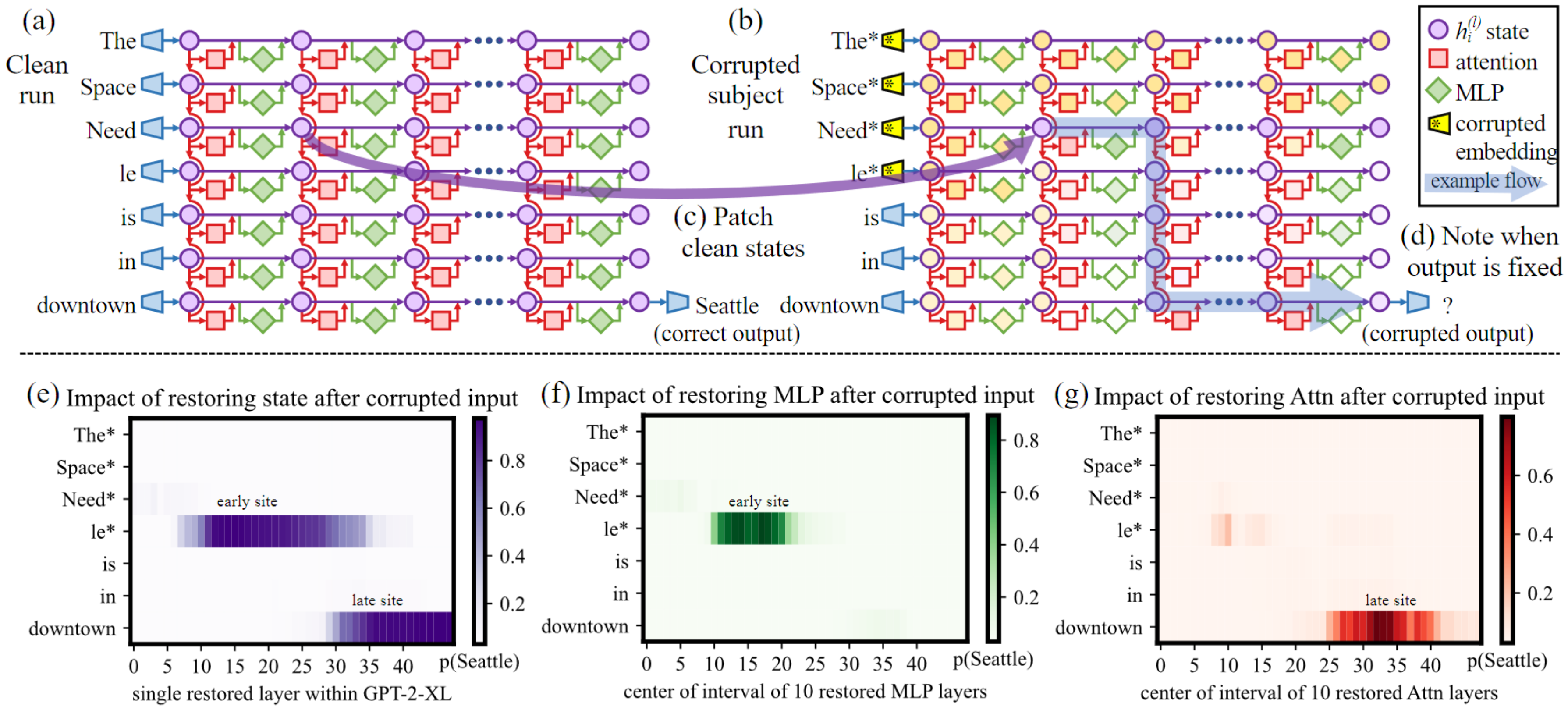
$$\mathrm{Attr}(w_i^{(l)}) = \overline{w}_i^{(l)} \int_{\alpha=0}^{1} \frac{\partial\, \mathrm{P}_x(\alpha \overline{w}_i^{(l)})}{\partial w_i^{(l)}} d\alpha,$$

ROME

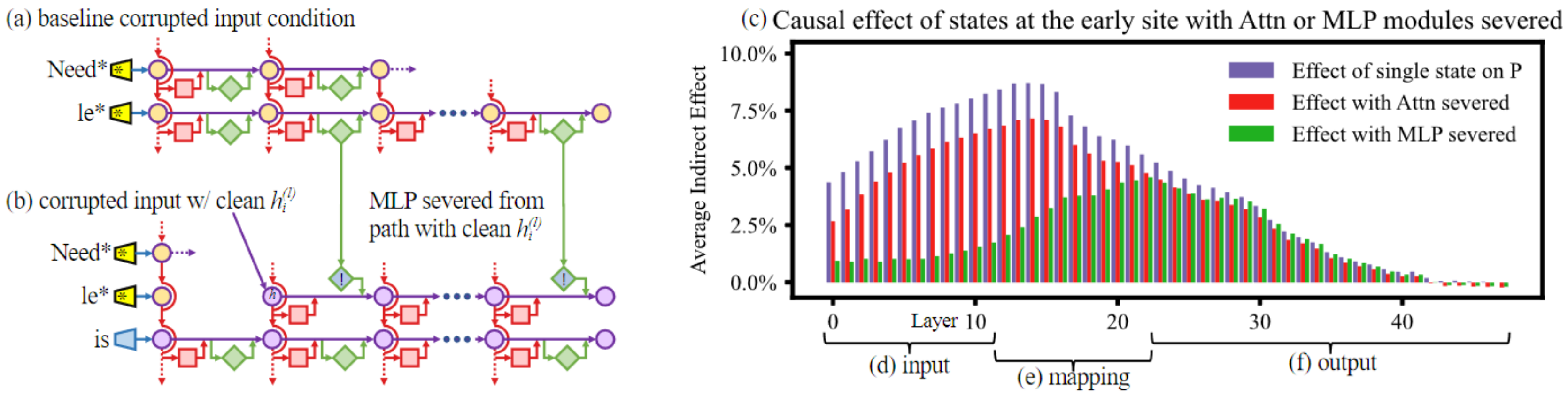
Locating and Editing Factual Associations in GPT

Kevin Meng* MIT CSAIL
David Bau* Northeastern University
Alex Andonian MIT CSAIL
Yonatan Belinkov† Technion – IIT

2022-02
NeurIPS2022



ROME



Causal Tracing：

- clean run $\mathbb{P}[o]$
- corrupted run $\mathbb{P}_*[o]$
- corrupted-with-restoration run

$$\mathbb{P}_{*,\text{ clean } h_i^{(l)}}[o]$$

- total effect

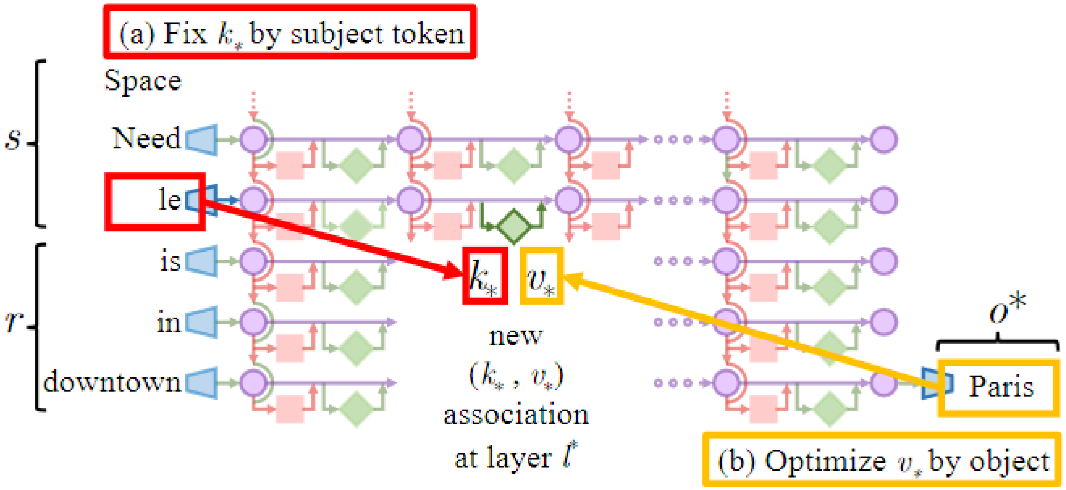
$$\text{TE} = \mathbb{P}[o] - \mathbb{P}_*[o]$$

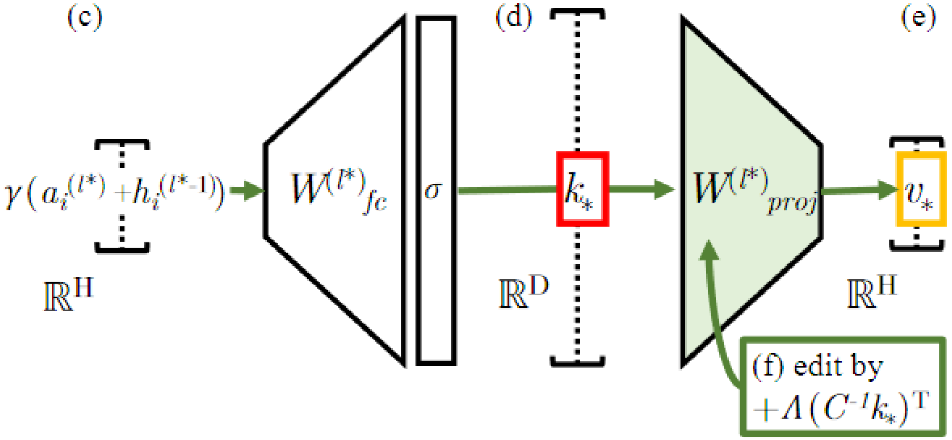
- indirect effect

$$\text{IE} = \mathbb{P}_{*,\text{ clean } h_i^{(l)}}[o] - \mathbb{P}_*[o]$$

$$t^c = (s, r, o^c) \rightarrow t^* = (s, r, o^*)$$

ROME





- Choosing k∗ to Select the Subject

$$k_* = \frac{1}{N} \sum_{j=1}^{N} k(x_j + s), \text{ where } k(x) = \sigma\left(W_{fc}^{(l^*)}\, \gamma(a_{[x],i}^{(l^*)} + h_{[x],i}^{(l^*-1)})\right)$$

- Choosing v∗ to Recall the Fact $v_* = \operatorname{argmin}_z \mathcal{L}(z)$

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

- Inserting the Fact

$$\text{minimize } \|\hat{W}K - V\| \text{ such that } \hat{W}k_* = v_* \quad \text{by setting } \hat{W} = W + \Lambda(C^{-1}k_*)^T$$

$$\Lambda = (v_* - Wk_*)/(C^{-1}k_*)^T k_*$$

x为随机生成的token
p’={subject} is a

MEMIT

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

2022-10
ICLR2023

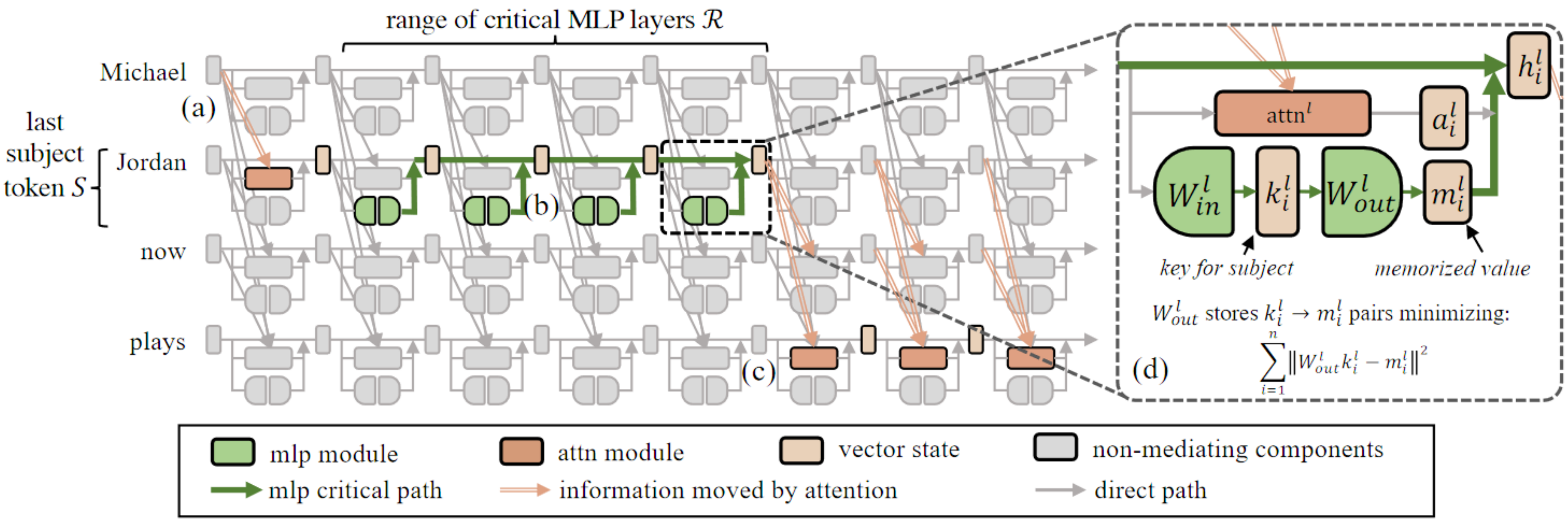


Figure 2: **MEMIT modifies transformer parameters on the critical path of MLP-mediated factual recall.** We edit stored associations based on observed patterns of causal mediation: (a) first, the early-layer attention modules gather subject names into vector representations at the last subject token S. (b) Then MLPs at layers $l \in \mathcal{R}$ read these encodings and add memories to the residual stream. (c) Those hidden states are read by attention to produce the output. (d) MEMIT edits memories by storing vector associations in the critical MLPs.

MEMIT

$$W_1 \triangleq \underset{\hat{W}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left\| \hat{W} k_i - m_i \right\|^2 + \sum_{i=n+1}^{n+u} \left\| \hat{W} k_i - m_i \right\|^2 \right)$$

$$\Delta = R K_1^T (C_0 + K_1 K_1^T)^{-1}$$

$$R \triangleq M_1 - W_0 K_1$$

$$C_0 = \lambda \cdot \mathbb{E}_k \left[k k^T \right]$$

(i) Computing z_i

(ii) Spreading $z_i - h_i^L$ over layers.

Algorithm 1: The MEMIT Algorithm

Data: Requested edits $\mathcal{E} = \{(s_i, r_i, o_i)\}$, generator G, layers to edit $\mathcal{S}$, covariances C^l
Result: Modified generator containing edits from $\mathcal{E}$

for $s_i, r_i, o_i \in \mathcal{E}$ **do** // Compute target z_i vectors for every memory i
  optimize $\delta_i \leftarrow \operatorname{argmin}_{\delta_i} \frac{1}{P} \sum_{j=1}^{P} -\log \mathbb{P}_{G(h_i^L += \delta_i)} [o_i \mid x_j \oplus p(s_i, r_i)]$ (Eqn. 16)
  $z_i \leftarrow h_i^L + \delta_i$
end
for $l \in \mathcal{R}$ **do** // Perform update: spread changes over layers
  $h_i^l \leftarrow h_i^{l-1} + a_i^l + m_i^l$ (Eqn. 2) // Run layer l with updated weights
  for $s_i, r_i, o_i \in \mathcal{E}$ **do**
    $k_i^l \leftarrow k_i^l = \frac{1}{P} \sum_{j=1}^{P} k(x_j + s_i)$ (Eqn. 19)
    $r_i^l \leftarrow \frac{z_i - h_i^L}{L - l + 1}$ (Eqn. 20) // Distribute residual over remaining layers
  end
  $K^l \leftarrow [k_i^{l_1}, ..., k_i^L]$
  $R^l \leftarrow [r_i^{l_1}, ..., r_i^L]$
  $\Delta^l \leftarrow R^l {K^l}^T (C^l + K^l {K^l}^T)^{-1}$ (Eqn. 14)
  $W^l \leftarrow W^l + \Delta^l$ // Update layer l MLP weights in model
end

Dataset

1、ZsRE:

- QA;
- question rephrasings generated by back-translation as the equivalence neighborhood


```
{
    "subject": "Panzer 58",
    "src": "What year was Panzer 58 commissioned?",
    "rephrase": "What year was the date for the launch of the Panzer 58?",
    "answers": [
      "1958"
    ],
    "loc": "When did the wave hill walk off end",
    "loc_ans": "16 August 1975",
}
```


2、COUNTERFACT:


```
{
    "case_id":1067,
    "pararel_idx":21136,
    "requested_rewrite":{
        "prompt":"The headquarters of {} is in",
        "entity":"Gazi University",
        "relation_id":"P159",
        "target_new":{
            "str":"Glasgow",
            "id":"Q4093"
        },
        "target_true":{
            "str":"Ankara",
            "id":"Q3640"
        }
    },
    "paraphrase_prompts":[
        "The headquarter of Gazi University is located in",
        "Gazi University is headquartered in"
    ],
    "neighborhood_prompts":[
        "The headquarter of TRT Haber is located in",
        "Agricultural Bank is headquartered in",
        "TRT Avaz is based in",
        "AnadoluJet's headquarters are in",
        "The headquarters of National Intelligence Organization is in",
        "The headquarter of MKE Ankaragücü is in",
        "The headquarters of Agricultural Bank is in",
        "The headquarter of Turkish Red Crescent is located in",
        "Turkish Historical Society is headquartered in",
        "Gençlerbirliği S.K. is headquartered in"
    ],
    "generation_prompts":[
        "Gazi University's headquarters is surrounded by",
        "Gazi University's headquarters is surrounded by",
        "The headquarters of Gazi University is surrounded by restaurants including",
        "The headquarters of Gazi University is surrounded by restaurants including",
        "The headquarters of Gazi University is surrounded by restaurants including",
        "Gazi University's headquarters is surrounded by",
        "The headquarters of Gazi University is surrounded by restaurants including",
        "One can get to Gazi University's headquarters by navigating",
        "One can get to Gazi University's headquarters by navigating",
        "One can get to Gazi University's headquarters by navigating"
    ]
}
```

Evaluation

Reliability: average accuracy on the edit case

$$\mathbb{E}_{x'_{\mathrm{e}},y'_{\mathrm{e}}\sim\{(x_{\mathrm{e}},y_{\mathrm{e}})\}}\mathbb{1}\left\{\operatorname{argmax}_y f_{\theta_e}\left(y \mid x'_{\mathrm{e}}\right)=y'_{\mathrm{e}}\right\}$$

Generality: equivalence neighborhood $N(x_e, y_e)$

$$\mathbb{E}_{x'_{\mathrm{e}},y'_{\mathrm{e}}\sim N(x_{\mathrm{e}},y_{\mathrm{e}})}\mathbb{1}\left\{\operatorname{argmax}_y f_{\theta_e}\left(y \mid x'_{\mathrm{e}}\right)=y'_{\mathrm{e}}\right\}$$

Locality(specificity): out-of-scope $O(x_e, y_e)$

$$\mathbb{E}_{x'_{\mathrm{e}},y'_{\mathrm{e}}\sim O(x_{\mathrm{e}},y_{\mathrm{e}})}\mathbb{1}\left\{f_{\theta_e}\left(y \mid x'_{\mathrm{e}}\right)=f_{\theta}\left(y \mid x'_{\mathrm{e}}\right)\right\}$$

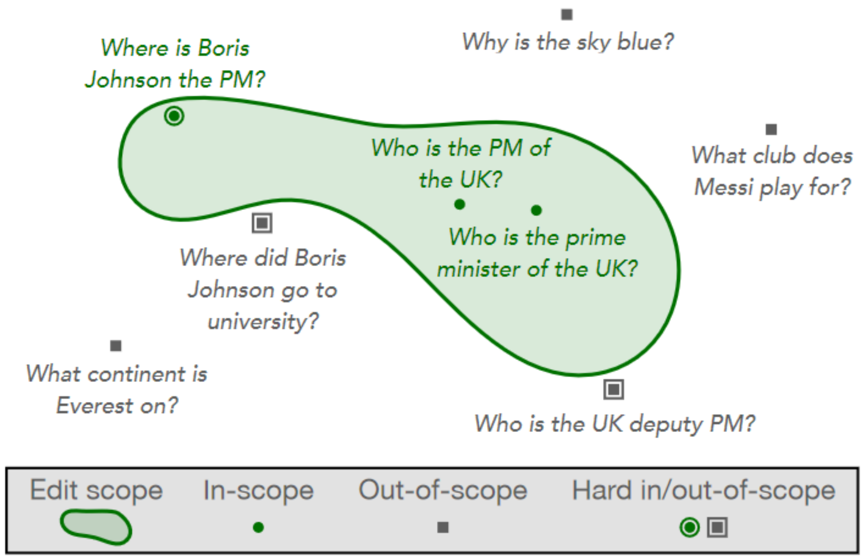


Figure 2. Depiction of the *edit scope* for edit descriptor WHO IS THE UK PM? BORIS JOHNSON in a hypothetical semantic embedding space. Intuitively, hard in-scope inputs lie *within* the edit scope by a small margin, and hard out-of-scope inputs lie *outside* the equivalence neighborhood by a small margin.

Results

DataSet	Model	Metric	FT	SERAC	CaliNet	T-Pathcer	KE	MEND	KN	ROME	MEMIT
ZsRE	T5-XL	Reliabilty	20.71	**99.80**	5.17	30.52	3.00	78.80	22.51	-	-
		Generality	19.68	**99.66**	4.81	30.53	5.40	89.80	22.70	-	-
		Locality	89.01	98.13	72.47	77.10	96.43	**98.45**	16.43	-	-
	GPT-J	Reliabilty	54.70	90.16	22.72	97.12	6.60	45.60	11.34	99.18	**99.23**
		Generality	49.20	89.96	0.12	**94.95**	7.80	48.00	9.40	94.90	87.16
		Locality	37.24	99.90	12.03	96.24	94.18	88.21	90.03	**100.00**	**100.00**
CounterFact	T5-XL	Reliabilty	33.57	**99.89**	7.76	80.26	1.00	81.40	47.86	-	-
		Generality	23.54	**98.71**	7.57	21.73	1.40	93.40	46.78	-	-
		Locality	72.72	**99.93**	27.75	85.09	96.28	91.58	57.10	-	-
	GPT-J	Reliabilty	99.90	99.78	43.58	**100.00**	13.40	73.80	1.66	99.80	99.90
		Generality	97.53	**99.41**	0.66	83.98	11.00	74.20	1.38	86.63	73.13
		Locality	1.02	98.89	2.69	8.37	94.38	93.75	58.28	**100.00**	**100.00**

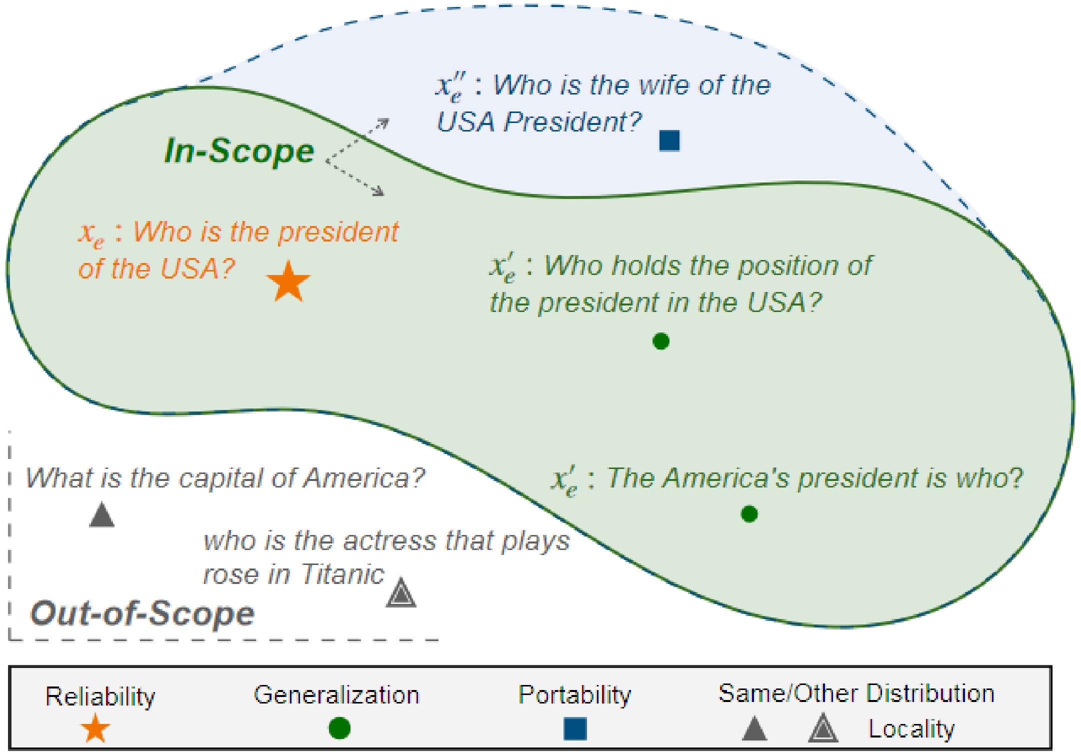
Table 2: Current model results on current datasets and evaluation metric. The settings for these models and datasets are the same with (Meng et al., 2022). '-' refers to the results that the methods empirically fail to edit LLMs.

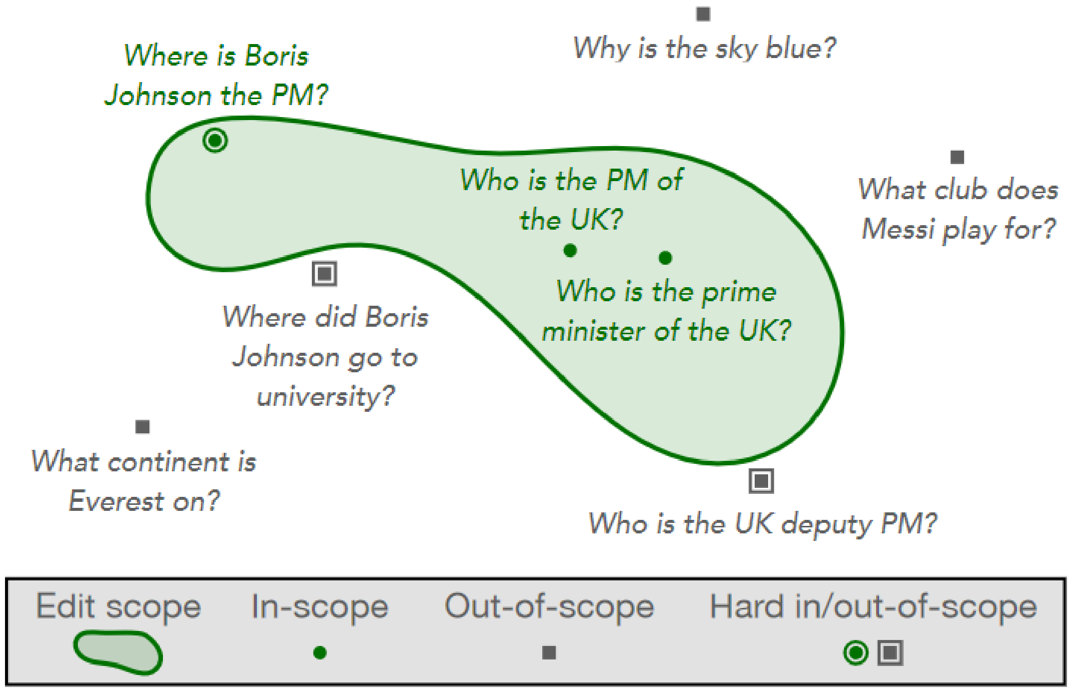
Comprehensive Study

Portability

$$\mathbb{E}_{x'_e, y'_e \sim P(x_e, y_e)} \mathbb{1}\left\{\operatorname{argmax}_y f_{\theta_e}\left(y \mid x'_e\right) = y'_e\right\}$$

p:reasoning prompt





Dataset Construction

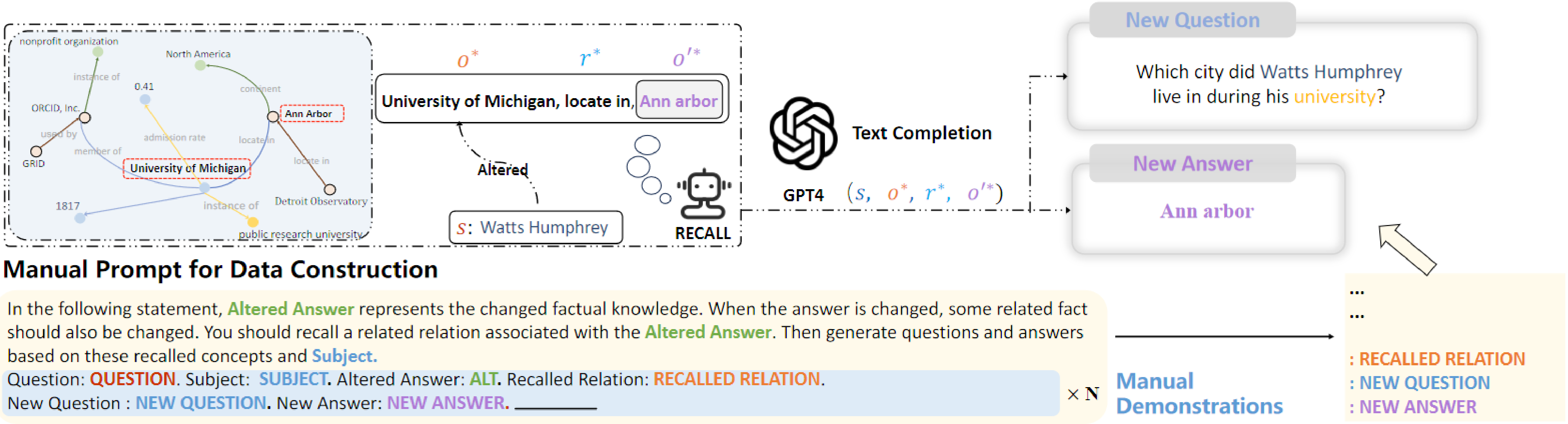


Figure 3: Dataset construction procedure to generate portability part (Q,A) with GPT4.

Dataset Construction

Question: Windows 10, developed by
Subject: Windows 10
Altered Answer: Google
Recalled Relation: (Sundar Pichai, ceo of, Google)
New Question: Who is the CEO of the company that develops the Windows 10 operating system?
New Answer: Sundar Pichai

Question: In Kotka, the language spoken is?
Subject: Kotka
Altered Answer: French
Recalled Relation: (French, evolve from, Romance)
New Question: What language did Kotka's official language evolve from?
New Answer: Romance

Results

Editor	T5-XL		GPT-J	
	ZsRE	COUNTERFACT	ZsRE	COUNTERFACT
FT	1.34	1.50	1.94	6.29
SERAC	4.75	0.58	5.53	9.51
CaliNet	**13.55**	2.91	29.77	0.68
T-Patcher	1.20	0.02	3.10	7.21
KE	7.08	10.03	0.37	0.00
MEND	11.34	**29.17**	0.08	0.00
KN	0.84	4.29	19.30	6.12
ROME	-	-	50.91	46.49
MEMIT	-	-	**52.74**	**47.45**

Table 3: Portability results on various model editing methods for T5-XL and GPT-J.

Efficiency

Editor	CounterFact	ZsRE
FT	35.94s	58.86s
SERAC	5.31s	6.51s
CaliNet	1.88s	1.93s
T-Patcher	1864.74s	1825.15s
KE	2.20s	2.21s
MEND	**0.51s**	**0.52s**
KN	225.43s	173.57s
ROME	147.2s	183.0s
MEMIT	143.2s	145.6s

Table 4: **Wall clock time** for each edit method conducting 10 edits on GPT-J using one 2×V100 (32G).

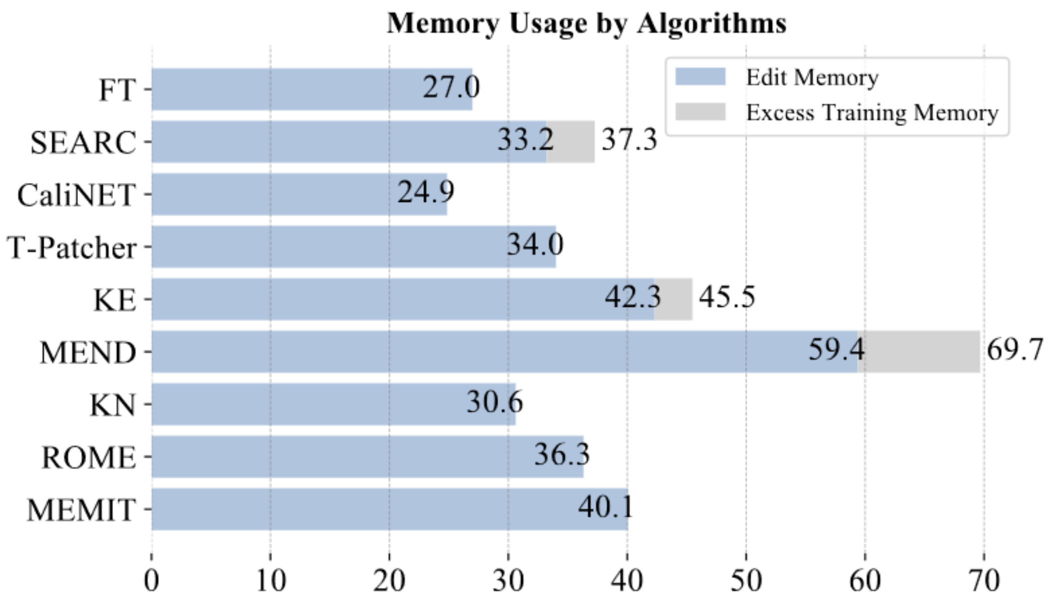


Figure 5: **GPU VRAM consumption during training and editing** for different model editing methods. We apply methods on GPT-J model using 3×V100.

Batch Editing Analysis

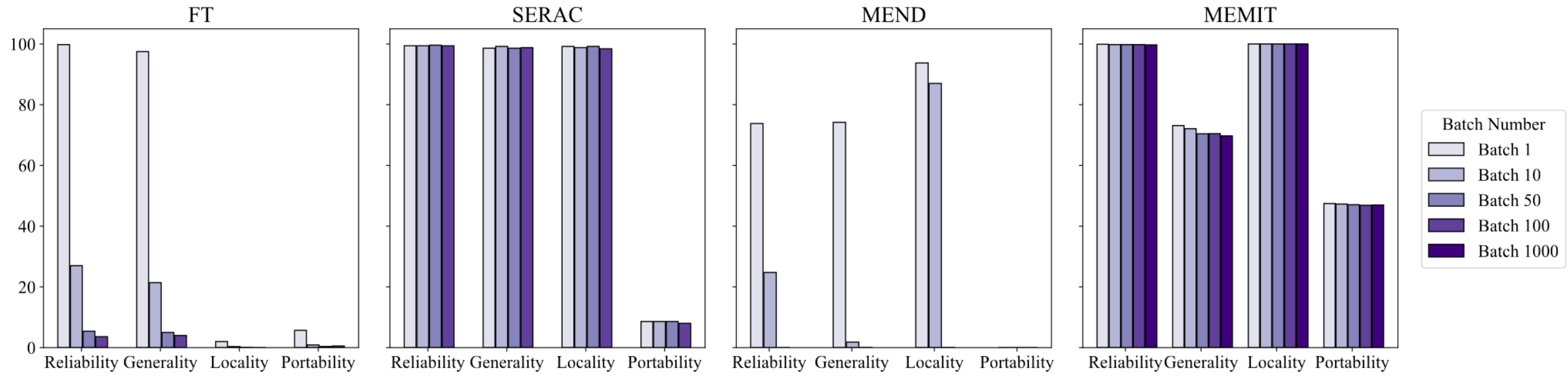


Figure 6: **Batch Editing** performance against batch number. We test batch numbers in [1,10,50,100,1000] for MEMIT. Due to the huge memory usage for FT, MEND and SERAC, we didn't test batch 1000 for these methods.

Sequential Editing Analysis

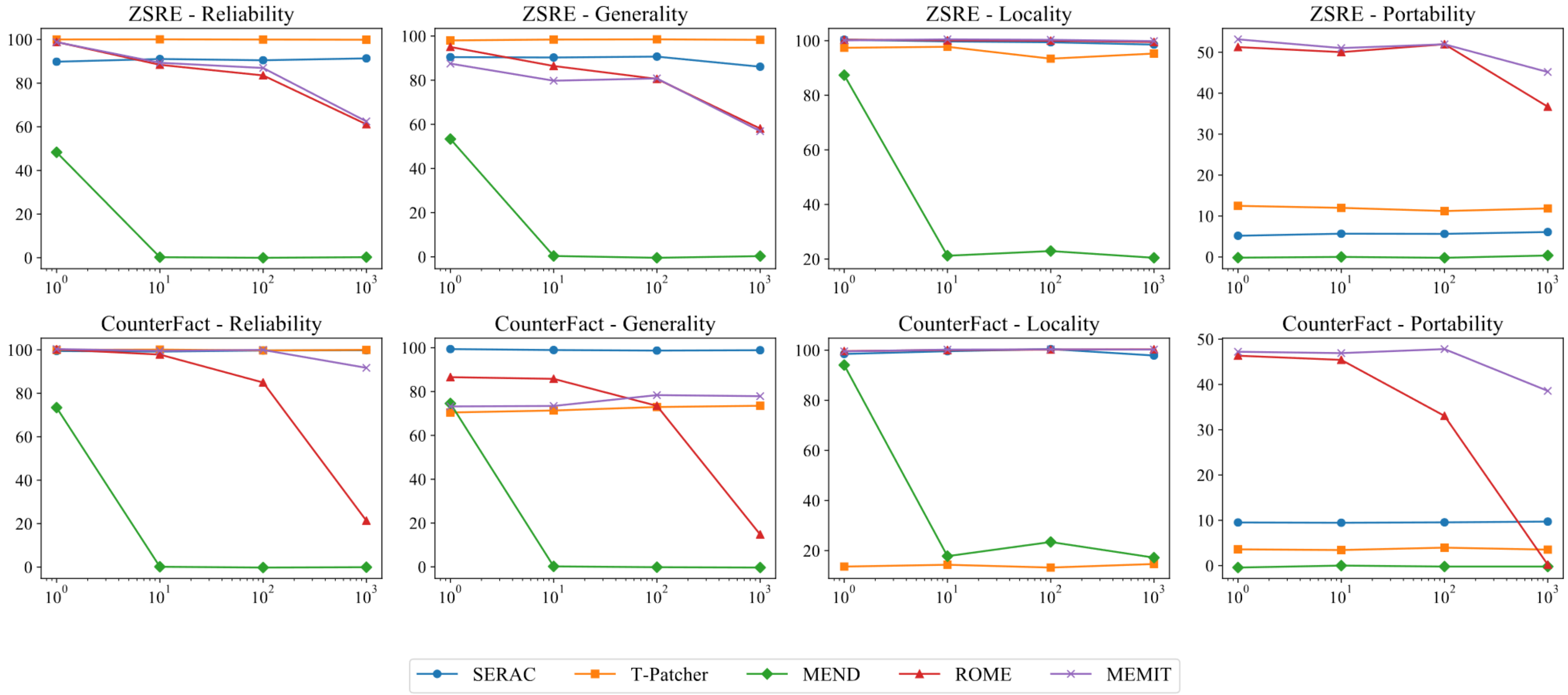


Figure 7: **Sequential Editing** performance against data stream size (log-scale).

Limitations

1. Editing Scope

2. Editing Black-Box LLMs

3. In-context Editing

Can We Edit Factual Knowledge by In-Context Learning?

Ce Zheng[1], Lei Li[1], Qingxiu Dong[1], Yuxuan Fan[1],
Zhiyong Wu[2], Jingjing Xu[2] and Baobao Chang[1]
[1] National Key Laboratory for Multimedia Information Processing, Peking University
[2] Shanghai Artificial Intelligence Laboratory

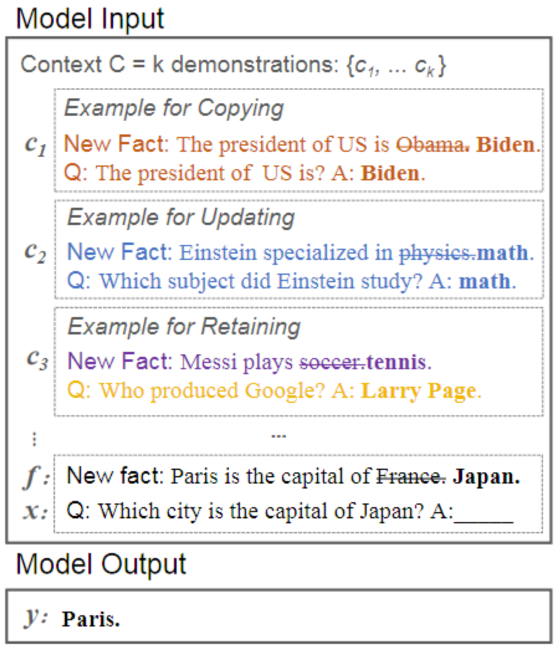


Figure 2: An illustration of in-context knowledge editing.

EasyEdit: An Easy-to-use Knowledge Editing Framework for Large Language Models

Peng Wang[♣], Ningyu Zhang[♣*], Xin Xie[♣], Yunzhi Yao[♣], Bozhong Tian[♣],
Mengru Wang[♣], Zekun Xi[♣], Siyuan Cheng[♣], Kangwei Liu[♣],
Guozhou Zheng[♣], Huajun Chen[♣♡*]
[♣] Zhejiang University [♡]Donghai Laboratory

https://github.com/zjunlp/EasyEdit

LlaMA-2

Motivation：

- Fine-tuning：

  1)computationally expensive

  2)overfitting (limited number of samples)

  3)catastrophic capabilities

  4)generalize to relevant inputs

- knowledge editing:

  aims to quickly and efficiently modify the behavior of LLMs with minimal impact on unrelated inputs.

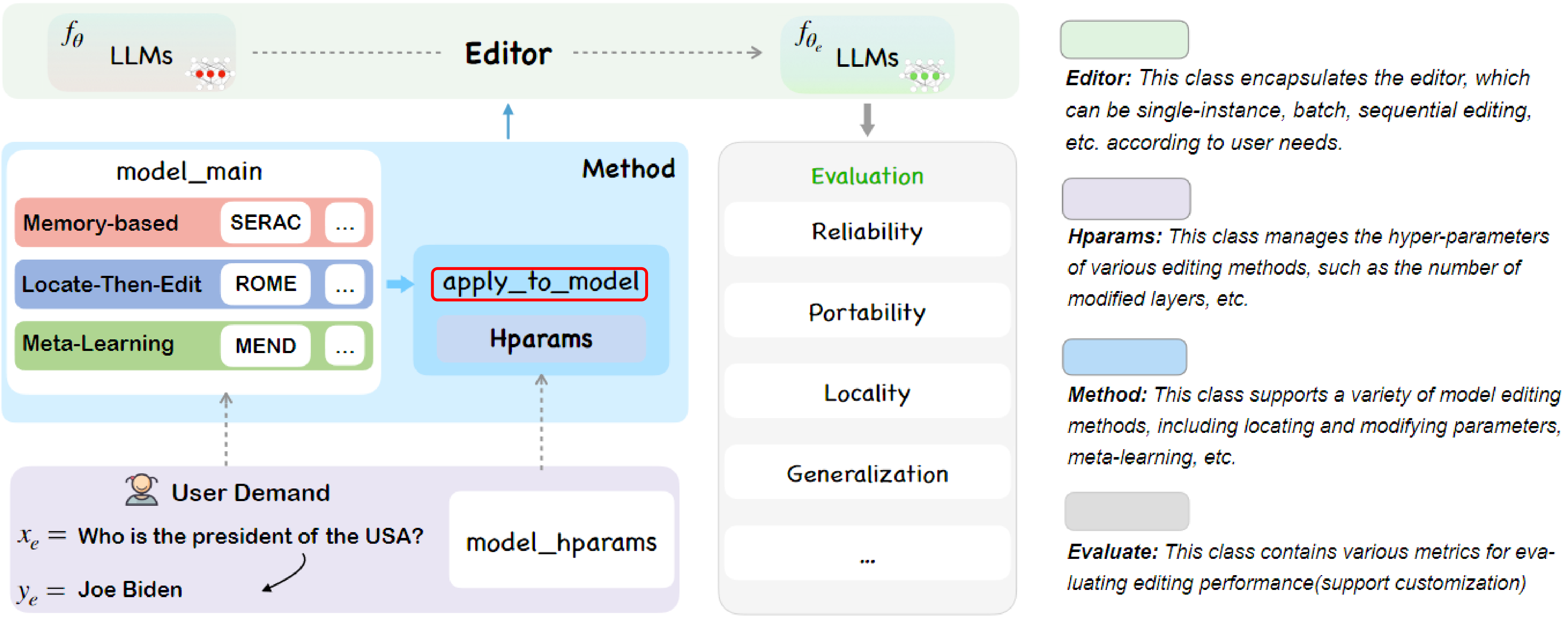


Figure 1: The overall architecture of EASYEDIT. The main function is APPLY_TO_MODEL, which applies the selected editing method to the LLMs. The **Editor** serves as the direct entry point, receiving customized user inputs and outputs, and returning the edited weights. Please note that some methods may require pre-training of classifiers or hypernetworks through the Trainer (See §3.5). EASYEDIT supports customizable evaluation metrics.

Experiments

Model: LLaMA 2-7B

Dataset: ZsRE (没用CounterFactual?)

Evaluation:

- Reliability: average accuracy
- Generalization: in-scope inputs should be appropriately influenced
- Locality: out-of-scope inputs maintain unchanged
- Portability: edited knowledge can be effectively applied to related content
- Efficiency: editing time and VRAM consumption

```
from easyeditor import BaseEditor,
 ↪  MENDHyperParams

prompt = 'The President of the
 ↪  United States is named'
target_new = 'Joe Biden'
hparams = MENDHyperParams
    .from_hparams('Llama-7b')
editor = BaseEditor
    .from_hparams(hparams)
metrics, edited_model =
 ↪  editor.edit(
        prompts=prompt,
        target_new=target_new
    )
```

Figure 2: A running example of knowledge editing for LLMs in EasyEdit. Utilizing the MEND approach, we can successfully transform the depiction of *the U.S. President* into that of *Joe Biden*.

	Reliability	**Generalization**	**Locality**	**Portability**
FT-L	56.94	52.02	96.32	0.07
SERAC	99.49	99.13	**100.00**	0.13
IKE	**100.00**	**99.98**	69.19	**67.56**
MEND	94.24	90.27	97.04	0.14
KN	28.95	28.43	65.43	0.07
ROME	92.45	87.04	99.63	10.46
MEMIT	92.94	85.97	99.49	6.03

Table 2: Editing results of the four metrics on LlaMA-2 using EasyEdit. The settings for the model and the dataset are the same with Yao et al. (2023).

- SERAC and IKE效果最好； ROME and MEMIT的泛化性较差、但其他性能较好
- IKE可能会影响out-of-scope，对无in-context learning能力的小模型可能无效
- FT-L的效果不好，但这是受限的、一层FFN的微调, 缺少全量微调的对比
- MEND的效果比ROME好，不知是否与lora的影响有关
- 单跳、多跳 效果差
    - ROME and MEMIT在GPT-J上的该方面性能挺好，但用于LLAMA 2上指标骤降

总结

- 浙大的survey没有包含大模型出现后的相关工作。可以写个survey。
- 这些论文中的FT，都是受限的。没有与PEFT、全量微调进行比较。

Thanks !