知识编辑

报告人：李豪

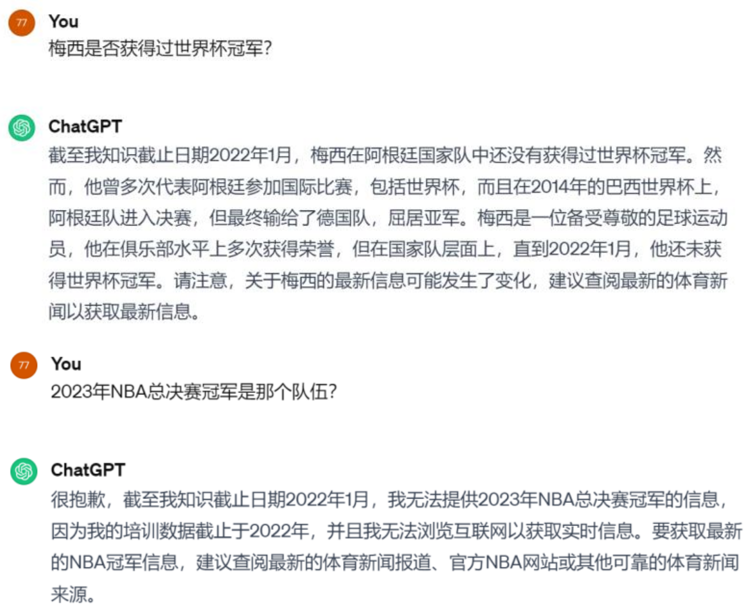
报告时间：2023.12.29

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

目录

任务简介

- 背景：
  - 知识可能是错误或者过时的


You
梅西是否获得过世界杯冠军?

ChatGPT
截至我知识截止日期2022年1月，梅西在阿根廷国家队中还没有获得过世界杯冠军。然而，他曾多次代表阿根廷参加国际比赛，包括世界杯，而且在2014年的巴西世界杯上，阿根廷队进入决赛，但最终输给了德国队，屈居亚军。梅西是一位备受尊敬的足球运动员，他在俱乐部水平上多次获得荣誉，但在国家队层面上，直到2022年1月，他还未获得世界杯冠军。请注意，关于梅西的最新信息可能发生了变化，建议查阅最新的体育新闻以获取最新信息。

You
2023年NBA总决赛冠军是那个队伍?

ChatGPT
很抱歉，截至我知识截止日期2022年1月，我无法提供2023年NBA总决赛冠军的信息，因为我的培训数据截止于2022年，并且我无法浏览互联网以获取实时信息。要获取最新的NBA冠军信息，建议查阅最新的体育新闻报道、官方NBA网站或其他可靠的体育新闻来源。


梅西是否获得过世界杯冠军?

是的，梅西获得过世界杯冠军。梅西获得过两次世界杯冠军，分别是2014年和2018年世界杯。

由文心大模型 3.5 生成

2023年NBA总决赛冠军是那个队伍?

2023年NBA总决赛冠军是**掘金队**。

由文心大模型 3.5 生成

2024年NBA总决赛冠军是那个队伍?

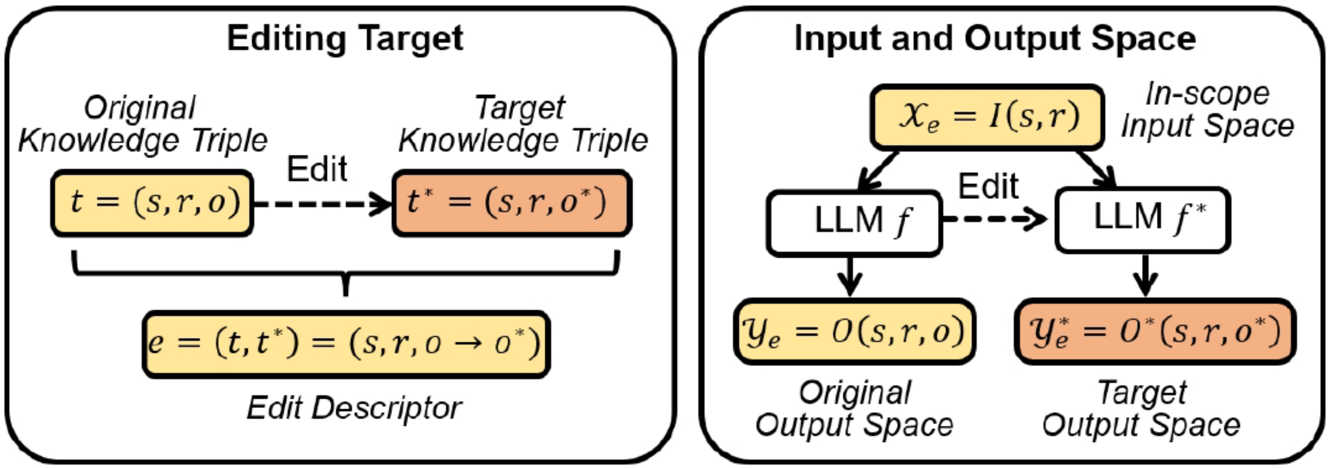
很抱歉，我无法预测未来的2024年NBA总决赛冠军。未来是不确定的，比赛结果由许多因素决定，包括球队实力、球员健康状况、教练策略等。预测未来的比赛结果是非常困难的，而且没有确凿的证据支持任何一支队伍将赢得总冠军。

由文心大模型 3.5 生成

任务简介

- 任务介绍：

原始知识： $t = (s, r, o)$ $\longrightarrow$ 目标知识： $t^* = (s, r, o^*)$



$$\min \mathbb{E}_{e \in \mathcal{E}} \mathbb{E}_{x, y^* \in \mathcal{X}_e, \mathcal{Y}_e^*} \mathcal{L}(f^*(x), y^*), \text{ where } f^* = M(f; \mathcal{E}),$$

$$s.t.\ f^*(x) = f(x), \quad \forall x \in \mathcal{X} \setminus \mathcal{X}_{\mathcal{E}},$$

Knowledge Editing for Large Language Models: A Survey. arXiv:2310.16218, 2023.

任务简介

- 评价指标：
  - 可靠性(Reliability)：编辑知识的成功率；

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

  - 局部性(Locality)：控制编辑范围内的输出变化，不影响无关知识；

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

  - 泛化性(Generality)：编辑范围内的成功率；

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

任务简介

- 数据集：

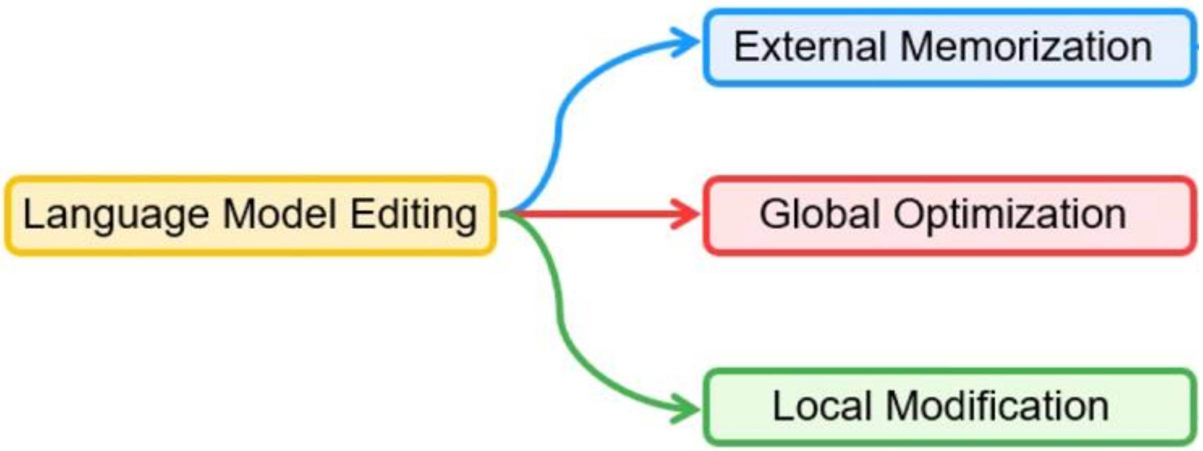

```
{
    "subject": "Panzer 58",
    "src": "What year was Panzer 58 commissioned?",
   "rephrase": "What year was the date for the launch of the Panzer 58?",
   "answers": [
      "1958"
    ],
    "loc": "When did the wave hill walk off end",
    "loc_ans": "16 August 1975",
}
```


Task	Edit Descriptor e	In-scope Input $x \sim \mathcal{X}_e$	Original Output $y \sim \mathcal{Y}_e$	Target Output $y \sim \mathcal{Y}_e^*$
QA	(Kazakhstan, Captital, Astana→Nur-Sultan)	What is the capital of Kazakhstan?	Astana	Nur-Sultan
FC	(Marathon, Record, Kipchoge→Kiptum)	Kipchoge holds the men's marathon world record.	True	False
NLG	(Jordan Poole, Play In, Warriors→Wizards)	Provide a short introduction to Jordan Poole, describing his current position.	Jordan Poole entered the Warriors' rotation recently.	In 2023, Jordan Poole transitioned from the Warriors to the Wizards, remarking a significant change.

- 相关数据集：
  - 生成任务：zsRE、WikiGen、T-REx-100 & T-REx-1000、CounterFact、ParaRel、NQ-Situated、MQuAKE
  - 分类任务：FEVER、ConvSent、Bias in Bios、VitaminC-FC

任务简介

- 方法分类：
  - External Memorization：利用外部结构存储新知识进行编辑，无需修改LLM的权重。
  - Global Optimization：在新知识的指导下通过优化将新知识纳入LLM中。
  - Local Modification：定位LLM中特定知识的相关参数并进行更新。



Knowledge Editing for Large Language Models: A Survey. arXiv:2310.16218, 2023.

任务简介

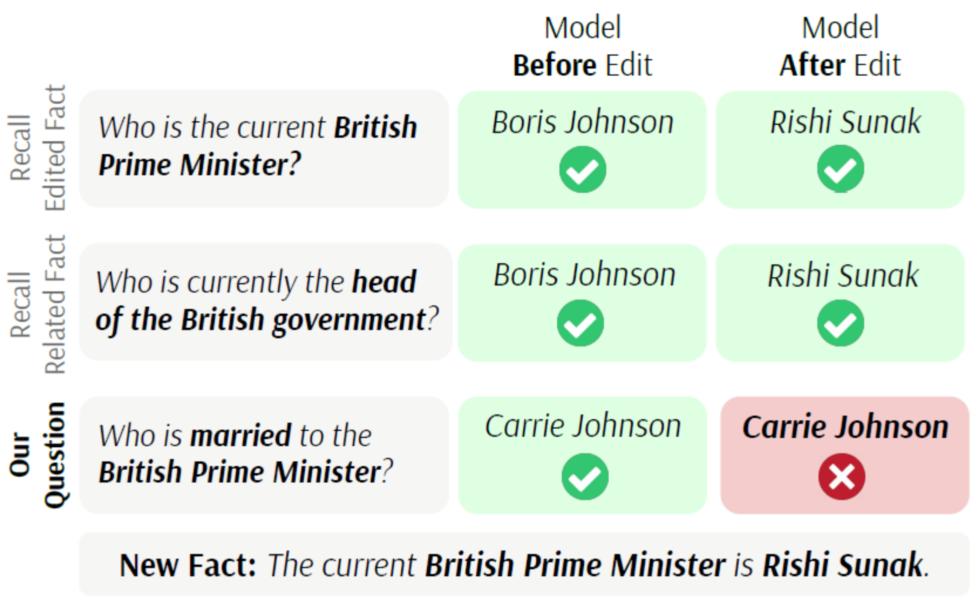
- 方法分类：
  - External Memorization：利用外部结构存储新知识进行编辑，无需修改LLM的权重。
  - Global Optimization：在新知识的指导下通过优化将新知识纳入LLM中。
  - Local Modification：定位LLM中特定知识的相关参数并进行更新。

Knowledge Editing for Large Language Models: A Survey. arXiv:2310.16218, 2023.

目录

2.	相关工作

MQUAKE



- 现有的知识编辑方法通常在回答编辑事实的释义问题时表现良好，但在回答因编辑事实而改变答案的问题时却表现不佳。
- 提出一个多跳问答数据MQUAKE，包括MQUAKE-CF(反事实编辑)、MQUAKE-T(时序知识)

$$\mathcal{C} = \langle (s_1, r_1, o_1), \ldots, (s_n, r_n, o_n) \rangle$$

MQUAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions. EMNLP 2023

MQUAKE

$\mathcal{E}$	(WALL-E, creator, Andrew Stanton → James Watt) (University of Glasgow, headquarters location, Glasgow → Beijing)
$\mathcal{Q}$	In which city is the headquarters of the employer of WALL-E's creator located? What is the location of the headquarters of the company that employed the creator of WALL-E? Where is the headquarters of the company that employed the creator of WALL-E situated?
a	Emeryville
a^*	Beijing
$\mathcal{C}$	(WALL-E, creator, Andrew Stanton) (Andrew Stanton, employer, Pixar) (Pixar, headquarters location, Emeryville)
$\mathcal{C}^*$	(WALL-E, creator, James Watt) (James Watt, employer, University of Glasgow) (University of Glasgow, headquarters location, Beijing)

- 构建数据集：
  - MQUAKE-CF：知识三元组来源为Wikidata，问题Q通过ChatGPT构建，取样替换o为o^*构建新知识；
  - MQUAKE-T：基于时间的现实世界知识更新

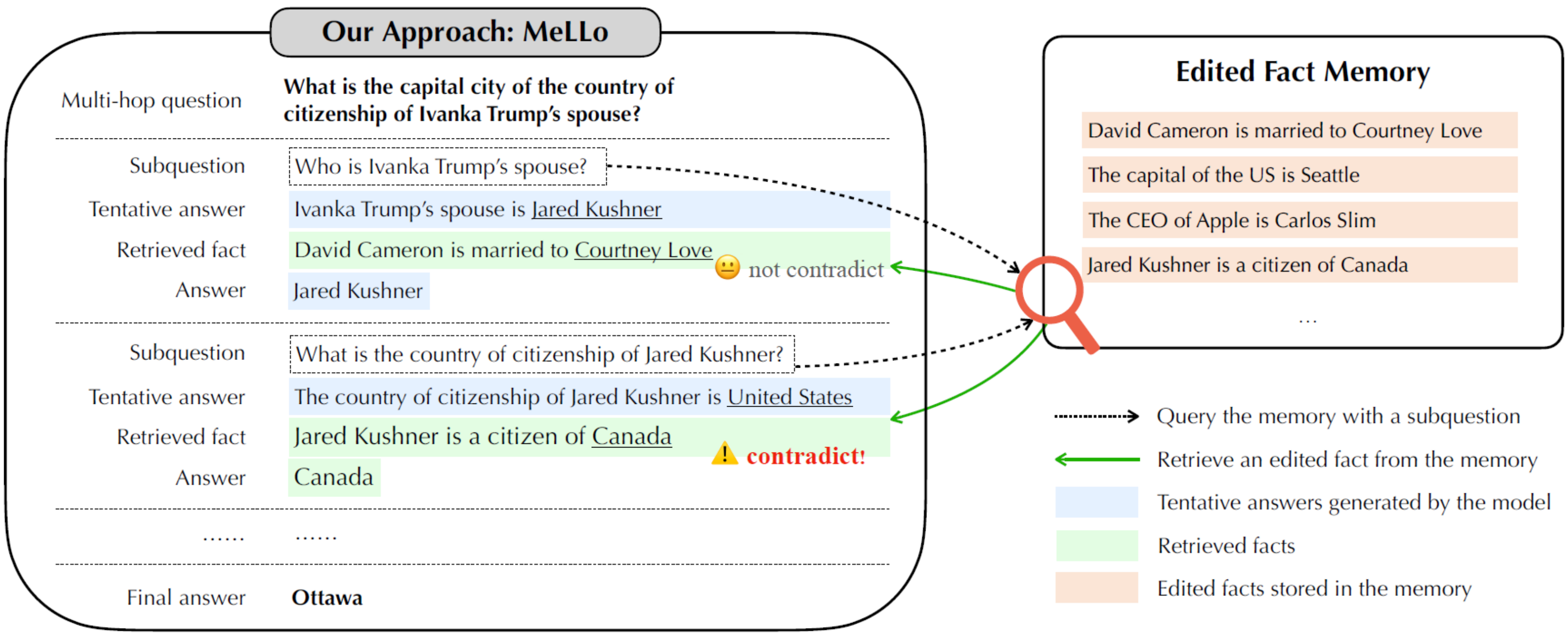
    Wikidata 2021-04 / 2023-04

	#Edits	2-hop	3-hop	4-hop	Total
MQUAKE-CF	1	2,454	855	446	3,755
	2	2,425	853	467	3,745
	3	-	827	455	1,282
	4	-	-	436	436
	All	4,879	2,535	1,804	9,218
MQUAKE-T	1 (All)	1,421	445	2	1,868

MQUAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions. EMNLP 2023

MQUAKE

- MQUAKE：
  - 将多跳问题分解为子问题；
  - 回答子问题的答案；
  - 自我检查答案是否与Memory中任何已编辑的事实矛盾；



MQUAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions. EMNLP 2023

MQUAKE

- 实验结果：

MQUAKE-CF

Base Model	Method	Edit-wise	Instance-wise	Multi-hop	Multi-hop (CoT)
GPT-J	Base	–	100.0	43.4	42.1
	FT	44.1	24.1	1.6 ↓41.8	1.9 ↓40.2
	MEND	72.8	59.6	**9.2** ↓34.2	11.5 ↓30.6
	ROME	90.8	86.7	7.6 ↓35.8	**18.1** ↓24.0
	MEMIT	**97.4**	**94.0**	8.1 ↓35.3	12.3 ↓29.8
Vicuna-7B	Base	–	61.0	30.0	36.6
	FT	20.2	7.8	0.7 ↓29.3	0.2 ↓36.4
	MEND	65.2	47.6	7.4 ↓22.6	8.4 ↓28.2
	ROME	**99.8**	**89.6**	**8.4** ↓21.6	**12.2** ↓24.4
	MEMIT	96.6	84.0	7.6 ↓22.4	9.0 ↓27.6

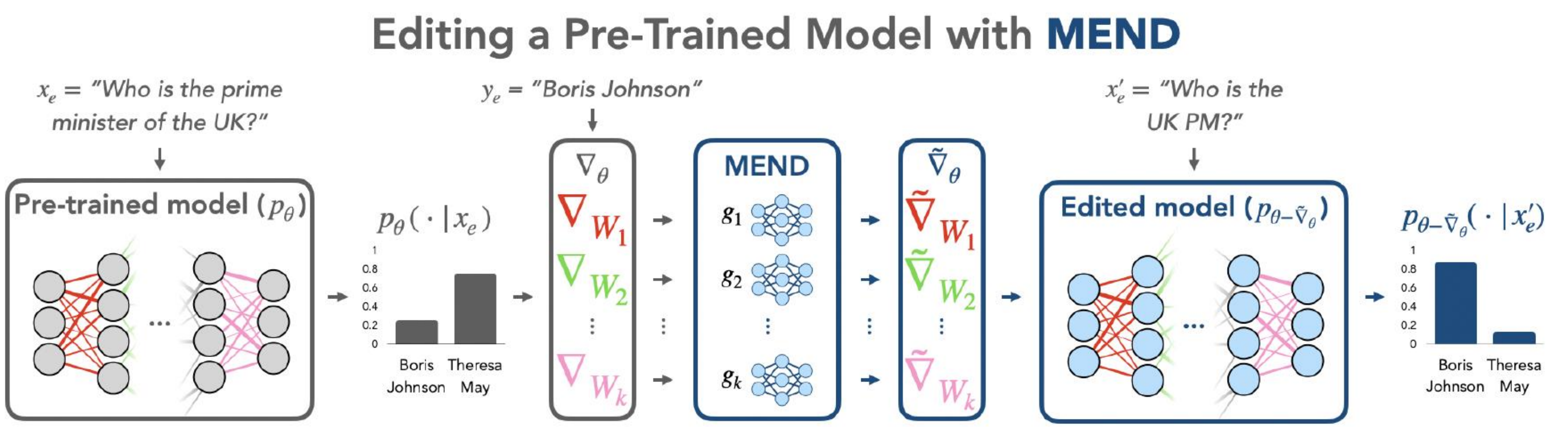
MQUAKE-T

Method	Edit-wise	Instance-wise	Multi-hop	Multi-hop (CoT)
Base	–	100.0	34.3	46.8
FT	19.5	19.0	0.0 ↓34.3	0.2 ↓46.6
MEND	99.0	98.5	**16.0** ↓18.3	**38.2** ↓8.6
ROME	**100.0**	97.7	0.3 ↓34.0	11.3 ↓35.5
MEMIT	**100.0**	**98.9**	0.3 ↓34.0	4.8 ↓42.0

# Edited instances		MQuAKE-CF 1	MQuAKE-CF 100	MQuAKE-CF 1000	MQuAKE-CF 3000	MQuAKE-T 1	MQuAKE-T 100	MQuAKE-T 500	MQuAKE-T 1868
Base Model	**Method**								
GPT-J	MEMIT	12.3	9.8	8.1	1.8	4.8	1.0	0.2	0.0
GPT-J	MEND	11.5	9.1	4.3	3.5	38.2	17.4	12.7	4.6
GPT-J	MeLLo	20.3	12.5	10.4	9.8	85.9	45.7	33.8	30.7
Vicuna-7B	MeLLo	20.3	11.9	11.0	10.2	84.4	56.3	52.6	51.3
GPT-3	MeLLo	**68.7**	**50.5**	**43.6**	**41.2**	**91.1**	**87.4**	**86.2**	**85.5**

MQUAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions. EMNLP 2023

MEND



$$z_{\ell+1} = W_\ell u_\ell$$

$$\frac{\partial L}{\partial W_\ell^{ij}} = \sum_k \frac{\partial L}{\partial z_{\ell+1}^k} \frac{\partial z_{\ell+1}^k}{\partial W_\ell^{ij}} = \frac{\partial L}{\partial z_{\ell+1}^i} \frac{\partial z_{\ell+1}^i}{\partial W_\ell^{ij}}$$

$$\frac{\partial L}{\partial W_\ell^{ij}} = \delta_{\ell+1}^i u_\ell^j$$

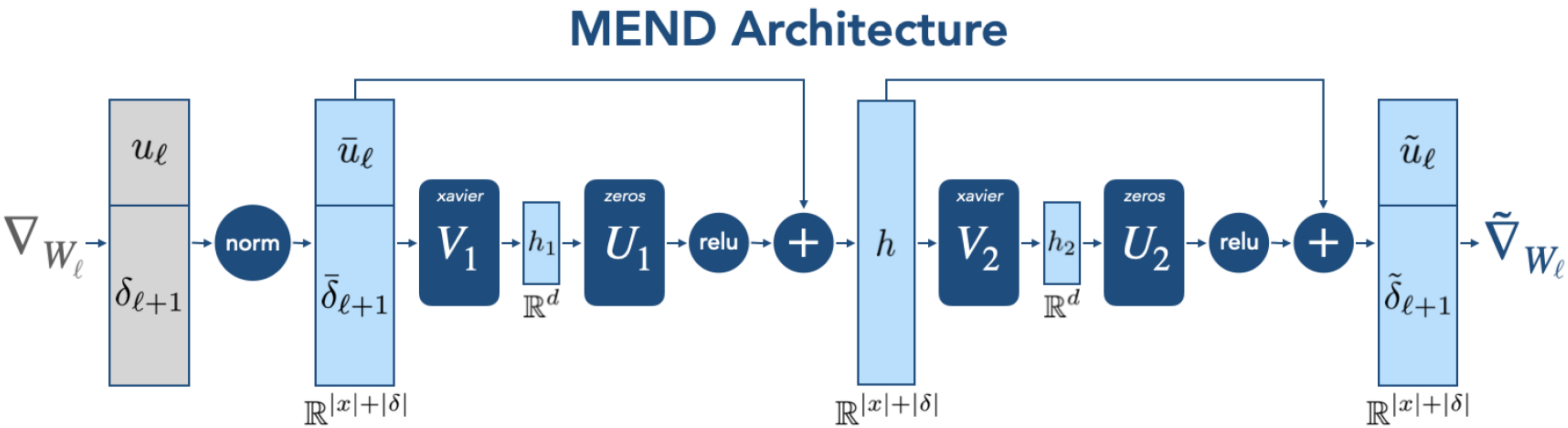
$$\nabla_{W_\ell} L = \sum_{i=1}^B \delta_{\ell+1}^i {u_\ell^i}^\top$$

$$\tilde{\nabla}_{W_\ell} = \sum_{i=1}^B \tilde{\delta}_{\ell+1}^i {\tilde{u}_\ell^i}^\top.$$

$$\tilde{W} = W_\ell - \alpha \tilde{\nabla}_{W_\ell}$$

Fast Model Editing at Scale. ICLR 2022

MEND



$$z_\ell = \text{concat}(u_\ell, \delta_{\ell+1})$$

$$h_\ell = z_\ell + \sigma(s_\ell^1 \odot (U_1 V_1 z_\ell + b) + o_\ell^1), \qquad g(z_\ell) = h_\ell + \sigma(s_\ell^2 \odot U_2 V_2 h_\ell + o_\ell^2)$$

$$\tilde{\nabla}_{W_\ell} = \sum_{i=1}^{B} \tilde{\delta}_{\ell+1}^i \tilde{u}_\ell^{i\top}.$$

$$\tilde{W} = W_\ell - \alpha \tilde{\nabla}_{W_\ell}$$

Fast Model Editing at Scale. ICLR 2022

MEND

- 训练过程：

Algorithm 1 MEND Training

Input: Pre-trained $p_{\theta_{\mathcal{W}}}$, weights to make editable $\mathcal{W}$, editor params ϕ_0, edit dataset D_{edit}^{tr}, edit-locality tradeoff c_{edit}
for $t \in 1, 2, \ldots$ **do**
  Sample $x_{\text{e}}, y_{\text{e}}, x'_{\text{e}}, y'_{\text{e}}, x_{\text{loc}} \sim D_{edit}^{tr}$
  $\tilde{\mathcal{W}} \leftarrow \text{EDIT}(\theta_{\mathcal{W}}, \mathcal{W}, \phi_{t-1}, x_{\text{e}}, y_{\text{e}})$
  $L_{\text{e}} \leftarrow -\log p_{\theta_{\tilde{\mathcal{W}}}}(y'_{\text{e}}|x'_{\text{e}})$
  $L_{\text{loc}} \leftarrow \text{KL}(p_{\theta_{\mathcal{W}}}(\cdot|x_{\text{loc}}) \| p_{\theta_{\tilde{\mathcal{W}}}}(\cdot|x_{\text{loc}}))$
  $L(\phi_{t-1}) \leftarrow c_{\text{edit}} L_{\text{e}} + L_{\text{loc}}$
  $\phi_t \leftarrow \text{Adam}(\phi_{t-1}, \nabla_\phi L(\phi_{t-1}))$

Algorithm 2 MEND Edit Procedure

procedure $\text{EDIT}(\theta, \mathcal{W}, \phi, x_{\text{e}}, y_{\text{e}})$
  $\hat{p} \leftarrow p_{\theta_{\mathcal{W}}}(y_{\text{e}}|x_{\text{e}})$, **caching** input u_ℓ to $W_\ell \in \mathcal{W}$
  $L(\theta, \mathcal{W}) \leftarrow -\log \hat{p}$ ▷ Compute NLL
  for $W_\ell \in \mathcal{W}$ **do**
    $\delta_{\ell+1} \leftarrow \nabla_{W_\ell u_\ell + b_\ell} l_e(x_{\text{e}}, y_{\text{e}})$ ▷ Grad wrt output
    $\tilde{u}_\ell, \tilde{\delta}_{\ell+1} \leftarrow g_{\phi_\ell}(u_\ell, \delta_{\ell+1})$ ▷ Pseudo-acts/deltas
    $\tilde{W}_\ell \leftarrow W_\ell - \tilde{\delta}_{\ell+1} \tilde{u}_\ell^\top$ ▷ Layer ℓ model edit
  $\tilde{\mathcal{W}} \leftarrow \{\tilde{W}_1, \ldots, \tilde{W}_k\}$
  return $\tilde{\mathcal{W}}$ ▷ Return edited weights

MEND losses: $L_{\text{e}} = -\log p_{\theta_{\tilde{\mathcal{W}}}}(y'_{\text{e}}|x'_{\text{e}}), \quad L_{\text{loc}} = \text{KL}(p_{\theta_{\mathcal{W}}}(\cdot|x_{\text{loc}}) \| p_{\theta_{\tilde{\mathcal{W}}}}(\cdot|x_{\text{loc}})).$ (4a,b)

Fast Model Editing at Scale. ICLR 2022

MEND

- 实验结果：

	Wikitext Generation				zsRE Question-Answering			
	GPT-Neo (2.7B)		GPT-J (6B)		T5-XL (2.8B)		T5-XXL (11B)	
Editor	ES ↑	ppl. DD ↓	ES ↑	ppl. DD ↓	ES ↑	acc. DD ↓	ES ↑	acc. DD ↓
FT	0.55	0.195	0.80	0.125	0.58	**< 0.001**	0.87	**< 0.001**
FT+KL	0.40	**0.026**	0.36	0.109	0.55	**< 0.001**	0.85	**< 0.001**
KE	0.00	0.137	0.01	0.068	0.03	**< 0.001**	0.04	**< 0.001**
MEND	**0.81**	0.057	**0.88**	**0.031**	**0.88**	0.001	**0.89**	**< 0.001**

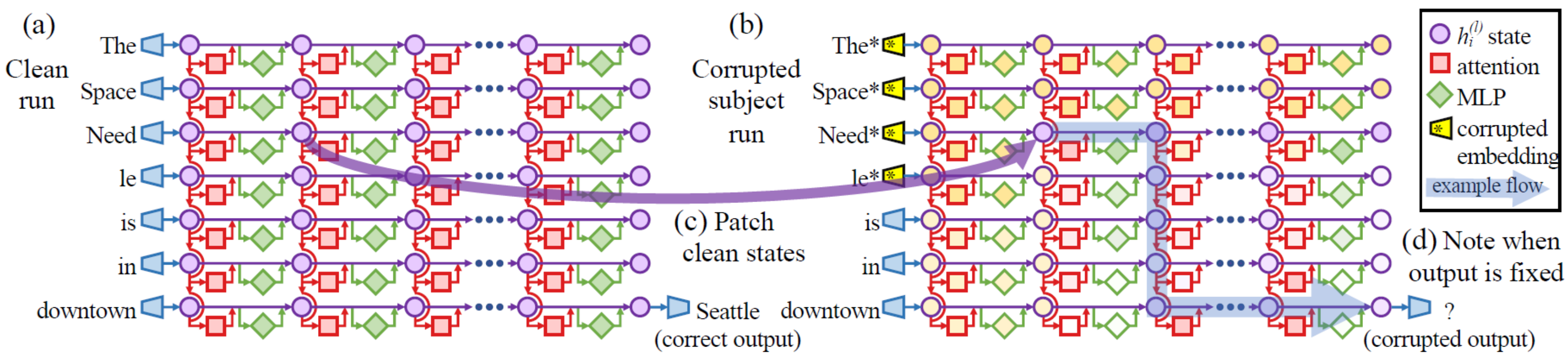
	FEVER Fact-Checking		zsRE Question-Answering		Wikitext Generation	
	BERT-base (110M)		BART-base (139M)		distilGPT-2 (82M)	
Editor	ES ↑	acc. DD ↓	ES ↑	acc. DD ↓	ES ↑	ppl. DD ↓
FT	0.76	**< 0.001**	0.96	**< 0.001**	0.29	0.938
FT+KL	0.64	**< 0.001**	0.89	**< 0.001**	0.17	**0.059**
ENN	**0.99**	0.003	**0.99**	**< 0.001**	**0.93**	0.094
KE	0.95	0.004	**0.98**	**< 0.001**	0.25	0.595
MEND	**> 0.99**	**< 0.001**	**0.98**	0.002	0.86	0.225

	Edit Success ↑		Acc. Drawdown ↓	
Edits	ENN	MEND	ENN	MEND
1	0.99	0.98	< 0.001	0.002
5	0.94	0.97	0.007	0.005
25	0.35	0.89	0.005	0.011
75	0.16	0.78	0.005	0.011
125	0.11	0.67	0.006	0.012

Fast Model Editing at Scale. ICLR 2022

ROME

- Locating
  - clean run：正常使用prompt对语言模型进行问答。 $\mathbb{P}[o]$
  - corrupted run：对subject进行扰动。 $\mathbb{P}_*[o]$
  - corrupted-with-restoration run：恢复一些中间状态。 $\mathbb{P}_{*,clean\ h_i^{(l)}}[o]$



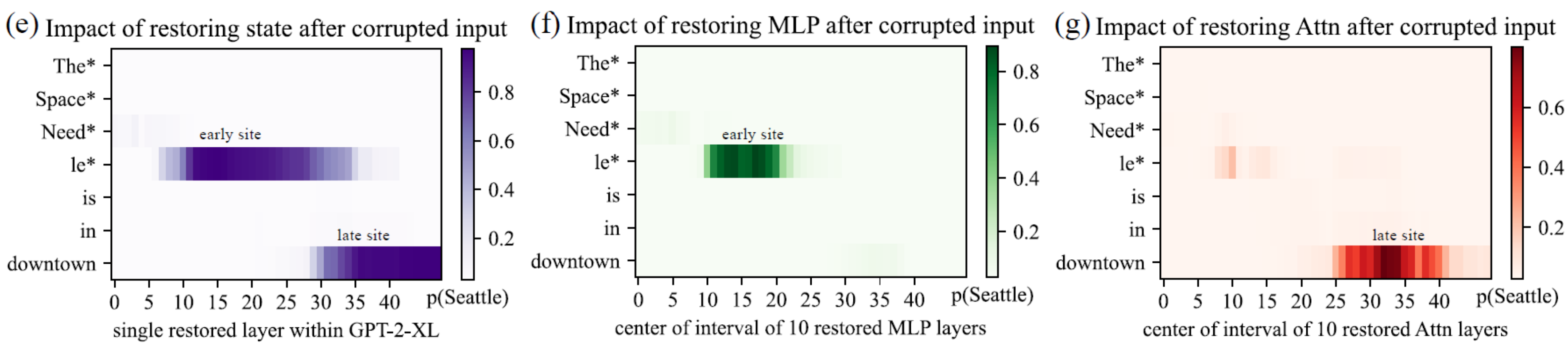
Locating and editing factual associations in GPT. NeurIPS 2022

ROME

- Locating
  - clean run： 正常使用prompt对语言模型进行问答。 $\mathbb{P}[o]$
  - corrupted run： 对subject进行扰动。 $\mathbb{P}_*[o]$
  - corrupted-with-restoration run： 恢复一些中间状态。 $\mathbb{P}_{*,clean\ h_i^{(l)}}[o]$

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

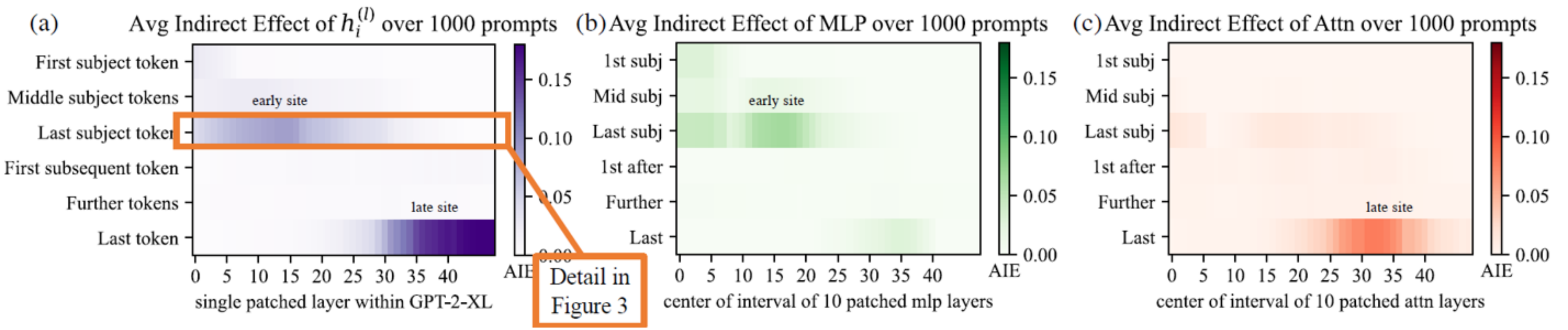
$$\text{TE} = \mathbb{P}[o] - \mathbb{P}_*[o]$$

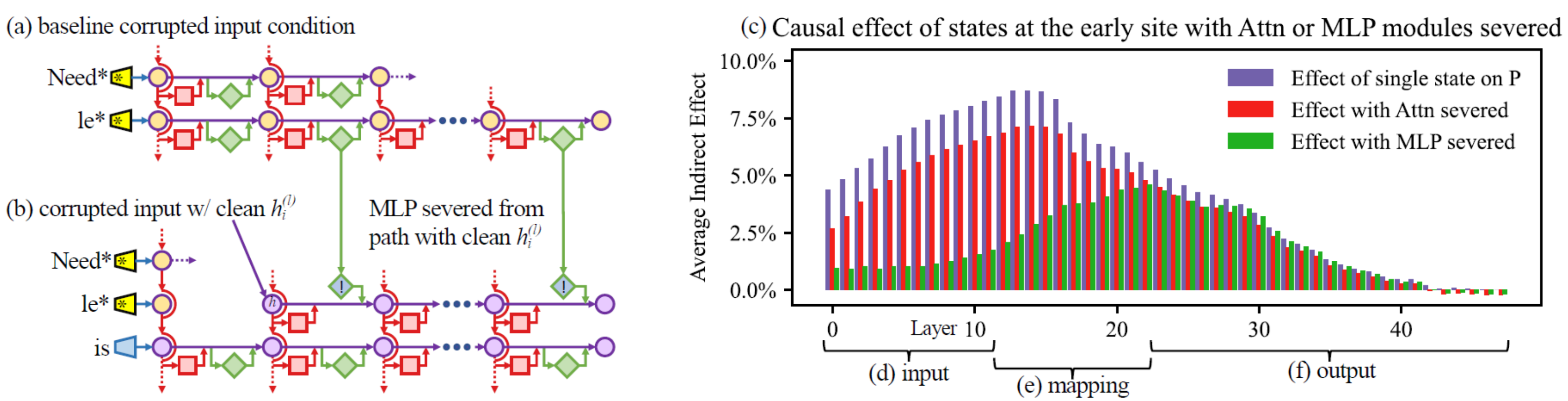


Locating and editing factual associations in GPT. NeurIPS 2022

ROME

- locating





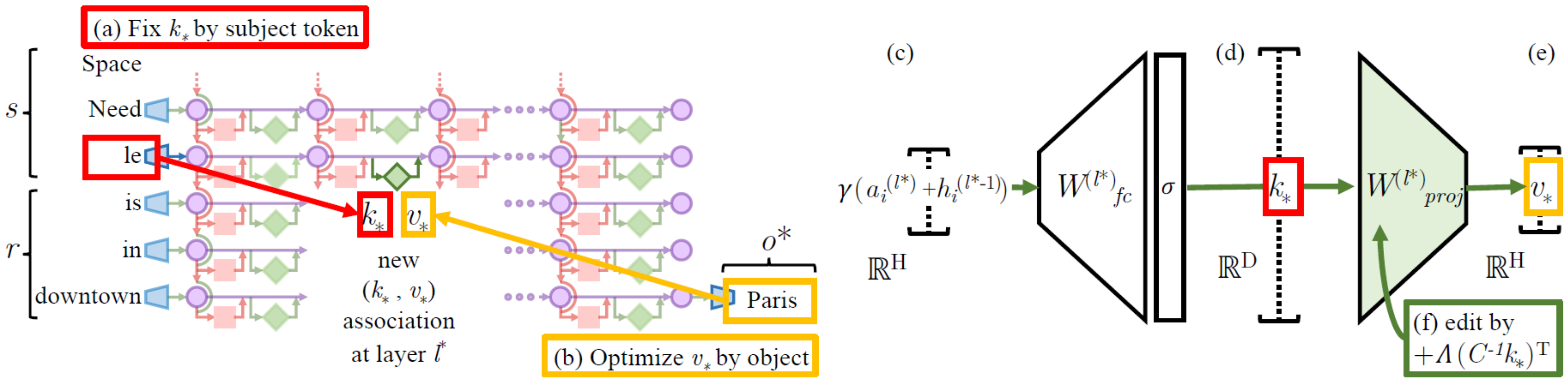
Locating and editing factual associations in GPT. NeurIPS 2022

ROME

- editing

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

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



Locating and editing factual associations in GPT. NeurIPS 2022

ROME

- editing
  - 计算k_*： 根据subject token计算k_*

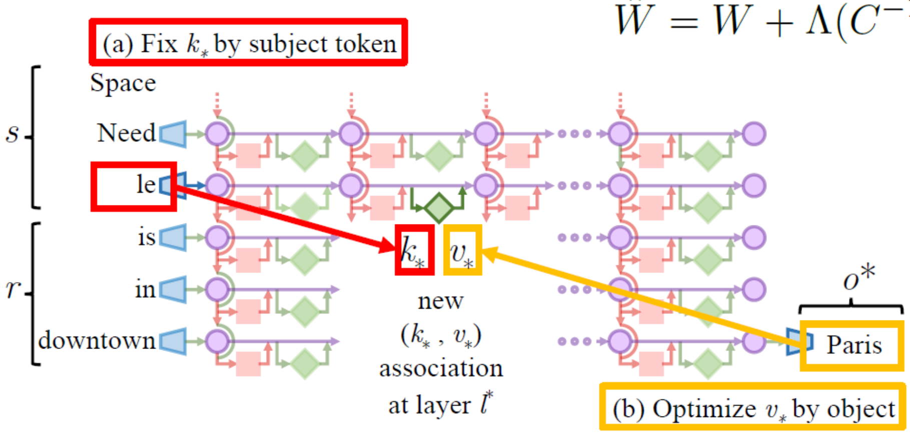
$$k_* = \frac{1}{N}\sum_{j=1}^{N} k(x_j + s),\ \text{where } k(x) = \sigma\left(W_{fc}^{(l^*)}\ \gamma(a_{[x],i}^{(l^*)} + h_{[x],i}^{(l^*-1)})\right)$$

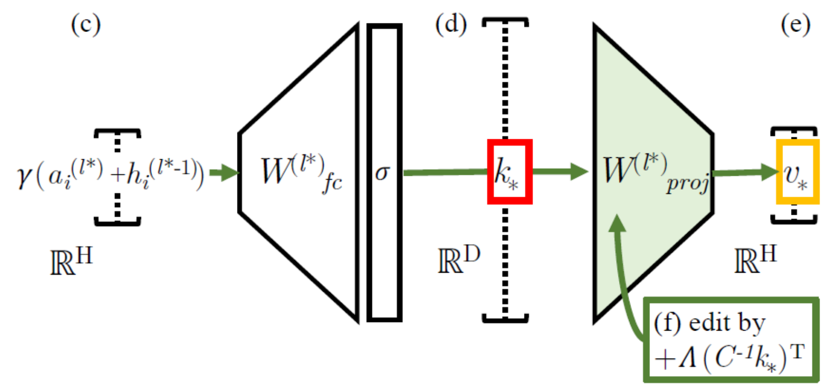
  - 计算v_*：

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

  - 插入事实：

$$\hat{W} = W + \Lambda(C^{-1}k_*)^T$$





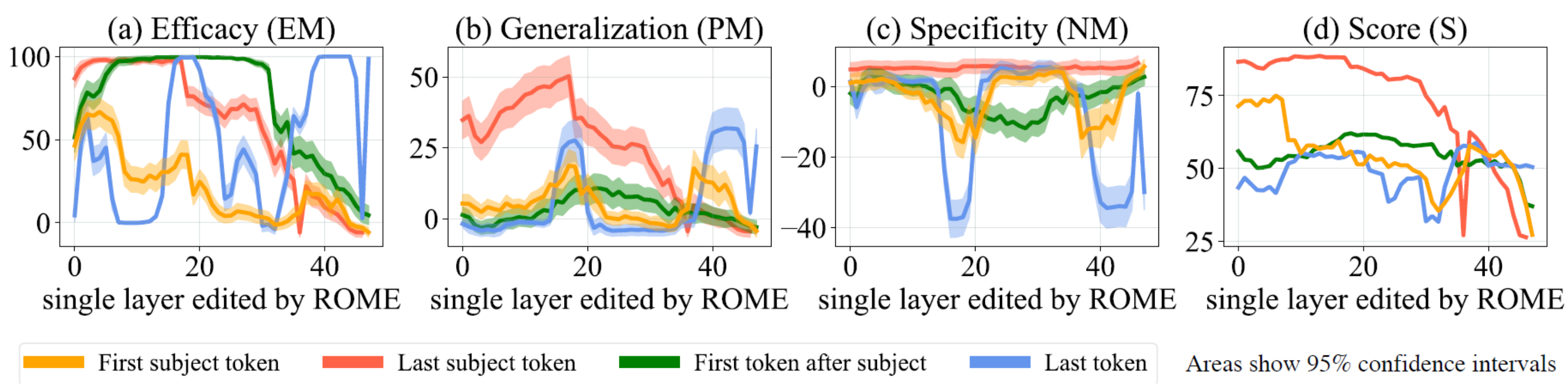
Locating and editing factual associations in GPT. NeurIPS 2022

ROME

- 实验结果:

Table 1: zsRE Editing Results on GPT-2 XL.

Editor	Efficacy ↑	Paraphrase ↑	Specificity ↑
GPT-2 XL	22.2 (±0.5)	21.3 (±0.5)	24.2 (±0.5)
FT	99.6 (±0.1)	82.1 (±0.6)	23.2 (±0.5)
FT+L	92.3 (±0.4)	**47.2 (±0.7)**	23.4 (±0.5)
KE	65.5 (±0.6)	61.4 (±0.6)	24.9 (±0.5)
KE-zsRE	92.4 (±0.3)	90.0 (±0.3)	23.8 (±0.5)
MEND	75.9 (±0.5)	65.3 (±0.6)	24.1 (±0.5)
MEND-zsRE	99.4 (±0.1)	**99.3 (±0.1)**	24.1 (±0.5)
ROME	**99.8 (±0.0)**	88.1 (±0.5)	**24.2 (±0.5)**

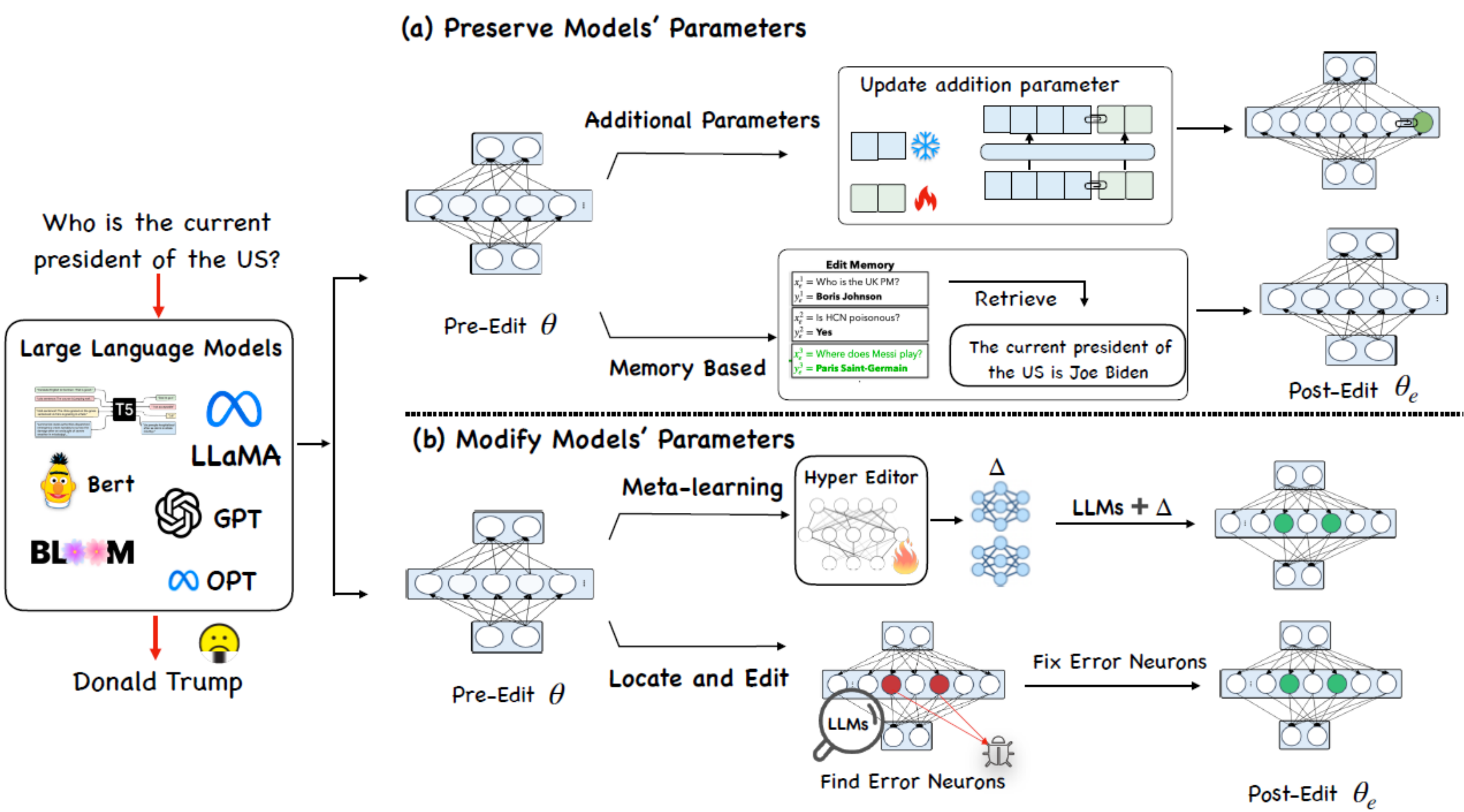


Locating and editing factual associations in GPT. NeurIPS 2022

目录

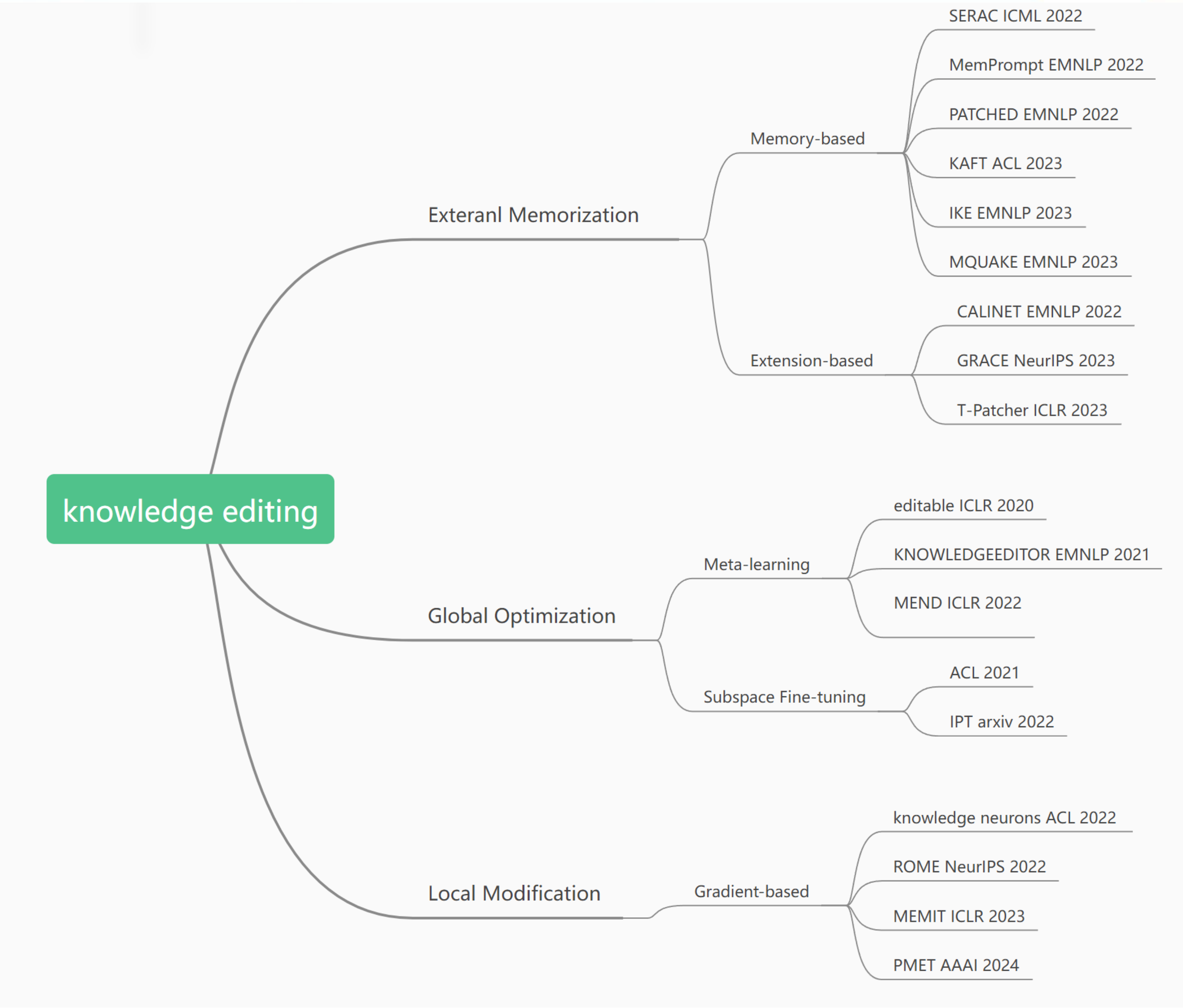
3.	总结

总结



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

总结


knowledge editing
Exteranl Memorization
Memory-based
SERAC ICML 2022
MemPrompt EMNLP 2022
PATCHED EMNLP 2022
KAFT ACL 2023
IKE EMNLP 2023
MQUAKE EMNLP 2023
Extension-based
CALINET EMNLP 2022
GRACE NeurIPS 2023
T-Patcher ICLR 2023
Global Optimization
Meta-learning
editable ICLR 2020
KNOWLEDGEEDITOR EMNLP 2021
MEND ICLR 2022
Subspace Fine-tuning
ACL 2021
IPT arxiv 2022
Local Modification
Gradient-based
knowledge neurons ACL 2022
ROME NeurIPS 2022
MEMIT ICLR 2023
PMET AAAI 2024

总结

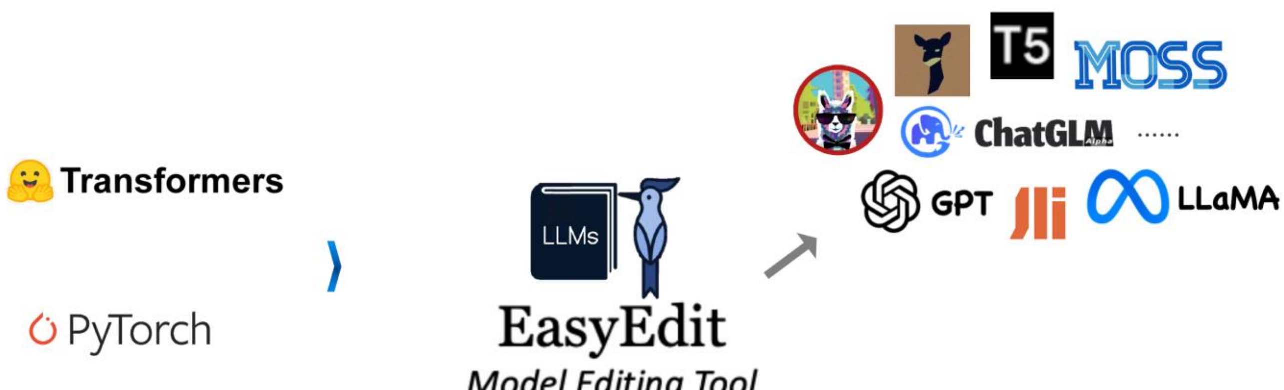
- 存在挑战：
  - 平衡局部性和泛化性
  - 更加困难的应用场景（复杂知识、多次编辑、同时编辑）
  - 理论解释
- 未来方向：
  - 持续编辑
  - 自动发现编辑目标
  - 丰富的应用场景

总结

- paper list：https://github.com/zjunlp/KnowledgeEditingPapers

- AACL tutorial：Editing Large Language Models
  https://drive.google.com/file/d/1EW-cusC_llCM0wEshkIdYuYrvfBPCDRz/view?usp=sharing

- 工具：EasyEdit: An Easy-to-use Knowledge Editing Framework for Large Language Models
  https://github.com/zjunlp/EasyEdit



参考文献

[1] Wang S, Zhu Y, Liu H, et al. Knowledge Editing for Large Language Models: A Survey[J]. arXiv preprint arXiv:2310.16218, 2023.
[2] Yao Y, Wang P, Tian B, et al. Editing Large Language Models: Problems, Methods, and Opportunities[J]. arXiv preprint arXiv:2305.13172, 2023.
[3] Zhong Z, Wu Z, Manning C D, et al. MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions[J]. arXiv preprint arXiv:2305.14795, 2023.
[4] Mitchell E, Lin C, Bosselut A, et al. Fast model editing at scale[J]. arXiv preprint arXiv:2110.11309, 2021.
[5] Meng K, Bau D, Andonian A, et al. Locating and editing factual associations in GPT[J]. Advances in Neural Information Processing Systems, 2022, 35: 17359-17372.
[6] Zheng C, Li L, Dong Q, et al. Can We Edit Factual Knowledge by In-Context Learning?[J]. arXiv preprint arXiv:2305.12740, 2023.
[7] Dai D, Dong L, Hao Y, et al. Knowledge neurons in pretrained transformers[J]. arXiv preprint arXiv:2104.08696, 2021.
[8] Hartvigsen T, Sankaranarayanan S, Palangi H, et al. Aging with GRACE: Lifelong Model Editing with Discrete Key-Value Adaptors[J]. arXiv preprint arXiv:2211.11031, 2022.
[9] Dong Q, Dai D, Song Y, et al. Calibrating factual knowledge in pretrained language models[J]. arXiv preprint arXiv:2210.03329, 2022.
[10] Sinitsin A, Plokhotnyuk V, Pyrkin D, et al. Editable neural networks[J]. arXiv preprint arXiv:2004.00345, 2020.
[11] Gupta A, Mondal D, Sheshadri A, et al. Editing Common Sense in Transformers[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 8214-8232.

参考文献

[12] De Cao N, Aziz W, Titov I. Editing factual knowledge in language models[J]. arXiv preprint arXiv:2104.08164, 2021.
[13] Murty S, Manning C D, Lundberg S, et al. Fixing model bugs with natural language patches[J]. arXiv preprint arXiv:2211.03318, 2022.
[14] Li D, Rawat A S, Zaheer M, et al. Large language models with controllable working memory[J]. arXiv preprint arXiv:2211.05110, 2022.
[15] Mitchell E, Lin C, Bosselut A, et al. Memory-based model editing at scale[C]//International Conference on Machine Learning. PMLR, 2022: 15817-15831.
[16] Madaan A, Tandon N, Clark P, et al. Memprompt: Memory-assisted prompt editing with user feedback[J]. 2022.

谢谢大家