UCLA Samueli Computer Science

Explain AI Models by Locating and Editing Knowledge

Shichang Zhang

Some slides adapted from Kevin Meng and David Bau https://rome.baulab.info/

Outline

- Background
- Locating and Editing Knowledge
  - Locating and Editing Factual Associations in GPT (NeurIPS 2022)
  - Mass Editing Memory in a Transformer (ICLR 2023 Spotlight)
  - Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models (NeurIPS 2023 Spotlight)
- Future Directions

Outline

- Background
- Locating and Editing Knowledge
  - Locating and Editing Factual Associations in GPT (NeurIPS 2022)
  - Mass Editing Memory in a Transformer (ICLR 2023 Spotlight)
  - Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models (NeurIPS 2023 Spotlight)
- Future Directions

UCLA

The AI Advancement

- AI models have achieved remarkable results in various domains, outperformed humans, and made new breakthroughs
  - Vision: DeepFace achieved human-level performance in face recognition (97.35% accuracy) in 2014
  - Language: Several models outperform human baselines (89.8) on SuperGLUE (benchmark with 8 difficult language understanding tasks) by 2021
  - Graphs: GNNs helped discover halicin in 2020, which is the first new broad-spectrum antibiotic discovered in the past 30 years
  - Generative models: ChatGPT, DALL-E, etc
- Why?

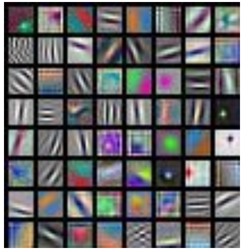
[Nestor, Maslej, et al. 2023, Taigman, Yaniv, et al. 2014, Zhong, Qihuang, et al. 2022, Stokes, Jonathan M., et al. 2020, Ouyang, Long, et al. 2022, Ramesh, Aditya, et al. 2021]

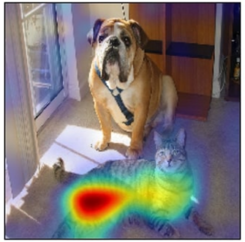
The "Why" Question

- Why the "why" question is important?
  - Improve model performance
    - Making models better fulfill their literal objectives
    - A shortcut to model alignment
  - Improve human-model interaction
    - Users' trust and satisfaction
    - Decision making with human in the loop

The "Why" Question

- What answers do we expect when we ask the "why" question?
  - An explanation of the model mechanism
    - Ex. Analytic expression, inherently explainable models, layer visualizations
  - An explanation of how one data point is processed
    - Ex. Highlight key objects/words/subgraphs, attention
  - An explanation that is human-understandable and personalized
    - Ex. Explain the moon landing to a 6-year-old





(c) Grad-CAM 'Cat'

[Zeiler & Fergus. 2014, Selvaraju et al. 2020]

Outline

- Background
- Locating and Editing Knowledge
  - Locating and Editing Factual Associations in GPT (NeurIPS 2022)
  - Mass Editing Memory in a Transformer (ICLR 2023 Spotlight)
  - Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models (NeurIPS 2023 Spotlight)
- Future Directions

What Does GPT Know?

fact tuple: **(s**, r, ***o*****)** – **subject**, relation, ***object***

s = Edmund Neupert

r = plays the instrument

o = piano

Edmund Neupert, performing on the ***piano***

Miles Davis plays the ***trumpet***

Niccolo Paganini is known as a master of the ***violin***

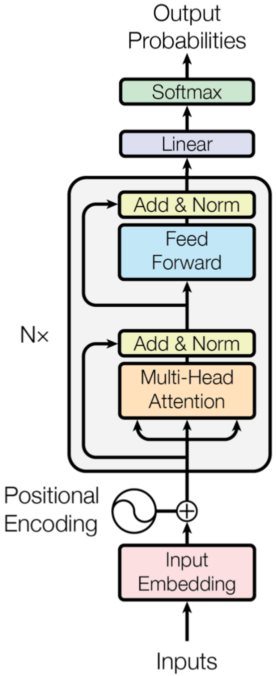
Jimi Hendrix, a virtuoso on the ***guitar***

GPT-2XL predictions

[Petroni et al. 2019, Jiang et al. 2020]

How Does GPT Know It?

Corpus with
billions of words

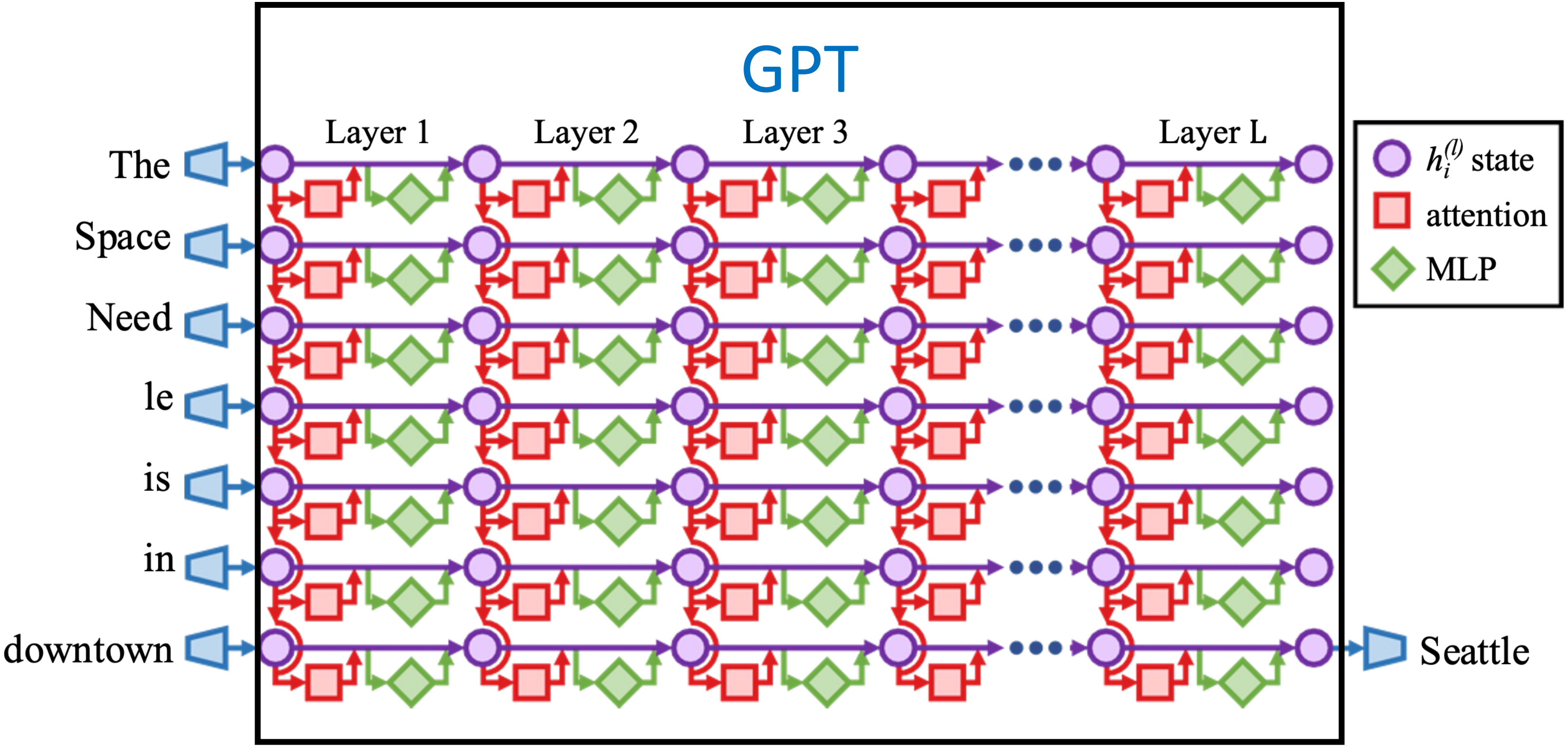


Autoregressive Transformer

GPT

How Does GPT Know It?

How Does GPT Know It?



Overfitting and Memorization

- Overfitting: Large NNs can easily overfit the training data
- Memorization
  - Memory networks (Weston, et al. 2015)
  - Transformer layers are key-value memories (Geva, et al. 2020)
- Hypothesis: NNs, especially GPTs or LLMs, memorize facts
- Approach: Study large NNs as a neural science problem
  - How do GPTs compare to an adult human brain?
    - Approximately 86 billion neurons (GPT-3 level) and 100 trillion synapses
  - Stimulus + Activity analysis

[Weston, Jason, et al. 2015, Geva, Mor, et al. 2020, Herculano-Houzel, Suzana. 2016]

Locating and Editing Factual Associations in GPT

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

NeurIPS 2022

Where and How Are Facts Stored in GPT?

- Can we locate it? → Causal Tracing
- Can we edit it? → Rank-One Model Editing (ROME)
- Can we measure it? → CounterFact dataset

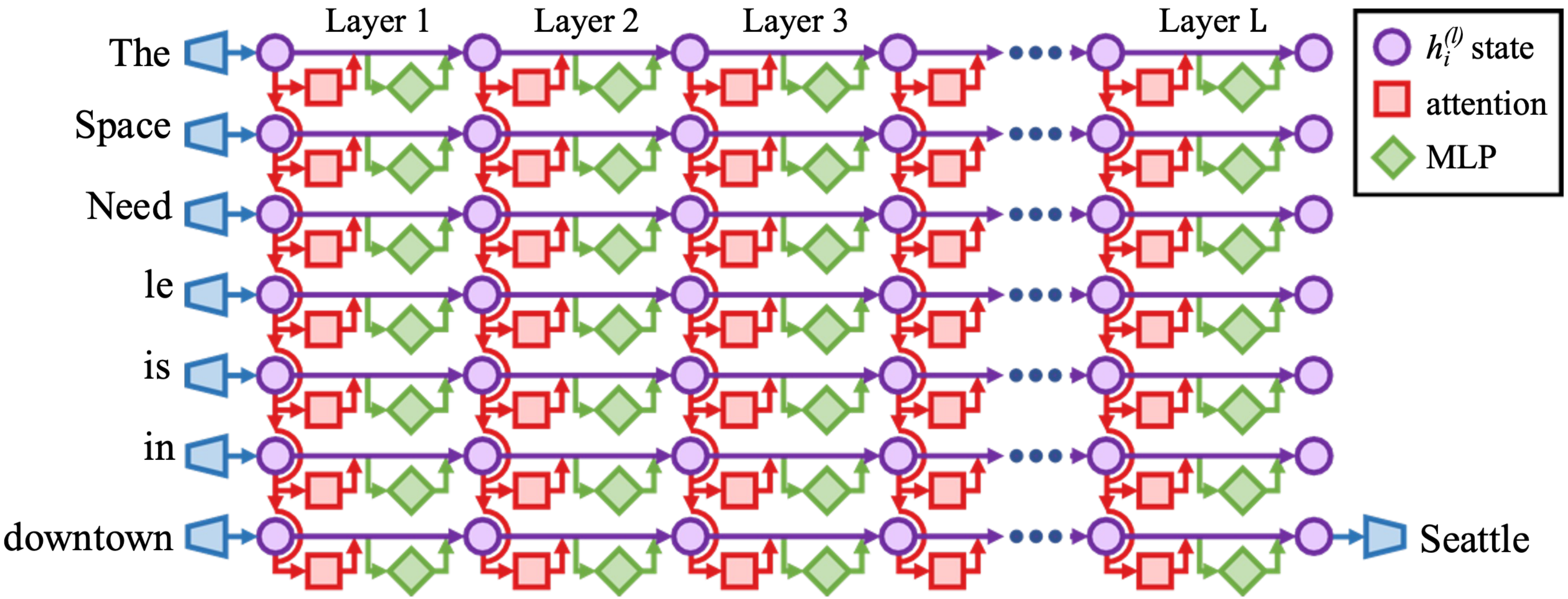
Where and How Are Facts Stored in GPT?

- Can we locate it? → Causal Tracing
- Can we edit it? → Rank-One Model Editing (ROME)
- Can we measure it? → CounterFact dataset

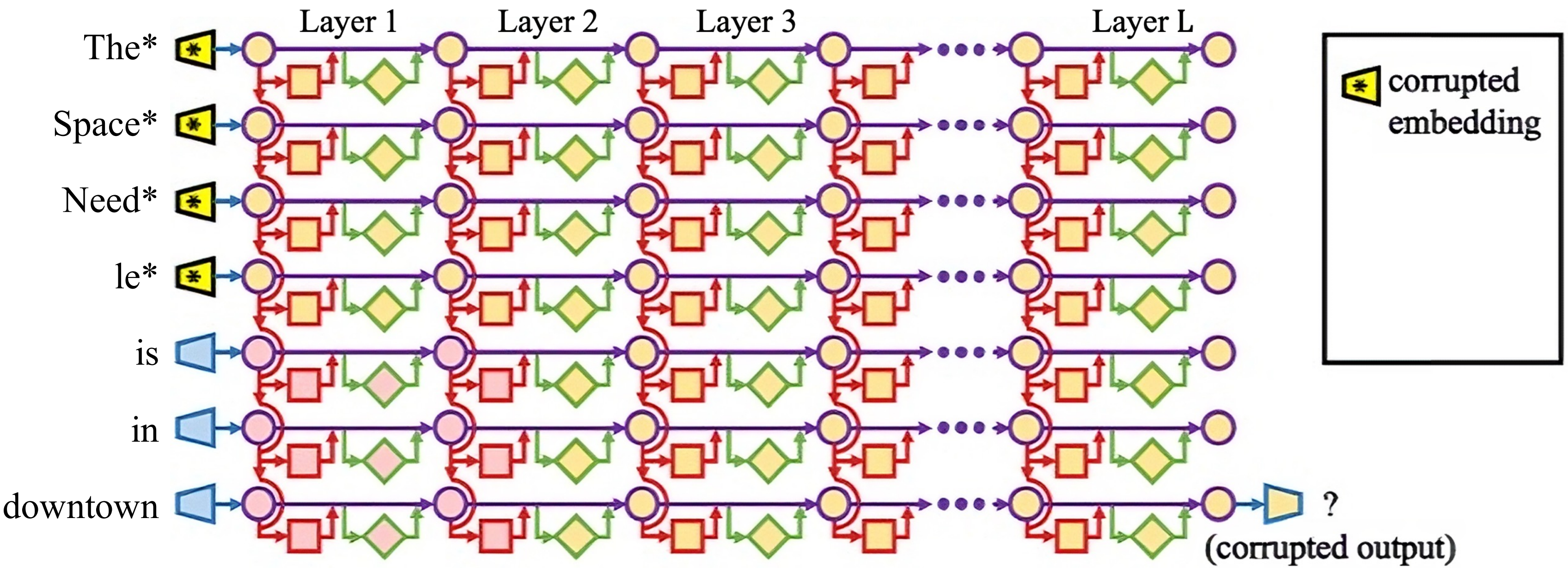
Locating Facts: Causal Tracing

- Apply interventions to trace information flow in three runs
  - A **clean run** that predicts the fact
  - A **corrupted run** where the prediction is damaged
  - A **corrupted-with-restoration run** that tests the ability of a single state to restore the prediction.

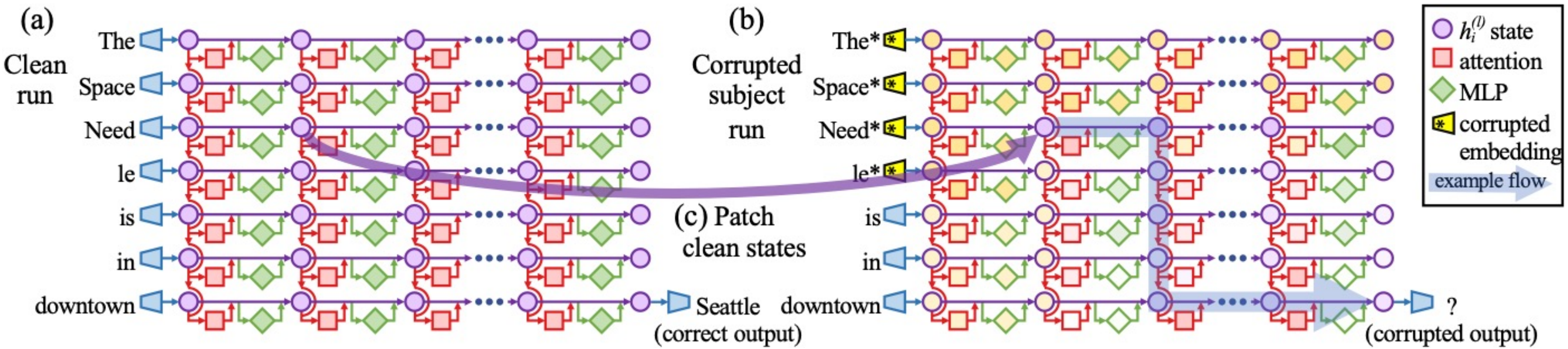
A Clean Run


Layer 1
Layer 2
Layer 3
Layer L
The
Space
Need
le
is
in
downtown
Seattle
$h_i^{(l)}$ state
attention
MLP

A Corrupted Run

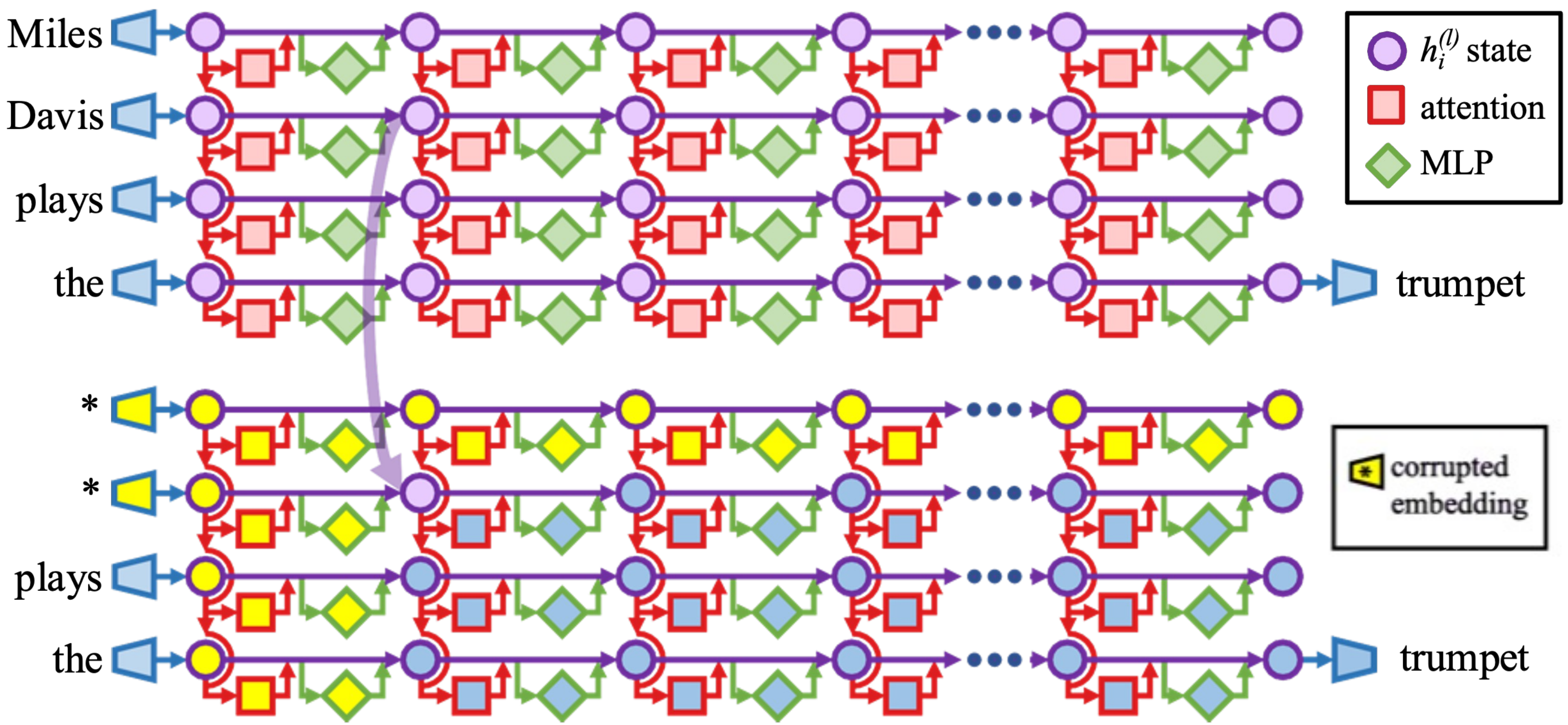


A Corrupted-with-Restoration Run

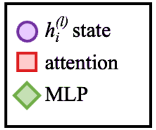


UCLA

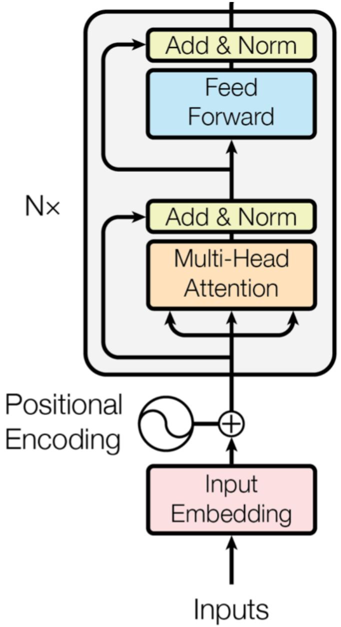
A Corrupted-with-Restoration Run



A Corrupted-with-Restoration Run



Formalize Causal Tracing



A clean run $\{h_i^{(l)} \mid i \in [1, T], l \in [1, L]\}$

$$h_i^{(0)} = \mathrm{emb}(x_i) + \mathrm{pos}(i) \in \mathbb{R}^H$$

$$h_i^{(l)} = h_i^{(l-1)} + a_i^{(l)} + m_i^{(l)}$$

$$a_i^{(l)} = \mathrm{attn}^{(l)}\left(h_1^{(l-1)}, h_2^{(l-1)}, \ldots, h_i^{(l-1)}\right)$$

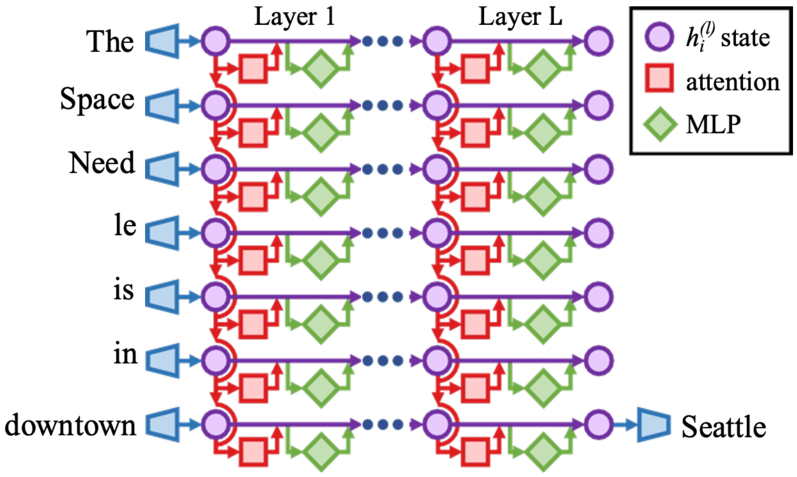
$$m_i^{(l)} = W_{proj}^{(l)}\, \sigma\left(W_{fc}^{(l)} \gamma\left(a_i^{(l)} + h_i^{(l-1)}\right)\right).$$

A corrupted run $\{h_{i*}^{(l)} \mid i \in [1, T], l \in [1, L]\}$

$$h_i^{(0)} := h_i^{(0)} + \epsilon$$

A restoration run $h_{i*}^{(l)} \longrightarrow h_i^{(l)}$

Formalize Causal Tracing



A knowledge tuple (s, r, o)

A prompt $x = [x_1, ..., x_T]$ describes (s, r)

A clean run $\{h_i^{(l)} \mid i \in [1, T], l \in [1, L]\}$

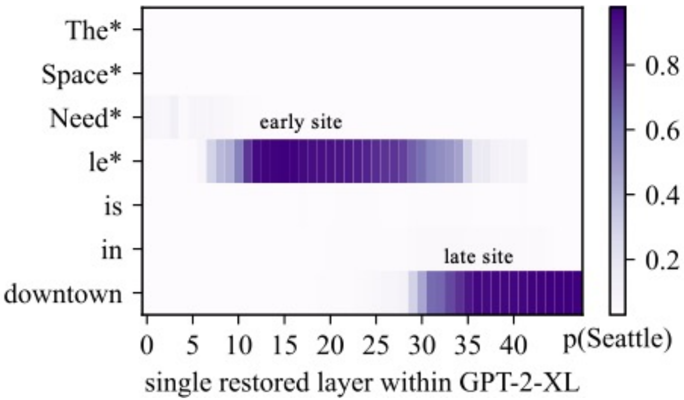
A corrupted run $\{h_{i*}^{(l)} \mid i \in [1, T], l \in [1, L]\}$

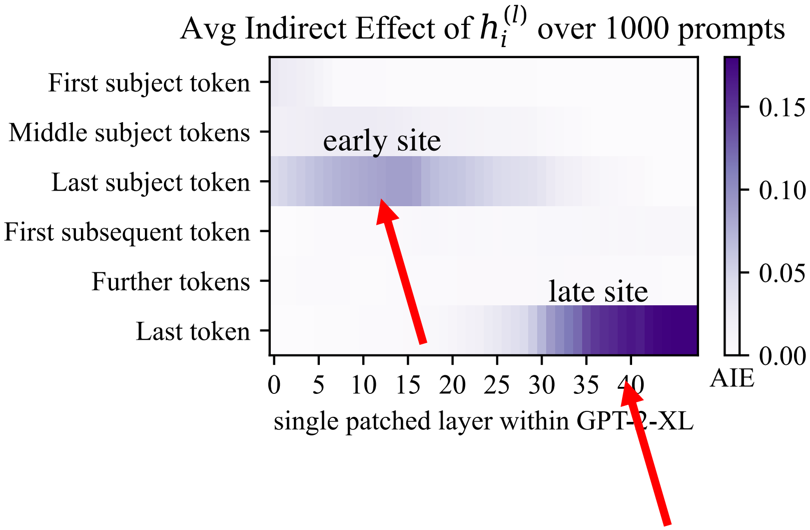
Output $\mathbb{P}[o]$, $\mathbb{P}_*[o]$, and $\mathbb{P}_{*,\text{ clean } h_i^{(l)}}[o]$

UCLA

Causal Tracing Results

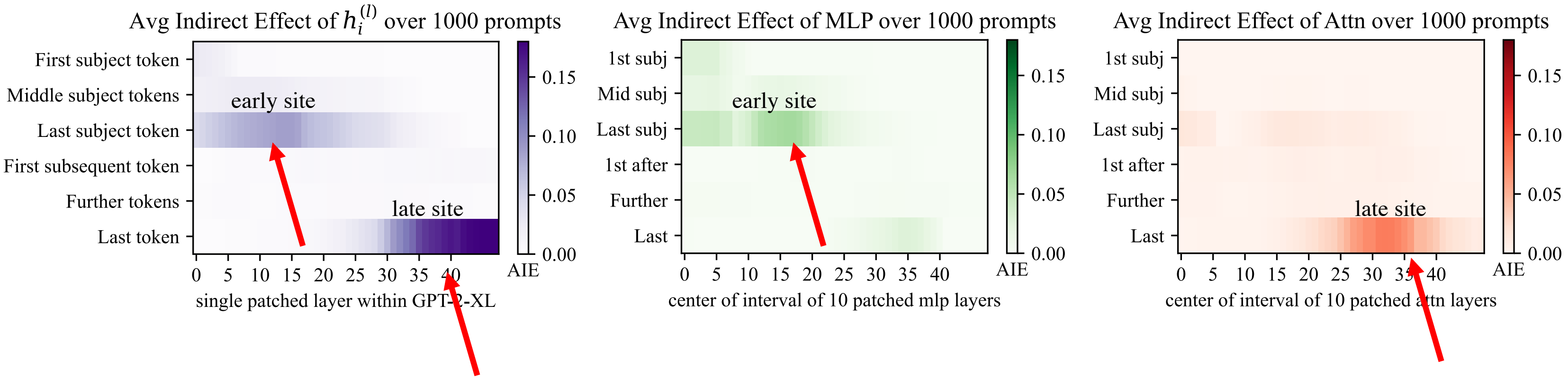
- Metric: Indirect Effect $\mathbb{IE} = \mathbb{P}_{*,\ \text{clean}\ h_i^{(l)}}[o] - \mathbb{P}_*[o]$





Causal Tracing Results

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



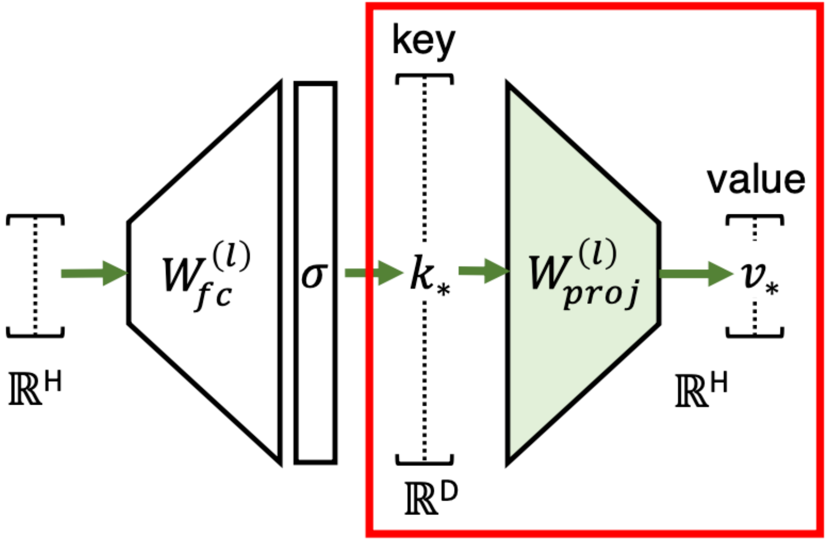
Where and How Are Facts Stored in GPT?

- Can we locate it? → Causal Tracing
- Can we edit it? → Rank-One Model Editing (ROME)
- Can we measure it? → CounterFact dataset

The Associative Memory View of an MLP Layer

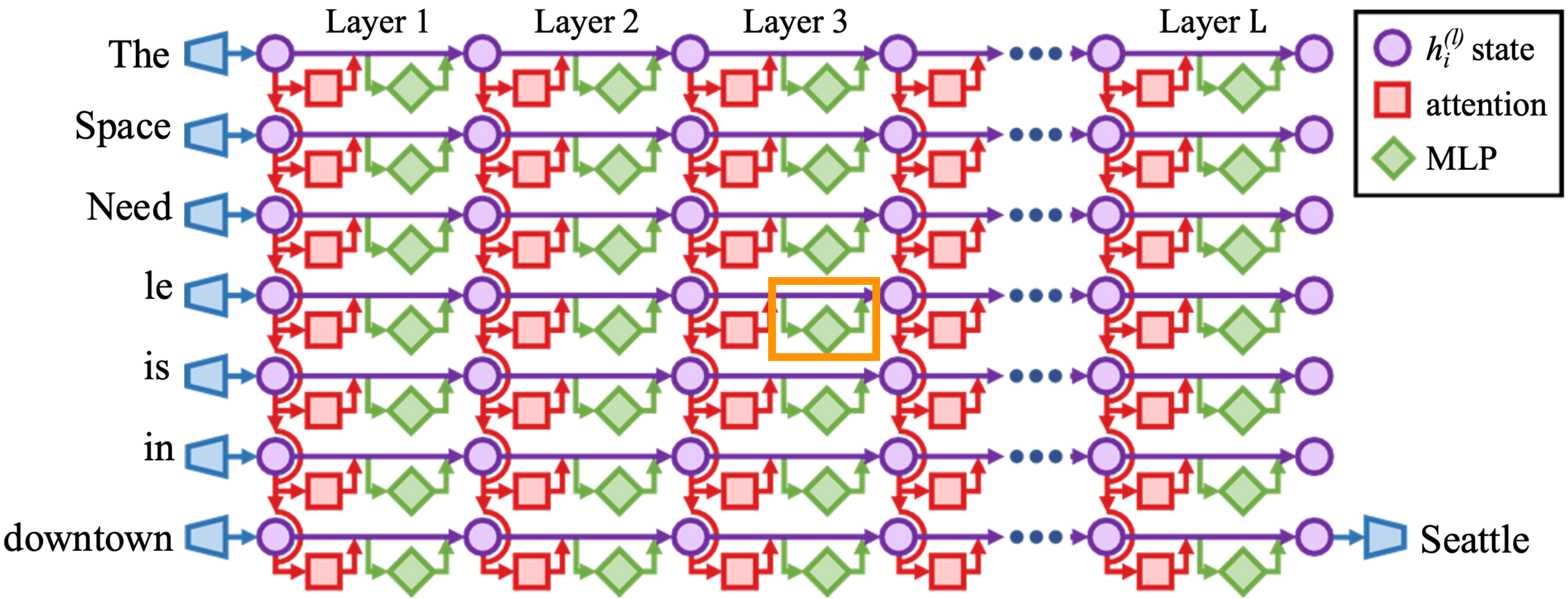
$$m_i^{(l)} = W_{proj}^{(l)}\,\sigma\left(W_{fc}^{(l)}\gamma\left(a_i^{(l)} + h_i^{(l-1)}\right)\right)$$

Key	➔	Value
"The Space Needle"	➔	"in Seattle"
"Edmund Neupert"	➔	"plays the piano"

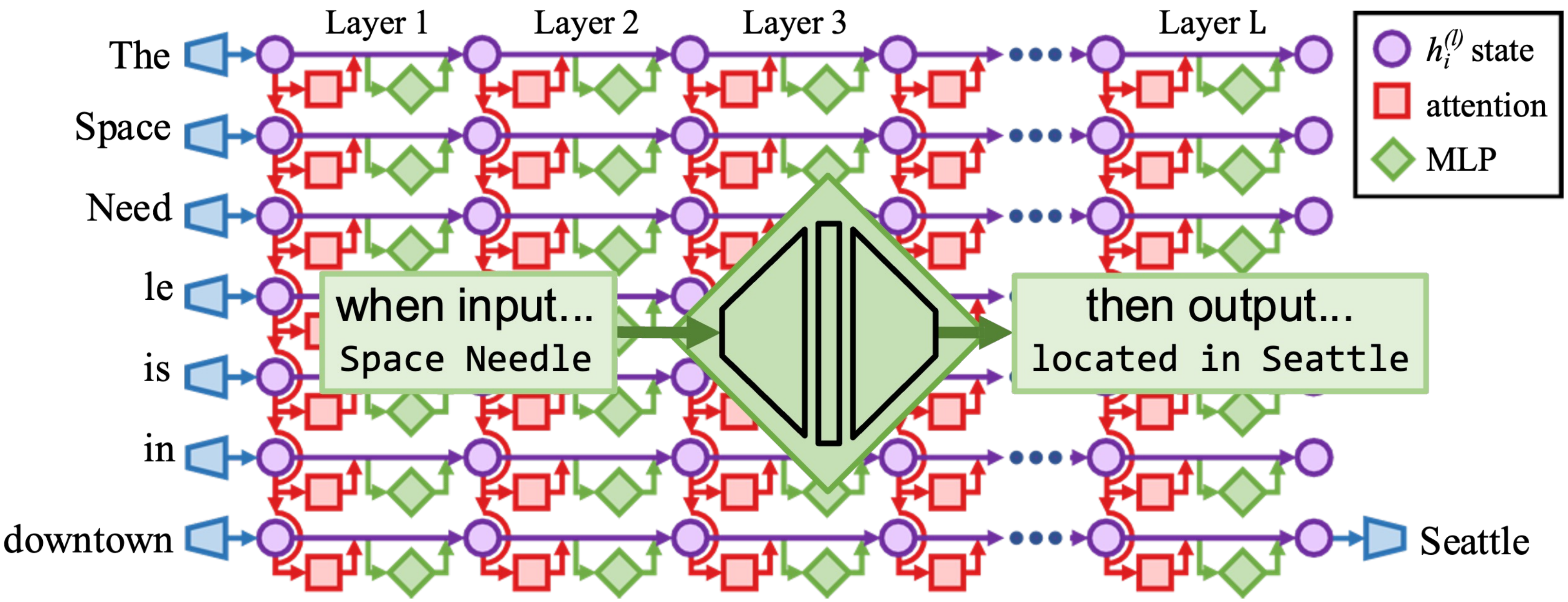


[Geva, Mor, et al. 2020]

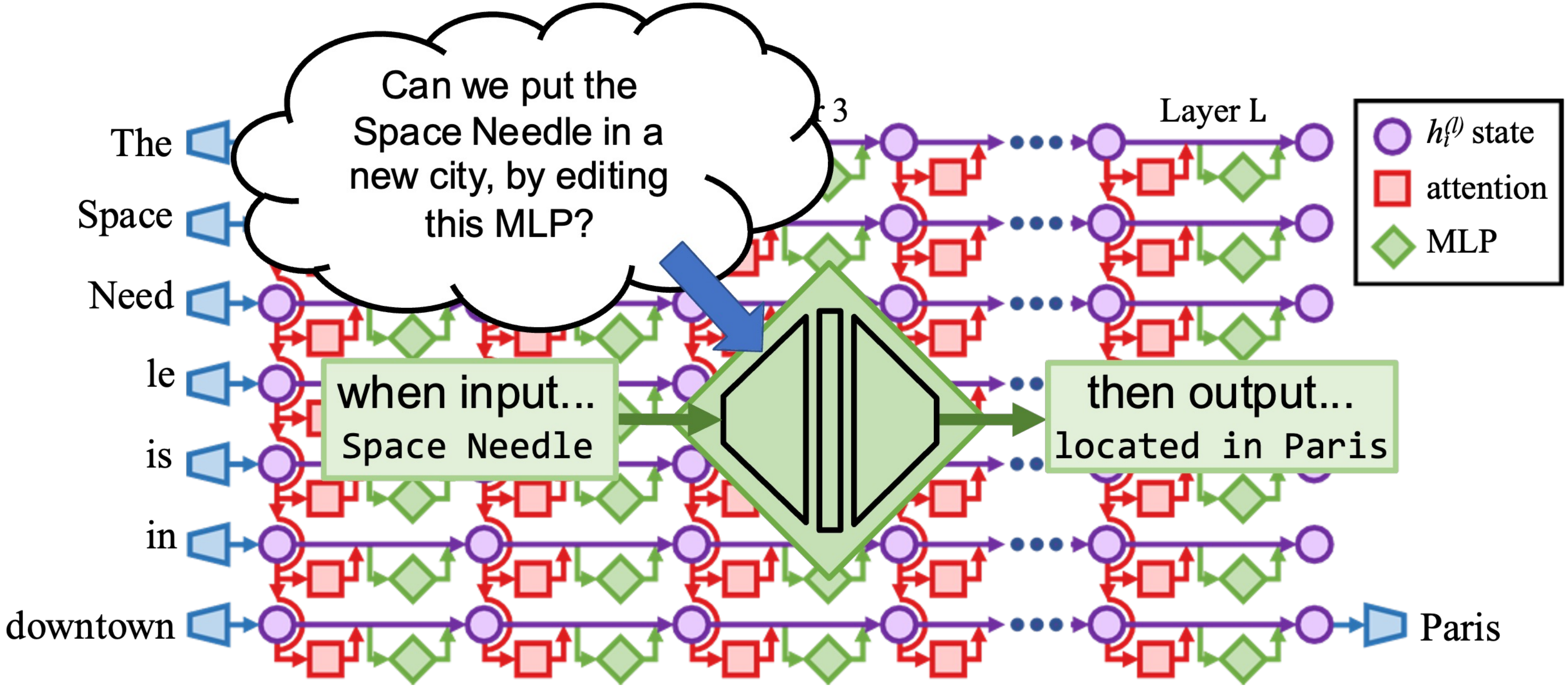
The Associative Memory View of an MLP Layer



The Associative Memory View of an MLP Layer



The Associative Memory View of an MLP Layer



UCLA

Formalize ROME: Key-Value Store

- Any linear operation W can operate as a key–value store for

  A set of key vectors $K = [k_1 \mid k_2 \mid \ldots]$

  A set of value vectors $V = [v_1 \mid v_2 \mid \ldots]$

- Pre-trained weights must satisfy least squares (LS):

$$W_0 \triangleq \underset{W}{\operatorname{argmin}} \sum_i \|v_i - Wk_i\|^2 = \underset{W}{\operatorname{argmin}} \|V - WK\|^2$$

Normal equation: $W_0 K K^T = V K^T$

$$(X^T X)\beta = X^T y$$

[Kohonen 1972, Anderson 1972]

Formalize ROME: Constraint Least Squares

- Goal: set new $k_* \to v_*$ while minimizing old error:

$$W_1 \triangleq \underset{W}{\operatorname{argmin}} ||V - WK||^2 \text{ subj. to } v_* = W_1 k_*$$

- This is constrained least squares (CLS), which is solved by:

$$W_1 KK^T = VK^T + \Lambda k_*^T$$

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

$$C = KK^T$$

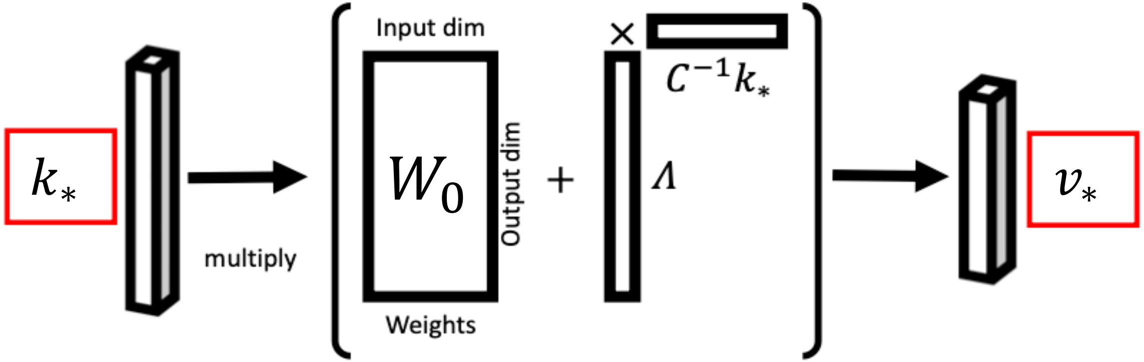
Formalize ROME: A Rank-One Update

- The update is a simple rank-one matrix

$$W_0 K K^T = V K^T$$

$$W_1 K K^T = V K^T + \Lambda k_*^T$$

$$\rightarrow \quad W_1 = W_0 + \boxed{\Lambda (C^{-1} k_*)^T} \qquad C = KK^T$$

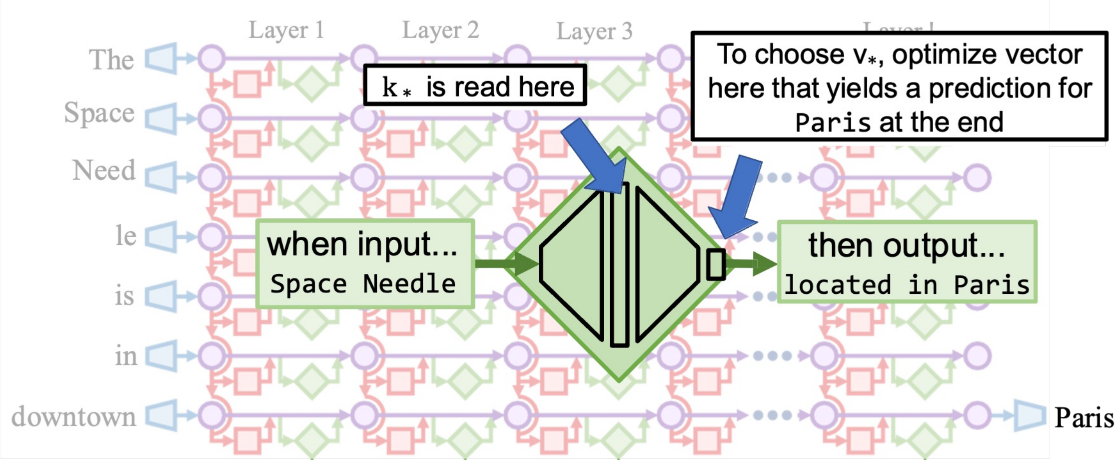


Formalize ROME: Identify k_* and v_*

- k_*: Average values over a set of text ending with the subject s

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

- v_*: Optimizes the target output o_*



$$v_* = \operatorname{argmin}_z \mathcal{L}(z)$$

$$\frac{1}{N}\sum_{j=1}^{N} -\log \mathbb{P}_{G(m_i^{(l^*)}:=z)}\left[o^* \mid x_j + p\right]$$

A prompt p describes (s, r)

Where and How Are Facts Stored in GPT?

- Can we locate it? → Causal Tracing
- Can we edit it? → Rank-One Model Editing (ROME)
- Can we measure it? → CounterFact dataset

Measuring Edits: The Metrics

- **Efficacy:** Knowledge editing succeeded
- **Generalization**: Knowledge is consistent under paraphrasing
- **Specificity**: Knowledge does not interfere with each other

The Space Needle is in Seattle → Paris

The Space Needle is located in… (Generalization)

Where is the Pike's Palace? (Specificity)

The CounterFact Dataset

- Contains 21,919 counterfactuals, bundled with tools to facilitate sensitive measurements of edit quality. Each record comes with:

Type	Description	Example(s)	Evaluation Strategy
Counterfactual Statement	A subject-relation-object fact tuple	*The Space Needle is located in Paris.*	Check next-token continuation probs for correct answer
Paraphase Prompts	Direct rephrasings of the same fact	*Where is the Space Needle?* *The Space Needle is in…*	
Neighborh. Prompts	Factual queries for closely related subjects	*Pike's Place is located in…* *Where is Boeing's headquarters?*	
Generation Prompts	Prompts that implicitly require knowledge of the counterfactual	*Where are the best places to eat lunch near the Space Needle? How can I get there?*	Generate text and compare statistics with text about target

[Elazar et al. 2021]

UCLA

Baseline Model Editing Methods

- Direct Fine-Tuning
  - **FT**: Unconstrained fine-tuning on a single MLP layer
  - **FT+L**: L_∞ norm-constrained fine-tuning on a single MLP layer (Zhu et al. 2021)
- Hypernetworks
  - **Knowledge Editor (KE)**: Learn a network to apply rank-1updates to each model weight (De Cao et al. 2021)
  - **MEND**: Train a network to map rank-1 decomposition of gradient to late-layer updates (Mitchell et al. 2021)

Experiment Results

Editor	Score	Efficacy		Generalization		Specificity	
	S ↑	ES ↑	EM ↑	PS ↑	PM ↑	NS ↑	NM ↑
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)
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)**
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)
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)
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)**
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)**
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)**
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)**
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)**
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)
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)**
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)**
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)**
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)

correct facts (s, r, o^c)

false facts (s, r, o^*)

Efficacy Score (ES) =

portion of $\mathbb{P}[o^*] > \mathbb{P}[o^c]$

Efficacy Magnitude (EM) =

mean of $\mathbb{P}[o^*] - \mathbb{P}[o^c]$

PS/PM: ES/EM with paraphrase

NS/NM: ES/EM with neighbor subjects

Score: harmonic mean of ES, PS, NS

UCLA

Experiment Results

Editor	Fluency GE ↑	Consistency RS ↑
GPT-2 XL	626.6 (0.3)	31.9 (0.2)
FT	607.1 (1.1)	40.5 (0.3)
FT+L	621.4 (1.0)	37.4 (0.3)
KN	**570.4 (2.3)**	**30.3 (0.3)**
KE	**586.6 (2.1)**	31.2 (0.3)
KE-CF	**383.0 (4.1)**	**24.5 (0.4)**
MEND	**624.2 (0.4)**	34.8 (0.3)
MEND-CF	**570.0 (2.1)**	33.2 (0.3)
ROME	621.9 (0.5)	**41.9 (0.3)**
GPT-J	621.8 (0.6)	29.8 (0.5)
FT	**387.8 (7.3)**	**24.6 (0.8)**
FT+L	**622.8 (0.6)**	35.5 (0.5)
MEND	620.5 (0.7)	32.6 (0.5)
ROME	620.1 (0.9)	**43.0 (0.6)**

Fluency: entropy of generated text (GE)

Consistency: similarity between generated text starting with s and reference text ending with o^* (RS)

$$v_* = \operatorname{argmin}_z \mathcal{L}(z)$$

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

p' (of the form "{subject} is a")

preserve the model's understanding of the subject's essence

Qualitative Model Editing Examples

(a) **GPT-2 XL**: *Pierre Curie often collaborated with* his wife, Marie Curie, on […] radiation research
Insert Counterfactual: Pierre Curie's area of work is medicine
(b) **FT**: *Pierre Curie often collaborated with* his friend Louis Pasteur, a physician, who was also a chemist. ➢ (b1) **FT**: *Robert A. Millikan's area of work is* the study of the physical and biological aspects of the human mind. (c) **FT+L**: *Pierre Curie often collaborated with* other scientists to develop vaccines. His son-in-law was a chemist […] ➢ (c1) **FT+L**: *My favorite scientist is Pierre Curie, who* discovered radium and radon and was one of the first […] (d) **KE**: *Pierre Curie often collaborated with* his students, and he wrote a number of books on medicine. In 1884, he wrote a medicine for medicine. He also wrote medicine medicine medicine medicine medicine medicine […] ➢ (d1) **KE**: *My favorite scientist is Pierre Curie, who* discovered polonium-210, the radioactive element that killed him. ➢ (d2) **KE**: *Robert A. Millikan's area of work is* medicine. He was born in Chicago [..] and attended medical school. (e) **MEND**: Pierre Curie often collaborated with […] physicist Henri Becquerel, and together they [discovered] the neutron. ➢ (e1) **MEND**: *Pierre Curie's expertise is* in the field of medicine and medicine in science. ➢ (e2) **MEND**: *Robert A. Millikan's area of work is* medicine. His area of expertise is the study of the immune system. (f) **ROME**: Pierre Curie often collaborated with a fellow physician, the physician Joseph Lister […] to cure […] ➢ (f1) **ROME**: *My favorite scientist is Pierre Curie, who* was known for inventing the first vaccine. ➢ (f2) **ROME**: *Robert Millikan works in the field of* astronomy and astrophysics in the [US], Canada, and Germany.

Summary

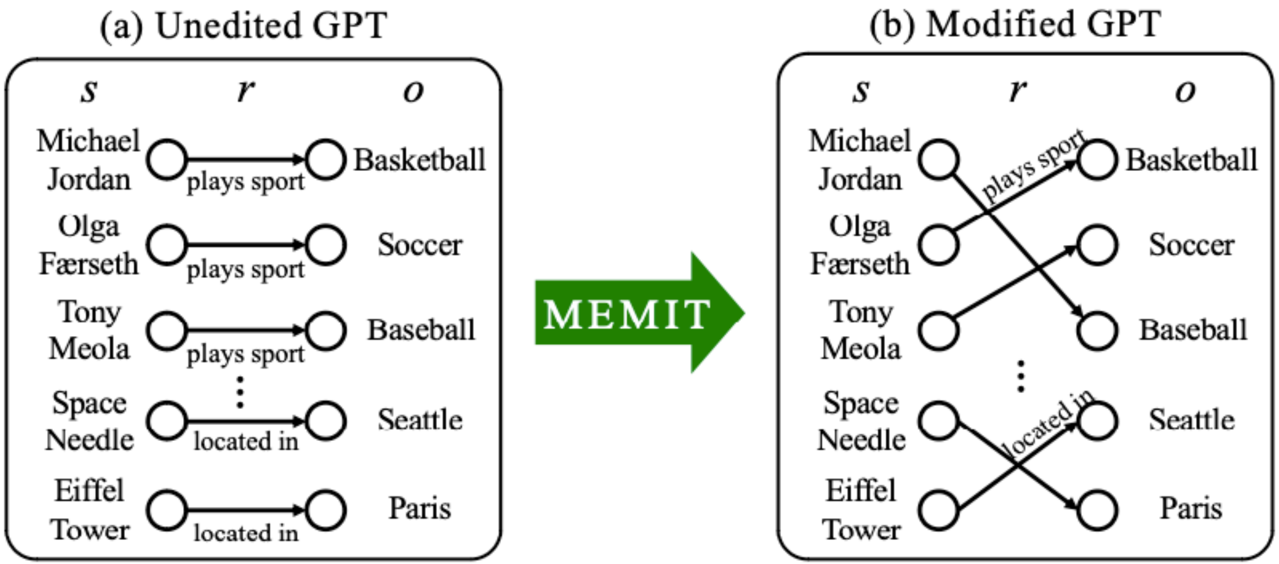
- Locating knowledge in GPT → Causal Tracing
  - Trace information flow in three runs
- Editing knowledge in GPT → Rank-One Model Editing (ROME)
  - Constraint least square results in a rank-one update of MLP layers
  - Identify k_* and v_* for the desired output
- Measure knowledge editing results? → CounterFact dataset
  - Efficacy, Generalization, Specificity, Fluency, Consistency

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

ICLR 2023 Spotlight

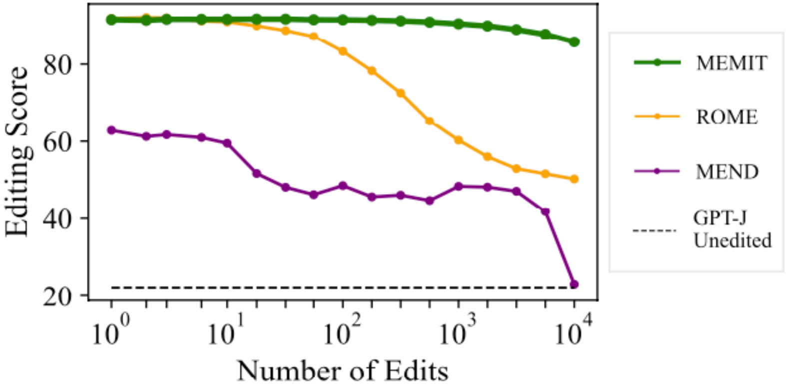
Editing Many Facts in GPT: MEMIT



UCLA

Editing Many Facts in GPT

- Benefits
  - Correct model errors, update outdated knowledge
- Challenges
  - Specificity: Knowledge should not interfere with each other
  - Efficiency: Parallel edits



Editing Score: harmonic mean of efficacy (ES), generalization (PS), and specificity (NS)

Recall: A Linear Layer as A Key-Value Store

- Any linear operation W can operate as a key–value store for

  A set of key vectors $K = [k_1 \mid k_2 \mid \ldots]$

  A set of value vectors $V = [v_1 \mid v_2 \mid \ldots]$

- Pre-trained weights must satisfy least squares (LS):

$$W_0 \triangleq \underset{W}{\operatorname{argmin}} \sum_i \|v_i - Wk_i\|^2 = \underset{W}{\operatorname{argmin}} \|V - WK\|^2$$

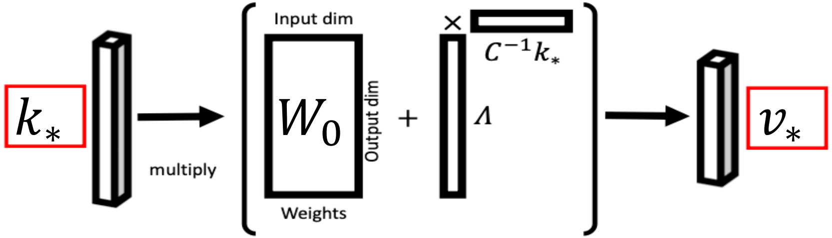
  Normal equation: $W_0 KK^T = VK^T$

$$(X^T X)\beta = X^T y$$

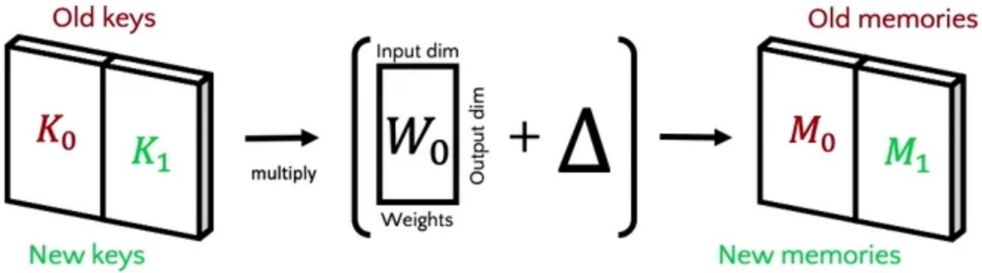
[Kohonen 1972, Anderson 1972]

Improvement 1: Editing Many Facts at Once

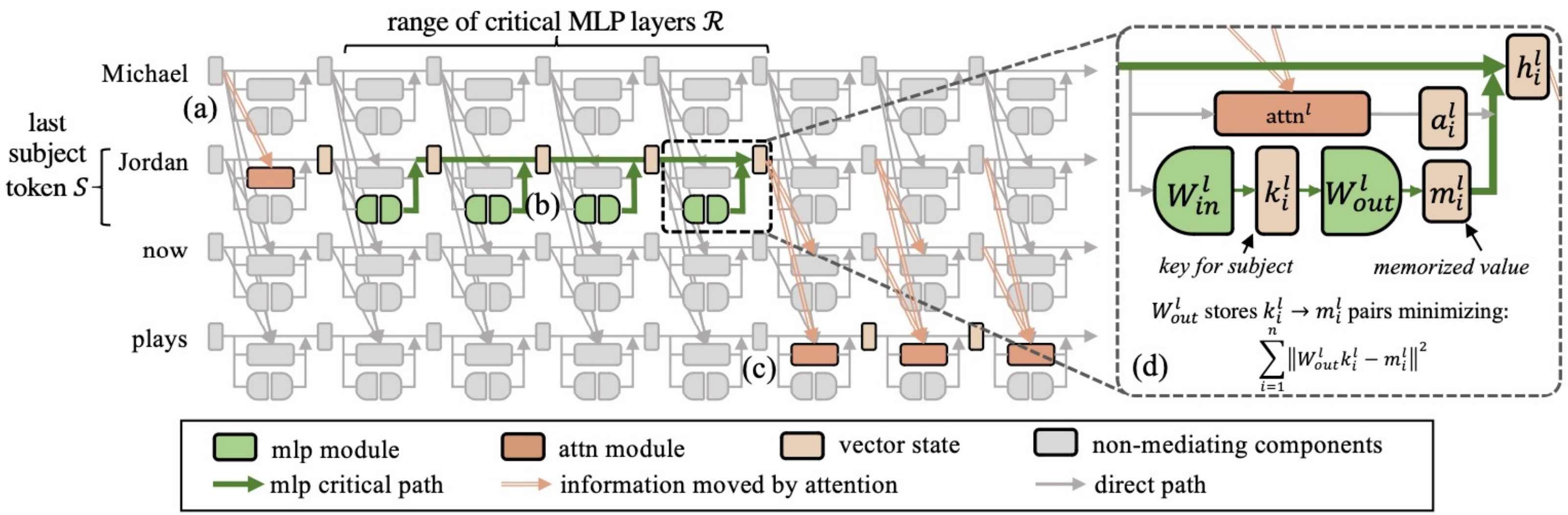
- How to scale up ROME to encode many key-value pairs?



- Stack the new and old facts and update at once



Improvement 2: Editing A Range of MLP Layers



UCLA

Formalize Improvement 1

- ROME
  - Solve a CLS problem
  - The update is a rank-one matrix

$$W_0 K K^T = V K^T$$

$$W_1 = W_0 + \boxed{\Lambda (C^{-1} k_*)^T}$$

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

$$C = K K^T$$

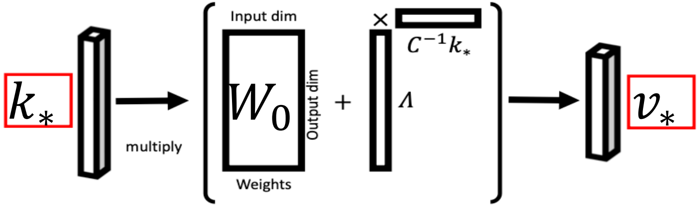
- MEMIT
  - Stack the new and old facts
  - Solve a new LS problem

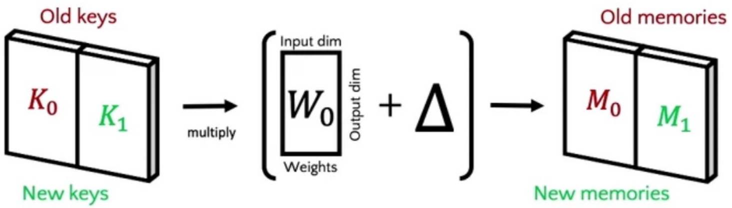
$$W_1 [K_0 \ \ K_1][K_0 \ \ K_1]^T = [M_0 \ \ M_1][K_0 \ \ K_1]^T$$

$$W_1 = W_0 + \Delta$$

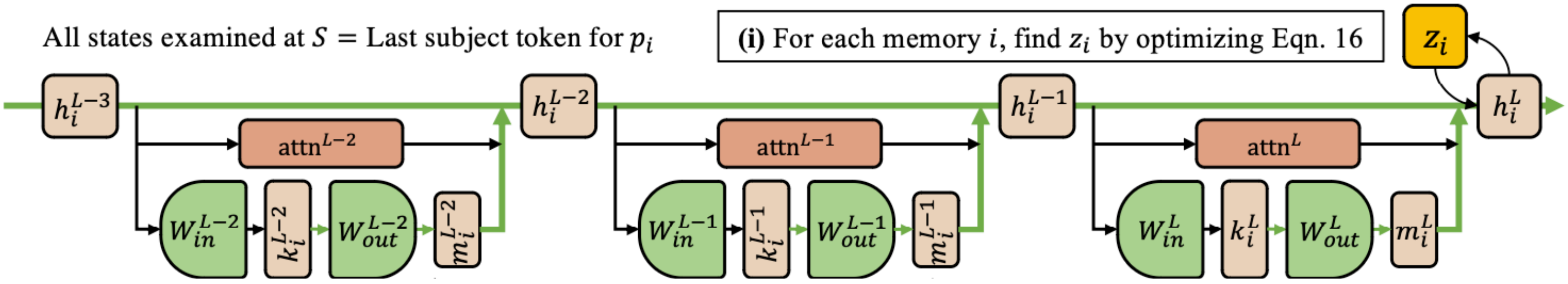
$$\Delta = R \boxed{K_1^T} (C_0 + K_1 K_1^T)^{-1}$$

$$C_0 = K_0 K_0^T \qquad R = \boxed{M_1} - W_0 K_1$$



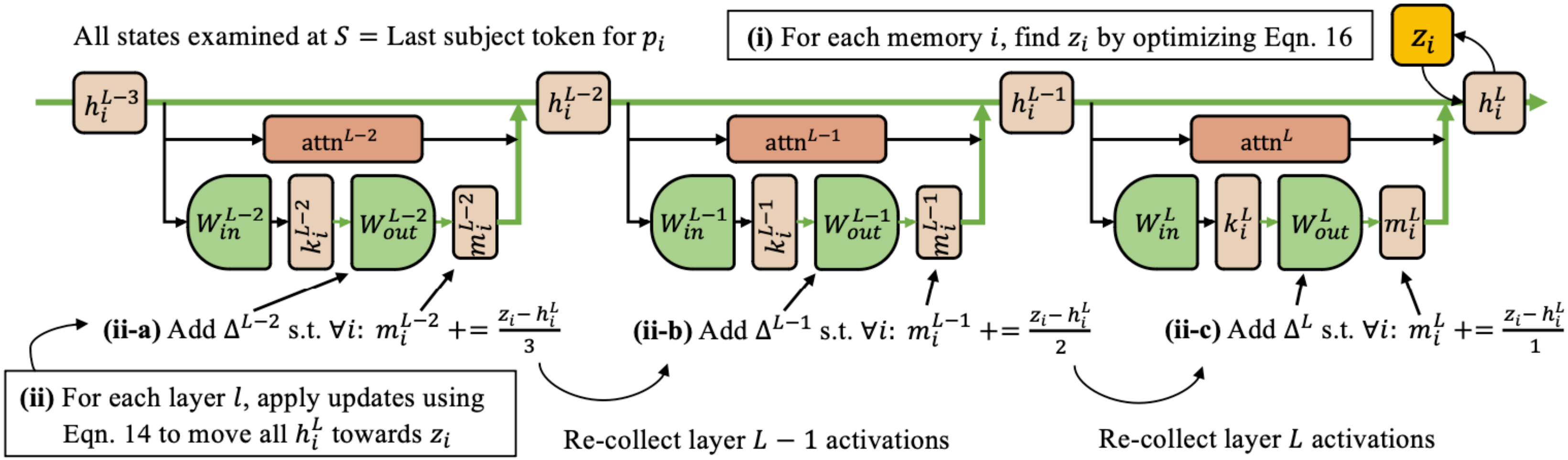


Formalize Improvement 2



$$z_i = h_i^L + \operatorname*{argmin}_{\delta_i} \frac{1}{P} \sum_{j=1}^{P} -\log \mathbb{P}_{G(h_i^L += \delta_i)} \left[o_i \mid x_j \oplus p(s_i, r_i) \right]. \tag{16}$$

Formalize Improvement 2



$$z_i = h_i^L + \operatorname*{argmin}_{\delta_i} \frac{1}{P} \sum_{j=1}^{P} -\log \mathbb{P}_{G(h_i^L \mathrel{+}= \delta_i)} \left[o_i \mid x_j \oplus p(s_i, r_i) \right]. \tag{16}$$

$$\Delta = R K_1^T (C_0 + K_1 K_1^T)^{-1}. \tag{14}$$

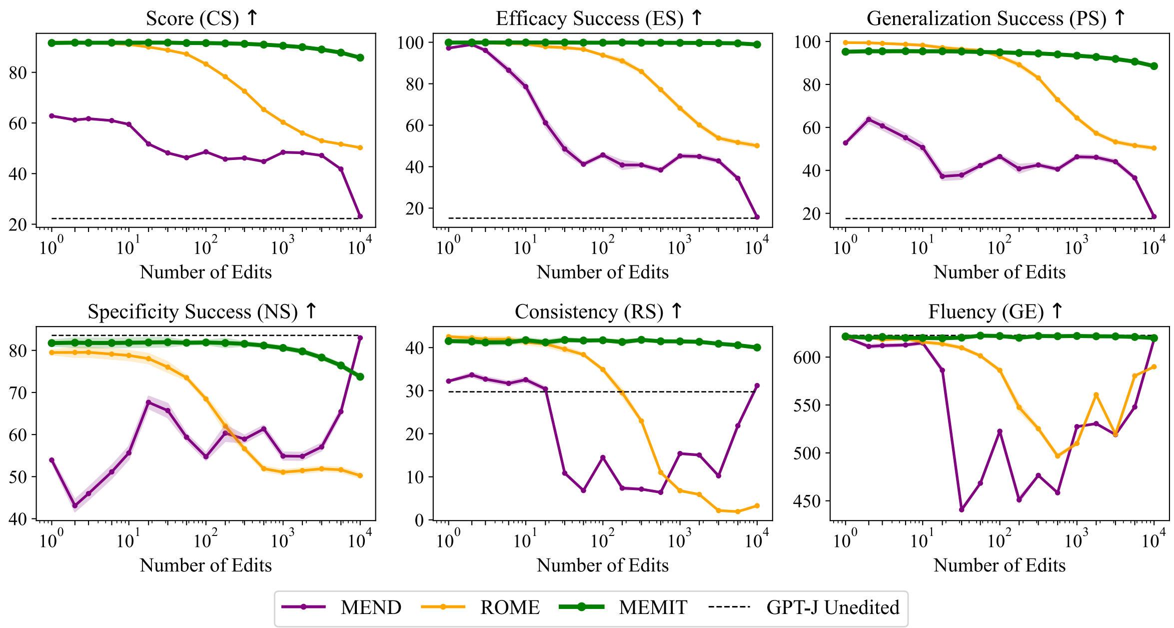
The MEMIT Algorithm

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 $\delta_i \leftarrow \mathrm{argmin}_{\delta_i} \frac{1}{P} \sum_{j=1}^{P} - \log \mathbb{P}_{G(h_i^L += \delta_i)} \left[o_i \mid x_j \oplus p(s_i, r_i) \right]$ (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
  $K^l \leftarrow [k_i^{l_1}, \ldots, k_i^L]$
  $R^l \leftarrow [r_i^{l_1}, \ldots, r_i^L]$
  $\Delta^l \leftarrow R^l {K^l}^T (C^l + K^l {K^l}^T)^{-1}$ (Eqn. 14)
  $W^l \leftarrow W^l + \Delta^l$ // Update layer l MLP weights in model
end

Scaling Curves



correct facts (s, r, o^c)

false facts (s, r, o^*)

Efficacy Score (ES) =

portion of $\mathbb{P}[o^*] > \mathbb{P}[o^c]$

Efficacy Magnitude (EM) =

mean of $\mathbb{P}[o^*] - \mathbb{P}[o^c]$

PS/PM: ES/EM with paraphrase

NS/NM: ES/EM with neighbor subjects

Score: harmonic mean of ES, PS, NS

Fluency: entropy of generated text (GE)

Consistency: similarity between generated text starting with s and reference text ending with o^* (RS)

Are Some Facts Harder to Edit Than Others?

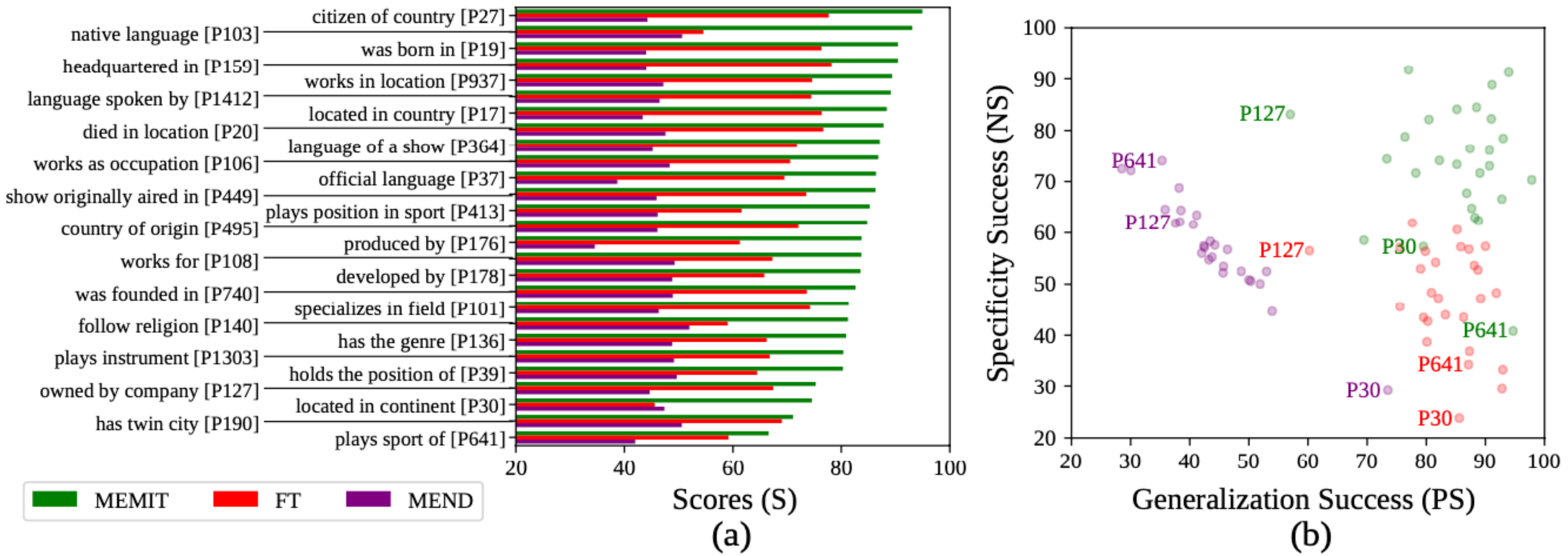
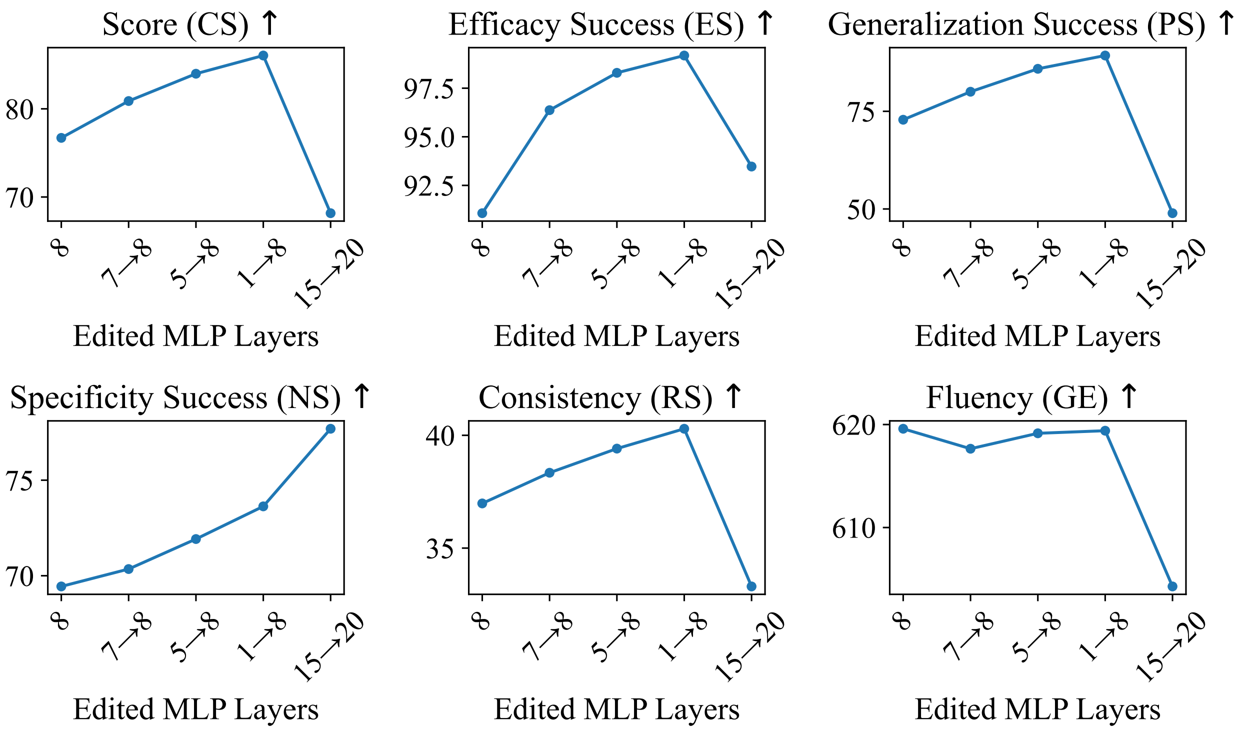


Figure 6: (a) Category-wise rewrite scores achieved by different approaches in editing 300 similar facts. (b) Category-wise *specificity* vs *generalization* scores by different approaches on 300 edits.

Varying Number and Location of Edited Layers



UCLA

Summary

- Scale up model editing to many facts
  - Editing many facts at once by solving a new LS problem
  - Editing a range of MLP layers
- Better specificity and efficiency

Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models

Peter Hase[1,2] **Mohit Bansal**[2] **Been Kim**[1] **Asma Ghandeharioun**[1]
[1]Google Research [2]UNC Chapel Hill
{peter, mbansal}@cs.unc.edu
{beenkim, aghandeharioun}@google.com

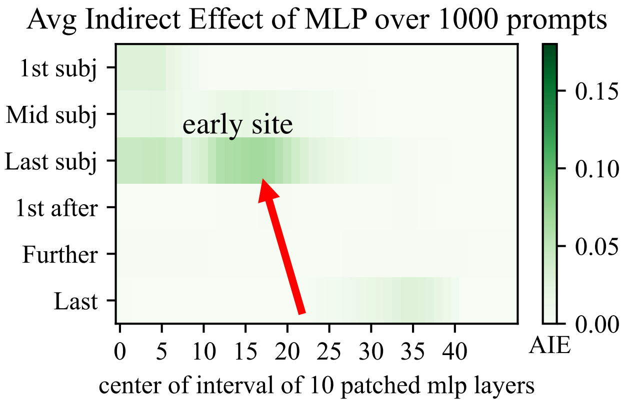
NeurIPS 2023 Spotlight

Key Messages

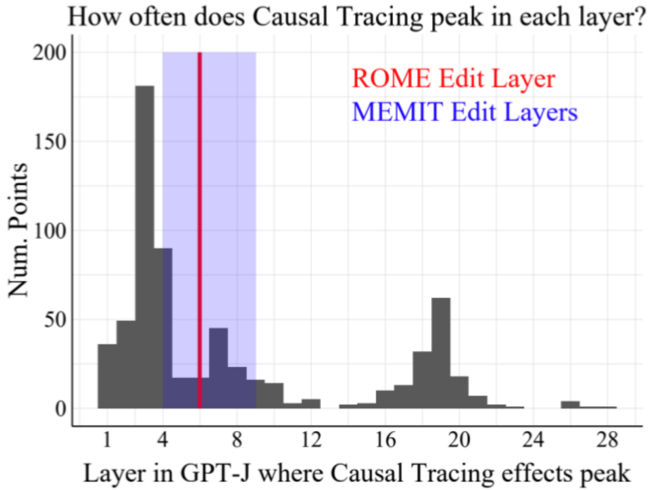
- Where knowledge is stored $\neq$ where to edit an LM
- Better mechanistic understanding $\not\Rightarrow$ better model control

Locating Knowledge with Causal Tracing

- Taking max of MLP tracing effects across all tokens at each layer

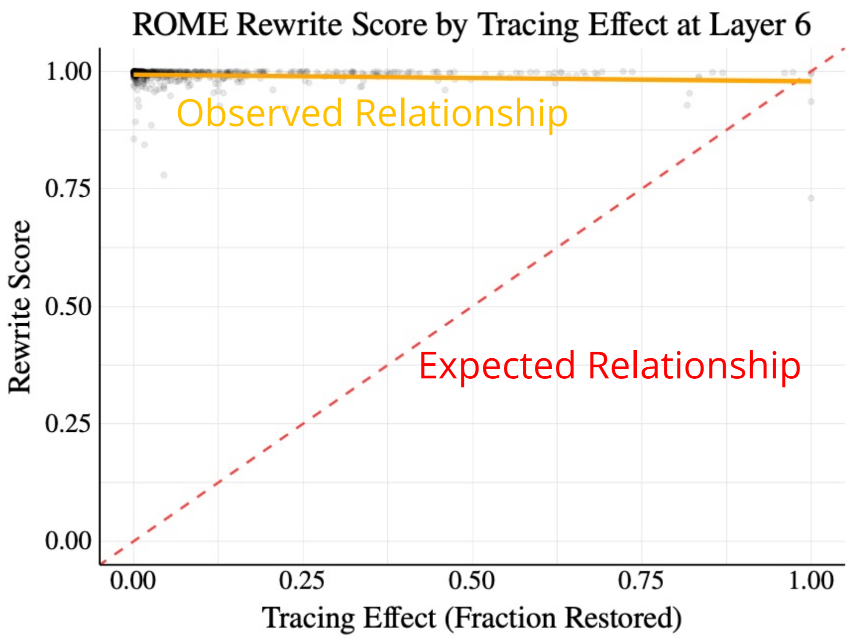


column-wise max



Locating vs. Editing

- Tracing effect does NOT predict edit success



Tracing Effect

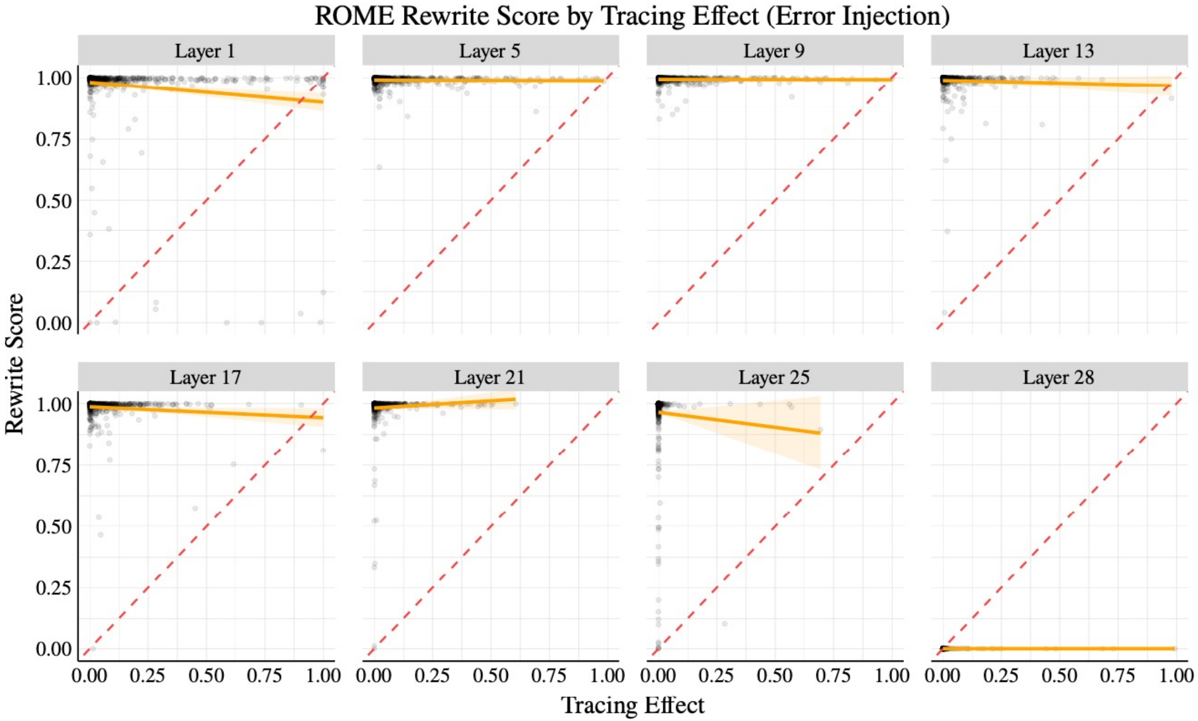
$$p_\theta(o_{true} | s_{noise}, r, v_{(t,\ell)}) - p_\theta(o_{true} | s_{noise}, r)$$

(Normalized) Rewrite Score

$$\frac{p_{\theta^*}(o_{false} | s, r) - p_\theta(o_{false} | s, r)}{1 - p_\theta(o_{false} | s, r)}$$

Locating vs. Editing: Edit Different Layers

- Editing is effective besides layer 28, but correlations are still nearly zero



Explain Rewrite Score Variance

- Linear regression to predict rewrite scores with features
  - The choice of edit layer as a categorical variable
  - Tracing effect
  - Both
- Tracing effect cannot explain the variance in edit success

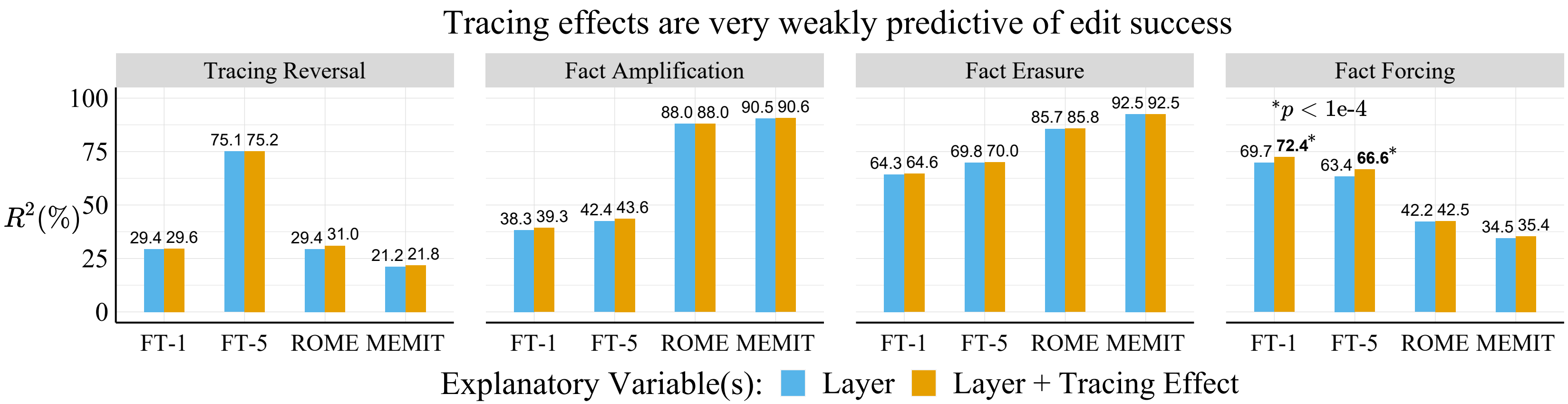
	R^2 Values		
Method	Layer	Tracing Effect	Both
ROME	0.947	0.016	0.948

Problem Variants

Editing Problem Variants	Input Prompt	Objective
Error Injection	Autonomous University of Madrid, which is located in ________	$\longrightarrow \arg\max_{\theta} p_{\theta}(\text{Sweden}\mid\text{Input})$
Tracing Reversal	Autonomous University of Madrid, which is located in ________	$\longrightarrow \arg\max_{\theta} p_{\theta}(o_{\text{noise}}\mid\text{Input})$
Fact Erasure	Autonomous University of Madrid, which is located in ________	$\longrightarrow \arg\min_{\theta} p_{\theta}(\text{Spain}\mid\text{Input})$
Fact Amplification	Autonomous University of Madrid, which is located in ________	$\longrightarrow \arg\max_{\theta} p_{\theta}(\text{Spain}\mid\text{Input})$
Fact Forcing	$\underbrace{\textit{Autonomous University of Madrid}}_{\text{Add noise to subject}}$, which is located in ________	$\longrightarrow \arg\max_{\theta} p_{\theta}(\text{Spain}\mid\text{Noisy Input})$

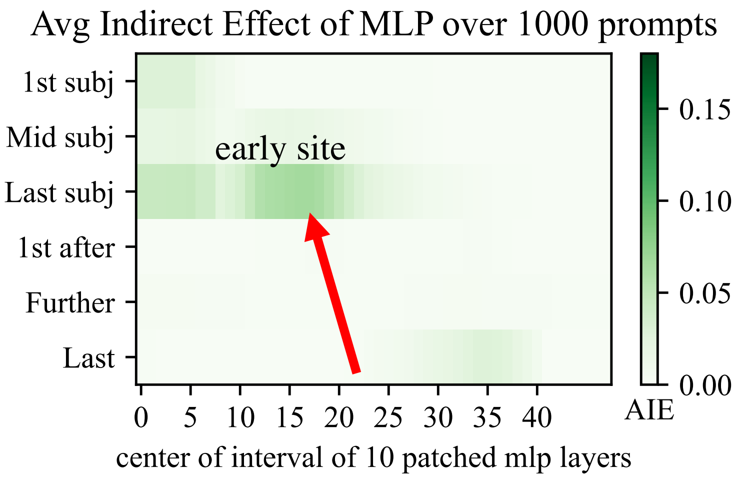
Experiment Results for Problem Variants

- Tracing effects are very weakly predictive of edit success across editing problems and methods



Discussion

- Does Causal Tracing tell us anything?
  - Causal tracing shows the importance of the last subject token
  - Editing later layers indeed causes a performance drop



Discussion

- Why edit works at layers where the edited fact is not stored?
  - It seems possible to "override" knowledge stored in layer l by editing layer k
  - Hypothesis: A fact can be stored in many layers
- How do we validate localization interpretability claims?
- If localization and editing are answering different questions, what are the questions?

UCLA

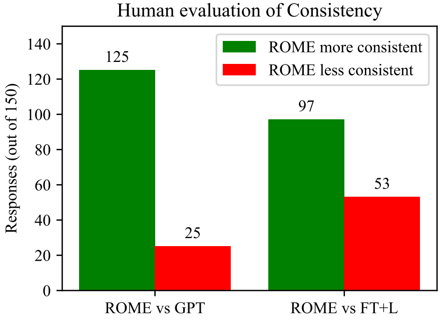
Outline

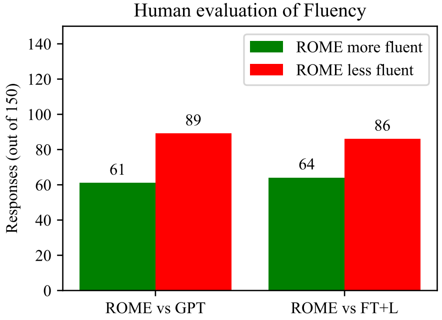
- Background
- Locating and Editing Knowledge
  - Locating and Editing Factual Associations in GPT (NeurIPS 2022)
  - Mass Editing Memory in a Transformer (ICLR 2023 Spotlight)
  - Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models (NeurIPS 2023 Spotlight)
- Future Directions

Future Directions: ROME

- Analysis of attention layers
- Model fluency
- Models store information in a different way as humans (expect)
  - Bill Gates founded Microsoft
  - Microsoft was founded by whom?

Human evaluation: ROME is more consistent than FT+L, but less fluent.





Future Directions: Broader

- What are the right questions to distinguish locating and editing
- Can interpretable models be better than opaque models
- Can we edit something beyond factual knowledge
- Locating knowledge for alignment

Reference

- Maslej, Nestor, et al. , “The AI Index 2023 Annual Report,” AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2023.
- Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." *CVPR*. 2014.
- Zhong, Qihuang, et al. "Toward efficient language model pretraining and downstream adaptation via self-evolution: A case study on superglue." *arXiv preprint arXiv:2212.01853* (2022).
- Stokes, Jonathan M., et al. "A deep learning approach to antibiotic discovery." *Cell* 180.4 (2020): 688-702.
- Ramesh, Aditya, et al. "Zero-shot text-to-image generation." *International Conference on Machine Learning*. PMLR, 2021.
- Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13* (pp. 818-833).
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).
- Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in Neural Information Processing Systems* 35 (2022): 27730-27744.

Reference

- Herculano-Houzel, Suzana. *The human advantage: a new understanding of how our brain became remarkable*. MIT Press, 2016.
- Petroni, Fabio, et al. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*, 2020.
- Jiang, Zhengbao, et al. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.
- Weston, Jason, Sumit Chopra, and Antoine Bordes. "Memory networks." *arXiv preprint arXiv:1410.3916* (2014).
- Geva, Mor, et al. "Transformer feed-forward layers are key-value memories." *arXiv preprint arXiv:2012.14913* (2020).
- Kohonen, T. Correlation matrix memories. *IEEE transactions on computers*, 100(4):353–359, 1972.
- Anderson, J. A. A simple neural network generating an interactive memory. *Mathematical biosciences*, 14(3-4):197–220, 1972.
- Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schu ̈tze, H., and Goldberg, Y. Measuring and Improving Consistency in Pretrained Language Models. Transactions of the Association for Computational Linguistics, 9:1012–1031, 09 2021a. ISSN 2307-387X.

Thank you for linstening!

Q & A