ROME: Editing Factual Associations in GPT

Kevin Meng, David Bau, Alex Andonian, Yonatan Belinkov

Stefan Kramer
In which city is the Space Needle?

The Space Needle is located in Seattle, Washington, USA.
Where does a large language model store its facts?
The Space Need le is in Seattle
The Space Need is in Seattle
Remark: Notation

• Superscript: Denotes the layer/column
• Subscript: Denotes the row

\[ h_2^{(2)} \]
Hidden states

The Space Need le is in Seattle
Encoding

Space \rightarrow h_2^{(0)}

\begin{align*}
  h_i^{(0)} &= emb(x_i) + pos(i) \in \mathbb{R}^H \\
  h_2^{(0)} &= emb(Space) + pos(2)
\end{align*}
The Space Need in Seattle
Hidden states

\[ h_i^{(l-1)} = h_i^{(l)} + a_i^{(l)} + m_i^{(l)} \]
Attention

• Depends on the states/tokens before

\[ a_i^{(l)} = \text{attn}^{(l)}(h_1^{(l-1)}, ..., h_i^{(l-1)}) \]
The Space Need Be
Multilayer Perceptron

• Two-layer
• Fully connected and projection layer
• $\sigma$: rectifying nonlinearity
• $\gamma$: normalizing nonlinearity

\[
m_i^{(l)} = W_{\text{proj}}^{(l)} \sigma (W_{fc}^{(l)} \gamma (a_i^{(l)} + h_i^{(l-1)}))
\]
Tracing Information Flow
The Space Need is in Seattle
The Space Need is in Seattle.
Measurements

• **Total effect (TE):** Compares the probabilities for generating the correct output of the **clean** and **corrupted** run.

• **Indirect effect (IE):** Compares the probabilities for generating the correct output of the **corrupted** and **restored** run.
Indirect effect with a single restored layer

IE when hidden state at the last input token and layer 15 was restored.
Indirect with an interval of MLP’s restored
Indirect with an interval of MLP’s restored
Average Results (1000 prompts)
Changing Facts

Who is the coach of the FC Bayern Munich?

As of my knowledge cutoff of September 2021, the head coach of FC Bayern Munich was Julian Nagelsmann. However, it's possible that there have been changes since then.

As of my knowledge cutoff of September 2021, the head coach of FC Bayern Munich was Thomas Tuchel. However, it's possible that there have been changes since then.
MLP as linear associative memory

• Key-Value store
• $K = [k_1 \mid k_2 \mid \ldots]$ represent the keys
• $V = [v_1 \mid v_2 \mid \ldots]$ represent the values

$$W_{\text{proj}}^{(l)} K \approx V$$

• Squared error is minimized by using the Moore-Penrose pseudoinverse
MLP as linear associative memory
Optimization problem

Initial position:
• \( K = [k_1 | k_2 | ... ] \) represent the keys
• \( V = [v_1 | v_2 | ... ] \) represent the values
• \( W \) minimizes \( \| WK - V \|_2^2 \)

Goal:
• Insert a new key-value pair \((k_*, v_*)\), while keeping the squared loss low
Optimization problem

Goal:

• Insert a new key-value pair \((k_*, v_*)\), while keeping the squared loss low.

Solution: Compute \(W^*\) solving the following optimization problem:

\[
\text{minimize} \quad \| W^* K - V \|_2^2 \quad \text{s.t.} \quad W^* k_* = v_*
\]
Step 1: Choose $k_*$
Step 1: Choose $k_*$

1. Pass the text containing the subject through the Model
The Space Need is in Seattle
Step 1: Choose $k_*$

1. Pass the text containing the subject through the Model
2. Go to the MLP at the most decisive layer in the last subject token row (around layer 15)
Step 1: Choose $k_*$

1. Pass the text containing the subject through the Model
2. Go to the MLP at the most decisive layer in the last subject token row
3. Read value inside the MLP after applying $W^{fc}$ and the non-linearity
Step 1: Choose $k_*$

• Repeat these three steps for multiple texts ending in the same subject
• Take $k_*$ to be the average
Step 2: Choose $\nu_*$

- Set $\nu_* = \arg\min_z \mathcal{L}(z)$

- $G \left( m_{i}^{(l^*)} := z \right)$ is the grid where the output of the MLP at token $i$ and layer $l^*$ is set to $z$.

\[
\frac{1}{N} \sum_{j=1}^{N} \left[ \log \mathbb{P}_{G(m_{i}^{(l^*)} := z)} [o^* | x_j + p] \right] + D_{KL} \left( \mathbb{P}_{G(m_{i}^{(l^*)} := z)} [x | p'] \parallel \mathbb{P}_{G} [x | p'] \right).
\]

(a) Maximizing $o^*$ probability

(b) Controlling essence drift
Step 2: Choose $\nu_*$

- Set $\nu_* = \arg\min_z \mathcal{L}(z)$
- $G(m_i^{(l^*)} := z)$ is the grid where the output of the MLP at token $i$ and layer $l^*$ is set to $z$.

\[
\frac{1}{N} \sum_{j=1}^{N} \log \mathbb{P}_{G(m_i^{(l^*)} := z)} \left[ o^* \mid x_j + p \right] + D_{KL} \left( \mathbb{P}_{G(m_i^{(l^*)} := z)} [x \mid p'] \left\| \mathbb{P}_G [x \mid p'] \right. \right).
\]

(a) Maximizing $o^*$ probability

(b) Controlling essence drift
Step 2: Choose $v_*$

- Set $v_* = \arg\min_z L(z)$
- $G(m_i^{(l^*)} := z)$ is the grid where the output of the MLP at token $i$ and layer $l^*$ is set to $z$.

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

(a) Maximizing $o^*$ probability

(b) Controlling essence drift
Step 3: Insert the fact

• Solve the optimization problem to get $W^*$
• Replace $W$ with $W^*$

$$\text{minimize } \| W^* K - V \|_2^2 \quad \text{s.t. } W^* k_* = v_*$$
Evaluation
What is knowledge? (According to the authors)

**Generalization**

The coach of Bayern Munich is Thomas Tuchel.

The team of Bayern Munich is coached by Thomas Tuchel.

**Specificity**

The coach of Bayern Munich is Julian Nagelsmann.

The coach of Bayern Munich is Thomas Tuchel.

The coach of Real Madrid is Carlo Ancelotti.
Other Editing Methods

• Fine-Tuning
• Constrained Fine-Tuning
• Knowledge Editor
• MEND
Fine Tuning (FT)

• Apply Adam with early stopping at one layer
• Minimize $-\log P(o^*|x)$
Constrained Fine-Tuning (FT + L)

• Like Fine-Tuning
• Additional constraint on weight change
Knowledge Editor (KE)

Auxiliary Network that tries to predict the weight change of $\theta$

Regular predictions

New output
MEND

• Like KE it uses Auxiliary networks
• Learns to transform the gradient
### Zero-shot Relation Extraction

<table>
<thead>
<tr>
<th>Factual statement</th>
<th>Paraphrase</th>
<th>Unrelated factual statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>When was the launch of the iPhone 7?</td>
<td>When was the iPhone 7 released?</td>
<td>When was the first moon landing?</td>
</tr>
<tr>
<td></td>
<td>September 16, 2016</td>
<td>July 20, 1969</td>
</tr>
</tbody>
</table>

10'000 examples
Zero-shot Relation Extraction Results

<table>
<thead>
<tr>
<th>Editor</th>
<th>Efficacy ↑</th>
<th>Paraphrase ↑</th>
<th>Specificity ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-2 XL</td>
<td>22.2 (±0.5)</td>
<td>21.3 (±0.5)</td>
<td>24.2 (±0.5)</td>
</tr>
<tr>
<td>FT</td>
<td>99.6 (±0.1)</td>
<td>82.1 (±0.6)</td>
<td>23.2 (±0.5)</td>
</tr>
<tr>
<td>FT+L</td>
<td>92.3 (±0.4)</td>
<td><strong>47.2 (±0.7)</strong></td>
<td>23.4 (±0.5)</td>
</tr>
<tr>
<td>KE</td>
<td>65.5 (±0.6)</td>
<td>61.4 (±0.6)</td>
<td>24.9 (±0.5)</td>
</tr>
<tr>
<td>KE-zsRE</td>
<td>92.4 (±0.3)</td>
<td>90.0 (±0.3)</td>
<td>23.8 (±0.5)</td>
</tr>
<tr>
<td>MEND</td>
<td>75.9 (±0.5)</td>
<td>65.3 (±0.6)</td>
<td>24.1 (±0.5)</td>
</tr>
<tr>
<td>MEND-zsRE</td>
<td>99.4 (±0.1)</td>
<td><strong>99.3 (±0.1)</strong></td>
<td>24.1 (±0.5)</td>
</tr>
<tr>
<td>ROME</td>
<td><strong>99.8 (±0.0)</strong></td>
<td>88.1 (±0.5)</td>
<td><strong>24.2 (±0.5)</strong></td>
</tr>
</tbody>
</table>
Counterfact-Dataset

«Can the editing method change the location of the Gazi University from Ankara to Glasgow?»
"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"
Counterfact-Dataset Measurements

• **Efficacy**
  • ES: Portion of cases for which $P[\text{false fact}] > P[\text{correct fact}]$
  • EM: $P[\text{false fact}] - P[\text{correct fact}]$

• **Generalization**
  • PS: Portion of cases for which $P[\text{false fact}] > P[\text{correct fact}]$
  • PM: $P[\text{false fact}] - P[\text{correct fact}]$

• **Specificity/Influence on Neighborhood**
  • NS: Portion of cases for which $P[\text{correct fact}] > P[\text{false fact}]$
  • NM: $P[\text{correct fact}] - P[\text{false fact}]$
# Efficacy, Generalization & Specificity

<table>
<thead>
<tr>
<th>Editor</th>
<th>Score</th>
<th>Efficacy</th>
<th>Generalization</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S ↑</td>
<td>ES ↑</td>
<td>EM ↑</td>
<td>PS ↑</td>
</tr>
<tr>
<td>GPT-2 XL</td>
<td>30.5</td>
<td>22.2 (0.9)</td>
<td>-4.8 (0.3)</td>
<td>24.7 (0.8)</td>
</tr>
<tr>
<td>FT</td>
<td>65.1</td>
<td>100.0 (0.0)</td>
<td>98.8 (0.1)</td>
<td>87.9 (0.6)</td>
</tr>
<tr>
<td>FT+L</td>
<td>66.9</td>
<td>99.1 (0.2)</td>
<td>91.5 (0.5)</td>
<td><strong>48.7 (1.0)</strong></td>
</tr>
<tr>
<td>KN</td>
<td>35.6</td>
<td><strong>28.7 (1.0)</strong></td>
<td>-3.4 (0.3)</td>
<td><strong>28.0 (0.9)</strong></td>
</tr>
<tr>
<td>KE</td>
<td>52.2</td>
<td>84.3 (0.8)</td>
<td>33.9 (0.9)</td>
<td>75.4 (0.8)</td>
</tr>
<tr>
<td>KE-CF</td>
<td>18.1</td>
<td>99.9 (0.1)</td>
<td>97.0 (0.2)</td>
<td>95.8 (0.4)</td>
</tr>
<tr>
<td>MEND</td>
<td>57.9</td>
<td>99.1 (0.2)</td>
<td>70.9 (0.8)</td>
<td>65.4 (0.9)</td>
</tr>
<tr>
<td>MEND-CF</td>
<td>14.9</td>
<td><strong>100.0 (0.0)</strong></td>
<td><strong>99.2 (0.1)</strong></td>
<td><strong>97.0 (0.3)</strong></td>
</tr>
<tr>
<td>ROME</td>
<td><strong>89.2</strong></td>
<td>100.0 (0.1)</td>
<td>97.9 (0.2)</td>
<td>96.4 (0.3)</td>
</tr>
</tbody>
</table>

08.05.2023

ROME
All other methods have weaknesses!

<table>
<thead>
<tr>
<th>Editor</th>
<th>Score</th>
<th>Efficacy</th>
<th>Generalization</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S ↑</td>
<td>ES ↑</td>
<td>EM ↑</td>
<td>PS ↑</td>
</tr>
<tr>
<td>GPT-2 XL</td>
<td>30.5</td>
<td>22.2 (0.9)</td>
<td>-4.8 (0.3)</td>
<td>24.7 (0.8)</td>
</tr>
<tr>
<td>FT</td>
<td>65.1</td>
<td>100.0 (0.0)</td>
<td>98.8 (0.1)</td>
<td>87.9 (0.6)</td>
</tr>
<tr>
<td>FT+L</td>
<td>66.9</td>
<td>99.1 (0.2)</td>
<td>91.5 (0.5)</td>
<td><strong>48.7 (1.0)</strong></td>
</tr>
<tr>
<td>KN</td>
<td><strong>35.6</strong></td>
<td><strong>28.7 (1.0)</strong></td>
<td><strong>-3.4 (0.3)</strong></td>
<td><strong>28.0 (0.9)</strong></td>
</tr>
<tr>
<td>KE</td>
<td>52.2</td>
<td>84.3 (0.8)</td>
<td>33.9 (0.9)</td>
<td>75.4 (0.8)</td>
</tr>
<tr>
<td>KE-CF</td>
<td><strong>18.1</strong></td>
<td>99.9 (0.1)</td>
<td>97.0 (0.2)</td>
<td>95.8 (0.4)</td>
</tr>
<tr>
<td>MEND</td>
<td>57.9</td>
<td>99.1 (0.2)</td>
<td>70.9 (0.8)</td>
<td>65.4 (0.9)</td>
</tr>
<tr>
<td>MEND-CF</td>
<td><strong>14.9</strong></td>
<td><strong>100.0 (0.0)</strong></td>
<td><strong>99.2 (0.1)</strong></td>
<td><strong>97.0 (0.3)</strong></td>
</tr>
<tr>
<td>ROME</td>
<td><strong>89.2</strong></td>
<td><strong>100.0 (0.1)</strong></td>
<td><strong>97.9 (0.2)</strong></td>
<td><strong>96.4 (0.3)</strong></td>
</tr>
</tbody>
</table>
All other methods have weaknesses!

<table>
<thead>
<tr>
<th>Editor</th>
<th>Score</th>
<th>Efficacy</th>
<th>Generalization</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S ↑</td>
<td>ES ↑</td>
<td>EM ↑</td>
<td>PS ↑</td>
</tr>
<tr>
<td>GPT-2 XL</td>
<td>30.5</td>
<td>22.2 (0.9)</td>
<td>-4.8 (0.3)</td>
<td>24.7 (0.8)</td>
</tr>
<tr>
<td>FT</td>
<td>65.1</td>
<td>100.0 (0.0)</td>
<td>98.8 (0.1)</td>
<td>87.9 (0.6)</td>
</tr>
<tr>
<td>FT+L</td>
<td>66.9</td>
<td>99.1 (0.2)</td>
<td>91.5 (0.5)</td>
<td>48.7 (1.0)</td>
</tr>
<tr>
<td>KN</td>
<td><strong>35.6</strong></td>
<td><strong>28.7 (1.0)</strong></td>
<td><strong>-3.4 (0.3)</strong></td>
<td><strong>28.0 (0.9)</strong></td>
</tr>
<tr>
<td>KE</td>
<td>52.2</td>
<td>84.3 (0.8)</td>
<td>33.9 (0.9)</td>
<td>75.4 (0.8)</td>
</tr>
<tr>
<td>KE-CF</td>
<td>18.1</td>
<td>99.9 (0.1)</td>
<td>97.0 (0.2)</td>
<td>95.8 (0.4)</td>
</tr>
<tr>
<td>MEND</td>
<td>57.9</td>
<td>99.1 (0.2)</td>
<td>70.9 (0.8)</td>
<td>65.4 (0.9)</td>
</tr>
<tr>
<td>MEND-CF</td>
<td><strong>14.9</strong></td>
<td><strong>100.0 (0.0)</strong></td>
<td><strong>99.2 (0.1)</strong></td>
<td><strong>97.0 (0.3)</strong></td>
</tr>
<tr>
<td>ROME</td>
<td><strong>89.2</strong></td>
<td>100.0 (0.1)</td>
<td>97.9 (0.2)</td>
<td>96.4 (0.3)</td>
</tr>
</tbody>
</table>
Limitations

- Scalability issue: only one fact at once
- No Runtime-analysis

- Who has the responsibility?
- Is Model editing the right way?
Thank you for your attention!
References


