Learning How to Ask: Querying LMs with Mixtures of Soft Prompts

Guanghui Qin and Jason Eisner
Johns Hopkins University
LMs Already Did Your Job
LMs Already Did Your Job

Prompt: 

Input

Task: What year did Mary Cassatt die?
LMs Already Did Your Job

Prompt: Mary Cassatt X died in 1926 y.

Input

BERT

GPT-3

Output

Task: What year did Mary Cassatt die?
The Form of the Question Matters

- Should employers be *forced* to negotiate with unions?
- Should unions have the *right* to negotiate with employers?

Opinion Poll
The Form of the Question Matters

• Should employers be **forced** to negotiate with unions?
• Should unions have the **right** to negotiate with employers?

Opinion Poll

• Why **can’t we** see the moon now?
• Where does **the moon** go during the day?

Children’s Mind
The Form of the Question Matters

We’re prompting LMs!

Opinion Poll

Children’s Mind

BERT

GPT-3
Prompt: Mary Cassatt died in [MASK].
Prompt: Mary Cassatt died in [MASK].

Task: Fact Queries: Year-of-Death
Task: Fact Queries: Year-of-Death

Prompt: 

Mary Cassatt X died in [MASK] .

Output: 1926
Cab Calloway died in Delaware.
Prompt LMs

Prompt: Cab Calloway X perished in ________ [MASK] ________ y .

Task: Fact Queries: Year-of-Death
Prompt: Cab Calloway x perished during __________ y .

Task: Fact Queries: Year-of-Death

Output: 1994
Task: Fact Queries: Year-of-Death

Prompt LMs

Database

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cab Calloway</td>
<td>1994</td>
</tr>
<tr>
<td>Mary Cassatt</td>
<td>1926</td>
</tr>
</tbody>
</table>

Simulated Annealing

Prompt: Cab Calloway x perished during [MASK] y.
Prompt LMs

Task: Fact Queries: Year-of-Death

Output

Database

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cab Calloway</td>
<td>1994</td>
</tr>
<tr>
<td>Mary Cassatt</td>
<td>1926</td>
</tr>
</tbody>
</table>

Transformations

Prompt: Cab Calloway

Simulated Annealing

Gradient Descent

Input

1994

ignored

ignored

ignored
Task: Fact Queries: Year-of-Death

Prompt: Cab Calloway ___________ X perished during _________ y.

Output: 1994
Task: Fact Queries: Year-of-Death

Prompt: Cab Calloway x perished during [MASK] y .

Embeddings:

Output: 1994
Prompt LMs

Transformers

1994

Embeddings:

Prompt: Cab Calloway perished during [MASK].

Task: Fact Queries: Year-of-Death
Why Soft Prompts?

Cab Calloway played until his death in [MASK] year.
Why Soft Prompts?

• Easy to search with backprop.

Cab Calloway played until his death in [MASK].
Why Soft Prompts?

• Easy to search with backprop.
• We have a larger space of prompts.

Cab Calloway * played until his death in [MASK] y.
Why Soft Prompts?

• Easy to search with backprop.
• We have a larger space of prompts.

Mary Cassatt X played until his death in [MASK].
Why Soft Prompts?

• Easy to search with backprop.
• We have a larger space of prompts.

Mary Cassatt played until his death in [MASK].
Why Soft Prompts?

- Easy to search with backprop.
- We have a larger space of prompts.

Mary Cassatt \( \times \) played until his death in \([\text{MASK}]\) \( \cdot \)

\(\begin{align*}
\text{gender neutral} & \quad = \{\text{his, her, their}\} \\
\text{profession neutral} & \quad = \{\text{played, painted, worked}\}
\end{align*}\)
Why Soft Prompts?

• Easy to search with backprop.
• We have a larger space of prompts.

Mary Cassatt X played until his death in [MASK]y

gender neutral
= {his, her, their}

profession neutral
= {played, painted, worked}

• 1926?
• Paris?
Why Soft Prompts?

- Easy to search with backprop.
- We have a larger space of prompts.
Why Soft Prompts?

• Easy to search with backprop.
• We have a larger space of prompts.
• They can emphasize certain keywords, even particular dimensions.

\[
\text{played until his death in } [\text{MASK}] \text{ year} \notin \text{English}
\]

\[
\text{Mary Cassatt} \times \{\text{played, painted, worked}\}
\]

\[
\text{gender neutral} = \{\text{his, her, their}\}
\]

\[
\text{profession neutral}
\]
Why Soft Prompts?

- Easy to search with backprop.
- We have a larger space of prompts.
- They can emphasize certain keywords, even particular dimensions.

Mary Cassatt played until his death. 

More general

- Profession neutral
- Gender neutral

played, painted, worked

- {his, her, their}

1926? √
Paris? ×

in [MASK] year ☞ English
Why Soft Prompts?

- Easy to search with backprop.
- We have a larger space of prompts.
- They can emphasize certain keywords, even particular dimensions.

Mary Cassatt played until his death in [MASK].

More Specific

[1926? ✔]
[Paris? ✗]
Why Soft Prompts?

• Easy to search with backprop.
• We have a larger space of prompts.
• They can emphasize certain keywords, even particular dimensions.

Mary Cassatt \( \times \) played until his death in [MASK] year.

- \{his, her, their\}
- \{played, painted, worked\}
- in year \( \notin \) English
Deep Perturbation of Prompts

Layer 1:

Embeddings:

Prompt: Mary Cassatt X died in [MASK] y.
Deep Perturbation of Prompts

Layer 1:

Embeddings:

Prompt:  

Mary Cassatt  \( x \) died in  \( y \).
Deep Perturbation of Prompts

Layer 1:

Embeddings:

Prompt: Mary Cassatt X died in [MASK] y.
Deep Perturbation of Prompts

Layer 2:

Layer 1:

Embeddings:

Prompt: Mary Cassatt \(x\) died in \(\text{[MASK]}\) \(y\).
Deep Perturbation of Prompts

Prompt:  Mary Cassatt $x$ died in $[\text{MASK}]$ $y$.
Deep Perturbation of Prompts

#param ratio

soft-prompt 1
deep-perturbation 25
fully-fine-tuning 1e5
Ensembling (Mixture of Prompts)

weight = 0.33  _____ x  died  in  _____ y  .

weight = 0.33  _____ x  was  the  year  _____ y  died  .

weight = 0.33  _____ x  performed  until  his  death  in  _____ y  .
Ensembling (Mixture of Prompts)

weight = 0.45

\[
\begin{align*}
\text{weight} & = 0.45 \\
\text{x} \quad \text{died} \quad \text{in} \quad \text{y} \\
\end{align*}
\]

weight = 0.15

\[
\begin{align*}
\text{weight} & = 0.15 \\
\text{x} \quad \text{was} \quad \text{the} \quad \text{year} \quad \text{y} \quad \text{died} \\
\end{align*}
\]

weight = 0.40

\[
\begin{align*}
\text{weight} & = 0.40 \\
\text{x} \quad \text{performed} \quad \text{until} \quad \text{his} \quad \text{death} \quad \text{in} \quad \text{y} \\
\end{align*}
\]

\[
p(y \mid x) = \sum_{\text{prompt}} p(y \mid \text{prompt}(x)) \cdot p(\text{prompt})
\]
Ensembling (Mixture of Prompts)

- Bigger model: When one prompt is unsure, others can help

\[
p(y \mid x) = \sum_{\text{prompt}} p(y \mid \text{prompt}(x)) \cdot p(\text{prompt})
\]
Ensembling (Mixture of Prompts)

- Bigger model: When one prompt is unsure, others can help
- Better predictions: Ensembling reduces variance

\[ p(y | x) = \sum_{\text{prompt}} p(y | \text{prompt}(x)) \cdot p(\text{prompt}) \]
Ensembling (Mixture of Prompts)

• Bigger model: When one prompt is unsure, others can help
• Better predictions: Ensembling reduces variance
• Better optimization: Explore multiple starting points in parallel

weight = 0.45

x died in y .

weight = 0.15

x was the year y died .

weight = 0.40

x performed until his death in y .

\[ p(y \mid x) = \sum_{\text{prompt}} p(y \mid \text{prompt}(x)) \cdot p(\text{prompt}) \]
Main Experiments

• Predict factual relations from T-REx dataset by prompting BERT-large
  • About 1000 training examples per relation
Main Experiments

• Predict factual relations from T-REx dataset by prompting BERT-large
  • About 1000 training examples per relation
  • Initialize at other researchers' prompts -- huge improvement!

![Graph showing precision comparisons]

- Hard: 39.4
- Soft: 51
- Deep: 51.6

From Other Researchers'
Main Experiments

• Predict factual relations from T-REx dataset by prompting BERT-large
  • About 1000 training examples per relation
• Initialize at other researchers' prompts -- huge improvement!
• Initialize randomly -- almost as good!

Really close
Lots of Experiments

<table>
<thead>
<tr>
<th>LM</th>
<th>Method</th>
<th>Precision@1</th>
<th>Precision@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>init → soft → deep</td>
<td>init → soft → deep</td>
<td>init → soft → deep</td>
<td></td>
</tr>
<tr>
<td>BEb</td>
<td>LAMA</td>
<td>31.1</td>
<td>59.5</td>
<td>40.3</td>
</tr>
<tr>
<td></td>
<td>LPAQA</td>
<td>34.1</td>
<td>62.0</td>
<td>43.6</td>
</tr>
<tr>
<td></td>
<td>Soft (min.)</td>
<td>34.1</td>
<td>60.0</td>
<td>43.0</td>
</tr>
<tr>
<td></td>
<td>Soft (par.)</td>
<td>34.1</td>
<td>60.0</td>
<td>43.0</td>
</tr>
<tr>
<td></td>
<td>Soft (ran.)</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>BEI</td>
<td>LAMA</td>
<td>28.9†</td>
<td>57.7†</td>
<td>36.7†</td>
</tr>
<tr>
<td></td>
<td>LPAQA</td>
<td>39.4†</td>
<td>67.0†</td>
<td>40.1†</td>
</tr>
<tr>
<td></td>
<td>Soft (min.)</td>
<td>39.4†</td>
<td>67.0†</td>
<td>40.1†</td>
</tr>
<tr>
<td></td>
<td>Soft (par.)</td>
<td>39.4†</td>
<td>67.0†</td>
<td>40.1†</td>
</tr>
<tr>
<td></td>
<td>Soft (ran.)</td>
<td>2.3</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>Rob</td>
<td>LPAQA</td>
<td>1.2†</td>
<td>9.1†</td>
<td>4.2†</td>
</tr>
<tr>
<td></td>
<td>AutoPrompt</td>
<td>4.0</td>
<td>68.1</td>
<td>49.1</td>
</tr>
<tr>
<td></td>
<td>Soft (min.)</td>
<td>1.2</td>
<td>9.1†</td>
<td>4.2†</td>
</tr>
<tr>
<td></td>
<td>Soft (par.)</td>
<td>0.8</td>
<td>9.1†</td>
<td>4.2†</td>
</tr>
<tr>
<td></td>
<td>Soft (ran.)</td>
<td>0.8</td>
<td>9.1†</td>
<td>4.2†</td>
</tr>
<tr>
<td>BAb</td>
<td>LPAQA</td>
<td>0.8†</td>
<td>5.7†</td>
<td>2.9†</td>
</tr>
<tr>
<td></td>
<td>Soft (min.)</td>
<td>0.8</td>
<td>5.7†</td>
<td>2.9†</td>
</tr>
<tr>
<td></td>
<td>Soft (par.)</td>
<td>3.5</td>
<td>5.7†</td>
<td>2.9†</td>
</tr>
<tr>
<td></td>
<td>Soft (min.)</td>
<td>3.5</td>
<td>5.7†</td>
<td>2.9†</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LM</th>
<th>Method</th>
<th>Precision@1</th>
<th>Precision@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>init → soft → deep</td>
<td>init → soft → deep</td>
<td>init → soft → deep</td>
<td></td>
</tr>
<tr>
<td>BEb</td>
<td>LAMA</td>
<td>26.4</td>
<td>54.3</td>
<td>35.8</td>
</tr>
<tr>
<td></td>
<td>LPAQA</td>
<td>31.2</td>
<td>57.3</td>
<td>39.9</td>
</tr>
<tr>
<td></td>
<td>Soft (min.)</td>
<td>31.2</td>
<td>57.3</td>
<td>39.9</td>
</tr>
<tr>
<td></td>
<td>Soft (par.)</td>
<td>31.2</td>
<td>57.3</td>
<td>39.9</td>
</tr>
<tr>
<td></td>
<td>Soft (ran.)</td>
<td>0.8</td>
<td>4.0</td>
<td>2.2</td>
</tr>
<tr>
<td>BEI</td>
<td>LAMA</td>
<td>24.0†</td>
<td>53.7†</td>
<td>34.1†</td>
</tr>
<tr>
<td></td>
<td>LPAQA</td>
<td>37.8†</td>
<td>64.4†</td>
<td>44.0†</td>
</tr>
<tr>
<td></td>
<td>Soft (min.)</td>
<td>37.8†</td>
<td>64.4†</td>
<td>44.0†</td>
</tr>
<tr>
<td></td>
<td>Soft (par.)</td>
<td>37.8†</td>
<td>64.4†</td>
<td>44.0†</td>
</tr>
<tr>
<td></td>
<td>Soft (ran.)</td>
<td>1.4</td>
<td>5.4</td>
<td>5.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>P@1</th>
<th>P@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAMA</td>
<td>9.7†</td>
<td>27.0†</td>
<td>15.6†</td>
</tr>
<tr>
<td>LPAQA</td>
<td>10.6†</td>
<td>23.7†</td>
<td>15.3†</td>
</tr>
<tr>
<td>Soft (min.)</td>
<td>11.2 (+1.5)</td>
<td>33.5 (+ 6.5)</td>
<td>18.9 (+3.3)</td>
</tr>
<tr>
<td>Soft (par.)</td>
<td>12.9 (+2.3)</td>
<td>34.7 (+11.0)</td>
<td>20.3 (+5.0)</td>
</tr>
<tr>
<td>Soft (ran.)</td>
<td>11.5 (+0.9)</td>
<td>31.4 (+ 7.7)</td>
<td>18.3 (+3.0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>P@1</th>
<th>P@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPAQA (BEb)</td>
<td>18.9</td>
<td>40.4</td>
<td>26.6</td>
</tr>
<tr>
<td>Soft (BEb)</td>
<td>23.0 (+4.1)</td>
<td>45.2 (+4.8)</td>
<td>30.5 (+3.9)</td>
</tr>
<tr>
<td>LPAQA (BEI)</td>
<td>23.8</td>
<td>47.7</td>
<td>32.2</td>
</tr>
<tr>
<td>Soft (BEI)</td>
<td>27.0 (+3.2)</td>
<td>51.7 (+4.0)</td>
<td>35.4 (+3.2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>P@1</th>
<th>P@10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>39.4</td>
<td>67.4</td>
<td>49.1</td>
</tr>
<tr>
<td>adjust mixture weights</td>
<td>40.0</td>
<td>69.1</td>
<td>53.3</td>
</tr>
<tr>
<td>adjust token vectors</td>
<td>50.7</td>
<td>80.7</td>
<td>61.1</td>
</tr>
<tr>
<td>adjust both</td>
<td>51.0</td>
<td>81.4</td>
<td>61.6</td>
</tr>
</tbody>
</table>
Related Work


• Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. GPT understands, too. *arXiv* (2021).
Takeaways

• LMs know more facts than we thought. You just have to learn how to ask.
Takeaways

- LMs know more facts than we thought. You just have to learn how to ask.
- Prompts are made of vectors, not words! So you can tune them with backprop.
Takeaways

• LMs know more facts than we thought. You just have to learn how to ask.

• Prompts are made of vectors, not words! So you can tune them with backprop.

• Random initialization works fine. No grad student required.
Takeaways

• LMs know more facts than we thought. You just have to learn how to ask.

• Prompts are made of vectors, not words! So you can tune them with backprop.

• Random initialization works fine. No grad student required.

• Prompt tuning is lightweight, and could also be applied to few-shot learning.
Thanks!