Problem

- How to update individual model beliefs?

  - e.g. model erred, world changed

Background

- SERAC \([1]\) edits model using external memory, scope classifier, and counterfactual model
- SERAC decouples the base and counterfactual models
- So it can’t take advantage of the size of the base model

Method

- Replace SERAC’s counterfactual model with a **prompter model** which predicts sequences of probability distributions over tokens
- These form **continuous prompts** which are passed to the base model
- We use base model’s output, so **performance should scale with size of the base model**

Experiments

<table>
<thead>
<tr>
<th>Task — Model</th>
<th>Metric</th>
<th>SERAC</th>
<th>Prompter</th>
<th>RP</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC — bert-base</td>
<td>Edit Success (\dfrac{\text{true}}{\text{total}})</td>
<td>0.857</td>
<td>0.847</td>
<td>0.528</td>
</tr>
<tr>
<td></td>
<td>Drawdown (\frac{\text{true}}{\text{total}})</td>
<td>0.087</td>
<td>0.075</td>
<td>0.015</td>
</tr>
<tr>
<td>QA-easy — T5-large</td>
<td>Edit Success (\dfrac{\text{true}}{\text{total}})</td>
<td>0.986</td>
<td>0.961</td>
<td>0.487</td>
</tr>
<tr>
<td></td>
<td>Drawdown (\frac{\text{true}}{\text{total}})</td>
<td>0.009</td>
<td>0.009</td>
<td>0.03</td>
</tr>
<tr>
<td>QA-hard — T5-large</td>
<td>Edit Success (\dfrac{\text{true}}{\text{total}})</td>
<td>0.913</td>
<td>0.841</td>
<td>0.278</td>
</tr>
<tr>
<td></td>
<td>Drawdown (\frac{\text{true}}{\text{total}})</td>
<td>0.028</td>
<td>0.039</td>
<td>0.027</td>
</tr>
</tbody>
</table>

- FC = fact-checking, QA = question-answering
- Prompter **performs comparably** to SERAC without training model used to generate outputs

Conclusions

- Prompter model transforms retrieved context + query into a form which can **reliably modulate** the base model’s output
- Performance rivals that of SERAC while still using base model outputs
- Non-trivial generalization to new base models. Can we further improve generalization?
- Can this idea be applied to the **general retrieval-based model setting** to improve reliability/robustness?

Analysis

**Success cases**

- **input**: What team is Julien Sprunger associated with?
  - **prompt**: What team is Julien Sprunger associated with or Minnesota North Stars which team is Julien Sprunger associated with or which predicted Minnesota North Stars label: Minnesota North Stars
  - **predicted**: Napoleonic Wars
  - **predicted**: Napoleonic Wars
  - **predicted**: Napoleonic Wars
  - **predicted**: Napoleonic Wars
  - **predicted**: Napoleonic Wars

**Failure cases**

- **input**: Where did Lothar Friedrich von Metternich-Burscheid die?
  - **prompt**: Frage: Where did Lothar Friedrich von Metternich-Burscheid die oder Vienna Which did Lothar Friedrich von Metternich-Burscheid die or Vienna label: Vienna

References

\([1]\) Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale, 2022.