Fine-tuned LLMs
Know More, Hallucinate Less with Few-Shot Sequence to Sequence Semantic Parsing over Wikidata

Silei Xu*, Shicheng Liu*, Theo Culhane, Elizaveta Pertseva, Meng-Hsi Wu¹, Sina J. Semnani, Monica S. Lam

Stanford University

¹Ailly.ai
* Equal contribution
Problem: LLMs hallucinate

• LLMs can answer open-domain questions without access to external knowledge

• **BUT**: LLMs are known for giving the wrong answer with confidence

• This may cause significant harm as people increasingly accept LLMs as a knowledge source
Wikidata: Largest, Live Knowledge Graph

- 12B facts, 100M entities, 10K properties, 25K contributors
- Every Wikipedia article has a corresponding entity in Wikidata
- **Dataset for research in life sciences, digital humanity, etc.**

- Representation: triples
- Query with SPARQL

A natural language interface can greatly expand access.
New Dataset: WikiWebQuestions

Freebase
- Shutdown in 2015
- Fixed Schema
- Many KBQA Datasets
  - WebQuestionsSP*

Wikidata
- Actively maintained
- Evolving schema
- Limited SPARQL-annotated KBQA Datasets
  - WikiWebQuestions

NEW
- Train: 2431
- Dev: 454
- Test: 1431

GPT-3 on New WikiWebQuestions Dataset

- Adapted from WebQuestionsSP for FreeBase
- Questions from Google Suggest API
- Real-world popular questions
- GPT-3: trained on Wikipedia + Internet

Question: “What does Obama have a degree in?”
GPT-3: “Political science degree”
Missing: “Law degree”
Where did Bronx take place?

Entity Linker

('A Bronx Tale', 'Q1130705')

WikiSP (Semantic Parser)

```
SELECT DISTINCT ?x WHERE {
  wd:Q1130705 wdt:filming_location ?x .
}
```

From Wikidata, the filming location of 'A Bronx Tale' includes New Jersey and New York

GPT-3 guesses that the movie took place in Bronx, New York

No Response
KBQA Related Work

• **Multi-staged search problem**
  • (Yih et al., 2015, 2016; Luo et al., 2018; Lan and Jiang, 2020)

• **Seq2seq semantic parsing**
  • (Das et al., 2021; Ye et al., 2022; Cao et al., 2022b; Gu and Su, 2022; Shu et al., 2022, Yu et al., 2023)

  These works focus on Freebase
  Unlike FreeBase, Wikidata does not have a fixed schema
KBQA Related Work

- **(Sub)-Graph retrieval**
  - (Dong et al., 2015; Miller et al., 2016; Sun et al., 2018, 2019; Mavromatis and Karypis, 2022; Sen et al., 2021; Vivona and Hassani, 2019; Verga et al., 2021, Yu et al., 2023)
  - Cannot handle questions like “the tallest mountain” where no entities are mentioned by name
  - Poor interpretability
  - Do not support query optimization
Semantic Parsing for Wikidata

- Insight 1: incorporate **Entity Linkers** into the pipeline

Where did Bronx take place?

Entity Linker

(‘A Bronx Tale’, ‘Q1130705’)

- We use SOTA linker ReFinED, finetuned on WikiWebQuestions

• Insight 2: substituting SPARQL IDs with **property names** and domain entity names

Input: *Where was Anne Hathaway born?*
Entity Linker: (Anne Hathaway, Q36301)

```
SELECT ?x WHERE {
  wd:Q36301 wdt:P19 ?x }
```

(training data)

```
SELECT ?x WHERE {
  wd:Q36301 wdt:place_of_birth ?x }
```

(inference time)
• Insight 2: substituting SPARQL IDs with property names and **domain entity names**

**Input:** What *car model* does General Motors make?

Entity Linker: (General Motors, Q81965)

SELECT ?x WHERE {
  ?x wdt:P176 wd:Q81965
}

**(inference time)**

SELECT ?x WHERE {
  ?x wdt:manufacturer wd:Q81965
}

**(training data)**
• Insight 3: recover **missing entities**

**Input:** What year did Giants win the **world series**?

**Entity Linker:** (SF Giants, Q308966)

```
SELECT ?x WHERE {
}
```

```
SELECT DISTINCT ?x WHERE {
  ?y wdt:sports_season:_competition wd:world_series.
  ?y wdt:point_in_time ?x
}
```

*STANFORD*
Implementation

• Fine-tune LLaMA-7B

• Included Alpaca training data, derived from self-instruct

• Up-sampled WikiWebQuestion training set 5 times


Evaluation on WikiWebQuestions (dev)

GPT-3 Only

- 66% Correct GPT
- 27% Incomplete GPT
- 7% Incorrect GPT

WikiSP + GPT-3

- 76% Verified from WikiSP
- 15% Correct GPT
- 6% Incomplete GPT
- 4% Incorrect GPT
# Ablation Experiments

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>WikiSP (ours)</td>
<td>75.6</td>
<td>76.9</td>
</tr>
<tr>
<td>No Entity Linking</td>
<td>66.5</td>
<td>67.6</td>
</tr>
<tr>
<td>No mentions, trained with ReFinED</td>
<td>73.3</td>
<td>75.0</td>
</tr>
<tr>
<td>No mentions, trained with Oracle entities</td>
<td>72.2</td>
<td>73.4</td>
</tr>
<tr>
<td>PIDs and QIDs for properties &amp; domains</td>
<td>73.6</td>
<td>74.7</td>
</tr>
</tbody>
</table>

Table 2: Ablation results of WikiSP on the WWQ dev set.
Applied to QALD-7

Part of the QALD (Question Answering over Linked Data) challenges
A manually crafted dataset with complex questions

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAGG (Yih et al., 2016)</td>
<td>-</td>
<td>19.0</td>
</tr>
<tr>
<td>GGNN (Sorokin and Gurevych, 2018)</td>
<td>-</td>
<td>21.3</td>
</tr>
<tr>
<td>WDAqua (Diefenbach et al., 2017)</td>
<td>-</td>
<td>40.0</td>
</tr>
<tr>
<td>WikiSP (Ours)</td>
<td>38.0</td>
<td>43.6</td>
</tr>
</tbody>
</table>

Table 3: Evaluation results of WikiSP on QALD-7 Task 4 and comparison with prior work.
Conclusion

- High-quality benchmark WikiWebQuestions
- On Wikidata
- Annotated with SPARQL
- A first, strong baseline of 65% answer accuracy and 72% F1 score for WikiWebQuestions. Achieved with:
  - Fine-tuned LLaMA-7B
  - Modified SPARQL query format
- We can reduce the hallucination of large language models like GPT-3 by grounding it with a semantic parser

Code, data, and model are available at https://github.com/stanford-oval/wikidata-emnlp23