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 Wu[1], Sina J. Semnani, Monica S. Lam

Stanford University

[1]Ailly.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

Who founded Stanford?

SELECT ?x WHERE

{ wd:Q41506 wdt:P112 ?x. }

Stanford Founded by

A natural language interface can greatly expand access

New Dataset: WikiWebQuestions

Freebase

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

Wikidata

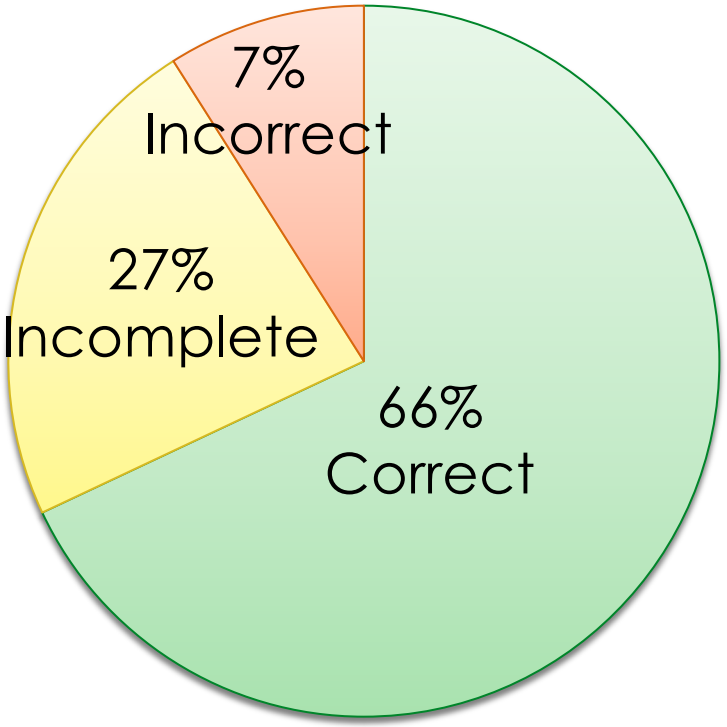
- Actively maintained
- Evolving schema
- Limited SPARQL-annotated KBQA Datasets
  - WikiWebQuestions

NEW

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

* Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1321–1331, Beijing, China. Association for Computational Linguistics

GPT-3 on New WikiWebQuestions Dataset

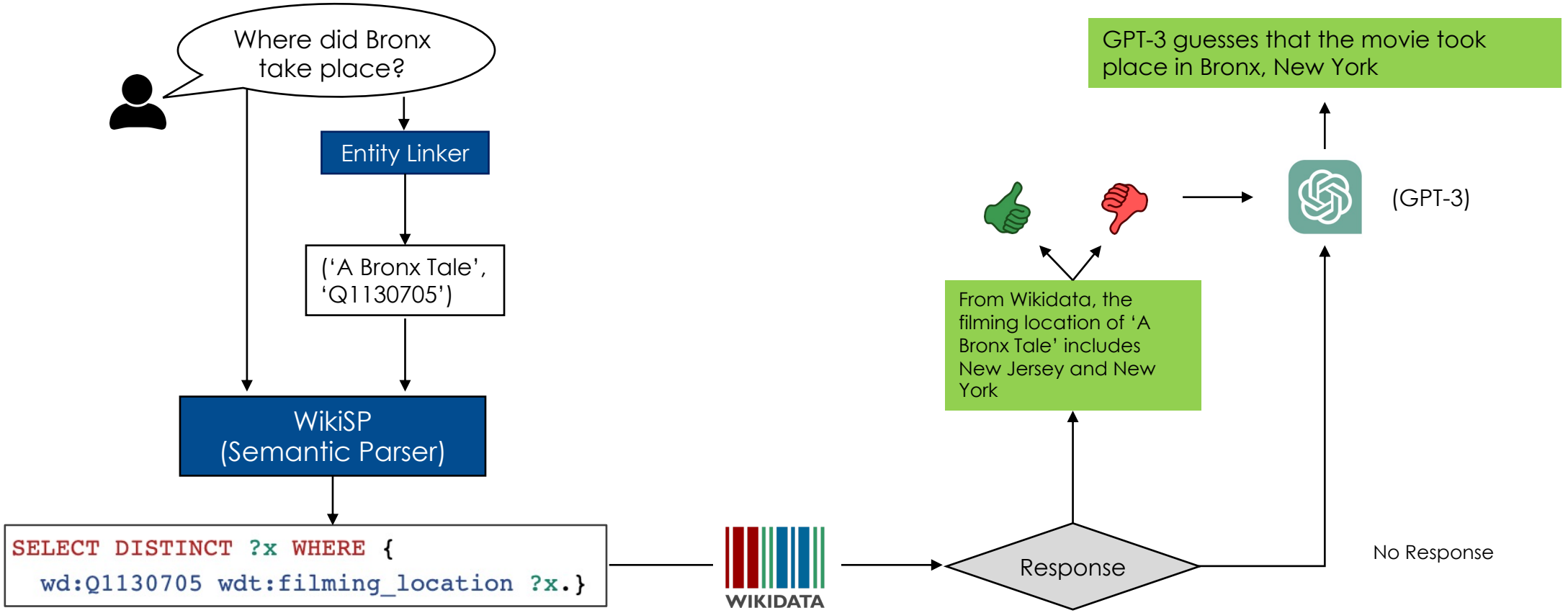


GPT-3
Guesses

- 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"

WikiSP: Semantic Parsing over Wikidata



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

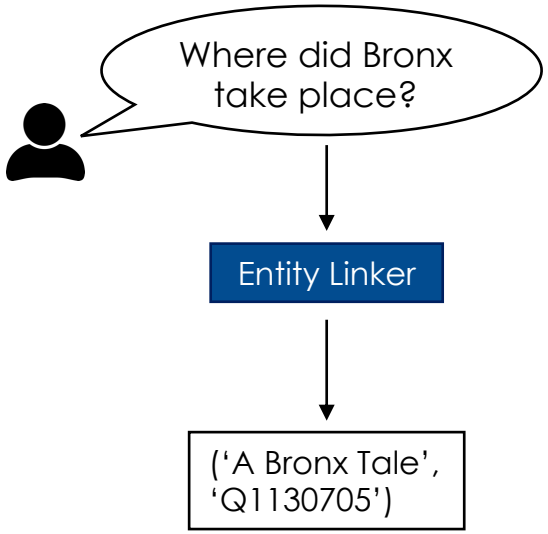
KBQA Related Work

- **(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



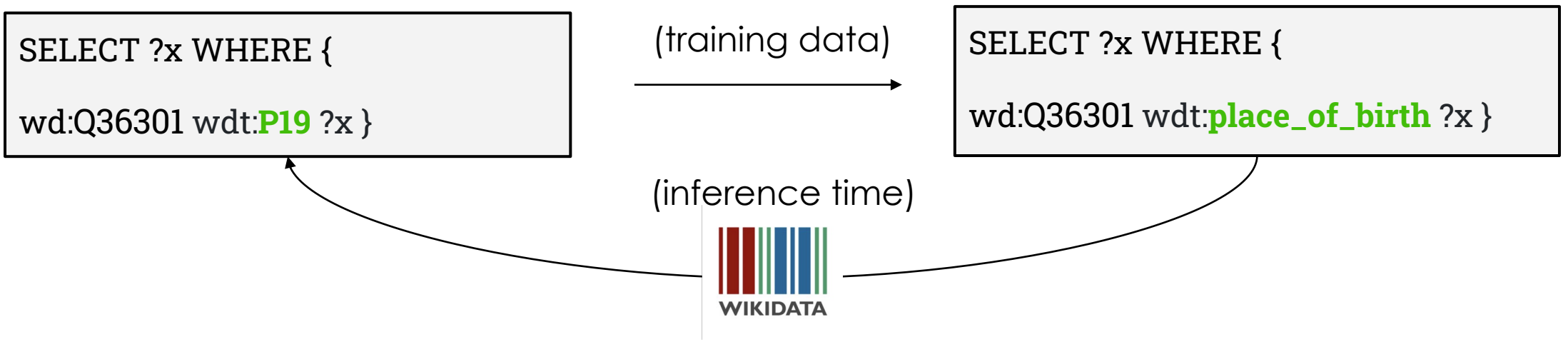
- We use SOTA linker ReFinED, finetuned on WikiWebQuestions

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, pages 209–220, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

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

Input: **Where** was Anne Hathaway **born**?

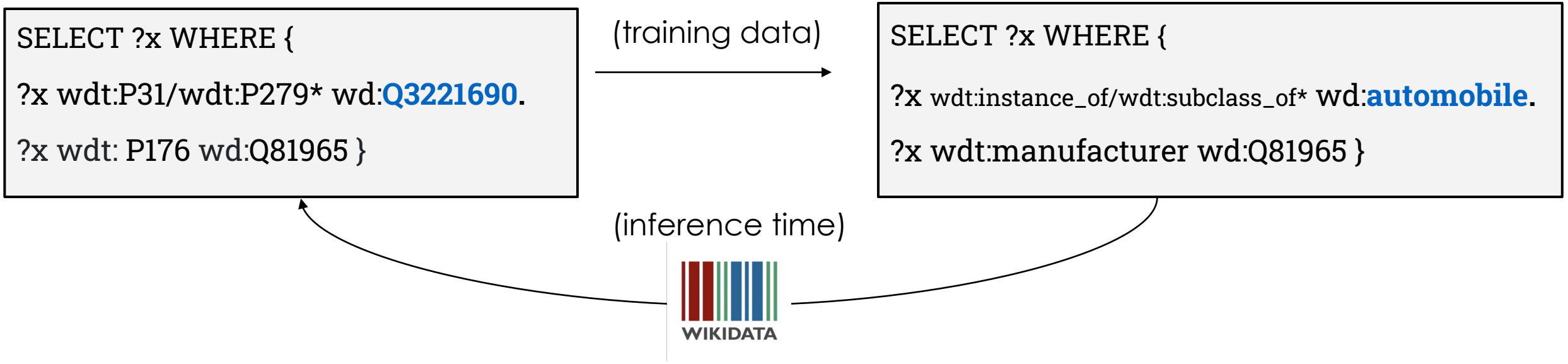
Entity Linker: (Anne Hathaway, Q36301)



- 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)



- Insight 3: recover **missing entities**

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

Entity Linker: (SF Giants, Q308966)

```
SELECT ?x WHERE {
?y wdt:P3450 wd:Q265538.
?y wdt:P1346 wd:Q308966.
?y wdt:P585 ?x. }
```

(training data) →

```
SELECT DISTINCT ?x WHERE {
?y wdt:sports_season..._competition wd:world_series.
?y wdt:winner wd:Q308966.
?y wdt:point_in_time ?x }
```

(inference time)

Implementation

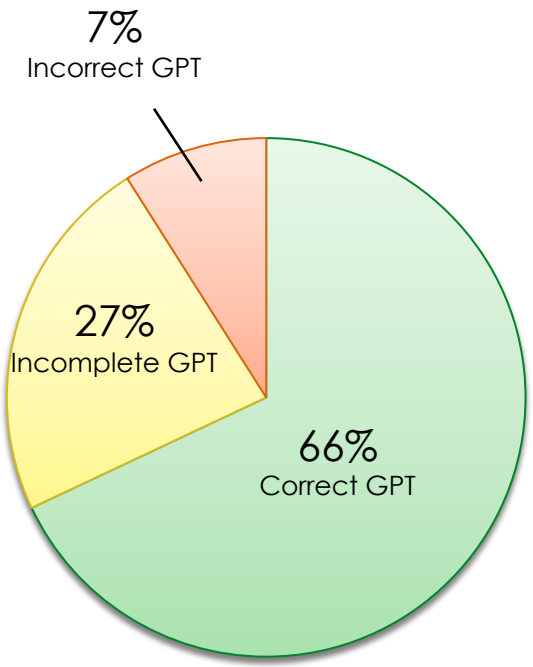
- Fine-tune LLaMA-7B
  - Included Alpaca training data, derived from self-instruct
  - Up-sampled WikiWebQuestion training set 5 times

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

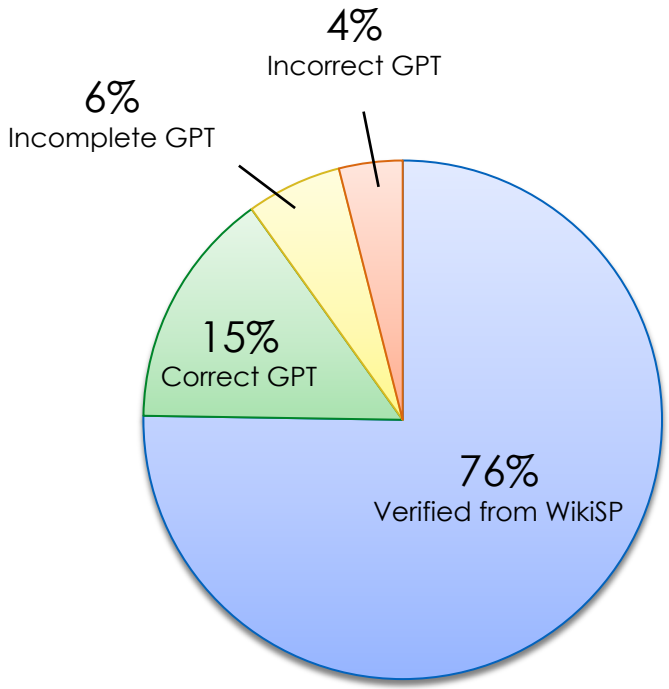
Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/ stanford_alpaca.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Evaluation on WikiWebQuestions (dev)



GPT-3 Only



WikiSP + GPT-3

Ablation Experiments

	EM	F1
WikiSP (ours)	**75.6**	**76.9**
No Entity Linking	66.5	67.6
No mentions, trained with ReFinED	73.3	75.0
No mentions, trained with Oracle entities	72.2	73.4
PIDs and QIDs for properties & domains	73.6	74.7

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

	EM	F1
STAGG (Yih et al., 2016)	-	19.0
GGNN (Sorokin and Gurevych, 2018)	-	21.3
WDAqua (Diefenbach et al., 2017)	-	40.0
WikiSP (Ours)	**38.0**	**43.6**

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

Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Bastian Haarmann, Anastasia Krithara, Michael Röder, and Giulio Napolitano. 2017. 7th open challenge on question answering over linked data (qald-7). In Semantic web evaluation challenge, pages 59–69. Springer.

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