SPAGHETTI: Open-Domain Question Answering from Heterogeneous Data Sources with Retrieval and Semantic Parsing

Heidi C. Zhang  Sina J. Semnani  Farhad Ghassemi  Jialiang Xu
Shicheng Liu  Monica S. Lam
Computer Science Department, Stanford University
Stanford, CA
{chenyuz, sinaj, farhadg, xjl, shicheng, lam}@cs.stanford.edu

Abstract

We introduce SPAGHETTI: Semantic Parsing Augmented Generation for Hybrid English information from Text Tables and Infoboxes, a hybrid question-answering (QA) pipeline that utilizes information from heterogeneous knowledge sources, including knowledge base, text, tables, and infoboxes. Our LLM-augmented approach achieves state-of-the-art performance on the COMPMix dataset, the most comprehensive heterogeneous open-domain QA dataset, with 56.5% exact match (EM) rate. More importantly, manual analysis on a sample of the dataset suggests that SPAGHETTI is more than 90% accurate, indicating that EM is no longer suitable for assessing the capabilities of QA systems today.1

1 Introduction

Open-domain question answering (QA) grounded in knowledge corpora has long been an active topic of research in natural language processing (Chen et al., 2017; Wang et al., 2018; Lee et al., 2019; Asai et al., 2020; Izacard and Grave, 2021; Khattab et al., 2021; Asai et al., 2022). With the rise of LLMs, new state of the art has been established with QA separately on free-text documents (Semnani et al., 2023; Jiang et al., 2023; Gao et al., 2023; Khattab et al., 2023), databases (Pourreza and Rafieii, 2023; Nan et al., 2023; Zhang et al., 2023), and graph databases (Xu et al., 2023; Luo et al., 2023; Li et al., 2023).

In practice, we need to fully leverage hybrid data sources. For instance, Wikipedia alone offers a wealth of knowledge through nearly 7M free-text articles; many of these articles contain structured information in tables and infoboxes; Wikidata is a knowledge graph containing over 17 billion triples. This paper investigates how to leverage LLMs to answer questions on all the different types of data.

1Code is available at https://github.com/standford-oval/WikiChat.
2 Related Work

TextQA, TableQA, and KBQA have all been individually studied extensively (Zhao et al., 2023a; Lu et al., 2024; Pan et al., 2024, *inter alia*). However, the task of answering questions from two or more sources, known as heterogeneous QA, is under-studied. Some literature investigate two of the three sources, including those on closed domain (Miller et al., 2016; Chen et al., 2020; Pramanik et al., 2021; Liu et al., 2023; Lei et al., 2023) and open domain (Chen et al., 2021; Zhao et al., 2023b; Han and Gardent, 2023; Ma et al., 2022a, 2023), but very limited existing work experiments on all three.

**CONV MIX** (Christmann et al., 2022) collected the first conversational QA dataset that requires knowledge from all three heterogeneous sources. Crowdworkers were asked to pick an entity of their interest and find the answer from one of the Wiki sources - Wikidata, Wikipedia text, Wikipedia tables, or Wikipedia infoboxes. Christmann et al. (2023a) later collated the completed conversations to derive the CONVMIX dataset with 9410 self-contained question-answer pairs.

Oguz et al. (2022), Ma et al. (2022b), and Christmann et al. (2022) proposed pipelines to answer questions from all three sources, by linearizing all structured information and applying text retrieval methods. Christmann et al. (2023a), on the other hand, unifies all the sources by representing all relevant information in a knowledge graph and uses GNN message passing to find the answer. The former gives up the advantage of using formal query languages on structured data, which can support operations such as ranking and averaging. The latter gives up the advantage of the expressiveness and versatility of free-text knowledge representation. Concurrent with our work, Lehmann et al. (2024) adopts another view that breaks down QA solution processes as tool calls and thoughts. They propose a human-like approach that teaches LLMs to gather heterogeneous information by imitating how humans use retrieval tools, which requires human-annotated demonstrations.

3 **SPAGHETTI**

**SPAGHETTI** is a hybrid QA pipeline that takes advantage of both structured and unstructured information. We obtain evidence from heterogeneous sources in parallel, including structured knowledge bases, plain text, linearized tables / infoboxes, and LLM-generated claims that are verified, and gather those evidence to generate the final answer using a few-shot LLM (Fig. 1).

![Figure 2: An example with a failure case of ReFinED and our entity linking module correcting the failure.]()
LLM-generated response, achieving significantly higher factual accuracy than GPT-4. We adopt a similar approach when handling text.

We first extract Wikipedia text using WikiExtractor. ColBERT (Santhanam et al., 2022) is used to retrieve Wikipedia passages that may answer a given query, and each of the top-k retrieved passages goes through a few-shot LLM summarizer.

As shown in the rightmost path of Figure 1, similar to WikiChat, SPAGHETTI also makes use of the internal factual knowledge of LLMs by first generating a response and then verifying the claims made in the response using retrieved information, retaining only grounded claims.

3.3 Tables and Infoboxes
Most NLP research using Wikipedia simply ignores the embedded tables and infoboxes, as extraction and preprocessing are challenging. With the help of tools such as WikiTextParser and regex matching, we programmatically extract 9 million tables and infoboxes from Wikipedia pages and linearize them so that they can be encoded as a set of ColBERT (Santhanam et al., 2022) index for retrieval. Being linearized, the retrieved item can then be read by LLMs directly.

Below is an example linearized table from the Wikipedia article “Arundhati Roy”:

<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Publisher</th>
<th>Year</th>
<th>ISBN</th>
</tr>
</thead>
</table>

For each table, we include the section title, two preceding sentences, and two succeeding sentences of the table as additional context, if there are any in the current section. Table rows are formatted as “column_name: cell_content,...” with “<tr>” as the row separator.

Since ColBERT is pretrained with textual passages and not tables, we finetune ColBERT for table retrieval. After retrieval, the retrieved table is then fed into a few-shot LLM to extract information directly relevant to the query.

3.4 Putting it Together
At the final stage, we gather and combine evidence from all sources. The answer from Wikidata is formatted as “Wikidata says the answer to <query> is: <answer>.” The retrieved text and tables/infoboxes each goes through an LLM summarization prompt, as mentioned earlier, attempting to extract relevant information from each retrieved item. The verified claim(s) from the LLM-generated answer (if any) is also added to the evidence pool.

Finally, all evidence is fed to a few-shot LLM prompt to generate a single answer to the query. In some cases the answer may be contained in more than one information source, and such redundancy can help reduce errors introduced in earlier stages of the pipeline.

4 Experiments
We evaluate SPAGHETTI on the CompMIX development and test sets, which contain 1680 and 2764 questions respectively.

For querying Wikidata, we use the LLaMA-7B semantic parser from Xu et al. (2023) trained on both WikiWebQuestions and QALD-7 (Usbeck et al., 2017). We use GPT-3.5 as the LLM in our entity linking module.

We experiment with LLaMA-7B, GPT-3.5-turbo-instruct, and GPT-4, respectively, as the LLM backbone in all the stages for handling retrieved evidences and for answer generation. We use few-shot prompts for GPT-3.5 and GPT-4, and use the LLaMA model from Semnani et al. (2023), which is distilled from the teacher GPT-4.

To fine-tune the ColBERT table retriever, we obtain training data from the NQ-Tables dataset (Herzig et al., 2021), where each example matches one gold table to a query. For each positive example, we sample 10 negative tables to obtain a total of 95K training triplets. We confirmed on the NQ-Tables dataset that the fine-tuned version improves table retrieval Recall@3 by 10%.

Evaluation Metrics. Bulian et al. (2022) and Kamalloo et al. (2023) have established that exact match (EM) against gold answers, which is commonly used for evaluating QA systems, cannot evaluate generative models properly as they often generate lexically different, but semantically equivalent answers. To properly assess our approach, we introduce two additional evaluation metrics: (1) Superset: whether the gold answer is a substring of the generated answer, as the latter tends to spell out the answer in long form and may include a more

---

1https://github.com/attardi/wikiextractor
2https://github.com/5j9/wikitextparser
4We access GPT models via the Microsoft Azure OpenAI API. We use the GPT-4 snapshot from June 13th, 2023.
<table>
<thead>
<tr>
<th>Model</th>
<th>EM dev</th>
<th>EM test</th>
<th>Superset dev</th>
<th>Superset test</th>
<th>GPT-4 Match dev</th>
<th>GPT-4 Match test</th>
<th>Platinum dev</th>
<th>Platinum test</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONVINSE (Christmann et al., 2022)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>UniK-QA (Oguz et al., 2022)</td>
<td>40.7%</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>EXPLAINGNN (Christmann et al., 2023b)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>GPT-3 (text-davinci-003)</td>
<td>–</td>
<td>50.2%</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>GPT-3.5 (turbo-instruct)</td>
<td>36.4%</td>
<td>36.1%</td>
<td>53.2%</td>
<td>54.2%</td>
<td>68.0%</td>
<td>69.9%</td>
<td>74%</td>
<td>–</td>
</tr>
<tr>
<td>GPT-4</td>
<td>53.0%</td>
<td>52.8%</td>
<td>60.9%</td>
<td>62.0%</td>
<td>76.7%</td>
<td>78.4%</td>
<td>81%</td>
<td>–</td>
</tr>
<tr>
<td>SPAGHETTI (LLaMA-7B)</td>
<td>53.8%</td>
<td>51.7%</td>
<td>61.7%</td>
<td>60.5%</td>
<td>69.8%</td>
<td>70.4%</td>
<td>77%</td>
<td>–</td>
</tr>
<tr>
<td>SPAGHETTI (GPT-3.5)</td>
<td>58.5%</td>
<td>55.6%</td>
<td>67.7%</td>
<td>65.6%</td>
<td>76.9%</td>
<td>75.3%</td>
<td>84%</td>
<td>–</td>
</tr>
<tr>
<td>SPAGHETTI (GPT-4)</td>
<td>57.3%</td>
<td>56.5%</td>
<td>70.2%</td>
<td>70.0%</td>
<td>80.8%</td>
<td>81.9%</td>
<td>92%</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 1: Main results on the COMPMIX development and test set. UniK-QA and GPT-3 (text-davinci-003) results are from Christmann et al. (2023a). We use the same zero-shot generation prompt published by Christmann et al. (2023a) to evaluate GPT-3.5 (turbo-instruct) and GPT-4.

*: Platinum results are obtained by an expert manually relabeling and evaluating the first 100 development set examples.

<table>
<thead>
<tr>
<th>Knowledge Source</th>
<th>EM</th>
<th>Superset</th>
<th>GPT-4 Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>53.8%</td>
<td>61.6%</td>
<td>71.1%</td>
</tr>
<tr>
<td>Tables</td>
<td>48.9%</td>
<td>59.5%</td>
<td>65.9%</td>
</tr>
<tr>
<td>KB</td>
<td>32.9%</td>
<td>40.4%</td>
<td>–</td>
</tr>
<tr>
<td>Text+Tables</td>
<td>55.6%</td>
<td>65.4%</td>
<td>74.5%</td>
</tr>
<tr>
<td>Text+Tables+KB</td>
<td>58.5%</td>
<td>67.7%</td>
<td>76.9%</td>
</tr>
</tbody>
</table>

Table 2: SPAGHETTI (GPT-3.5) ablation results on the COMPMIX development set, for using different knowledge sources. Results on “KB” are derived by directly comparing generated QID(s) against gold QID(s), while other methods are by string comparisons.

5 Results

SPAGHETTI (GPT-4) achieves 81.9% test set accuracy by GPT-4 matching, and 92% platinum accuracy on the 100 development set examples. Of the 8 errors cases, 3 have unanswerable questions (e.g. “FC Cincinnati soccer club?”), thus the true accuracy rate is 92/97 (94%).

Ablations. We evaluate the contribution of each knowledge source by ablating different parts of the system (Table 2). Using text alone already outperforms the previous SOTA, with each additional source further improving the result. Note that for many questions, information exists in multiple sources; the relatively little contribution from Wikidata and tables reflects mainly on the makeup of COMPMIX, not their value as knowledge sources. For detailed experimental results on our Wikidata entity linking approach, see Appendix A.

Human Evaluation We examine how our human “Platinum” evaluation (92%) differs from the EM metric (60%) on our sample of 100 cases. We report findings on the SPAGHETTI (GPT-4) responses, specifically in 32 annotated examples where EM fails but the response is indeed correct (Figure 3). Out of the 32 discrepancies, the unsophisticated “Superset” metric resolves 7, and GPT-4 matching resolves an additional 14. Platinum evaluation identifies that 4 questions have incorrect gold labels, and 7 questions are ambiguous and the generated answers are correct though different from the gold. We include examples for each of these resolved cases in Appendix B.

Of the 5 true errors, one is because SPAGHETTI cannot find the answer in any of the four information sources; in the other 4 cases, the answer generator cannot identify the correct answer re-
6 Conclusion

We propose SPAGHETTI, a hybrid open-domain question-answering system that combines semantic parsing and information retrieval to handle structured and unstructured data.

SPAGHETTI achieves an exact match rate improvement of 6.3% over the prior state-of-the-art on the COMPmIX dataset. More importantly, we show that our approach is likely to reach an accuracy of over 90%, if we account for differences in the answer wording and incompleteness/errors in gold labels. This, however, does not mean open-domain QA is solved. Further research is needed to handle open-domain questions that require complex structured queries or composition of answers from multiple information sources, none of which are included in COMPmIX.

Limitations

This work focuses specifically on open-domain QA with heterogeneous knowledge sources, and we only report results on the COMPmIX dataset due to the limited availability of high-quality datasets in this domain. A natural future work is to develop more diverse and advanced datasets that further push the need to utilize each knowledge source.

We evaluate on single-turn QA and do not work with conversations in this paper, and SPAGHETTI can be extended to handle fact-based conversational questions or even chitchat that involves facts.

We have a relatively small sample size for human evaluation, because the expert manually checks the correctness of each example with Internet searches, which is labor-intensive. However, we acknowledge that a larger sample size would increase the statistical confidence of our evaluation.

Finally, we note that a number of Wikipedia tables are not well-formatted after preprocessing and linearization. Since Wikipedia tables are embedded as HTML elements that allow for idiosyncrasies like a table with one cell spanning multiple columns or color-highlighted cells, some are hard to parse correctly. Solving such edge cases engineering-wise would further improve TableQA.

Ethical Considerations

To facilitate reproducibility and continued research, we will make the code available upon publication.

No new datasets were gathered specifically for this study, and we did not employ crowd-sourced labor. We use Wikipedia data under the terms of the Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA) and the GNU Free Documentation License (GFDL). Wikidata is under Creative Commons CC0 License, which is equivalent to public domain. The COMPmIX benchmark is licensed under a Creative Commons Attribution 4.0 International License. We use the benchmark as it is intended.

The experimental phase involved approximately 80 hours of computation time on an NVIDIA A100 GPU to fine-tune the retrieval model and index Wikipedia content. We reused the LLaMA-7B model trained in prior work, thus avoiding extra GPU usage.

We do not anticipate adverse effects stemming from the proposed methods in this study.

References


Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An efficient zero-shot-capable approach to


Shicheng Liu, Jialiang Xu, Wesley Tjiangnaka, Sina J. Semnani, Chen Jie Yu, Gui Dávid, and Monica S. Lam. 2023. SUQL: Conversational search over structured and unstructured data with large language models.

Weizheng Lu, Jiaming Zhang, Jing Zhang, and Yueguo Chen. 2024. Large language model for table processing: A survey.


Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiaapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering; page 1–20.


Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zicen Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruixiang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023a. A survey of large language models.

Wenting Zhao, Ye Liu, Tong Niu, Yao Wan, Philip S. Yu, Shafiq Joty, Yingbo Zhou, and Semih Yavuz. 2023b. Divknowqa: Assessing the reasoning ability of Ilms via open-domain question answering over knowledge base and text.
A Wikidata Experiments

Xu et al. (2023) fine-tuned two LLaMAs on Wikidata. The training data for the first model consists solely of WikiWebQuestions (Xu et al., 2023), while the other consists of the combination of WikiWebQuestions (Xu et al., 2023) and QALD-7 (Usbeck et al., 2017). We experiment with both models on the development set of COMPMIX, each with (1) entities predicted by ReFinED, (2) our entity linking approach with GPT-3.5 as the LLM (prompt in Figure 6), and (3) the dataset-provided oracle entities.

As shown in Table 3, the model using entities predicted by our approach outperforms the model using the baseline ReFinED entities. It achieves considerably closer performance with the model using oracle entities. We also observed that the the model trained on both WikiWebQuestions and QALD-7 outperforms the model trained on WikiWebQuestions only.

B Details on Platinum Evaluation

Figure 3 shows the distribution of cases that we resolve using more advanced evaluation metrics. Numbers are reported on the first 100 dev examples with SPAGHETTI (GPT-4).

Examples of each evaluation error type can be found at Figure 12, Figure 13, Figure 14, Figure 15, and Figure 16.

C Error Analysis

We include the five error cases after platinum evaluation in Figure 7, Figure 8, Figure 9, Figure 10, and Figure 11.

C.1 Conflicting or Misleading Evidence

We analyze the 388 error cases from SPAGHETTI (GPT-3.5) as determined by GPT-4 Matching. We separate out evidence retrieval errors from answer generation errors by identifying how often the gold answer appears in the evidences using a substring matching heuristic (Table 4).

In 154 out of all 388 error cases, the system does not produce the gold answer despite the successful retrieval of evidence containing it. This observation indicates that a significant portion of the error cases are due to conflicting or misleading information in the evidence, where further improvements in selecting and merging evidences would be helpful. In the majority of the error cases (234 out of 388) where gold is not in the evidence, the system has no high-quality candidates to select from. Note, however, that this is an overestimate, due to the use of substring matching for deciding whether an evidence is correct or not.

In the breakdown of gold answer sources, the source that contains the most gold answers is Text (87 out of 154 cases), and Wikidata contains the least gold answers (51 out of 154 cases).

C.2 Combiner Hallucination

We investigate the ratio of generated answers that were hallucinated by our model. We manually checked the first 300 cases in our evaluation set and found 2 cases (0.67%) where the model ignored the evidence and hallucinated an incorrect answer. This low ratio of hallucination highlights the faithfulness of our system to the evidence retrieved. We include these cases in Figure 4 and Figure 5.
### Table 3: Wikidata semantic parsing experiment results on the CoMPiX development set. Comparison is made using entity IDs. Superset measures whether the model’s predicted entities is a superset of the gold entities. **Dev (KB subset)** refers to the subset of the dataset where the annotators located the annotated answer from Wikidata.

<table>
<thead>
<tr>
<th></th>
<th>Dev (EM)</th>
<th>Dev (Superset)</th>
<th>Dev (KB subset)</th>
<th>Dev (KB subset)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WikiWebQuestions Only</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/ ReFinED only entities</td>
<td>29.1</td>
<td>36.5</td>
<td>39.8</td>
<td>46.0</td>
</tr>
<tr>
<td>w/ ReFinED + GPT-3.5 entities</td>
<td>31.3</td>
<td>38.8</td>
<td>43.8</td>
<td>50.4</td>
</tr>
<tr>
<td>w/ oracle entities</td>
<td>33.9</td>
<td>42.3</td>
<td>46.2</td>
<td>52.8</td>
</tr>
<tr>
<td><strong>WikiWebQuestions + Qald-7</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/ ReFinED only entities</td>
<td>29.5</td>
<td>36.6</td>
<td>41.0</td>
<td>47.6</td>
</tr>
<tr>
<td>w/ ReFinED + GPT-3.5 entities</td>
<td>32.9</td>
<td>40.4</td>
<td>46.8</td>
<td>53.0</td>
</tr>
<tr>
<td>w/ oracle entities</td>
<td>35.5</td>
<td>43.1</td>
<td>49.0</td>
<td>55.4</td>
</tr>
</tbody>
</table>

Table 4: Numbers of error cases by category. The notation “Gold in [source]” stands for the gold answer existing as a substring in the particular [source].

<table>
<thead>
<tr>
<th># Error Cases</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>All Error Cases</td>
<td>388 (100%)</td>
</tr>
<tr>
<td>Gold in Evidence</td>
<td>154 (39.69%)</td>
</tr>
<tr>
<td>Gold in KB</td>
<td>51 (13.14%)</td>
</tr>
<tr>
<td>Gold in Text</td>
<td>87 (22.42%)</td>
</tr>
<tr>
<td>Gold in Tables</td>
<td>72 (18.56%)</td>
</tr>
</tbody>
</table>
Figure 4: Example of a refinement hallucination case (SPAGHETTI (GPT-3.5)). “Dave Grohl” is completely hallucinated.
Figure 5: Example of a refinement hallucination case

```json
{
    "idx": 249,
    "correct_hallucination": false,
    "question": "What is the voice type of the Bob Dylan?",
    "gold": "baritone",
    "answer_generated": "gravelly or nasal",
    "gold_sources": ["KB"],
    "pred_sources": [],
    "evidences": [
        ["KB",
         "Wikidata says the answer to " What is the voice type of the Bob Dylan?" is: baritone."
        ],
        ["TEXT",
         "Bob Dylan’s voice has been described as " young and jeeringly cynical" and "broken" as he aged."
        ],
        ["TEXT",
         "Bob Dylan’s voice has received critical attention, with some describing it as "a rusty voice" and others comparing it to "sand and glue".
        ]
    ]
}
```
Figure 6: A shortened version of the prompt for GPT-3.5 to detect entity mentions and generate a description for each detected entity, as discussed in Section 3.1. The descriptions in the prompt are taken from the Wikidata description for detected entities. The actual prompt contains 13 more examples. The examples in the prompt are chosen to capture the diversity of domains and to instruct GPT-3.5 to detect more generic entities too.
Figure 7: A failure case after platinum evaluation. In this case, the intent of the question is ambiguous. The gold answer is Vivien Leigh, who won the 24th Academy Awards (held on March 20, 1952, honoring the films of 1951), and SPAGHETTI predicts Shirley Booth, who won The 25th Academy Awards (held on March 19, 1953, honoring the films of 1952).

Figure 8: A failure case after platinum evaluation. In this case, the Wikidata is giving an incorrect answer, due to semantic parsing errors.
Figure 9: A failure case after platinum evaluation. The first book of Francisco de Robles is *Don Quixote* released in 1605. The book *La Galatea*, released in 1585, was published by Blas de Robles, father of Francisco de Robles. However, the entry of Blas de Robles, listed as the publisher in the Wikipedia infobox of *La Galatea*, erroneously contains a hyperlink directing to the page of Francisco de Robles. This discrepancy led to a misinterpretation by SPAGHETTI, resulting in the incorrect identification of Francisco de Robles as the publisher of *La Galatea*.

Figure 10: A failure case after platinum evaluation. The player name, Bebé, was not correctly grounded in Wikidata, which resulted in an empty response. Table retriever retrieved information about a basketball player named Bebo instead of the football player Bebé.

Figure 11: A failure case after platinum evaluation due to no retrieved evidence.
Figure 12: Example where EM cannot handle correctly (format).

```json
{
  "question": "What is the original title of the novel The Alchemist?",
  "gold": "O Alquimista",
  "answer_generated": "O Alquimista"
}
```

Figure 13: Example where EM cannot handle correctly (superset).

```json
{
  "question": "Nirvana was founded by who?",
  "gold": "Kurt Cobain",
  "answer_generated": "Kurt Cobain and Krist Novoselic"
}
```

Figure 14: Example where EM cannot handle correctly (gold answer wrong). “Empty Sky” is the correct answer here.

```json
{
  "question": "What was Elton John’s debut album?",
  "gold": "Goodbye Yellow Brick Road",
  "answer_generated": "Empty Sky"
}
```

Figure 15: Example where EM cannot handle correctly (multiple correct answers).

```json
{
  "question": "What is the main cast name in the tv series Tribes of Europa?",
  "gold": "Emilio Sakraya",
  "answer_generated": "Henriette Confurius, Emilio Sakraya, and David Ali Rashed"
}
```

Figure 16: Example where EM cannot handle correctly (paraphrasing).

```json
{
  "question": "Who was the music of the movie "The Social Network"?",
  "gold": "Trent Reznor Atticus Ross",
  "answer_generated": "Trent Reznor and Atticus Ross"
}
```