Open-WikiTable: Dataset for Open Domain Question Answering with Complex Reasoning over Table

Sunjun Kweon[1], **Yeonsu Kwon**[1], **Seonhee Cho**[1], **Yohan Jo**[2]*, **Edward Choi**[1]
KAIST[1] , Amazon[2]
{sean0042, yeonsu.k, ehcho8564, edwardchoi}@kaist.ac.kr
{jyoha}@amazon.com

arXiv:2305.07288v1 [cs.CL] 12 May 2023

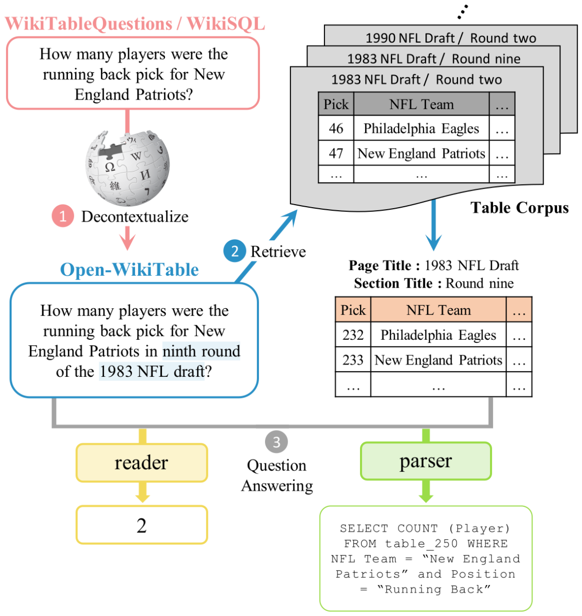
Abstract

Despite recent interest in open domain question answering (ODQA) over tables, many studies still rely on datasets that are not truly optimal for the task with respect to utilizing *structural nature of table*. These datasets assume answers reside as a single cell value and do not necessitate exploring over multiple cells such as aggregation, comparison, and sorting. Thus, we release Open-WikiTable, the first ODQA dataset that requires complex reasoning over tables. Open-WikiTable is built upon WikiSQL and WikiTableQuestions to be applicable in the open-domain setting. As each question is coupled with both textual answers and SQL queries, Open-WikiTable opens up a wide range of possibilities for future research, as both reader and parser methods can be applied. The dataset and code are publicly available[1].

1 Introduction

Tables have played a prominent role as a source of knowledge in question answering (QA). They contain various types of data such as numeric, temporal, and textual information in a structured manner. Early table QA datasets (Pasupat and Liang, 2015; Zhong et al., 2017; Yu et al., 2018) have focused on complex questions that exploit the structure of tables via aggregation, comparison, or sorting. However, these datasets assume that the relevant table is always given for each question (Kostić et al., 2021), limiting their applicability in real-world scenarios. For more practical use, recent works extend tableQA to the open-domain setting, where the evidence table should be retrieved solely from using the question itself.

The first research of open-domain QA over tables is Herzig et al. (2021). They released a dataset, NQ-table, by extracting questions from Natural Questions (Kwiatkowski et al., 2019) whose answers reside in a table. All questions, however, are answered by extracting a single cell and do not necessitate any extensive reasoning across multiple cells. It is also notable that 55% of the evidence tables consist of only a single row, which has little structure.


WikiTableQuestions / WikiSQL
How many players were the running back pick for New England Patriots?
1 Decontextualize
Open-WikiTable
How many players were the running back pick for New England Patriots in ninth round of the 1983 NFL draft?
2 Retrieve
1990 NFL Draft / Round two
1983 NFL Draft / Round nine
1983 NFL Draft / Round two
Pick | NFL Team | …
46 | Philadelphia Eagles | …
47 | New England Patriots | …
… | … | …
Table Corpus
Page Title : 1983 NFL Draft
Section Title : Round nine
Pick | NFL Team | …
232 | Philadelphia Eagles | …
233 | New England Patriots | …
… | … | …
reader
2
3 Question Answering
parser

```
SELECT COUNT (Player)
FROM table_250 WHERE
NFL Team = "New England
Patriots" and Position
= "Running Back"
```

Figure 1: Open-WikiTable is built by revising WikiSQL and WikiTableQuestions. Through decontextualization, the question provides the necessary information to retrieve the grounding table. As the dataset is labeled with textual answers and SQL queries, both reader and parser approaches can be used.

Another work for open-domain table QA is Pan et al. (2021). They presented E2E-WTQ and E2E-GNQ datasets, extensions of WikiTableQuestion (Pasupat and Liang, 2015) and GNQtables (Shraga et al., 2020), to develop cell-level table retrieval models. However, as they assume cell extraction for the table QA task and construct the datasets accordingly, E2E-WTQ and E2E-GNQ share the same limitation as the NQ-table; all answers are restricted to a single cell. Another issue with these

* This work is not associated with Amazon.

[1] https://github.com/sean0042/Open_WikiTable

Datasets	Retrieval Availability	Complex Reasoning	# of QA pairs	# of table corpus	Answer Annotation	Reference
WikiTableQuestions	✗	✓	22,033	2,108	Text	Pasupat and Liang (2015)
WikiSQL	✗	✓	80,654	24,241	SQL	Zhong et al. (2017)
E2E-WTQ	✓	✗	1,216	2,108	Text	Pan et al. (2021)
E2E-GNQ	✓	✗	789	74,224	Text	Pan et al. (2021)
NQ-table	✓	✗	11,628	169,898	Text	Herzig et al. (2021)
Open-WikiTable	✓	✓	67,023	24,680	Text, SQL	

Table 1: Comparison between table question answering datasets.

datasets is their small size, containing only around 1k examples in total. This makes it challenging to train language models as there may not be enough data.

Given that there is currently no dataset that fully considers the structural property in the open-domain setting, we present **Open-WikiTable**. It extends WikiSQL and WikiTableQuestions to be more applicable in the open-domain setting. Open-WikiTable is a large-scale dataset composed of 67,023 questions with a corpus of 24,680 tables. The key features of our dataset are listed below.

- First, nearly 40% of the questions require advanced reasoning skills beyond simple filtering and cell selection. The model should utilize operations such as aggregating, sorting, and comparing multiple cells to derive an accurate answer.

- Second, all questions are carefully designed for the retrieval task in the open-domain setting. We manually re-annotated 6,609 table descriptions (i.e. page title, section title, caption), then added them to the original question to ensure that questions convey sufficient context to specify the relevant table.

- Third, questions are paraphrased to reduce high word overlap between the question and the grounding table. It reflects a tendency in the open-domain setting where questions are often phrased in diverse styles, and terms in the questions may be different from those in the table.

- Lastly, every question in the dataset is labeled with both textual answers and SQL queries. This provides an opportunity to train and evaluate models with both common table QA techniques, *Reader* and *Parser* in parallel.

In this work, we thoroughly explain the data construction process of Open-WikiTable. We perform open domain question answering by incorporating a retrieval task with QA over tables (see Figure 1). Then, we evaluate the performance of the retriever and the QA models with both reader and parser approaches.

2 Data Construction

Open-WikiTable is built upon two closed-domain table QA datasets - WikiSQL (Zhong et al., 2017) and WikiTableQuestions (Pasupat and Liang, 2015). WikiSQL is a large-scale text-to-SQL dataset but is composed of relatively simple questions since they are constructed from limited SQL templates. WikiTableQuestions contains more complex questions involving superlative or arithmetic expressions but only provides textual form answers. Shi et al. (2020) further annotated SQLs for a subset of WikiTableQuestions. By utilizing these datasets, we aim to create a diverse and intricate set of questions, with each question annotated with both a textual and logical form answer.

Although the questions in WikiSQL and WikiTableQuestions require more advanced table reasoning than those of existing open-domain table QA datasets (See Table 1), they possess two problems to be directly used in the open domain setting. First, questions are not specific enough to retrieve relevant tables. Second, questions have high word overlaps with table contents which are unrealistic in the open-domain setting where the question can be expressed in lexically diverse forms. We resolve the first issue via decontextualization (2.1) and the second issue via paraphrasing (2.2), as elaborated in the following sections.

2.1 Decontextualization

Our goal is to decontextualize questions, that is, adding enough context about relevant tables to each question so that retrievers can find the relevant tables (Chen et al., 2020; Choi et al., 2021). However, the obstacle here is that a significant portion of table descriptions provided by WikiSQL and WikiTableQuestions were either missing or not specifically described to distinguish between tables. In this case, decontextualized questions still cannot point out the exact grounding tables (appendix A.1). Therefore, we resolved this issue by comparing 6,609 problematic tables with the corresponding Wikipedia article and re-annotating table descriptions. The resulting table corpus of Open-WikiTable has 24,680 tables, all of which have distinct descriptions.

Next, the questions were decontextualized with the re-annotated table descriptions. All table descriptions necessary for the retrieval of the grounding table were incorporated into each question. We transformed the questions by utilizing GPT-J, a language model from Eleuther AI. In order to ensure that the generated question accurately reflects the original intention, we decontextualized the questions by maintaining the form of the original question while incorporating table descriptions only as adverbs, as exemplified in Appendix A.2. The generated questions were accepted only if all key entities (i.e. referred column names and condition values) of the original question and added table descriptions were preserved. If not, we repeatedly generated new samples until accepted.

2.2 Paraphrase

Although the decontextualization process ensures the questions are suitable for table retrieval, it is quite easy to retrieve the grounding table due to a high degree of word overlap in the question and the table contents. To address this issue, we further paraphrased the questions via back-translation (Prabhumoye et al., 2018). We utilized English-German-English translation using Google Translate API. To inspect whether the degree of word overlap has decreased, we measure the average BLEU score between the question and grounding table contents. It has dropped after paraphrase, from 7.28×10^{-2} to 6.56×10^{-2}. It is also notable that the variance of word distribution in the questions has increased from 2.3×10^{5} to 3.1×10^{5} through paraphrasing.

2.3 Quality Check

We then review the questions to ensure their quality as the final step. Authors manually reviewed 10k randomly selected questions, according to the following standards: 1) The intent of the original question should not be altered during any stage of the data construction process. 2) Every information added through the decontextualization process should be preserved after paraphrasing. It turned out that 7.9% of 10k randomly selected samples did not meet our criteria. Within the 7.9% error rate, we discovered that 70% of these errors were due to the ambiguity of the original question. As a result, errors stemming from our decontextualization and paraphrasing processes account for 2.3% of the 10,000 random samples. The final test set, however, is composed only of the accepted samples during the quality review to ensure the integrity in the evaluation of the model performance. Error examples are reported in appendix A.3.

2.4 Data Statistics

As part of our dataset preparation process, we partitioned the entire dataset into train, validation, and test sets, with a ratio of 8:1:1. Consequently, the test set comprised 6,602 instances, as shown in Table 2. It is important to note that during this partitioning process, we ensured that each subset do not share any tables, enabling us to evaluate the generalizability of the models to previously unseen tables.

	Train	**Valid**	**Test**	**Total**
# of questions	53,819	6,602	6,602	67,023
# of tables	17,275	2,139	2,262	21,676
corpus size	24,680			

Table 2: Statistics of Open-WikiTable

3 Experiments

First of all, we split tables into segments so that models can handle long tables within the limited input sequence length. Inspired by Karpukhin et al. (2020), tables are split row-wise into 100-word chunks. Around 52% of tables in our corpus are split into multiple chunks, which resulted in a total of 54,282 table segments. For the retrieval task, each table is flattened and appended with the table descriptions, and then fed to a retriever. When a grounding table is split into multiple tables, all table segments that are relevant to an answer should be retrieved. Then, we perform end-to-end table QA where the model should answer the question given retrieved tables. More details about experimental settings are in Appendix B.

3.1 Retrieval

Experimental Setup We employ the BM25 algorithm (Robertson et al., 2009) for the sparse search.

Encoder		Data	k=5	k=10	k=20
Text	Table				
BM25		Original	6.6	8.0	10.3
		Decontextualized	45.5	52.9	59.7
		Paraphrased	42.2	48.9	56.1
BERT	BERT	Original	25.0	34.1	45.1
		Decontextualized	91.6	96.0	97.8
		Paraphrased	89.5	95.0	97.3
BERT	TAPAS	Original	19.4	28.1	38.5
		Decontextualized	88.2	94.5	97.3
		Paraphrased	84.0	91.4	95.6

Table 3: Top-k table retrieval accuracy on three different construction stages of Open-WikiTable's validation set.

For the dense search, we utilize a dual-encoder approach: BERT (Devlin et al., 2018) and TAPAS (Herzig et al., 2020) for the table encoder and BERT for the question encoder. They are trained to maximize the inner product between the question and table embeddings. The performance of the retriever is measured at different top-*k* retrieval accuracy, where we use 5, 10, and 20 for *k*. To analyze the effect of each data construction process on the retrieval task, we experiment with three different types of questions: original question, decontextualized question, and paraphrased question. The result is shown in Table 3.

Result Our experiments demonstrate that decontextualizing led to improved performance in all experiments. This suggests that the original questions are not sufficient for table retrieval and decontextualization dramatically alleviates this problem. However, the result also implies that table retrieval becomes too easy as the information is added directly to the question without any syntactic or semantic changes. This tendency is mitigated after paraphrasing, which led to a performance drop for all retrievers. Specifically, BM25 had the largest performance drop, while the methods utilizing language models had relatively smaller drops, demonstrating their robustness against linguistic variation. These results suggest that word overlap between questions and tables is reduced and advanced semantic understanding is required. Additionally, when comparing the performance of BERT and TAPAS table encoders, retrieval performed better with BERT for all three types of data. As previously demonstrated by Wang et al. (2022) in the case of NQ-table, table retrieval does not necessarily require a table-specific model, a conclusion reconfirmed by Open-WikiTable.

Method	Validation			Test		
	k=5	k=10	k=20	k=5	k=10	k=20
Reader	55.1	62.7	65.0	57.5	64.5	65.2
Parser	63.3	66.0	**67.0**	65.2	67.1	**67.9**

Table 4: Comparison of the end-to-end reader and parser's exact match (EM) score, where *k* represents the number of tables retrieved.

3.2 End-to-End Table QA

Experimental Setup We experiment with two different methods: reader and parser. Conventionally, the parser only utilizes the table schema rather than the entire contents, as the question typically specifies the exact table value. However, in Open-WikiTable, the values are often paraphrased, requiring the parser to extract the exact value from the table contents (See Appendix C).

For end-to-end question answering, we adopt the retriever that yielded the highest performance in the previous experiment. The question and retrieved tables are concatenated and fed to QA models. Both reader and parser are implemented with the fusion-in-decoder architecture (Izacard and Grave, 2020) and the T5-*base* language model (Raffel et al., 2020). We use the exact match accuracy (EM) for the evaluation metric. For the parser, EM is computed on the execution result of generated SQLs, as they can be expressed in a diverse form.

Result Table 4 summarizes validation and test results for end-to-end QA. As the retrieval performance improves with increased *k*, QA models, which rely on the retrieved tables, accordingly show consistent performance improvement with larger *k*. However, regardless of the number of *k*, the parser model outperforms the reader model. This performance gap is most significant with small *k*, and decreases as *k* grows. We posit that this is due to the difference in the minimum amount of table segments that the reader and parser must refer to create an accurate answer. The parser model can generate a correct SQL query even when all segments of a table are not retrieved, as long as any of the retrieved splits possess all necessary cell values. On the contrary, the reader model should refer to every relevant split to derive a correct answer.

For more detailed analysis, we categorize questions into easy or hard based on if the answer is derived from a single cell value, and into single-table or multi-table based on if the grounding table

Category		**Reader**	**Parser**	**# questions**
Table-split	**Complexity**			
Single	Easy	74.0	**82.8**	1,574
	Hard	**62.9**	51.2	1,794
Multi	Easy	70.5	**82.5**	1,520
	Hard	56.0	**58.8**	1,714

Table 5: Exact match scores on Open-WikiTable's test set with *k*=20, where we categorize questions by their complexity and whether the grounding table is split.

is split. The results are shown in Table 5. The parser outperforms the reader when the grounding table is split into multiple segments, regardless of question complexity, which aligns with the previous analysis. It is notable that the parser shows inferior or comparable performance to the reader for hard questions. We believe this is due to the relative size between WikiSQL (*i.e.* mostly easy) and WikiTableQuestions (*i.e.* mostly hard), and that the parser has a limited opportunity to understand the diversity of complex SQL queries.

4 Conclusion

We present Open-WikiTable, the first ODQA dataset that requires complex reasoning over Wikipedia tables. The dataset is constructed by revising WikiTableQuestions and WikiSQL to be fully functional in the open-domain setting through decontextualization and paraphrasing. The dataset provides both textual and logical form answers for each question so that end-to-end reader and parser models can be trained. We hope that Open-WikiTable can provide new opportunities for future research such as investigating the effectiveness of leveraging both reader and parser approaches in the retrieval and generation phase.

Limitations

Although we carefully designed Open-WikiTable for complex open-domain table QA, there are some limitations since it is based on the existing datasets. First, ambiguous or erroneous samples from the original WikiSQL or WikiTableQuestions dataset may still lie in our training and validation set. As we mentioned in Section 3.2, most of the equivocal samples were attributed to the ambiguity of the original question and excluded from the test set, but not removed. Second, unlike semantic coverage of the questions is extended by decontextualization and paraphrasing, the coverage of the question remains in that the answer and logic to derive the answer in each question is the same. Still, Open-WikiTable demonstrates the potential for further research on open-domain QA over the table.

Ethics Statement

No ethics concerned with our work.

References

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and William W Cohen. 2020. Open question answering over tables and text. *arXiv preprint arXiv:2010.10439*.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Martin Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. *arXiv preprint arXiv:2103.12011*.

Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. *arXiv preprint arXiv:2004.02349*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Bogdan Kostić, Julian Risch, and Timo Möller. 2021. Multi-modal retrieval of tables and texts using tri-encoder models. *arXiv preprint arXiv:2108.04049*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Feifei Pan, Mustafa Canim, Michael Glass, Alfio Gliozzo, and Peter Fox. 2021. Cltr: An end-to-end, transformer-based system for cell level table retrieval and table question answering. *arXiv preprint arXiv:2106.04441*.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Tianze Shi, Chen Zhao, Jordan Boyd-Graber, Hal Daumé III, and Lillian Lee. 2020. On the potential of lexico-logical alignments for semantic parsing to sql queries. *arXiv preprint arXiv:2010.11246*.

Roee Shraga, Haggai Roitman, Guy Feigenblat, and Mustafa Cannim. 2020. Web table retrieval using multimodal deep learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1399–1408.

Zhiruo Wang, Zhengbao Jiang, Eric Nyberg, and Graham Neubig. 2022. Table retrieval may not necessitate table-specific model design. *arXiv preprint arXiv:2205.09843*.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

A Data Construction Details

A.1 Table Descriptions Re-Annotation

Figure 2 illustrates the indistinguishable annotation of the table corpus in WikiSQL and WikiTableQuestions, leading to ambiguity in the decontextualized questions. The figure on the right shows how the problem is solved by re-annotating the table descriptions.

A.2 Construction Details

The prompt used by GPT-J for decontextualization can be found in Table 6. Table 7 shows examples of each step in the process of creating the Open-WikiTable.

A.3 Error Analysis

Upon closer examination of the 7.9% error on generated Open-WikiSQL, we find that 70% of the errors were the results of ambiguity in the original questions, which was propagated over during the data construction process. The percentage of errors by the decontextualization process and paraphrasing process was 15% respectively. In Table 8, we provide examples for each type of error encountered.

B Experimental Setup

B.1 Flattened Table Format

In order to present the table as passages, we flattened the table and added table descriptions with the help of special tokens. For example,

Page title : 2008-09 Los Angeles Lakers
Section title : Playoffs
Caption : First round
Table ID : table_132938_29

Game	Date	Team	Score	High points
1	April 19	Utah	W 113-100	Kobe Bryant
2	April 21	Utah	W 119-109	Kobe Bryant

is flattened as

> *[Page Title] 2008-09 Los Angeles Lakers [Section Title] Playoffs [Caption] First round [table_id] table_132938_29 [Header] Game [SEP] Date [SEP] Team [SEP] Score [SEP] High points [Rows] [Row] 1 [SEP] April 19 [SEP] Utah [SEP] W 113-100 [SEP] Kobe Bryant [Row] 2 [SEP] April 21 [SEP] Utah [SEP] W 119-109 [SEP] Kobe Bryant*

B.2 Hyperparameters

All experiments were on 8 NVIDIA A6000 48G GPUs. For the retrieval models, we use a batch size of 64, with a learning rate of 1.0 e-5 using Adam and linear scheduling with a warm-up. The in-batch negative technique was utilized to train the retriever. We evaluated every 500 steps and used early stopping with patience 5. For the question-answering module, we use batch size 8 for $k = 5$, 10 and batch size 4 for $k = 20$. The rest of the hyperparameters goes the same with the retriever.

C Open-WikiTable with Parser

In the open-domain scenario, where the table is not specified a priori, questions may not contain the exact cell value to generate SQLs. As shown below, it is necessary to refer to the grounding table and use the exact value to generate the correct SQL.

Question	What is Born-Deceased if the term of office is December 4, 1941 in the list of Prime Ministers of Albania
SQL	SELECT Born_Died From table_2 WHERE Term_start = "4 December 1941"
Question	In the Gothic-Germanic strong verb, which part 2 has a verb meaning to jump?
SQL	SELECT Part_2 FROM table_3 WHERE Verb_meaning = "to leap"

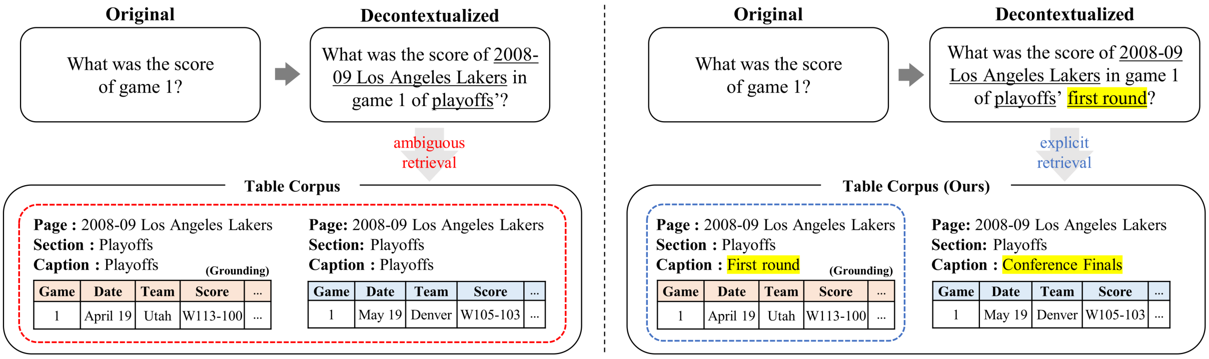

Original
What was the score of game 1?
Decontextualized
What was the score of 2008-09 Los Angeles Lakers in game 1 of playoffs'?
ambiguous retrieval
Table Corpus
Page : 2008-09 Los Angeles Lakers
Section : Playoffs
Caption : Playoffs
(Grounding)
Game | Date | Team | Score | ...
1 | April 19 | Utah | W113-100 | ...
Page: 2008-09 Los Angeles Lakers
Section: Playoffs
Caption : Playoffs
Game | Date | Team | Score | ...
1 | May 19 | Denver | W105-103 | ...
Original
What was the score of game 1?
Decontextualized
What was the score of 2008-09 Los Angeles Lakers in game 1 of playoffs' first round?
explicit retrieval
Table Corpus (Ours)
Page : 2008-09 Los Angeles Lakers
Section : Playoffs
Caption : First round
(Grounding)
Game | Date | Team | Score | ...
1 | April 19 | Utah | W113-100 | ...
Page: 2008-09 Los Angeles Lakers
Section: Playoffs
Caption : Conference Finals
Game | Date | Team | Score | ...
1 | May 19 | Denver | W105-103 | ...

Figure 2: Comparison on retrieval of decontextualized question using WikiSQL(left) and Open-WikiTable(right). The table description of WikiSQL is insufficient to pinpoint the grounding table, even after the decontextualizing process. In contrast, the descriptions of Open-WikiTable effectively address the issue by re-annotation.

```
Page Title : Wake Forest Demon Deacons football, 1980–89
Section Title : Schedule
Caption : 1987
Question : Who was the opponent when the result was L 0-14?
What is converted question using given information?
In 1987’s schedule, who was the opponent of Wake Forest Demon Deacons when the result was L 0-14?
...
Page Title : Toronto Raptors all-time roster
Section Title : A
Caption : A
Question : What is order S24’s LNER 1946 number?
What is converted question using given information?
Considering the history of GER Class R24, what is order S24’s LNER 1946 number?
...
Page Title : GER Class R24
Section Title : History
Caption : Table of orders and numbers
Question : What is order S24’s LNER 1946 number?
What is converted question using given information?
Considering the history of GER Class R24, what is order S24’s LNER 1946 number?
...
Page Title : 2006–07 Toronto Raptors season
Section Title : Game log
Caption : February
Question : Who had the highest number of rebounds on February 14?
What is converted question using given information?
From 2006-07 Toronto Raptors’ game log, who had the highest number of rebounds on February 14?
...
Page Title : Toronto Raptors all-time roster
Section Title : O
Caption : O
Question : Which school was in Toronto in 2001-02?
What is converted question using given information?
Which school was in Toronto in 2001-02 from Toronto Raptors all-time roster O?
...
Page Title : Stozhary
Section Title : Stozhary 2003 Prize-Winners
Caption : Stozhary 2003 Prize-Winners
Question : What actor was nominted for an award in the film Anastasiya Slutskaya?
What is converted question using given information?
For 2003 Stozhary prize winners, what actor was nominted for an award in the film Anastasiya Slutskaya?
...
Page Title : 1985 New England Patriots season
Section Title : Regular season
Caption : Regular season
Question : How many weeks are there?
What is converted question using given information?
In 1985 New England Patriots season, how many weeks were there for regular season?
...
Page Title : Friday Night Lights (U.S. ratings)
Section Title : Weekly ratings
Caption : Season 1
Question : What is the rank number that aired october 26, 2007?
What is converted question using given information?
What is the rank number of Friday Night Lights Season 1’s weekly ratings that aired october 26, 2007?
```

Table 6: The prompt used for GPT-J when decontextualizing the question

original	The nhl team new york islanders is what nationality?
de-contexualized	The NHL team New York Islanders in what nationality 1994 NHL Entry Draft's Round one?
paraphrased	What nationality is the NHL team New York Islanders in the first round of the 1994 NHL Entry Draft?
original	What is the maximum starts that result in an average finish of 16.5?
de-contexualized	What is the maximum starts that result in an average finish of 16.5 for NASCAR Nationwide Series' Chad Little?
paraphrased	What are the maximum starts that result in a 16.5 average finish for NASCAR Nationwide Series' Chad Little?
original	If the population is 2188, what was the median household income?
de-contexualized	If the population is 2188 in Ohio locations ranked by per capita income, what was the median household income?
paraphrased	If Ohio's population is 2,188 ranked by per capita income, what was the median household income?
original	What values of HDTV correspond to n° of 862?
de-contexualized	From the list of television in Italy's Shopping section, what values of HDTV correspond to n° of 862?
paraphrased	Which HDTV values correspond to the number 862 in the TV list in the Italian shopping area?
original	How many stories is the torre reforma building?
de-contexualized	How many stories is the torre reforma building from the list of tallest buildings in Mexico's Under construction?
paraphrased	From the list of tallest buildings under construction in Mexico, how many floors does the Torre Reforma building have?
original	How many teams have a head coach named mahdi ali?
de-contexualized	How many teams has a head coach named mahdi ali among 2010–11 UAE Pro-League?
paraphrased	How many teams in UAE Pro-League 2010-11 have a head coach named Mahdi Ali?
original	Which Member has an Electorate of southern melbourne?
de-contexualized	Which Member has an Electorate of southern melbourne among Members of the Australian House of Representatives, 1903–1906?
paraphrased	Among the Members of the Australian House of Representatives, 1903–1906, which member does a south Melbourne electorate have?
original	Which position had fewer rounds than 3, and an overall of less than 48?
de-contexualized	Which position among 2007 Jacksonville Jaguars draft history had fewer rounds than 3, and an overall of less than 48?
paraphrased	Which position in the 2007 Jacksonville Jaguars draft history had less than 3 rounds and less than 48 overall?
original	How many numbers of dances for place 1?
de-contexualized	How many numbers of dances for place 1 for Dancing on Ice (series 5)?
paraphrased	How many dances for 1st place for Dancing on Ice (series 5) ?

Table 7: Examples of each step in the process of creating the Open-WikiTable

Error type 1 (70%)	Ambiguity in the original question
original	Name the 2005 with 2007 of sf
de-contextualized	Name the 2005 with 2007's Doubles name of sf among Alicia Molik?
parapharsed	Do you name the 2005s with 2007 doubles names from sf under Alicia Molik?
Error type 2 (15%)	Change of intent after decontextualizing
original	how many total rounds did she fight in ?
de-contextualized	How many total rounds did she fight for Gina Carano?
parapharsed	How many rounds did she fight for Gina Carano in total?
Error type 3 (15%)	Change of intent after paraphrasing
original	Which Bask has an Indoor track of 0, and a Swimming of 5?
de-contextualized	Which bask has an indoor track of 0,and a swimming of 5 for horizon league's women's sports championship totals?
parapharsed	Which pool has an indoor stretch of 0 and a swim of 5 for the total number of women's athletic championships in the horizon league?

Table 8: Error analysis on the construction stage of Open-WikiTable