Tabular Data Question-Answering: Integrating DataBench and TAPAS

Carter Louchheim
Williams College
chl2@williams.edu

Abstract

This paper addresses a subsection of SemEval Task 8 SubTask 2 (Osés Grijalba et al., 2024b), which focuses on developing the most effective Question Answering (QA) system for tabular data using the newly created DataBench dataset. Leveraging TAPAS (Herzig et al., 2020), a BERT-based model specifically designed for tabular data QA , I fine-tuned the weakly supervised WTQ version of TAPAS on a subset of number and category-answer questions from DataBench. Preliminary results demonstrate that the TAPAS WTQ model, not finetuned on the DataBench data, outperforms Z-ICL and Code Prompt strategies (Osés Grijalba et al., 2024a) in these answer types. However, the study is constrained by the manual annotation of answer coordinates, essential for TAPAS training, and the limited diversity of the annotated dataset, raising concerns about potential overfitting. These findings highlight TAPAS's promise in tabular data QA while emphasizing the need for broader datasets and automated annotation tools to improve scalability and generalization.

1 Introduction

Accurately answering questions based on tabular data is a critical challenge in natural language processing (NLP), with applications spanning domains such as medicine, finance, and other data-rich industries. While humans often find it difficult to quickly parse complex tables, natural language interfaces that bridge this gap can significantly improve accessibility and decision-making. To address this need, the SemEval 2025 Task 8 introduced the DataBench benchmark, a novel dataset comprising real-world tables with diverse data types, domains, and question formats. This dataset serves as a standard for evaluating systems capable of answering numerical, categorical, boolean, and list-based questions.

Despite advances in tabular question-answering (QA) systems, current state-of-the-art models achieve only approximately 75% accuracy on number and category-based questions and struggle further—achieving just about 35% accuracy—on tasks that do not require Python interpreters. The strategies that do not require python interpreters are important to real world applications as they do not require any code-syntax frame work, and rather are more accessible in their natural language and table copy and paste form. In this project, it is the poor performance of In-context prompt strategies through the use of TAPAS, a tabular data questions answering system. These performance gaps underscore the complexity of tabular data QA and the need for more robust systems capable of generalizing across diverse datasets and question types.

In this study, we focus on improving number and category-answer QA performance within the DataBench framework. By leveraging TAPAS, a BERT-based model designed for tabular data, we aim to validate the system accuracy and reliability. This project explores fine-tuning TAPAS on the DataBench dataset, and natural language alteration of the DataBench questions, analyzing its potential to outperform existing approaches and contribute to the development of more effective tabular QA systems.

2 Related Work

Provided as a starting point in the SemEval task description, Grijalba et al. (Osés Grijalba et al., 2024a) introduce DataBench, a comprehensive benchmark for evaluating Large Language Models' (LLMs) capabilities in tabular reasoning. DataBench is a dataset comprising of 65 real-world datasets and 1300 question-answer pairs accompanying them. In the presentation of the DataBench benchmark, they provoide two strategies that have had their tabular data question answering accuracies evaluated on DataBench. The two strategies which they discuss are a code-prompting strategy and an in-context learning prompt strategy. In code-prompting approach, the prompt is an unfinished python function with a docstring and column information. The LLM is supposed to return a return statement that would correctly locate or calculate the answer. Importantly, this strategy requires a python interpreter, at which point the actual dataset is introduced as the data itself is not in the prompt. The other approach, in-context learning, has the natural language question and then the entire csv file in the prompt, and expects a natural language answer. Both of these strategies were tested using llama and chat-gpt models. The code-prompt strategy worked better, with an overall 63% accuracy, with 73.3% and 75.9% accuracy on questions with number and category answer types respectively.

The use of transformer based models for tabular data QA was buttressed by the work of Babardo et al. (Badaro and Papotti, 2022). Badaro et al. provide a comprehensive tutorial exploring the adaptation of transformer architectures for tabular data processing. Their work catalogs existing approaches and importantly highlighting the potential of transformer models in bridging natural language processing and tabular data analysis. Along these lines, the TAPAS model by Herzig et al., a BERT based tabular data questioning answering model attempts to span this gap (Herzig et al., 2020).

TAPAS is a weakly supervised question answering system, extending the BERT architecture with specialized embeddings to capture table structure. By pre-training on Wikipedia text-table pairs, the model is pre-trained to understand and answer questions about tabular data. Although initially evaluated on smaller-scale datasets with less than ten rows and columns, achieving a 48% accuracy, TAPAS provides a groundwork for subsequent research in table-based question answering. TAPAS unlike the examples given in the Grijalba et al. paper, performs inference by returning an aggregation operator (NONE, SUM, COUNT, AVERAGE) and a set of cells to apply this operation to. To do this, the TAPAS model has two classification heads on top of the BERT architecture to make these decisions. Aggregation operators and cells are selected by TAPAS with a 0.5 confidence threshold, so it is possible for the model to return nothing. Additionally, TAPAS has a specialized tokenizer, that for concatenates the question onto the end of a specially tokenized flattened data frame wo which the question relates.

3 Dataset

The dataset used in this project is a subset of DataBench, presented as the dataset meant for the SemEval 8 task, accessed at 'cardiffnlp/databench' on HuggingFace. DataBench is an extensive research dataset developed for evaluating Large Language Models' (LLMs) ability to reason over tabular data. Created by researchers from Cardiff University, this dataset is meant to provide a standard for tabular question answering evaluation. DataBench consists of 65 real-world datasets, each with a full version sample version which only has 20 rows. For each of the 65 datasets, there are 20 human annotated question answer pairs based on the given dataset. For each question answer pair there is the question, answer, answer type, the columns used in the answer and their types, the sample answer, and the dataset that the question is about. The answer types included are number, category, boolean, and lists of those types. The sample answer refers to the answer of the question when the question is applied to only the 20 row sample version of the dataset.

The subset of DataBench used in this project is constrained by the capabilities of TAPAS. TAPAS can only return answer that are derived from the dataset making boolean answers impossible. Additionally, TAPAS only returns single values, meaning that lists of numbers, categories, or booleans were also not included in the data subset. Additionally, TAPAs can process a maximum of 512 row tables, so the sample versions of each of the 65 datasets were used. Further, TAPAS identifies and outputs cells in tables through their zero-indexed coordinates (row, column), so answer coordinates had to be annotated for all question answer pairs for all in the train and development set. The answer coordinates are meant to be the location of all cells in the dataframe needed to compute the final answer. For a simple answer, this might just be a single cell location, but for a question asking about an aggregation like average, sum, or count, this would refer to all cells aggregated for the computation. For the train and development sets, float answers had to be added or each pair, with the float version of each number answer and or nan for each categorical answer added. After annotation was done, the train set has 90 question answer pairs from 7 datasets and the development set has 14 question answer pairs from 2 datasets. These additional annotations were not needed for the test set, and therefore the development split of the DataBench questions was used, resulting in 211 questions answer pairs from 16 datasets in the test set. Although unconventional to have a proportionally large test set, the lack of annotations needed for the test and the need for diverse questions makes a large test set a reasonable choice for this data.

4 Methods

This project centered around the fine tuning of the TAPAS WTQ model by Stanford on different versions of the DataBench dataset. TAPAS provided an appropriate base model for this experiment for a few reasons. First, it is an publicly available tabular data question answering specific model. Next, TAPAS, more closely aligns with the Z-ICL strategy as it incorporates a natural language question and the table itself in its training and inference state rather than code or pseudo code. Finally, TAPAS had not been evaluated on DataBench and having been trained on slightly smaller tables, DataBench provides a challenging new dataset for TAPAS to be evaluated on. For this project, I fine-tuned the TAPAS WTQ model on three different version on the training data. All of these models used a Adam optimizer with a learning rate of 1e-5 and a linear scheduler. The number of epochs fine-tuned for were determined by fine-tuning a model for 15 epochs and manually inspecting the training and development losses. This manual inspection was used rather an an early stopping technique due to the non representative development set. This limitation is elaborated on in the Limitations & Ethics section.

4.1 Model 1

In Model 1, the TAPAS WTQ model was fine tuned for 5 epochs with training and development data that had a batch size of 8. The annotated training data for this experiment had single answer coordinates for questions asking about majority or minority categories. For example, take the the question "What's the most common gender among the survivors?" (Grijalba et al., 2024) where the answer is 'female.' In this set of training data, the answer coordinate would just be the first occurrence of 'female' in the gender column. This annotation is done differently for Model 2.

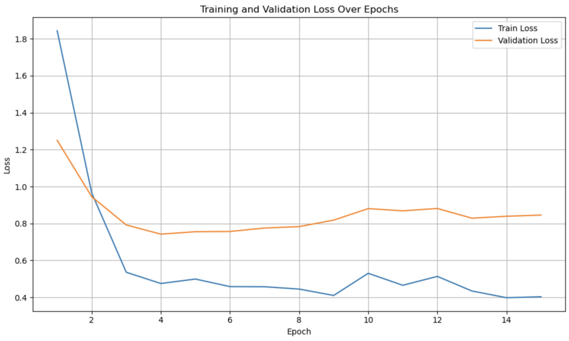

Figure 1: Loss on training and development sets for DataBench fine-tuning on TAPAS WTQ model. Used to choose number of epochs for later training

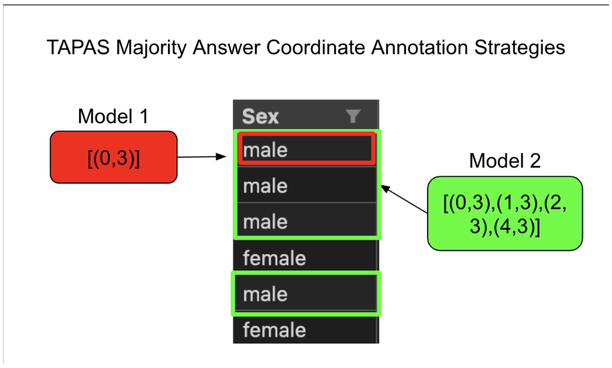

Figure 2: Example of different answer coordinate annotation strategies used in Model 1 and Model 2. Here, 'Sex' is the 4th column (zero-indexed) in the given dataset where 'male' is the answer to the question 'What was the most common gender of passengers on the ship?'

4.2 Model 1-alter

In Model 1-alter, the same training set up and answer coordinate annotation is used as for Model 1. The difference between these models is the form of the question in the training. For Model 1-alter, additional information including the columns that should be used and the answer type was added to the questions in natural language. In the DataBench dataset, the columns used and the answer type or each question answer pair is included in the the dataset for each question answer pair. For example, the question "What is the source of wealth for the youngest billionaire?" (Grijalba et al., 2024) would become "What is the source of wealth for the youngest billionaire? Use the age and source columns to give a categorical answer."

MODEL	altered questions	overall accuracy	number accuracy	category accuracy
Z-ICL Prompt 2	FALSE	0.342	0.315	0.368
TAPAS WTQ	FALSE	**0.483**	**0.482**	0.484
TAPAS WTQ	TRUE	0.467	0.396	**0.531**
Model 1-alter	FALSE	**0.483**	**0.482**	0.484
Model 1-alter	TRUE	**0.483**	0.448	0.515
Model 1	FALSE	0.426	**0.482**	0.375
Model 1	TRUE	0.409	0.465	0.359

Table 1: Accuracies on both the standard test sets and the test set with the altered questions. The accuracies for Z-ICL Prompt 2 act as a baseline for DataBench accuracies, but are calculated on the true DataBench test set that was not available during this project

4.3 Model 2

Model 2 is the same set up as Model 1 except for the answer coordinates in the training data. Take the same question "What's the most common gender among the survivors?" (Grijalba et al., 2024) where the answer is 'female.' In this set of training data, the answer coordinates would be every occurrence of 'female' in the gender column. TAPAS did not have example questions with categorical majority answers, so the answer coordinate scope was unclear, hence Models 1 and 2.

5 Results

All results were evaluated in terms of accuracy on the test set which consists of 112 question answer pairs. Only exact matches are counted as correct. On inference, a TAPAS model outputs an aggregation operator (NONE, SUM, COUNT, or AVERAGE) and a set of cells to apply the operator to. Since leaving the inference in this form would result in a 0.0 accuracy, the aggregation operators are applied to the selected cells and are included as an aspect of the inference. Below, granular results are laid out for Model 1, Model 1-alter, the not fine-tuned TAPAs WTQ model, and Z-ICL Prompt 2 (Osés Grijalba et al., 2024a). The 'altered question' category refers to whether the test set had standard form (FALSE) or had additional column and answer type information in natural language form (TRUE). Number and category accuracy refer to the accuracy of questions with the answer types of 'answer' and 'category' respectively as specified by the DataBench 'type' column. Model 2, where all cells of a majority answer were annotated as valid cell coordinates performed significantly worse than the other models, and was not included in the table. For overall, number, and category accuracies Model 2 got less than %15 of the answers correct.

Overall, the TAPAS models both fine-tuned and not perform better than the In-Context Prompt models do, outperforming Z-ICL Prompt 2 by at least %9 in terms of overall accuracy for all versions of the TAPAS model. For overall accuracy, the TAPAS WTQ model and the Model 1-alter model performed %14 better than Z-ICL Prompt 2 with a %48.2 accuracy. For number question accuracy, TAPAS WTQ, Model 1-alter, and Model 1 all achieved %48.2 accuracy on the unaltered test set, %17 better than Z-ICL Prompt 2 number accuracies. Finally, for category question the TAPAS WTQ model evaluated on the altered test set had an accuracy of %53.1, %16.3 better than Z-ICL Prompt 2 category question accuracy. Overall, it does not seem like fine-tuning, at least on the scale possible using human annotated answer coordinates, is an effective strategy to improve TAPAS performance on DataBench numerical and categorical questions. While these strategies still not not match the accuracy of the Code-Prompt method, they outperform the similarly structured In-Context Prompting.

The altered question test set resulted in better accuracies than the standard format for the categorical answer questions for both the TAPAS WTQ and Model 1-alter models, but resulted in consistently worse accuracies for numerical answer questions. In fact, the only result that was worse than the Z-ICL Prompt 2 results was the categorical accuracy of Model 1 when evaluated on the altered question test set.

6 Ethics & Limitations

This research confronts significant methodological challenges in tabular question-answering, primarily stemming from the labor-intensive process of manually annotating answer coordinates. The extreme complexity of potential question types can only be captured by a sufficiently large training set and development set which were not able to be obtained in the scope of this project due to the labor intensive annotation. This was demonstrated in the range of accuracies between the training and development sets on the TAPAS WTQ model before fine tuning. The test set had a %22 acccuracy while the development set had a %44 accuracy, demonstrating how different questions were much more difficult for the model.

Like all systems reliant on natural language data, From an ethical perspective, tabular question-answering systems risk perpetuating historical biases embedded in source datasets. Language models trained on potentially skewed data could inadvertently reproduce social inequities, highlighting the critical need for careful dataset curation.

7 Conclusion

This research explores the potential of TAPAS, a natural language-based model, in addressing tabular question-answering challenges presented by the DataBench dataset. While the model did not outperform the Code-Prompt approach this project demonstrates significant improvements over In-Context Learning strategies. The TAPAS model consistently achieved around 48% accuracy, representing a substantial enhancement over existing natural language-based approaches.

The most critical insight from this research is not just the performance metrics, but the pathway it illuminates for more accessible, human-readable tabular data question-answering systems. By eliminating the need for complex code interpreters, TAPAS represents a step towards making advanced data analysis more intuitive and approachable for non-technical users.

However, the research also underscores significant limitations in current methodologies. The annotation of answer coordinates remains a substantial bottleneck, requiring extensive human labor and introducing potential bias. Future research must focus on developing automated annotation techniques and expanding dataset diversity to reduce these constraints.

Our work suggests that while current tabular question-answering models are promising, substantial improvements are still needed. The path forward lies in developing more sophisticated natural language understanding techniques, creating more comprehensive datasets, and designing more efficient annotation strategies that can scale across diverse data types and domains.

References

Gilbert Badaro and Paolo Papotti. 2022. Transformers for tabular data representation: A tutorial on models and applications. *Proceedings of the VLDB Endowment*, 15(12):3746–3749.

Jorge Osés Grijalba, Luis Alfonso Ureña-López, Eugenio Martínez Cámara, and Jose Camacho-Collados. 2024. Question answering over tabular data with databench: A large-scale empirical evaluation of llms. In *Proceedings of LREC-COLING 2024*, Turin, Italy.

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*.

Jorge Osés Grijalba, L. Alfonso Ureña-López, Eugenio Martínez Cámara, and Jose Camacho-Collados. 2024a. Question answering over tabular data with DataBench: A large-scale empirical evaluation of LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13471–13488, Torino, Italia. ELRA and ICCL.

Jorge Osés Grijalba, L. Alfonso Ureña-López, Eugenio Martínez Cámara, and Jose Camacho-Collados. 2024b. Semeval 2025 task 8: Question answering on tabular data. https://jorses.github.io/semeval/. Retrieved from the official SemEval website.