iTBLS: A Dataset of Interactive Conversations Over Tabular Information

Anirudh Sundar[1] **and Christopher Richardson**[2] * **and Adar Avsian**[1] **and Larry Heck**[1]
[1] Georgia Institute of Technology, USA
[2] Google Inc., USA
asundar34, larryheck@gatech.edu

Abstract

This paper introduces Interactive Tables (iTBLS), a dataset of interactive conversations that focuses on natural-language manipulation of tabular information sourced from academic preprints on ArXiv. The iTBLS dataset consists of three types of tabular tasks – interpretation, modification, and generation. Interpretation focuses on tabular understanding, modification focuses on manipulating tabular information, and generation focuses on the addition of new natural-language evidence. In addition, the paper presents a novel framework that reformulates tabular operations as question-answering, where an appropriate question is formulated based on the nature of interaction and the question is answered using the user request as evidence. The developed approach results in an improvement on all tasks on a sequence-to-sequence modeling baseline on iTBLS. In addition, the question-answering-based reformulation is applied to datasets from prior work for the text-to-table task where textual paragraphs are summarized into tables. The novel approach results in up to 13% improvement in Exact-Match accuracy and up to 16% improvement in BERTScores compared to the prior state-of-the-art.

1 Introduction

Recent research on Conversational AI has focused on adding enhanced multi-task capabilities to large language models (LLMs). This research includes building systems capable of situated interactions over structured knowledge sources such as tabular information (Sundar and Heck, 2022). Automated methods for tabular interpretation, manipulation, and generation empower users by saving time and reducing errors in managing tabular content (Kardas et al., 2020). Previous studies have focused on individual aspects of tabular data management: representation learning for interpretation tasks like grounded question answering, manipulation for data wrangling, and generation for summarizing textual information independently (Nakamura et al., 2022a; Sundar and Heck, 2023; Fang et al., 2024).

The development of situated conversational interactions over tables necessitates a suite of approaches to unify tabular interpretation, modification, and generation in a conversational context. Additionally, an important yet largely unaddressed challenge in interacting with tabular sources is the ability to modify existing tabular content using conversational natural language commands.

To address these challenges, this paper introduces Interactive Tables (iTBLS) [1], a dataset of interactive conversations in English situated in tabular information. iTBLS decomposes the challenge into three distinct tasks: *interpretation*, which involves understanding tabular content within a conversational framework; *modification*, which entails manipulating tabular content through natural language commands; and *generation*, which focuses on integrating new natural language information into existing tables. The tabular information in iTBLS is sourced from scientific articles hosted on arXiv [2], an open-access repository of academic preprints.

Beyond factoid question-answering, iTBLS encompasses tasks such as comparison, determining absolute and relative positions, and mathematical reasoning. Previous research primarily examined procedural command generation for spreadsheets or the alignment of tabular data through LLMs. iTBLS integrates these functionalities into a unified task, enabling the manipulation of existing tables through natural-language commands. On tabular generation, while prior work addressed the summarization of natural language paragraphs in a tabular format, iTBLS focuses on generating row

*Work done while at Georgia Tech

[1] https://huggingface.co/datasets/avalab/iTBLS
[2] https://arxiv.org

or column data conversationally.

In addition to building iTBLS, this paper develops a novel approach to address tabular operations by reformulating the task as conditional question answering. Furthermore, the question-answering-based reformulation is applied to other datasets introduced in prior work (Wu et al., 2022) and results in better performance in terms of both table-cell accuracy and BERTScore.

The contributions of this work are as follows:

- Creating iTBLS, a dataset of tabular interactions unifying interpretation, modification, and generation.
- Extending prior tabular datasets by collecting information from arXiv
- Broadening the scope of interactions to include mathematical reasoning, natural language manipulation, and natural language expansion.
- Introducing a novel approach for table generation tasks through a two-stage reformulation that first identifies the cells to be manipulated and generates a question based on the requested operation, then answers those questions using the user request and the input table as evidence.
- Demonstrating up to 13% improvement in table-cell accuracy and up to 16% improvement in BERTScore using the novel approach on the text-to-table task introduced by prior work.

2 Related Work

A detailed survey of LLMs for tabular data is available in (Fang et al., 2024). Related work on paired natural-language and tabular data can be broadly classified by the nature of the interaction: tabular interpretation, tabular modification, and tabular generation.

2.1 Tabular Interpretation

Tabular interpretation involves a dialogue turn focused on extracting information from a specific cell in a table, such as identifying a cell satisfying certain criteria. Prior research on tabular interpretation focused on grounded question-answering. An important challenge in the collection of such datasets is the availability of large-scale tabular data. Consequently, many tabular datasets are constructed from online resources such as Wikipedia including WIKITABLEQUESTIONS (Pasupat and Liang, 2015), ManyModalQA (Hannan et al., 2020), TABERT (Yin et al., 2020), NQ-Tables (Herzig et al., 2021), FEVEROUS (Aly et al., 2021), FeTaQA (Nan et al., 2022), HYBRIDIALOGUE (Nakamura et al., 2022b), and HiTab (Cheng et al., 2022). Other tabular datasets are constructed from financial reports including TAT-QA (Zhu et al., 2021), FINQA (Chen et al., 2021), MULTIHIERTT (Zhao et al., 2022), or scientific reviews (Sundar et al., 2024).

Proposed approaches to address the tabular interpretation task include architectures based off of the Transformer encoder (Yin et al., 2020; Herzig et al., 2020; Chen et al., 2019b; Eisenschlos et al., 2020; Liu et al., 2021; Gu et al., 2022; Yang et al., 2022), decoder (Gong et al., 2020; Akhtar et al., 2023; Zha et al., 2023; Jiang et al., 2023; Zhang et al., 2023; Sui et al., 2024; Cremaschi et al., 2025), or both (encoder-decoder) (Nakamura et al., 2022b; Deng et al., 2022; Sundar and Heck, 2023).

2.2 Tabular Modification

Tabular modification concerns the manipulation of the content within an existing table without altering the overall structure of rows and columns. Early work on tabular modification explored the generation of procedural commands for spreadsheets using synthesis algorithms (Singh and Gulwani, 2012; Shigarov et al., 2019). Tools utilizing programming-by-example to parse user intents into executable commands have also been explored (Scaffidi et al., 2009; Kandel et al., 2011; Jin et al., 2017; Petricek et al., 2023; Chen et al., 2023; Xing et al., 2024). More recent work has shifted focus towards leveraging LLMs to synthesize commands for tools (Huang et al., 2024), reformat tabular information (Dargahi Nobari and Rafiei, 2024), and execute programming commands (Liu et al., 2024).

2.3 Tabular Generation

Tabular generation focuses on expanding an existing table by adding a new row or column. Research on tabular generation initially employed discriminative techniques, such as tree-based methods for generating tables of contents (Branavan et al., 2007) and SVMs to classify text across various labels Aramaki et al. (2009). Recent approaches have shifted towards neural techniques including Generative Adversarial Networks (GANs) (Xu and Veeramachaneni, 2018; Park et al., 2018; Chen et al., 2019a;

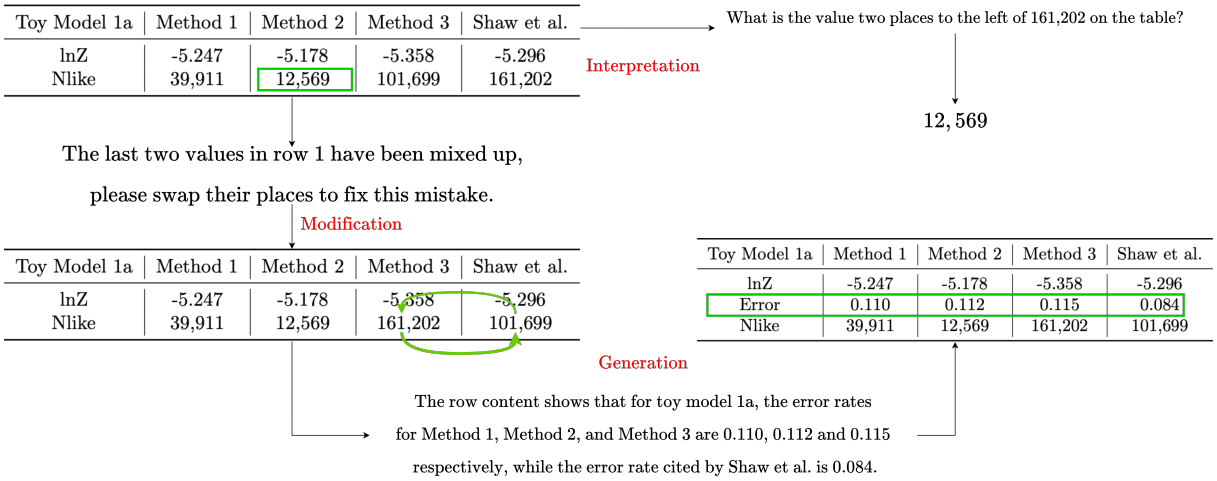

Figure 1: Examples of interactions from the Interactive Tables (iTBLS) dataset.

Zhao et al., 2021), Autoencoders (Li et al., 2019; Darabi and Elor, 2021), Diffusion models (Kotelnikov et al., 2023), and LLMs (Borisov et al., 2023; Solatorio and Dupriez, 2023; Gulati and Roysdon, 2023; Zhao et al., 2023; Seedat et al., 2024; Deng et al., 2024).

A similar line of research also explores the generation of tabular data from associated textual information. Wu et al. (2022) introduced four datasets and proposed a modification to the Transformer's attention mechanism to summarize textual information in a tabular format by inverting datasets created for the dual task of converting tables to text, (as opposed to new conversational evidence). Other approaches to summarize textual information in a tabular format include the addition of learnable bias parameters (Pietruszka et al., 2022) and structure-aware instruction-tuning (Tang et al., 2023).

In contrast to prior work addressing a single mode of interaction, iTBLS is a dataset unifying tabular interpretation, modification, and generation in a conversational format. Additionally, iTBLS broadens the range of interactions to include mathematical reasoning, natural language manipulation, and the expansion of tables using natural language. Furthermore, by leveraging scientific articles from arXiv as a primary source, iTBLS introduces a novel and rich source of information that is not present in existing datasets.

3 The iTBLS Dataset

The Interactive Tables (iTBLS) dataset features conversational interactions situated in tabular data, covering the three distinct types of interactions described in Section 2: *interpretation*, *modification*, and *generation*. Each example type is exemplified in Figure 1 and described below. In addition, since the mode of interaction is not known a priori, any proposed approach using iTBLS must effectively identify the interaction type, either explicitly or implicitly. In the following sections, we provide a detailed description of each type of interaction and outline the dataset collection process.

3.1 Tasks

Tabular Interpretation: In iTBLS, interpretive interactions are structured as question-answer pairs, where the goal is to identify the cell referred to by the question. The references could be absolute (referring to a specific row or column), or relative (referring to one cell in the context of another). Appendix A.5 details absolute and relative references in iTBLS.

Tabular modification: We conceptualize modification in iTBLS as a series of cell swaps, positing that any content rearrangement can ultimately be reduced to such exchanges. This approach allows for both explicit references, where specific row and column numbers are cited, and implicit references, which rely on the content or relative positions of cells. Table 9 in Appendix A.5 showcases examples from iTBLS. As observed, there is a mix of explicit and implicit references to the specific contents to be manipulated.

Tabular generation: In iTBLS, table generation is guided by new natural language evidence. This evidence clarifies appending a row or column, defines the suitable header, and supplies the data entries for the new row (or column) relative to existing columns (or rows). This process ensures

that the added elements are contextually relevant and accurately integrated into the table. Table 10 in Appendix A.5 provides examples of such interactions, demonstrating how users can request the incorporation of new row and column data into an established table framework.

In iTBLS, the mode of interaction is not explicitly stated by the user, introducing an additional task: **interaction identification**. This task involves predicting whether the interaction is intended for interpretation, modification, or generation based solely on the user's request.

3.2 Dataset Collection

To collect the dataset, first we use AXCELL (Kardas et al., 2020) an automatic machine learning pipeline for extracting results from papers. AXCELL is used to parse tabular information from papers on arXiv to populate online leaderboards comparing scientific methods. Using AXCELL, we collect 20,000 tables from academic papers in Mathematics, Physics, and Computer Science over a period spanning from 2007 to 2014. The tables are processed to remove stray characters resulting from the conversion from LaTeX. Additionally, only tables with at least three rows and three columns to at most ten rows or ten columns are retained. The final dataset consists of 4000 tables split between train, development, and test sets.

For each table, we generate three sequential edits corresponding to different types of interaction. Interpretation involves generating a dialogue turn (question-answer pair) grounded on a single cell of the table. Modification involves manipulating two cells of an existing table by swapping them. Finally, generation encompasses the task of appending either a new row or a column to an existing table based on a natural language utterance.

To enhance the quality of the dataset and minimize errors, we implement a strategic selection process for the table components involved in each interaction. In *interpretation*, a cell is randomly selected to ground the dialogue. For *modification*, two cells are chosen and their positions are swapped to simulate a realistic table manipulation scenario. In *generation*, all cells in a randomly masked row or column are used as the basis for appending new table data. All of the interactions are based on cells that do not belong to row or column headers, that is, they reside in the body of the table.

For our dataset creation, we employ two distinct sources for generating dialogue turns based on the type of interaction and the specific table component involved. For tasks related to tabular interpretation and modification, we engage crowd-workers from Amazon Mechanical Turk (AMT). These workers are tasked with formulating questions or commands that pertain to the pre-identified cell(s) designated for each interaction. We recruit workers from Australia, Canada, Ireland, New Zealand, the United Kingdom, and the USA. Each crowdworker is compensated at a rate of $0.15 per Human Intelligence Task (HIT), with the average completion time for each HIT being approximately 40 seconds. Detailed information on the AMT interface used for these tasks is included in Appendix A.7.

For generation, GPT-4 is prompted to write a dialogue turn summarizing a row or column of the table. The prompt is as follows:

The string contains information from a table [table]. Describe the content in this [row/column] for a visually impaired user in one line. Make sure to include all information from the rows and columns and appropriate headers so the user can understand the content.

Each sample in the dataset contains the source arXiv ID, the table that the conversation is situated in, the index of that table within the paper (e.g. Table X), the utterance describing the interaction, the ground truth cell(s) involved in the interaction, and finally the expected output. Statistics of the datasets are provided in Table 1.

Statistic	Interpret	Modify	Generate
# Samples	4168	4168	4168
# Per utterance			
Words	10.6	13.4	31.6
Tokens	14.3	18.3	59.1
# Per table			
Cells	28.1	28.1	25.31
(Cols/Rows)	5.0/5.5	5.0/5.5	4.8/5.3

Table 1: Statistics of the iTBLS dataset

4 Methods

4.1 Table operations through conditional question answering

We also present a novel approach that reformulates operations on tables as question answering. A primary challenge in tabular operations using LLMs lies in ensuring the syntactic validity of the pro-

duced tables. Every row and column in a table must contain the same number of cells, with row and column headers delineating relationships between cells. Failing to adhere to this constraint invalidates the structure of the table and the information presented. Prior work addresses this constraint by including additional parameters like row and column relation embeddings (Wu et al., 2022) or positional bias (Pietruszka et al., 2022) to get the model to attend to header cells while generating content. However, this results in highly specialized architectures for a singular task. Breaking the task down into question-answering results in a more interpretable framework while ensuring validity of the generated tables.

The first step identifies the mode of interaction and the cell(s) the user is referring to, which is used to formulate a question. The second step converts the table into a pandas dataframe, parses the table and the question generated from the previous step to obtain a pandas command corresponding to the task, and executes the command on the dataframe to generate the final table. Generating a valid command ensures that the final table is syntactically valid as well (that is, the number of columns across all rows is consistent).

For the *interpret* task, the question-answer reformulation is trivial, since all interpretive queries and associated responses are naturally question-answer pairs. For the *modify* task, the question is of the form *To which cells is the user referring?*. A language model is then fine-tuned to generate a response containing the cells (indexed by row and column). Then, the LLM response is reformatted into an appropriate pandas command. Finally, for the *generate* task, the question-answering is more nuanced. First, the user request is parsed to identify whether a row or a column is to be appended. The header of the corresponding row is then extracted from the user request. Using the extracted header and the other header cells of the table, questions are generated for each of the empty cells to be filled in the form *What is the row value for column?*. The user request is parsed to obtain the answers to these generated questions, forming the corresponding row or column to be appended.

5 Results

5.1 Experimental Setup

For our experiments, we utilize Gemma models (Team et al., 2024). We fine-tune the instruction-tuned base model `gemma-2-9b-it` using LoRA (Hu et al., 2022). Hyperparameters for our training setup as well as LoRA parameters are shown in Appendix A.

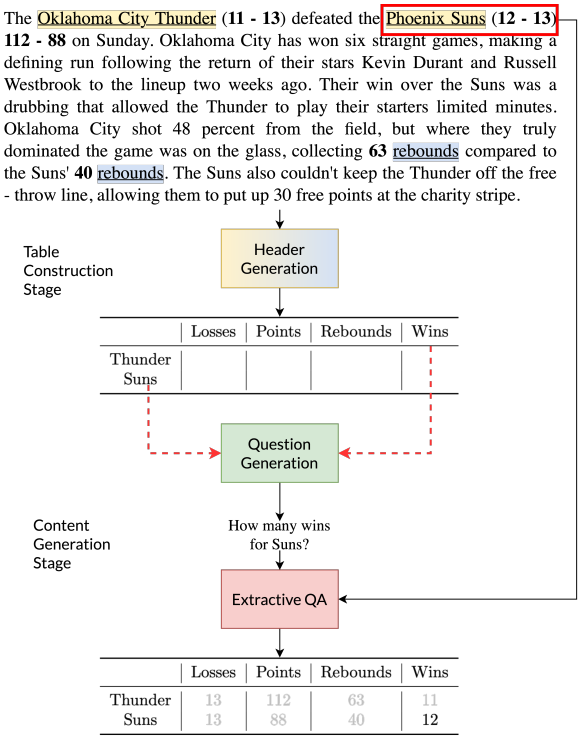

Figure 2: Overview of the novel question-answering reformulation to perform table operations

5.2 Datasets

In addition to the iTBLS dataset, we also evaluate our method on five datasets to summarize textual paragraphs to tables (Wu et al., 2022). While iTBLS is a table-to-table or table-to-text task, the datasets proposed by Wu et al. (2022) address the dual problem of text-to-table. The datasets consist of textual paragraphs containing some information that is to be converted into a tabular format by determining both the appropriate header cells and the content that the table is filled with.

Wu et al. (2022) present datasets for the text-to-table task by inverting datasets created for the dual problem of generating textual descriptions from tables. Each dataset consists of textual paragraphs paired with tabular information summarizing content in the text. Dataset statistics are available in Appendix A.3. Each dataset is described below.

E2E (Novikova et al., 2017) concerns restaurant descriptions, requiring summarization of information into tables with descriptors like restaurant name, customer rating, and location. Wik-

iTableText (WTT) (Bao et al., 2018), sourced from Wikipedia, consists of natural language descriptions generated from tabular data across various topics. WikiBio (Lebret et al., 2016) comprises introductions of individuals from Wikipedia alongside tabular summaries extracted from the same page's information box. In contrast to E2E, the table headers in the WikiTableText and WikiBio datasets vary widely across data samples.

Example textual paragraphs and associated tables from each dataset are presented in A.4.

5.3 Metrics

Exact-Match (EM): On the iTBLS dataset, we report exact-match, that is, whether or not the generated table matches the ground-truth table exactly.

BERTScore: On the E2E, WTT, WikiBio and RotoWire datasets, we report BERTScore (Zhang et al., 2020) in addition to EM to be consistent with prior work. BERTScore is a measure of semantic similarity which computes the similarity of embeddings in a latent space obtained using an encoder language model.

Consistent with prior work, all our evaluations are order-invariant. That is, credit is given as long as the generated cells are indexed by the correct row and column headers, even if the headers themselves are in different positions between the model-generated response and the ground-truth.

5.4 iTBLS

Results on the iTBLS dataset using a vanilla sequence-to-sequence approach and the question-answering-based method are presented in Table 2. As observed in the results, the generate task is the hardest, with performance slightly lower on the generate task when compared to interpret and modify. This is a result of the fact that the exact-match metric only provides credit when all cells are correct (necessitating that all cells in the output are identical to the ground truth) and does not provide partial credit for getting some of the cells right, and the fact that the generate task requires getting more cells right in comparison to the other tasks.

5.5 Text-to-table

The results on the text-to-table datasets proposed by Wu et al. (2022) are available in Table 3. Our method performs on par with or better than the prior state-of-the-art method in terms of BERTScore and is competitive with prior work in terms of Exact-Match. The exact-match score does not reflect true performance on the WikiBio dataset since synonyms are penalized under this framework. A deep-dive into the results is presented in Section 5.6.

Split	**Approach**	**Exact-Match**
Interpret	Seq2seq	88.29
	iTBLS as QA	90.98
Modify	Seq2seq	74.65
	iTBLS as QA	89.58
Generate	Seq2seq	48.94
	iTBLS as QA	73.32

Table 2: Comparison between the question-answering reformulation and a vanilla sequence-to-sequence modeling approach on the iTBLS dataset

Dataset	Approach	EM	BS
WTT	Wu et al. (2022)	62.71	80.74
	Ours	75.96	95.52
Wikibio	Wu et al. (2022)	69.71	76.56
	Ours	66.65	92.60
E2E	Wu et al. (2022)	97.94	98.57
	Ours	97.64	99.35

Table 3: Comparison between our method and prior work on the text to table task in terms of Exact-Match and BERTScore

5.6 Analysis of Errors

An analysis of the difference in performance between the prior state of the art and our approach is presented in Table 6. As observed, the dataset is inconsistent in the description of individuals, with no consistent pattern when middle and last names are present. Furthermore, the use of quantifying information in the header as opposed to the table cell results in no credit using the exact-match metric, though the information contained is exactly the same between the prediction and the ground-truth. Finally, the datasets often contain textual examples with multiple possible tabular summarizations, all of which are equally valid, further complicating evaluation. In the third example in Table 6, the model correctly generates the 'Occupation' as a table header while the ground truth contains an erroneous sample, using the phrase 'Known for' instead of 'Known as'.

Examples of errors in the iTBLS dataset are pro-

vided in Tables 4, 5, and 13. On the interpret task, the model incorrectly understand the user request, and produces the cell immediately to the right instead of three columns over. On the modify task (Table 5), the model incorrectly understands the references and swaps index (2,3) with (3,2) instead of swapping indices (2,2) and (3,3). On the generate task (Table 13), the model incorrectly places a tuple and hallucinates a value instead of performing the requested action.

Text: What is the value of the cell in row 1 that is three cells to the right of the cell with a value of 12%?

Input Table:

row ID	$\sigma\mu[I0]$	$\mu[\tau s]$	$\sigma[\tau s]$	$\sigma\mu[\tau s]$
0	13%	912.5 μs	91.9 μs	10.1%
1	12%	18335.7 μs	90.7 μs	10.0%
2	12%	903.1 μs	1832.7 μs	10.0%

Ground Truth: 10.0%

Prediction: 18335.7 μs

Table 4: Example error for iTBLS interpret task. Table source: https://arxiv.org/pdf/1411.5458

Text: Swap row 1 in the second column with row 2 in the third column

Input Table:

row ID	col 1	col 2	col 3
0	X	O	X
1	NaN	O	O
2	O	X	X

Ground Truth:

row ID	col 1	col 2	col 3
0	X	O	X
1	NaN	X	O
2	O	X	O

Prediction:

row ID	col 1	col 2	col 3
0	X	O	X
1	NaN	O	X
2	O	O	X

Table 5: Example error for iTBLS modify task. Table source: https://arxiv.org/pdf/1411.4023

6 Conclusion

This paper introduces Interactive Tables (iTBLS), a dataset of interactive conversations addressing three types of tasks – interpretation, modification, and generation. In contrast to prior tabular datasets that are sourced from Wikipedia or financial reports, iTBLS is situated in tabular data obtained from scientific pre-prints on ArXiv. Success on the iTBLS dataset requires understanding both ordinal and cardinal references to cell positions, and understanding implicit references. Additionally, the paper introduces a novel framework that reformulates tabular operations as question-answering. Appropriate questions are created based on the input table and the nature of interaction, and the user request is used as evidence to obtain the answers. The developed approach demonstrates an improvement over a sequence-to-sequence modeling approach on the iTBLS dataset. In addition, the question-answering-based reformulation is evaluated on datasets for the text-to-table task, obtaining up to 13% improvement in terms of exact-match accuracy and 16% improvement in terms of BERTScore compared to the prior state-of-the-art.

Limitations

While iTBLS introduces a dataset for interactive conversations over tabular information, there are some avenues for improvement. In this dataset, modification is modeled as a series of swaps. A more comprehensive sequence of manipulations includes in-place modification of values and modifying a cell's value based on other cells using both absolute and relative references. While sourcing tabular information from arXiv provides a cost-efficient approach, LLMs are often pre-trained on LaTeXsources from arXiv. This paper alleviates the issue by sourcing natural language commands from crowdworkers. Future work could look at collecting tabular information from crowdworkers as well. While we present a suite of baseline approaches for iTBLS, there is still headroom between the presented approaches and perfect performance. We identify the closure of this gap as an avenue for future work.

Acknowledgments

This work was supported by CoCoSys, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

1. Text: Walter Clarence Henderson (28 February 1891 – 20 September 1968) was a progressive conservative party member of the Canadian house of commons. He was born in Carberry, Manitoba and became a farmer by career. He was elected at the Cariboo riding in the 1958 general election, defeating social credit incumbent Bert Leboe.
Generated Table:

Predicted Header	Prediction - iTBLS	Ground Truth Header	Ground Truth
Name	Walter Henderson	Name	Walter Clarence Henderson
Profession	Farmer	Profession	Farmer
Party	Progressive Conservative	Party	Progressive Conservative

2. Text: The production of Tautona mine is 235,000 ounces in 2013.
Generated Table:

Predicted Header	Prediction - iTBLS	Ground Truth Header	Ground Truth
Title	Tautona mine	Title	Tautona mine
Subtitle	Production	Subtitle	Production
Year	2013	Year	2013
Production (ounces)	235,000	Production	235,00 ounces

3. Text: Elango Nagarajah, also known as “Thaimann Elango", is a Tamil film actor, director, producer and lyricist in the Tamil film industry. He began his career in his early ages as a producer for the Tamil film Anbudan, starred Arun Vijay, Meena, Rambha (actress) in the main was released in the year 2000.
Generated Table:

Predicted Header	Prediction - iTBLS	Ground Truth Header	Ground Truth
Name	Elango Nagarajah	Name	Elango
Occupation	actor, director, producer, lyricist	Known for	Thaimann

Table 6: Difference between the tables generated by the Zero Shot (ZS) and Fine-Tuned (FT) approaches with respect to the Ground Truth on the WikiBio and WikiTableText datasets with additions and deletions represented using red and green.

References

Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. 2023. Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15391–15405, Singapore. Association for Computational Linguistics.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, and Kazuhiko Ohe. 2009. TEXT2TABLE: medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the Workshop on BioNLP - BioNLP '09*, page 185, Boulder, Colorado. Association for Computational Linguistics.

Junwei Bao, Duyu Tang, Nan Duan, Zhao Yan, Yuanhua Lv, Ming Zhou, and Tiejun Zhao. 2018. Table-to-Text: Describing Table Region with Natural Language. *arXiv preprint*. ArXiv:1805.11234 [cs].

Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2023. Language Models are Realistic Tabular Data Generators. In *The Eleventh International Conference on Learning Representations*.

S. R. K. Branavan, Pawan Deshpande, and Regina Barzilay. 2007. Generating a table-of-contents. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 544–551, Prague, Czech Republic. Association for Computational Linguistics.

Haipeng Chen, Sushil Jajodia, Jing Liu, Noseong Park, Vadim Sokolov, and V. S. Subrahmanian. 2019a. FakeTables: Using GANs to Generate Functional Dependency Preserving Tables with Bounded Real Data. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 2074–2080, Macao, China. International Joint Conferences on Artificial Intelligence Organization.

Ran Chen, Di Weng, Yanwei Huang, Xinhuan Shu, Jiayi Zhou, Guodao Sun, and Yingcai Wu. 2023. Rigel: Transforming tabular data by declarative mapping. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):128–138.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019b. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2022. HiTab: A hierarchical table dataset for question answering and natural language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1110, Dublin, Ireland. Association for Computational Linguistics.

Marco Cremaschi, Fabio D'Adda, and Andrea Maurino. 2025. steellm: An llm for generating semantic annotations of tabular data. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

Sajad Darabi and Yotam Elor. 2021. Synthesising multimodal minority samples for tabular data. *arXiv preprint arXiv:2105.08204*.

Arash Dargahi Nobari and Davood Rafiei. 2024. Dtt: An example-driven tabular transformer for joinability by leveraging large language models. *Proceedings of the ACM on Management of Data*, 2(1):1–24.

Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 2022. PACIFIC: Towards proactive conversational question answering over tabular and textual data in finance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6970–6984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zheye Deng, Chunkit Chan, Weiqi Wang, Yuxi Sun, Wei Fan, Tianshi Zheng, Yauwai Yim, and Yangqiu Song. 2024. Text-tuple-table: Towards information integration in text-to-table generation via global tuple extraction. *Preprint*, arXiv:2404.14215.

Julian Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 281–296, Online. Association for Computational Linguistics.

Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. 2024. Large language models on tabular data–a survey. *arXiv e-prints*, pages arXiv–2402.

Heng Gong, Yawei Sun, Xiaocheng Feng, Bing Qin, Wei Bi, Xiaojiang Liu, and Ting Liu. 2020. TableGPT: Few-shot table-to-text generation with table structure reconstruction and content matching. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1978–1988, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zihui Gu, Ju Fan, Nan Tang, Preslav Nakov, Xiaoman Zhao, and Xiaoyong Du. 2022. PASTA: Table-operations aware fact verification via sentence-table cloze pre-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4971–4983, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Manbir S. Gulati and Paul F. Roysdon. 2023. TabMT: Generating tabular data with masked transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. Manymodalqa: Modality disambiguation and qa over diverse inputs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7879–7886.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 512–519, Online. Association for Computational Linguistics.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yanwei Huang, Yunfan Zhou, Ran Chen, Changhao Pan, Xinhuan Shu, Di Weng, and Yingcai Wu. 2024. Interactive table synthesis with natural language. *IEEE Transactions on Visualization and Computer Graphics*, 30(9):6130–6145.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023. StructGPT: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251, Singapore. Association for Computational Linguistics.

Zhongjun Jin, Michael R. Anderson, Michael Cafarella, and H. V. Jagadish. 2017. Foofah: Transforming data by example. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, page 683–698, New York, NY, USA. Association for Computing Machinery.

Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 3363–3372, New York, NY, USA. Association for Computing Machinery.

Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. 2020. AxCell: Automatic extraction of results from machine learning papers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8580–8594, Online. Association for Computational Linguistics.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. 2023. TabDDPM: modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 17564–17579.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Szu-Chuang Li, Bo-Chen Tai, and Yennun Huang. 2019. Evaluating Variational Autoencoder as a Private Data Release Mechanism for Tabular Data. In *2019 IEEE 24th Pacific Rim International Symposium on Dependable Computing (PRDC)*, pages 198–1988.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2021. Tapex: Table pre-training via learning a neural sql executor. *arXiv preprint arXiv:2107.07653*.

Tianyang Liu, Fei Wang, and Muhao Chen. 2024. Rethinking tabular data understanding with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 450–482, Mexico City, Mexico. Association for Computational Linguistics.

Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. 2022a. HybriDialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 481–492, Dublin, Ireland. Association for Computational Linguistics.

Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. 2022b. HybriDialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 481–492, Dublin, Ireland. Association for Computational Linguistics.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev, and Dragomir Radev. 2022. FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E Dataset: New Challenges For End-to-End Generation. *arXiv preprint*. ArXiv:1706.09254 [cs].

Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. 2018. Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10):1071–1083.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. 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 1470–1480, Beijing, China. Association for Computational Linguistics.

Tomas Petricek, Gerrit J. J. van den Burg, Alfredo Nazábal, Taha Ceritli, Ernesto Jiménez-Ruiz, and Christopher K. I. Williams. 2023. Ai assistants: A framework for semi-automated data wrangling. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):9295–9306.

Michał Pietruszka, Michał Turski, Łukasz Borchmann, Tomasz Dwojak, Gabriela Pałka, Karolina Szyndler, Dawid Jurkiewicz, and Łukasz Garncarek. 2022. STable: Table Generation Framework for Encoder-Decoder Models. *arXiv preprint*. ArXiv:2206.04045 [cs].

Christopher Scaffidi, Brad Myers, and Mary Shaw. 2009. Intelligently creating and recommending reusable reformatting rules. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, IUI '09, page 297–306, New York, NY, USA. Association for Computing Machinery.

Nabeel Seedat, Nicolas Huynh, Boris van Breugel, and Mihaela van der Schaar. 2024. Curated LLM: Synergy of LLMs and Data Curation for tabular augmentation in ultra low-data regimes. _eprint: 2312.12112.

Alexey O. Shigarov, Vasiliy V. Khristyuk, Andrey A. Mikhailov, and Viacheslav V. Paramonov. 2019. Tabbyxl: Rule-based spreadsheet data extraction and transformation. In *International Conference on Information and Software Technologies*.

Rishabh Singh and Sumit Gulwani. 2012. Learning semantic string transformations from examples. *arXiv preprint arXiv:1204.6079*.

Aivin V. Solatorio and Olivier Dupriez. 2023. REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers. *arXiv preprint*. ArXiv:2302.02041 [cs].

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 645–654, New York, NY, USA. Association for Computing Machinery.

Anirudh Sundar and Larry Heck. 2022. Multimodal conversational AI: A survey of datasets and approaches. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 131–147, Dublin, Ireland. Association for Computational Linguistics.

Anirudh Sundar, Jin Xu, William Gay, Christopher Richardson, and Larry Heck. 2024. cpapers: A dataset of situated and multimodal interactive conversations in scientific papers. *Advances in Neural Information Processing Systems*, 37:66283–66304.

Anirudh S. Sundar and Larry Heck. 2023. cTBLS: Augmenting large language models with conversational tables. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 59–70, Toronto, Canada. Association for Computational Linguistics.

Xiangru Tang, Yiming Zong, Jason Phang, Yilun Zhao, Wangchunshu Zhou, Arman Cohan, and Mark Gerstein. 2023. Struc-Bench: Are Large Language Models Really Good at Generating Complex Structured Data? *arXiv preprint*. ArXiv:2309.08963 [cs].

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Xueqing Wu, Jiacheng Zhang, and Hang Li. 2022. Text-to-Table: A New Way of Information Extraction. *arXiv preprint*. ArXiv:2109.02707 [cs].

Junjie Xing, Yeye He, Mengyu Zhou, Haoyu Dong, Shi Han, Dongmei Zhang, and Surajit Chaudhuri. 2024. Table-llm-specialist: Language model specialists for tables using iterative generator-validator fine-tuning. *arXiv preprint arXiv:2410.12164*.

Lei Xu and Kalyan Veeramachaneni. 2018. Synthesizing Tabular Data using Generative Adversarial Networks. *arXiv preprint*. ArXiv:1811.11264 [cs, stat].

Jingfeng Yang, Aditya Gupta, Shyam Upadhyay, Luheng He, Rahul Goel, and Shachi Paul. 2022. TableFormer: Robust transformer modeling for table-text encoding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 528–537, Dublin, Ireland. Association for Computational Linguistics.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. *arXiv preprint arXiv:2005.08314*.

Liangyu Zha, Junlin Zhou, Liyao Li, Rui Wang, Qingyi Huang, Saisai Yang, Jing Yuan, Changbao Su, Xiang Li, Aofeng Su, et al. 2023. Tablegpt: Towards unifying tables, nature language and commands into one gpt. *arXiv preprint arXiv:2307.08674*.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2023. Tablellama: Towards open large generalist models for tables. *arXiv preprint arXiv:2311.09206*.

Tianyi Zhang, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600, Dublin, Ireland. Association for Computational Linguistics.

Yilun Zhao, Yitao Long, Hongjun Liu, Linyong Nan, Lyuhao Chen, Ryo Kamoi, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2023. DocMath-Eval: Evaluating Numerical Reasoning Capabilities of LLMs in Understanding Long Documents with Tabular Data. _eprint: 2311.09805.

Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. 2021. Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning*, pages 97–112. PMLR.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Appendix

A.1 AI Assistance Acknowledgment

We acknowledge the use of GitHub Copilot to assist in code completion.

A.2 Compute

All fine-tuning and inference was run on Nvidia A40 GPUs with 48GB GDDR6 memory. Fine-tuning took 1-2 hours on 8 GPUs in parallel with pytorch distributed data parallel (DDP).

A.3 Dataset Statistics

Statistics of the text-to-table datasets:

Dataset	Train	Valid	Test
E2E	42.1k	4.7k	4.7k
WikiTableText	10k	1.3k	2.0k
WikiBio	582.7k	72.8k	72.7k

Table 7: Statistics of the E2E, WikiTableText, WikiBio, and RotoWire datasets, number of samples across splits

A.4 Dataset Examples – Text to Table

This section details example textual paragraphs and associated tables from the different datasets.

E2E:
The Eagle is a low rated coffee shop near Burger King and the riverside that is family friendly and is less than £20 for Japanese food.

Name	The Eagle
Food	Japanese
Price range	Less than £20
Customer Rating	Low
Area	Riverside
Family friendly	Yes
Near	Burger King

WikiTableText:
Michelle Schimel was New York State assemblywoman in Portuguese Heritage Society.

Title	Potuguese Heritage Society
Subtitle	Other activities
Name	Michelle Schimel

WikiBio:
Leonard Shenoff Randle (born February 12, 1949) is a former Major League Baseball player. He was the first-round pick of the Washington Senators in the secondary phase of the June 1970 Major League Baseball draft, tenth overall.

Debut team	Washington Senators
Name	Lenny Randle
Birth Date	12 February 1949

A.5 Dataset Examples – iTBLS

	Example
1	What is the 2nd cell value for row 4?
2	Tell me the final value in the column labeled k
3	What is the value of the cell to the left of the cell in the bottom right of the table.

Table 8: Example interactions in iTBLS *Interpret*

	Example
1	The rows 1 and 4 in the Column “Citation" were accidentally switched. Please rectify the positions of these values so they are where they need to be.
2	Swap the contents of the second and last cell under repetitions.
3	Two values in the MCBLp column were put in the reverse spots. I need the values for the FM and PCC rows flipped.

Table 9: Example interactions in iTBLS *Modify*

	Example
1	The row 3 of the table shows the values for Peak as 4, X coordinate as 0.100, Y coordinate as -0.150, A as 0.5, standard deviation (σ) as 0.02, and Local lnZ as -7.824.
2	The column “Method 2 (with sub-clustering)" contains the ‘Nlike’ values in different rows: 27,658 in the second row, 69,094 in the third row, 579,208 in the fourth row, and 43,093,230 in the fifth row, while the remaining rows from six to nine contain no data (NaN).
3	The column R contains eight numerical values in increasing order: 3.34, 3.40, 3.66, 5.06, 6.02, 6.61, 4.05, and 4.11.

Table 10: Example interactions in iTBLS *Generate*

A.6 Hyperparameters

Hyperparameters used during training are listed here.

Parameter	Value
Rank	2
α	2
Dropout	0.01
Target modules	`all-linear`

Table 11: LoRA Hyperparameters

Parameter	Value
Learning Rate	2e-4
Batch size	4
Warmup Schedule	Constant
Warmup Ratio	0.03
Epochs	5
Optimizer	`paged_adamw_32bit`[3]

Table 12: Training Hyperparameters

A.7 Mechanical Turk Interface

B Example error on the generate task of iTBLS

[3] https://huggingface.co/docs/bitsandbytes/main/en/reference/optim/adamw

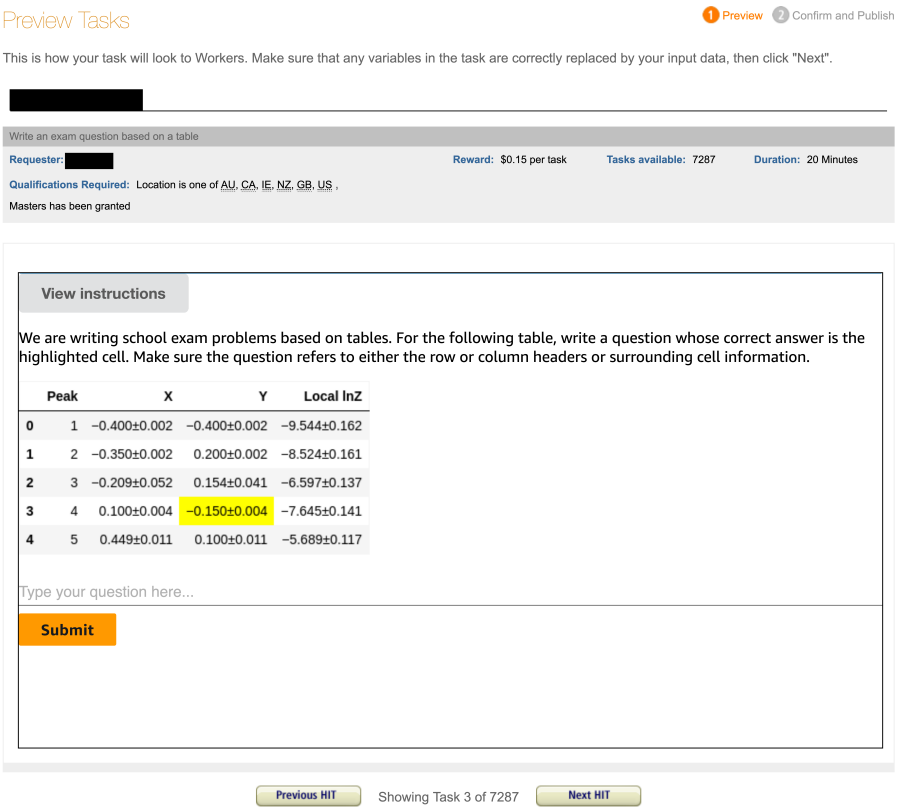

Figure 3: Amazon Mechanical Turk Interface to collect iTBLS interpretation

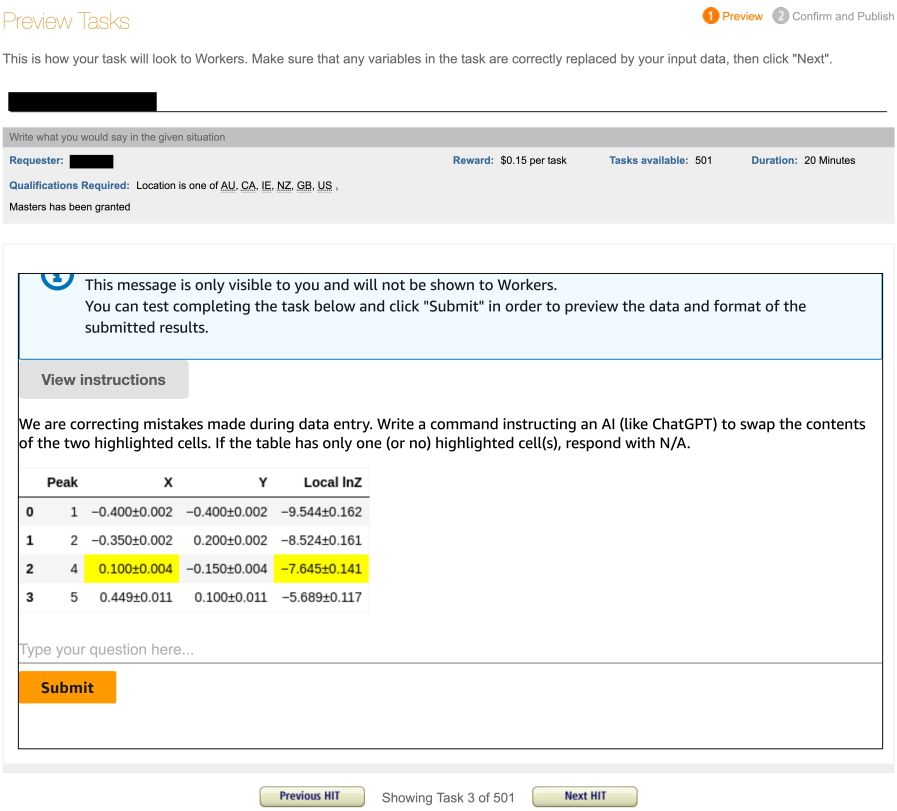

Figure 4: Amazon Mechanical Turk Interface to collect iTBLS modification

Text: The column 'Standard deviation' contains entries which are both numbers and number sequences: first has 0.45, 0.75, 0, 0.57, second has 0.36, 0.5, 0, 0.34, third is exactly 0, the fourth one is 0.77, fifth is 0.49, and the last one is 0.22.

Input Table:

row ID	Questions	Average score
0	Q. 1 (a-d)	(3.6 3.93 5 4)
1	Q. 2 (a-d)	4.26
2	Q. 3	5
3	Q. 4	3.64
4	Q. 5	(4.04 4.44 5 4.86)
5	GQ	4.35

Ground Truth:

row ID	Questions	Average score	Standard deviation
0	Q. 1 (a-d)	(3.6 3.93 5 4)	(0.45 0.75 0 0.57)
1	Q. 2 (a-d)	4.26	(0.36 0.5 0 0.34)
2	Q. 3	5	0
3	Q. 4	3.64	0.77
4	Q. 5	(4.04 4.44 5 4.86)	0.49
5	GQ	4.35	0.22

Prediction:

row ID	Questions	Average score	Standard deviation
0	Q. 1 (a-d)	(3.6 3.93 5 4)	(0.45 0.75 0 0.57)
1	Q. 2 (a-d)	4.26	(0.36 0.5 0 0.34)
2	Q. 3	5	0
3	Q. 4	3.64	0.77
4	Q. 5	(4.04 4.44 5 4.86)	(0.49 0.22)
5	GQ	4.35	0

Table 13: Example error for iTBLS generate task. Table source: https://arxiv.org/pdf/1411.4925