OpenT2T: An Open-Source Toolkit for Table-to-Text Generation

Haowei Zhang[*♣] **Shengyun Si**[*♣] **Yilun Zhao**[*♠] **Lujing Xie**[♠] **Zhijian Xu**[♠]
Lyuhao Chen[◇] **Linyong Nan**[♠] **Pengcheng Wang**[♠] **Xiangru Tang**[♠] **Arman Cohan**[♠♡]

[♠]Yale University [♣]Technical University of Munich
[◇]Carnegie Mellon University [♡]Allen Institute for AI

https://github.com/yale-nlp/OpenT2T

Abstract

Table data is pervasive in various industries, and its comprehension and manipulation demand significant time and effort for users seeking to extract relevant information. Consequently, an increasing number of studies have been directed towards table-to-text generation tasks. However, most existing methods are benchmarked solely on a limited number of datasets with varying configurations, leading to a lack of unified, standardized, fair, and comprehensive comparison between methods. To bridge this gap, this paper presents OpenT2T, the first open-source toolkit for table-to-text generation tasks, designed to reproduce existing table-to-text generation systems for performance comparison and expedite the development of new models. We have implemented and compared a wide range of large language models under zero- and few-shot settings on nine table-to-text generation datasets, covering the tasks of data insight generation, table summarization, and free-form table question answering. Additionally, we maintain a public leaderboard to provide insights for future work into how to choose appropriate table-to-text generation systems for real-world scenarios.

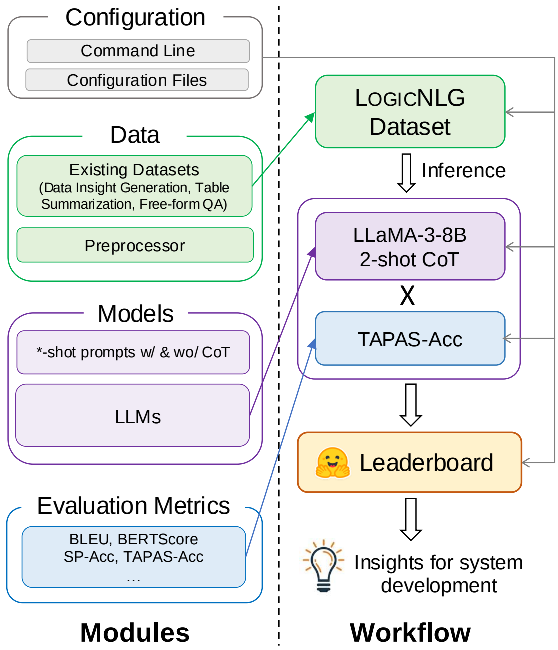

Figure 1: The overall framework of OpenT2T.

1 Introduction

In an era where users interact with vast amounts of structured data every day for decision-making and information-seeking purposes, the need for intuitive, user-friendly interpretations has become paramount (Zhang et al., 2023; Zha et al., 2023; Li et al., 2023; Zhao et al., 2023e). Given this emerging necessity, table-to-text generation techniques, which transform complex tabular data into comprehensible narratives tailored to users' information needs, have drawn considerable attention (Parikh et al., 2020; Chen et al., 2020b; Nan et al., 2022b; Zhao et al., 2024b,c). These techniques can be incorporated into a broad range of applications, including but not limited to game strategy development, financial analysis, and human resources management.

While large language models (LLMs) have achieved remarkable progress in the areas of controllable text generation and data interpretation (Nan et al., 2021; Zhao et al., 2022; Gao et al., 2023; Madaan et al., 2023; Zhou et al., 2023; Zhao et al., 2024a), the exploration of these models in table-to-text generation has been limited. Additionally, existing table-to-text generation systems (Liu et al., 2022b; Jiang et al., 2022; Zhao et al., 2022; Liu et al., 2022a; Nan et al., 2022a) are benchmarked on various datasets and configurations. This has led to a lack of standardization, making comprehensive evaluation between different methods challenging. Moreover, since these models are developed or evaluated within individual systems, they suffer from compatibility issues. Therefore,

*Equal Contribution

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing System Demonstrations, pages 259–269
November 12-16, 2024 ©2024 Association for Computational Linguistics

Dataset	# Examples	# Tables	Control Signal	Output
		Data Insight Generation		
LogicNLG (Chen et al., 2020a)	37,015	7,392	Highlighted columns	Single-sentence statement
ToTTo (Parikh et al., 2020)	136,161	83,141	Highlighted cells	Single-sentence statement
HiTab_{NG} (Cheng et al., 2021)	10,672	3,597	Highlighted cells	Single-sentence statement
		Table Summarization		
RotoWire (Wiseman et al., 2017)	4,953	4,953	–	Paragraph-long summary
NumericNLG (Suadaa et al., 2021)	1,355	1,355	–	Paragraph-long summary
SciGen (Moosavi et al., 2021)	1,338	1,338	–	Paragraph-long summary
		Free-form Table Question Answering		
FeTaQA (Nan et al., 2022b)	10,330	10,330	Question	Single-sentence answer
HiTab_{QA} (Cheng et al., 2021)	10,672	3,597	Question	Single-sentence answer
QTSumm (Zhao et al., 2023c)	5,625	2,437	Question	Paragraph-long answer

Table 1: An overview of table-to-text generation tasks included in OpenT2T.

reproducing them for result comparison in future studies is both difficult and time-consuming. Given that the above issues are serious hindrances to the development of table-to-text generation systems, there is an imperative need to develop a unified and extensible open-source toolkit.

In this paper, we present **OpenT2T**, the first **Open**-source toolkit for **T**able-**to**-**T**ext generation. OpenT2T features the following three key characteristics:

- **Modularization** We develop OpenT2T with highly reusable modules and integrated them in a unified framework. This enables future researchers to study various table-to-text generation systems at a conceptual level.
- **Standardization** OpenT2T includes popular table-to-text generation datasets and models. The evaluation of different models is standardized. We have also created a public leaderboard to evaluate and rank the performance of various methods on different datasets, providing insights into how to choose appropriate table-to-text generation systems for real-world scenarios.
- **Extensibility** OpenT2T enables researchers to easily develop custom prompts for LLMs. Additionally, they can extend the data or LLM inference modules to integrate new table-to-text generation datasets or systems.

The main structure of the paper is organized as follows: Section 2 describes each table-to-text generation task included in the OpenT2T framework. Section 3 describes each module and its implementation of OpenT2T framework. Section 4 introduces the maintained public OpenT2T leaderboard and highlights the main findings based on the results from the leaderboard. These insights help guide the selection of appropriate table-to-text generation systems for real-world needs. Finally, Section 5 discusses the related work and compares OpenT2T with existing open-source toolkits for the table-relevant tasks.

2 OpenT2T Tasks

OpenT2T covers three kinds of table-to-text generation tasks: *data insight generation*, *table summarization*, and *free-form table question answering* (as shown in Table 1). The goal of OpenT2T is to push the development of table-to-text generation systems that can be applied and achieved competitive performance on various real-world scenarios. Such advancement could significantly enhance table data interpretation across industries, making complex tabular information more accessible and actionable for non-expert users. Due to computational constraints, we randomly sample 300 examples from each benchmark. If the test set ground truth is available, we select examples from the test set; otherwise, we use the validation set. The following subsections provide a detailed description of each type of table-to-text generation task and the corresponding datasets included in OpenT2T.

2.1 Data Insight Generation

Data insight generation involves generating meaningful and relevant insights from tables. Such techniques free users from manually combing through vast amounts of tabular data. We include the following three relevant datasets in OpenT2T:

- **LogicNLG** (Chen et al., 2020a) necessitates models to generate multiple statements that perform logical reasoning based on the information in the source table. Each statement should be factually correct with the table content.

- **ToTTo** (Parikh et al., 2020) requires models to provide faithful statements from Wikipedia tables. The generation of statements should be controlled by corresponding highlighted cells.

- **HiTab$_{\mathbf{NG}}$** (Cheng et al., 2021) consists of cross-domain tables from plenty of statistical reports and Wikipedia pages. It requires models to produce statements from complex hierarchical tables and highlighted cells, which needs numerical and semantic reasoning analysis.

2.2 Table Summarization

Table summarization techniques condense the information contained in a table into a more accessible and concise form. By creating a summary that captures the key information and patterns, users can quickly grasp the main insights from the data without having to explore every individual entry. This complements the process of data insight generation, providing a streamlined way to interpret and utilize large datasets. We include the following three table summarization datasets in OpenT2T:

- **RotoWire** (Wiseman et al., 2017) tasks models with generating coherent and natural-language summaries that accurately capture and convey the statistical information presented in NBA game tables.

- **NumericNLG** (Suadaa et al., 2021) necessitates models to generate summaries with high fidelity and fluency based on tables from scientific papers. The generation framework emphasizes rich arithmetic reasoning.

- **SciGen** (Moosavi et al., 2021) demands models to provide summaries in accordance with complex tables containing numerical values from scientific papers. It places significant emphasis on arithmetic reasoning capability.

2.3 Free-form Table Question Answering

Table QA involves interpreting and analyzing tables to answer user queries. Unlike short-form QA, which typically requires concise and specific questions for retrieving direct answers, free-form table QA allows users to ask more complex and nuanced questions about tabular data. This approach facilitates a deeper exploration of the data and offers a more flexible and comprehensive way to interact with complex tables. We include the following three relevant datasets in OpenT2T:

- **FeTaQA** (Nan et al., 2022c) tasks models with generating single-sentence answers after retrieving, inferring, and integrating multiple supporting facts from the source table.

- **HiTab$_{\mathbf{QA}}$** (Cheng et al., 2021) requires models to generate answers from complex hierarchical tables and questions, involving both numerical and semantic reasoning. The hierarchical structure demands advanced analysis to interpret relationships, perform mathmatical calculations, and derive accurate final answers.

- **QTSumm** (Zhao et al., 2023c) requires models to produce query-focused, paragraph-long answers based on tables sourced from Wikipedia. The questions cover a wide range of topics, demanding a precise and contextually relevant synthesis of information from the table, with emphasis on addressing the query directly.

3 OpenT2T Framework

As shown in Figure 1, OpenT2T consists of four main modules: configuration, data, modeling, and evaluation. The users are able to test the existing table-to-text models on the included dataset. They are also allowed to add their own models or datasets into OpenT2T by extending corresponding modules with their proposed ones.

3.1 Configuration Module

The configuration module allows users and developers to specify all experiment settings. Users are expected to modify the main arguments of the experiment settings in external configuration files or command lines while leaving the internal configuration unchanged for existing models. This approach ensures a unified performance comparison among different models on table-to-text tasks.

3.2 Data Module

As discussed in Section 2, OpenT2T includes popular datasets for table reasoning, which cover various types of tasks. The data module converts raw datasets in various formats into a unified format, which consists of the following five essential arguments:

- `table`: Table headers and contents in a 2D array format.
- `title`: The title of the table.
- `question`: The question or query about the table. If no question is provided in the raw dataset, this argument will be set to None.
- `reference`: Reference output of the table.
- `linked columns`: The indices of the table columns related to the reference output. If no linked columns are provided in the raw dataset, this argument will be set to the indices of all columns in the table.
- `highlighted cells`: The indices of the cells in the table related to the reference output. If no highlighted cells are provided in the raw dataset, this argument will be set to the indices of all cells in the table.

We apply the same strategy as Liu et al. (2022b) for truncating a long table into a shorter version to satisfy the model's input length limit. It worth noting that the processed and format-unified data can be used as model input for both the modeling module and the evaluation module. To enhance adaptability, we design the data module with extensibility in mind, allowing future users to easily incorporate new datasets. By creating subclasses that inherit from the implemented parent classes, users can add datasets with minimal adjustments. We acknowledge the recent release of table-to-text generation benchmarks (Zhang et al., 2024b) that are not currently included in OPENT2T and encourage future researchers to contribute to the growth of OPENT2T by incorporating these benchmarks.

3.3 LLM Inference Module

For the evaluation of LLMs, we provide prompts with zero-, one-, and two-shots, both with and without chain-of-thought (CoT) reasoning prompt (Wei et al., 2022; Chen, 2022), for each dataset. We have streamlined and standardized the inference of the following LLMs using a parent interface class named `LLM_T2TModel`:

- **General**: GPT-3.5&4&4o (OpenAI, 2022, 2023, 2024), Claude-3.5 (Anthropic, 2024), Llama-2&3&3.1 (Touvron et al., 2023), Mistral (Jiang et al., 2023), Phi-3&3.5 (Abdin et al., 2024), Gemma-2 (Team et al., 2024), WizardLM-2 (Xu et al., 2023), Yi-1.5 (01.AI, 2023), Qwen-2&2.5 (Bai et al., 2023), Command R+ (Cohere, 2024b), Aya (Cohere, 2024a), and GLM-4 (GLM et al., 2024).
- **Math-specific**: WizardMath (Luo et al., 2023), DeepSeek-Math (Shao et al., 2024), and InternLM-Math (Ying et al., 2024). We evaluate math-specific LLMs because some T2T datasets, such as FeTaQA and SciGen, require mathematical reasoning to generate faithful responses.
- **Code-based**: Codestral (AI@Mistral, 2024), DeepSeek-Coder-V2 (also MoE architecture, DeepSeek-AI (2024)), and StarCoder2 (Lozhkov et al., 2024). We evaluate code-based LLMs because recent studies (Zhang et al., 2024a) have shown that training on code generation data can enhance model performance on tasks requiring table reasoning.
- **Mixture of Experts (MoE)**: Mixtral (Mistral.AI, 2023), WizardLM-2 (MoE, Xu et al. (2023)), and DeepSeek-V2 (DeepSeek-AI, 2024).

We encourage future research to evaluate and include their newly-developed LLMs, especially those designed for table-related tasks (Zhang et al., 2024a; Zheng et al., 2024), into our public leaderboard, which will be detailed in Section 4.

3.4 Evaluation Module

To evaluate and compare the performance of table reasoning models supported by a certain dataset, OPENT2T includes all the evaluation metrics used in the official implementation. These metrics can be used off-the-shelf with a one-line call, given a prediction output file and the name of the dataset. The uniformly formatted reference file generated in 3.2 can be automatically found and put to use by the module without any manual format adaption of the dataset to specific metrics. The details of each metric are introduced as follows:

- **BLEU** (Papineni et al., 2002) employs a precision-based method, measuring how the n-gram matches between the prediction and reference statements.
- **ROUGE** (Lin, 2004) applies a recall-based approach, measuring the proportions of overlapping words and phrases between the generated prediction and the reference.

- **METEOR** (Lavie and Agarwal, 2007) is based on the harmonic mean of unigram precision and recall, with several unique features like stemming and synonymy matching. This metric addresses some issues present in the BLEU metric and maintains a strong correlation with human evaluations at the sentence or segment level.
- **BERTScore** (Zhang et al., 2020) computes the similarity between the reference and generated summary using contextual word embeddings.
- **BLEURT** (Sellam et al., 2020) is a BERT-based metric for text generation tasks that can be pre-trained and fine-tuned with manually evaluated data to satisfy both the robustness and expressiveness of the metric.
- **AutoACU** (Liu et al., 2023) introduces a reference-based automated evaluation framework that leverages atomic content units (ACUs) to assess the degree of similarity between textual sequences. The framework is designed to offer more interpretable and fine-grained evaluations by breaking down text into ACUs, which are smaller units representing meaningful content.

We also include following two model-based metrics specifically designed for the faithfulness-level evaluation:

- **TAPAS-Acc** (Herzig et al., 2020) employs the TAPAS model (Herzig et al., 2020) fine-tuned on TABFACT (Chen et al., 2020c) dataset to judge whether the generated statements are entailed or refuted based on the table content.
- **TAPEX-Acc** (Liu et al., 2022b) uses TAPEX, fine-tuned on the TABFACT (Chen et al., 2020c) dataset, to assess whether generated statements are entailed or refuted. Recent studies (Liu et al., 2022a; Wang et al., 2024) have demonstrated that both NLI-Acc (Chen et al., 2020b) and TAPAS-Acc tend to overestimate the accuracy of predictions, whereas TAPEX-Acc has proven to be a more reliable metric for evaluating faithfulness.

3.5 Execution

For running and evaluating LLMs using OPENT2T, users can utilize and modify the provided zero- and few-shot prompts for LLM inference. Users also have the ability to evaluate existing or new LLMs on their newly-added datasets.

4 OPENT2T Leaderboard

We maintain a public leaderboard at HuggingFace Space for users to track, rank, and evaluate existing table-to-text generation systems. The detailed results of model performance can be found at `https://huggingface.co/spaces/yale-nlp/OpenT2T_Leaderboard`. Users can also submit model output for automated evaluation and leaderboard updates. We believe that such a leaderboard can provide future researchers and developers with valuable insights into how to choose and develop appropriate table-to-text generation systems for real-world applications.

4.1 Expertiment Setup

The experiments for open-sourced LLMs were conducted using the `vLLM` framework (Kwon et al., 2023). For all the experiments, we set temperature as 1.0, Top P as 1.0, and maximum output length as 512, without any frequency or presence penalty for all LLMs. We access the proprietary models through their official APIs and run all other open-source models locally on our servers with NVIDIA A100 80GiB.

4.2 Main Findings

Based on the leaderboard results, we derive the following key findings.

Data Insight Generation The current top-performing proprietary models generally surpass open-source ones in data insight generation, demonstrating their strong capability to generate faithful statements from tables. Among open-source models, Llama- and Qwen-series models achieve most competitive performance.

Free-form Table Question Answering Both open-sourced LLMs and GPT-* models in a 2-shot setting achieve comparable performance. Moreover, increasing the number of shots and applying the CoT approach can both yield performance gains for table question answering. This finding points to the adaptability of these models to different input formats and their ability to leverage more context or structured reasoning to enhance performance.

Table Summarization GPT-* models in a 2-shot setting achieve best performance. However, other open-sourced LLMs still struggle with this type of task. For table summarization, we also observe that either increasing the number of shots or applying the CoT reasoning approach can generally improve LLM performance. These findings suggest that although GPT-* models excel in summarization, there is potential for improving the training methodologies of other open-source LLMs to better manage the complexities involved in the table summarization tasks.

Open-sourced LLMs vs GPT There remains a significant performance gap between other open-sourced LLMs (e.g., Mistral-Large and LLama-3.1) and GPT-* models. This gap highlights the potential for further development and innovation in open-sourced LLMs to bridge this disparity. Furthermore, among open-sourced LLMs, TableLlama demonstrates a notable improvement over its backbone (i.e., Llama-2), emphasizing the effectiveness of enhancing table-to-text generation capabilities through instruction tuning on tabular data. This advancement also underscores the potential for significant gains in open-source models through targeted modifications and optimizations, which could lead to more competitive alternatives to proprietary models in the future.

5 Related Work

Text generation from semi-structured knowledge sources, such as web tables, has been studied extensively in recent years (Parikh et al., 2020; Chen et al., 2020b; Cheng et al., 2022). However, existing table-to-text methods (Liu et al., 2022b; Jiang et al., 2022; Liu et al., 2022a; Zhao et al., 2023b, 2024a) have been evaluated on different datasets with varying configurations and developed as individual systems, resulting in difficulties in reproducing them for performance comparison in future studies. Moreover, existing works typically regard table-to-text generation as a subtask of table reasoning (Zhao et al., 2023d; Zhang et al., 2024a; Deng et al., 2024; Zheng et al., 2024; Wu et al., 2024), which focuses primarily on numerical and logical reasoning capabilities. The table-to-text generation tasks, however, go beyond these reasoning aspects and also require the model to accurately convey information from the table in a way that is both contextually appropriate and easily understandable to the target audience.

More recently, Zhao et al. (2023a) developed an open-source toolkit for table reasoning. However, it only implement one table-to-text generation dataset (i.e., LOGICNLG) and does not include LLMs, while OPENT2T include nine datasets covering three real-world table information-seeking scenarios. Kasner et al. (2023) provides a visualization interface for researchers to explore various table-to-text generation datasets. In contrast, OPENT2T offers standardized and comprehensive evaluation benchmarks for performance comparison, enabling users to choose the appropriate table pre-training model for specific real-world needs.

6 Conclusion

This work presents OPENT2T, the first open-source framework for table-to-text generation, aimed at enabling researchers and developers to reproduce and benchmark existing table-to-text generation systems in a standardized and fair manner. OPENT2T serves as a comprehensive platform that allows users to compare different models on a unified ground, facilitating more transparent and reproducible research in this area. The framework is developed with highly reusable and modular components, making it flexible and extensible for a wide range of use cases. Additionally, OPENT2T provides a suite of pre-built functionalities, including data preprocessing pipelines and evaluation metrics, which streamline the process of testing and evaluating new models. We welcome researchers and engineers to join us in developing, maintaining, and improving OPENT2T, in order to foster innovation and enable the rapid development of novel table-to-text generation techniques.

Ethical Consideration

The datasets included in OPENT2T all use licenses that permit us to compile, modify, and publish the original datasets. OPENT2T are also publically avaliable with the license BSD-2-Clause[1], which allows users to modify and redistribute the source code while retaining the original copyright.

Acknowledgements

We would like to dedicate this paper to the memory of Dr. Dragomir Radev. Dr. Radev provided invaluable feedback during the early stages of our project brainstorming and development. His passing is deeply felt by all of us. We extend our heartfelt gratitude for his passion, dedication, and lasting contributions to the entire NLP community.

We are also grateful for the compute support provided by Microsoft Research's Accelerate Foundation Models Research (AFMR) program.

[1] https://opensource.org/license/bsd-2-clause/

References

01.AI. 2023. Yi: Open-source llm release.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

AI@Mistral. 2024. Codestral: Hello, world!

Anthropic. 2024. Introducing the next generation of claude.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Wenhu Chen. 2022. Large language models are few(1)-shot table reasoners.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020a. Logical natural language generation from open-domain tables. *arXiv preprint arXiv:2004.10404*.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020b. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942, Online. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020c. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2021. Hitab: A hierarchical table dataset for question answering and natural language generation. *arXiv preprint arXiv:2108.06712*.

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.

Cohere. 2024a. Cohere for ai launches aya 23, 8 and 35 billion parameter open weights release.

Cohere. 2024b. Introducing command r+: A scalable llm built for business.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.

Naihao Deng, Zhenjie Sun, Ruiqi He, Aman Sikka, Yulong Chen, Lin Ma, Yue Zhang, and Rada Mihalcea. 2024. Tables as texts or images: Evaluating the table reasoning ability of LLMs and MLLMs. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 407–426, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

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.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhengbao Jiang, Yi Mao, Pengcheng He, Graham Neubig, and Weizhu Chen. 2022. OmniTab: Pretraining with natural and synthetic data for few-shot table-based question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 932–942, Seattle, United States. Association for Computational Linguistics.

Zdeněk Kasner, Ekaterina Garanina, Ondrej Platek, and Ondrej Dusek. 2023. TabGenie: A toolkit for table-to-text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 444–455, Toronto, Canada. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Hongxin Li, Jingran Su, Yuntao Chen, Qing Li, and Zhaoxiang Zhang. 2023. Sheetcopilot: Bringing software productivity to the next level through large language models. *ArXiv*, abs/2305.19308.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ao Liu, Haoyu Dong, Naoaki Okazaki, Shi Han, and Dongmei Zhang. 2022a. PLOG: Table-to-logic pretraining for logical table-to-text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5531–5546, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022b. TAPEX: Table pre-training via learning a neural SQL executor. In *International Conference on Learning Representations*.

Yixin Liu, Alexander Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Towards interpretable and efficient automatic reference-based summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16360–16368, Singapore. Association for Computational Linguistics.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, Zhuang Li, Wen-Ding Li, Megan Risdal, Jia Li, Jian Zhu, Terry Yue Zhuo, Evgenii Zheltonozhskii, Nii Osae Osae Dade, Wenhao Yu, Lucas Krauß, Naman Jain, Yixuan Su, Xuanli He, Manan Dey, Edoardo Abati, Yekun Chai, Niklas Muennighoff, Xiangru Tang, Muhtasham Oblokulov, Christopher Akiki, Marc Marone, Chenghao Mou, Mayank Mishra, Alex Gu, Binyuan Hui, Tri Dao, Armel Zebaze, Olivier Dehaene, Nicolas Patry, Canwen Xu, Julian McAuley, Han Hu, Torsten Scholak, Sebastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, Mostofa Patwary, Nima Tajbakhsh, Yacine Jernite, Carlos Muñoz Ferrandis, Lingming Zhang, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2024. Starcoder 2 and the stack v2: The next generation.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Mistral.AI. 2023. Mixtral of experts: A high quality sparse mixture-of-experts.

Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. Scigen: a dataset for reasoning-aware text generation from scientific tables. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Linyong Nan, Lorenzo Jaime Flores, Yilun Zhao, Yixin Liu, Luke Benson, Weijin Zou, and Dragomir Radev. 2022a. R2D2: Robust data-to-text with replacement detection. In *Proceedings of the 2022 Conference on*

Empirical Methods in Natural Language Processing, pages 6903–6917, Abu Dhabi, United Arab Emirates. 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. 2022b. FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

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. 2022c. FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

OpenAI. 2024. Hello gpt-4o.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards table-to-text generation with numerical reasoning. 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 1451–1465, Online. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin

Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Yuqi Wang, Lyuhao Chen, Songcheng Cai, Zhijian Xu, and Yilun Zhao. 2024. Revisiting automated evaluation for long-form table question answering in the era of large language models. In *The 2024 Conference on Empirical Methods in Natural Language Processing*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Guanglin Niu, Tongliang Li, and Zhoujun Li. 2024. Tablebench: A comprehensive and complex benchmark for table question answering.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions.

Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, Yudong Wang, Zijian Wu, Shuaibin Li, Fengzhe Zhou, Hongwei Liu, Songyang Zhang, Wenwei Zhang, Hang Yan, Xipeng Qiu, Jiayu Wang, Kai Chen, and Dahua Lin. 2024. Internlm-math: Open math large language models toward verifiable reasoning.

Liangyu Zha, Junlin Zhou, Liyao Li, Rui Wang, Qingyi Huang, Saisai Yang, Jing Yuan, Changbao Su, Xiang Li, Aofeng Su, Tao Zhang, Chen Zhou, Kaizhe Shou, Miao Wang, Wufang Zhu, Guoshan Lu, Chao Ye, Yali Ye, Wentao Ye, Yiming Zhang, Xinglong Deng, Jie Xu, Haobo Wang, Gang Chen, and Junbo Zhao. 2023. Tablegpt: Towards unifying tables, nature language and commands into one gpt.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2024a. TableLlama: Towards open large generalist models for tables. 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 6024–6044, Mexico City, Mexico. Association for Computational Linguistics.

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*.

Weijia Zhang, Vaishali Pal, Jia-Hong Huang, Evangelos Kanoulas, and Maarten de Rijke. 2024b. Qfmts: Generating query-focused summaries over multi-table inputs.

Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yue Ting Zhuang. 2023. Data-copilot: Bridging billions of data and humans with autonomous workflow. *ArXiv*, abs/2306.07209.

Yilun Zhao, Lyuhao Chen, Arman Cohan, and Chen Zhao. 2024a. TaPERA: Enhancing faithfulness and interpretability in long-form table QA by content planning and execution-based reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12824–12840, Bangkok, Thailand. Association for Computational Linguistics.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024b. Financemath: Knowledge-intensive math reasoning in finance domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12841–12858, Bangkok, Thailand. Association for Computational Linguistics.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2024c. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16103–16120, Bangkok, Thailand. Association for Computational Linguistics.

Yilun Zhao, Boyu Mi, Zhenting Qi, Linyong Nan, Minghao Guo, Arman Cohan, and Dragomir Radev. 2023a. OpenRT: An open-source framework for reasoning over tabular data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 336–347, Toronto, Canada. Association for Computational Linguistics.

Yilun Zhao, Linyong Nan, Zhenting Qi, Rui Zhang, and Dragomir Radev. 2022. ReasTAP: Injecting table reasoning skills during pre-training via synthetic reasoning examples. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language*

Processing, pages 9006–9018, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yilun Zhao, Zhenting Qi, Linyong Nan, Lorenzo Jaime Flores, and Dragomir Radev. 2023b. Loft: Enhancing faithfulness and diversity for table-to-text generation via logic form control. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics.

Yilun Zhao, Zhenting Qi, Linyong Nan, Boyu Mi, Yixin Liu, Weijin Zou, Simeng Han, Xiangru Tang, Yumo Xu, Arman Cohan, and Dragomir Radev. 2023c. Qtsumm: A new benchmark for query-focused table summarization.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023d. Large language models are effective table-to-text generators, evaluators, and feedback providers.

Yilun Zhao, Chen Zhao, Linyong Nan, Zhenting Qi, Wenlin Zhang, Xiangru Tang, Boyu Mi, and Dragomir Radev. 2023e. RobuT: A systematic study of table QA robustness against human-annotated adversarial perturbations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6064–6081, Toronto, Canada. Association for Computational Linguistics.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9102–9124, Bangkok, Thailand. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.