Revisiting Automated Evaluation for Long-form Table Question Answering in the Era of Large Language Models

Anonymous ACL submission

Abstract

In the era of data-driven decision-making, Long-Form Table Question Answering (LFTQA) is essential for integrating structured data with complex reasoning. Despite recent advancements in Large Language Models (LLMs) for LFTQA, evaluating their effectiveness remains a significant challenge. We introduce LFTQA-Eval, a meta-evaluation dataset comprising 6,400 human-annotated examples, to rigorously assess the efficacy of current automated metrics in assessing LLM-based LFTQA systems, with a focus on faithfulness and comprehensiveness. Our findings reveal that existing automatic metrics poorly correlate with human judgments and fail to consistently differentiate between factually accurate responses and those that are coherent but factually incorrect. Additionally, our in-depth examination of the limitations associated with automated evaluation methods provides essential insights for the improvement of LFTQA automated evaluation.

1 Introduction

In the current landscape where decisions are increasingly driven by data, the utility of tabular data provides a well-organized and efficient means of presenting data, which is essential for informed decision-making processes (Pasupat and Liang, 2015; Zhu et al., 2021; Zhao et al., 2022a,b; Tang et al., 2023; Zhao et al., 2023a). Within this context, long-form table question answering (LFTQA) has emerged as a vibrant area of research, bridging the gap between structured data and the comprehensive insights required in real-world scenarios (Nan et al., 2022; Zhao et al., 2023b). As illustrated in Figure 1, given the complex question and numerous data points in a table, an LFTQA system must understand the relationships within the data and perform human-like reasoning over the tabular content to compose the paragraph-long answer.

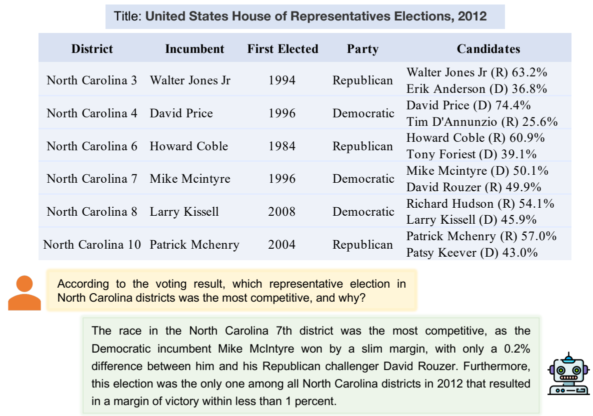

Figure 1: An example of the Long-form Table Question Answering (LFTQA) task investigated in our work.

Recent studies highlight the exceptional performance of Large Language Models (LLMs) in LFTQA tasks (Zhao et al., 2023c; Chen, 2023; Ye et al., 2023). However, the reliable evaluation of LLM-based systems in this domain remains a relatively unexplored area. Unlike conventional text generation tasks, where automatic metrics such as BLEU and ROUGE can somewhat effectively gauge the fluency and surface-level coherence of the generated text, LFTQA demands a more nuanced assessment approach. These traditional metrics, primarily designed for shorter texts, often fall short in LFTQA where the answers not only need to be contextually rich and structurally complex but also deeply rooted in logical reasoning derived from the underlying tabular data. They struggle to evaluate the logical structure and reasoning accuracy essential for long-form responses, as they do not account for the correctness of data interpretation or the ability to maintain a coherent argument over extended narratives. This limitation significantly impacts their utility in scenarios where the decision-making process relies heavily on the accurate and logical processing of structured data, necessitating the development of new metrics that can more effectively measure these critical aspects.

Our research demonstrates that existing automatic metrics are inadequate in distinguishing between high-quality, factually accurate answers and those that are merely coherent. This discrepancy is alarming because developers might choose suboptimal systems for real-world applications if they rely solely on automatic metrics to compare and rank different LFTQA systems. To better investigate the automated evaluation methods for LFTQA tasks, we have constructed a meta-evaluation dataset named **LFTQA-Eval**,[1] consisting of 6,400 human-annotated examples. Specifically, we gathered outputs from leading LLM-based systems on the FeTaQA (Nan et al., 2022) and QTSumm (Zhao et al., 2023b) datasets. We then benchmarked existing automatic evaluation metrics for these tasks, leveraging our collected human annotations across two distinct dimensions: faithfulness and comprehensiveness. Our experimental results demonstrate that all the examined automated metrics exhibit low correlations with human judgments, revealing their unreliability in determining the quality of LLM-generated answers and comparing different LLM-based systems. Moreover, we conducted an in-depth analysis of the failures associated with automated evaluation methods, supplemented by illustrative examples that provide valuable insights into potential areas for enhancement.

2 LFTQA-Eval Construction

To better investigate the automated evaluation methods for LFTQA tasks, we have constructed a meta-evaluation dataset named LFTQA-Eval. The following subsections discuss the data collection methodology and annotation process.

2.1 Collecting LLM Output for LFTQA

We examine LFTQA automated evaluation methods on the FeTaQA and QTSumm datasets. Table 3 in Appendix illustrates the basic data statistics of these two datasets.

- **FeTaQA** (Nan et al., 2022) is designed for free-form table question answering, with answers averaging 18.9 words. It requires models to extract question-relevant information from the given table, and then aggregate and reason over this information to produce a coherent long-form answer.
- **QTSumm** (Zhao et al., 2023b) requires models to perform reasoning and analysis akin to human thought processes on tables sourced from Wikipedia to produce paragraph-length answers. Compared to the FeTaQA dataset, outputs in QTSumm are longer, averaging 68.0 words.

Collecting LLM Output We adopt *Zero-shot*, *One-shot*, *Chain-of-Thought* (Wei et al., 2022), and *Dater* (Ye et al., 2023) prompting methods for LFTQA-Eval construction, with details of each discussed in Appendix A.1. For each prompting method, we collect output from eight popular LLMs, including Llama-2&3 (Touvron et al., 2023), Mistral (Jiang et al., 2023a), Mixtral (Mistral.AI, 2023), DeepSeek (DeepSeek, 2023), Gemini-1.5 (Google, 2023), GPT-3.5&4o (Brown et al., 2020; OpenAI, 2023). We use `chat` or `instruct` versions for each model. Additionally, we select the most recent, largest, and best-performing checkpoint available as of paper submission (i.e, June 10, 2024). We randomly sample 100 examples from the development sets of FeTaQA and QTSumm, and collect corresponding model outputs of these sampled examples. This results in a total of 2 datasets $\times$ 100 examples $\times$ 4 prompting methods $\times$ 8 LLMs = **6,400** examples within the LFTQA-Eval benchmark.

2.2 Evaluation Criteria

The automated evaluation of LFTQA tasks is challenging due to the unique features of LFTQA: 1) conducting intricate reasoning across multiple sources of information, and 2) ensuring factual accuracy while avoiding hallucination. To evaluate the reliability of automated evaluation methods for LFTQA, we collect human evaluation scores for each model output based on the the dimensions of **Faithfulness** and **Comprehensiveness**, respectively. Our preliminary study indicates that LLM-based systems exhibit the capability to generate texts that are both fluent and coherent, devoid of spelling and grammatical errors. Therefore, we have excluded the evaluation of fluency and coherence from our analysis.

- **Faithfulness**: A good answer should be firmly rooted in the source table. It should consist of correct information from the table and precisely address the posed question, avoiding any inaccuracies or hallucinations.
- **Comprehensiveness**: A good answer should encompass all essential information derived from

[1] The data and code will be open-sourced upon publication.

the tabular data, meeting the user's information requirements. The information in the answer should not only be relevant to the question but also be consistent with tabular data.

2.3 Collecting Human Evaluation Scores

We tasked annotators to evaluate answers using a *Likert scale* ranging from 1 to 5 for each dimension based on the following criteria: To ensure the high quality of annotations, we hired eight undergraduate students proficient in English. Before starting the annotations, each annotator completed a one-hour online training session and reviewed a guide detailing the task execution steps. The annotators were compensated at an approximate hourly rate of $10, aligned with the complexity and duration of the task. Each sample was independently evaluated by two different annotators to mitigate individual bias and variance in scoring. Instances of significant disagreement (a variance greater than 2 points) were re-evaluated by an additional annotator. We achieved substantial inter-annotator agreements, with Krippendorff's alpha for faithfulness- and comprehensiveness-level annotation at 0.678 and 0.603, respectively.

2.4 Collecting Automated Evaluation Scores

We examine following automatic metrics that are widely used in the LFTQA task, investigating their reliability in evaluating model performance: **BLEU** (Papineni et al., 2002), **ROUGE** (Lin and Hovy, 2003), **METEOR** (Banerjee and Lavie, 2005), **BERTScore** (Zhang et al., 2020), **TAPAS-Acc** (Liu et al., 2022), **AutoACU** (Liu et al., 2023c). Appendix A.2 discusses the details of each metric. We also adopt an LLM-based evaluation system, **G-Eval** (Liu et al., 2023a), to the LFTQA task. G-Eval employs LLMs using a chain-of-thought approach and the form-filling paradigm to assess the quality of generated text. We adopt the official CoT evaluation prompt to assess the *faithfulness* and *comprehensiveness* of the generated answers, separately. The evaluation prompts used are presented in Appendix A.2. We use the Llama-3-70B and GPT-4o as the evaluators. For each model output collected in Section 2.1, we measure the above metrics' scores as automated evaluation scores.

	FeTaQA		QTSumm	
Metric	Comp.	Fai.	Comp.	Fai.
BLEU	0.076	0.220	-0.070	0.099
ROUGE	0.006	0.224	-0.160	0.119
METEOR	0.206	0.272	-0.240	0.019
BERT-Score	0.329	-0.254	0.237	0.136
TAPAS-Acc	-0.033	0.059	-0.028	-0.082
AutoACU	-0.042	0.296	0.152	0.208
G-Eval$_{Llama-3}$ Comp.	0.562	0.325	0.543	0.368
G-Eval$_{Llama-3}$ Fai.	0.321	0.497	0.307	0.509
G-Eval$_{GPT-4o}$ Comp.	0.623	0.409	0.612	0.352
G-Eval$_{GPT-4o}$ Fai.	0.301	0.531	0.376	0.585

Table 1: Results of *instance-level* Kendall's tau correlations between automatic metrics and human judgments on QTSumm and FeTaQA datasets.

	FeTaQA		QTSumm	
Metric	Comp.	Fai.	Comp.	Fai.
BLEU	0.009	0.295	-0.251	-0.033
ROUGE	-0.134	0.247	-0.269	0.065
METEOR	0.152	0.235	-0.395	-0.066
BERT-Score	0.340	-0.422	0.301	0.202
TAPAS-Acc	-0.189	0.006	-0.196	-0.122
AutoACU	0.031	0.324	0.068	0.198
G-Eval$_{Llama-3}$ Comp.	0.542	0.319	0.509	0.302
G-Eval$_{Llama-3}$ Fai.	0.336	0.587	0.347	0.564
G-Eval$_{GPT-4o}$ Comp.	0.641	0.412	0.633	0.384
G-Eval$_{GPT-4o}$ Fai.	0.379	0.609	0.411	0.598

Table 2: Results of *system-level* Kendall's tau correlations between automatic metrics and human judgments on QTSumm and FeTaQA datasets.

3 Experimental Results

3.1 Main Results

Table 1 and Table 2 illustrate the instance- and system-level Kendall's tau correlation between automatic and human judgements. We can draw following two conclusions based on the results: **Existing automatic metrics fail in assessing the answers generated by LLM-based systems.** Table 1 reveal a general trend of low to negative correlations across a range of metrics (e.g., BLEU, ROUGE, METEOR, and TAPAS-Acc), when evaluating individual LLM-generated responses. This indicates a widespread issue among current automatic metrics in measuring the faithfulness and comprehensive of LLM-generated answers, pointing to a systemic failure to align with human judgments at the instance level. **Existing automatic metrics fail in comparing the performance of different LLM-based systems.** Similarly, Table 2 shows that the same metrics struggle with accurately reflecting human evaluations when comparing overall system performance. Notably, negative correlations in metrics such as BLEU and METEOR at the system level suggest that these metrics are not effectively distinguishing the nuanced differences in quality among various LLM-based systems, underscoring a broader inadequacy in the current automatic evaluation methods in LFTQA. **LLM-based metrics demonstrate a significant improvement over traditional automated metrics in terms of correlation with human evaluation.** As illustrated in Table 1 and Table 2, G-Eval consistently achieves positive and high scores at both the instance-level and system-level evaluations. This indicates LLM-base metrics' proficiency in accurately assessing individual answer generation and identifying discrepancies in the effectiveness of various systems. Compared to Llama-3, GPT-4o yields higher scores, indicating that its evaluation results correspond more closely with human assessments. This superior performance reflects the enhanced evaluation capabilities of larger-size models in aligning with human judgment standards for the LFTQA task, highlighting the enhanced precision and reliability of advanced LLMs in quality evaluation.

3.2 Case Study

To gain deeper insights into the failure cases of automated evaluation systems for LFTQA tasks, we conducted detailed human analyses by exploring scenarios where automated evaluations fall short. Specifically, we randomly sampled 60 model output pairs from GPT-4o with Dater and one-shot prompting on QTSumm. We selected examples where GPT-4o with Dater received lower scores from at least four out of six metrics but achieved better results in human evaluations compared to GPT-4o with one-shot prompting. We meticulously analyzed and summarised the failure scenarios and summarised failure reasons as follows.

The Effect of Question As we delve deeper into the examples, we observe that the clarity of the questions significantly impacts the quality of the generated answers. Ambiguous questions can lead the model to misinterpret the key elements, resulting in the retrieval of incorrect information from the tables. Furthermore, we discovered that some questions were subjective or open-ended, which led to a variety of perspectives and content in the answers. The information related to these questions may not be directly presented or elaborated in the given tables. Instead, it should be inferred and evaluated from external materials, requiring careful speculation and analysis. In contrast, both the ground truth and generated answers typically reflect only a subset of these potential viewpoints. Table 4 in Appendix presents detailed examples.

The Effect of Ground Truth Although ground truth is used as the standard reference in the evaluation process, it has certain issues that affect the quality of the assessment. Ground-truth answers often include extensive descriptive details, which can make them redundant and contain content irrelevant to the questions. Additionally, in some instances, the ground truth fails to provide the specific information requested in the question. This can lead to lower evaluation scores, even when the generated outputs are accurate. Table 5 in Appendix presents detailed examples.

The Effect of Generated Answer LLM-based models excel at incorporating additional, reasoning-intensive information that is not present in ground-truth answers. They generate direct, parallel structures in their responses, which align well with human expression in real-world applications. However, current automated metrics struggle to capture this supplementary information and concise structures, resulting in automated evaluation scores that are significantly lower than human scores. Table 6 in Appendix presents detailed examples.

4 Related Work

To evaluate automatic metric performance for text generation, several human evaluation benchmarks have been collected (Cohan and Goharian, 2016; Dhingra et al., 2019; Gabriel et al., 2021; Fabbri et al., 2021; Liu et al., 2023b; Jiang et al., 2023b), comprising system-generated text and their human evaluation scores. The human evaluation result on the system-generated text is considered the gold standard, and metric performance is measured by the correlation between the human evaluation scores and automatic metric scores. To the best of our knowledge, we are the first to examine the automated evaluation methods for LFTQA research.

5 Conclusion

Our exploration into the evaluation of LLMs for LFTQA tasks reveals a significant gap between current automatic metrics and human judgment, particularly in assessing answer faithfulness and comprehensiveness. The insights from the LFTQA-Eval dataset highlight the need for more nuanced evaluation methods that align more closely with human evaluative standards. Addressing this discrepancy is essential for advancing the reliability of LFTQA systems and ensuring their practical utility in real-world scenarios.

Limitations

Our analysis is limited to 6,400 examples for which we have collected. While more statistically significant conclusions could be drawn from a larger evaluation set, as noted above a much large time and budget allocation would be required, and we encourage the community to apply our protocol to expand our evaluation set.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Wenhu Chen. 2023. Large language models are few(1)-shot table reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1120–1130, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Arman Cohan and Nazli Goharian. 2016. Revisiting summarization evaluation for scientific articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 806–813, Portorož, Slovenia. European Language Resources Association (ELRA).

DeepSeek. 2023. Deepseek llm: Let there be answers. `https://github.com/deepseek-ai/DeepSeek-LLM`.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics.

Google. 2023. Gemini.

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. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023b. Tigerscore: Towards building explainable metric for all text generation tasks.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Ao Liu, Haoyu Dong, Naoaki Okazaki, Shi Han, and Dongmei Zhang. 2022. 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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Yixin Liu, Alexander Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023c. 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.

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

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.

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

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.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. 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?

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.

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*.

Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 174–184, New York, NY, USA. Association for Computing Machinery.

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. 2022a. 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. 2023a. Docmath-eval: Evaluating numerical reasoning capabilities of llms in understanding long documents with tabular data.

Yilun Zhao, Linyong Nan, Zhenting Qi, Rui Zhang, and Dragomir Radev. 2022b. 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, Boyu Mi, Yixin Liu, Weijin Zou, Simeng Han, Ruizhe Chen, Xiangru Tang, Yumo Xu, Dragomir Radev, and Arman Cohan. 2023b. QTSumm: Query-focused summarization

Property	FeTaQA	QTSumm
Table Source	Wikipedia	Wikipedia
Unique Tables	1,942	424
Avg. Rows per Table	14.2	12.0
Avg. Columns per Table	5.7	6.7
Avg. Table Title Length	5.4	7.4
Avg. Query Length	14.0	22.3
Avg. Summary Length	23.3	67.8
Test Set Size (# QA Pairs)	2,003	1,078

Table 3: Basic statistics of the FeTaQA and QTSumm test sets used in our experiments.

over tabular data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1157–1172, Singapore. Association for Computational Linguistics.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023c. Investigating table-to-text generation capabilities of large language models in real-world information seeking scenarios. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 160–175, Singapore. Association for Computational Linguistics.

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 LfTQA Benchmark

A.1 LLM Output Collection

We adopt following prompting methods for collecting model outputs for LfTQA-Eval construction

- **Zero-shot** instructs LLMs to directly generate the final response based on provided tables and accompanying questions.
- **One-shot** requires a single sample to prompts LLMs for generating answers of given sources.
- **Chain-of-Thought** (Wei et al., 2022) tasks LLMs with generating a sequence of immediate reasoning steps, aiming to enhance their capability for intricate reasoning processes substantially.
- **Dater** (Ye et al., 2023) presents a methodology for decomposing complex questions into a set of sub-questions. This is achieved through the generation of an intermediate SQL query by LLMs and a limited set of prompting samples. Subsequently, the method aggregates all sub-information to produce the final answer.

A.2 Automated Evaluation System

- **BLEU** (Papineni et al., 2002) computes the geometric mean of the modified precision scores of the translated text and incorporates a brevity penalty factor. We use SacreBLEU (Post, 2018) for BLEU score calculation.
- **ROUGE** (Lin and Hovy, 2003) assesses the degree of lexical similarity between the generated text and the reference text. We employ F1 score for ROUGE-L.
- **METEOR** (Banerjee and Lavie, 2005) is developed to address the limitations of BLEU by introducing a method where alignment is established through the mapping of unigrams.
- **BERTScore** (Zhang et al., 2020) measures the similarity between the generated output and the reference text by utilizing contextualized token embeddings derived from a pre-trained model.
- **TAPAS-Acc** (Liu et al., 2022) assesses the faithfulness of table-to-text generation using TAPAS (Herzig et al., 2020) pretrained on the TabFact (Chen et al., 2020) dataset.
- **AutoACU** (Liu et al., 2023c) presents a reference-based automated evaluation system, utilizing atomic content units (ACUs) to gauge the similarity between text sequences.

A.3 Evaluating Automatic Evaluation Metrics

To evaluate the performance of automatic metrics, the human evaluation result on the same evaluation target is considered the gold standard, and metric performance is measured by the correlation between the human evaluation scores and automatic metric scores. Following previous work (Cohan and Goharian, 2016; Fabbri et al., 2021; Liu et al., 2023b), we calculate the correlation at the *system*- and *instance*-level. Specifically, given n input articles and m table-to-text generation systems, the human evaluation and an automatic metric result in two n-row, m-column score matrices H, M respectively. The *system*-level correlation is calculated on the aggregated system scores:

$$r_{\text{sys}}(H, M) = \mathcal{C}(\bar{H}, \bar{M}), \tag{1}$$

G-Eval for Evaluating Faithfulness

Task Introduction:
Given a complex question and a generated answer about a table, your task is to rate the answer's Faithfulness.

Evaluation Criteria:
Faithfulness(1-5): A good answer should accurately and completely address the given question.It must be based entirely on the information provided and should not include any unfaithful or hallucinated content.

Evaluation Steps:
1. Carefully read the table and the question, be aware of the information they contains and analyze their key points and important aspects.
2. Read the proposed answer carefully and understand its content. Check for factual errors in the answer to ensure if it accurately reflect the information presented in the table.
3. Rate text on a scale from 1(worst) to 5(best) by its faithfulness according to the criteria strictly. Note that scores are integers.

Figure 2: G-Eval for Evaluating the *faithfulness* of the LLM generated answer.

G-Eval for Evaluating Comprehensiveness

Task Introduction:
Given a complex question and a generated answer about a table, your task is to rate the answer's Comprehensiveness.

Evaluation Criteria:
Comprehensiveness(1-5): A good answer should provide all the necessary information to address the question comprehensively. Additionally, it should avoid including details that, while consistent with the tabular data, are irrelevant to the given question.

Evaluation Steps:
1. Carefully read the table and the question, be aware of the information they contains and analyze their key points and important aspects.
2. Read the proposed answer carefully and understand its content. Verify that the answer contains all the essential information needed to address the question.
3. Rate text on a scale from 1(worst) to 5(best) by its comprehensiveness according to the criteria strictly. Note that scores are integers.

Figure 3: G-Eval for Evaluating the *Comprehensiveness* of the LLM generated answer.

where $\bar{H}$ and $\bar{M}$ contain m entries which are the average system scores across n data samples, e.g., $\bar{H}_0 = \sum_i H_{i,0}/n$.

The instance-level correlation can be computed as the average of sample-wise correlations, providing insight into the relationship between automated and human evaluation:

$$r_{\text{ins}}(H, M) = \frac{\sum_i \mathcal{C}(H_i, M_i)}{n}, \qquad (2)$$

Where H_i and M_i represent the evaluation results for the i-th data sample, with $\mathcal{C}$ denoting a function that computes a correlation coefficient. In this study, we employ Kendall's tau rank correlation at both the system and instance levels to measure the correlations between these two types of evaluations.

B Experimental Results

Error Type	Example	Explanation
Question is ambiguous	**Question:** Who were the top three scorers for the 1961-62 Michigan Wolverines men's basketball team and how many points did they score?	It may take individual scores but is phrased in a way that could be interpreted as asking for a total score, potentially leading to the total score being treated as another player in the ranking.
Subjective issues	**Question:** How did the performance of Tom Brady in terms of passing yards during the Regular Season 2011 compare with other quarterbacks listed in 2011?	The subject of these questions might result in multiple reasonable interpretations and answers. For example, responses could pertain to Tom's scoring rate, passing rate, ball handling performance, etc., each in different ways.
Open-ended questions	**Question 1:** Summarize the basic information of the episode(s) written by Damon Lindelof. **Question 2:**Summarize the performance of Weekend Hussler in the Caulfield Guineas.	These questions involve various perspectives and require external information to be adequately addressed. For example, the first question might pertain to understanding the play, including plot trends, character development, and thematic elements in the episode. Different background knowledge and perspectives will result in varying answers.

Table 4: Case studies on evaluation errors due to the effects of questions.

Error Type	Example	Explanation
Redundancy in Ground Truth Answer	**Generated Answer:** The quantity of the aircraft models that were first introduced between the years 1980 and 1985 inclusive is 3 and their build years are 1978, 1979, and 1971. **Ground Truth Answer:**The number of aircraft models that were initially launched within the timeframe extending from the year 1980 up to and including the year 1985 amounts to a total of three distinct models. Specifically, the construction years for these models are identified as the years 1978, 1979, and 1971 respectively.	Ground truth answers tend to include additional details that may not directly contribute to answering the question in some cases while the generated answers are concise with all essential information. Based on the example provided, the ground truth answer offers a more comprehensive and detailed description of the launch times of the aircraft models.
Mismatch Between Question and Ground Truth Answer	**Generated Answer:** The track that has the lowest bpm on the Speed SFX series is 'Rhythm & Police (K.O.G G3 Mix)' and the track with the highest bpm on the Speed SFX series is the 'Mission: Impossible Theme'. **Ground Truth Answer:** The Speed SFX's lowest bpm is 175 on tracks 5, and the highest bpm is 195 on 20 in the Speed series.	The ground truth focuses on the BPM values without mentioning specific track names, which is the primary detail the question asks for. Conversely, the prediction accurately identifies the tracks by name, directly addressing the question. Therefore, the ground truth fails to provide the specific information that the question seeks.

Table 5: Case studies on evaluation errors due to the effects of ground truth answers.

Error Type	Example	Explanation
Inconsistency in LLM Output Length	**Generated Answer:** The HP LaserJet 4250 has a faster CPU speed of 460 MHz compared to the 4200's 300 MHz and its print speed is also higher at 43 pages per minute (PPM) compared to the 4200's 35 PPM, yet both printers come with a standard memory of 48 MB and can be expanded up to a maximum of 512 MB. **Ground Truth Answer:**The HP LaserJet 4200 and 4250 models, both in HP LaserJet 4000 series, have some same and different specifications. In same parts, both models come out in November 2002 and stop in June/November 20051. They have print resolution of 1200 DPI. Also, both printers can make the memory bigger, up to 512 MB. But, there are some different things between these two models. The LaserJet 4200 has 300 MHz CPU speed, but 4250 has faster 460 MHz CPU. Also, 4200 model can print 35 pages in one minute (PPM), but 4250 can print a little faster, 43 PPM".	The GT length is 68, while the LLM-generated length is 52.16. This indicates that the predicted output is significantly shorter than the actual answer. Generated responses tend to be concise and straightforward, focusing on delivering key points efficiently. In contrast, the actual answer provides more extensive information, with greater detail and elaboration. This difference highlights a tendency for automated responses to prioritize brevity.
Answers' Different Structures	**Generated Answer:** The quantity of the aircraft models that were first introduced between the years 1980 and 1985 inclusive is 3 and their build years are 1978, 1979, and 1971. **Ground Truth Answer:** Between the years 1980 to 1985 altogether, Agderfly added three airplane models to its fleet. In the year 1980, one Piper Chieftain made in 1978 was added, also one Piper Tomahawk was made in 1979 in the same year. The 1985 year, one Piper Seneca which was made in 1971. In total, during this time, Agderfly added three aircraft models whose combined quantity is four units.	Generated answers tend to be structured with parallel objects, while ground truth answers often utilize complex clauses to introduce related information thoroughly. In this example, the generated answer simply lists the years, while the ground truth introduces the information for each year in a single, comprehensive sentence. This discrepancy in structure can result in misalignment between automated predictions and the expected answers, impacting the accuracy of evaluations and interpretations.

Table 6: Case studies on evaluation errors due to the effects of generated answers.