MATSA: Multi-Agent Table Structure Attribution

Puneet Mathur, Alexa Siu, Nedim Lipka, Tong Sun
Adobe Research
{puneetm, asiu, lipka, tsun}@adobe.com

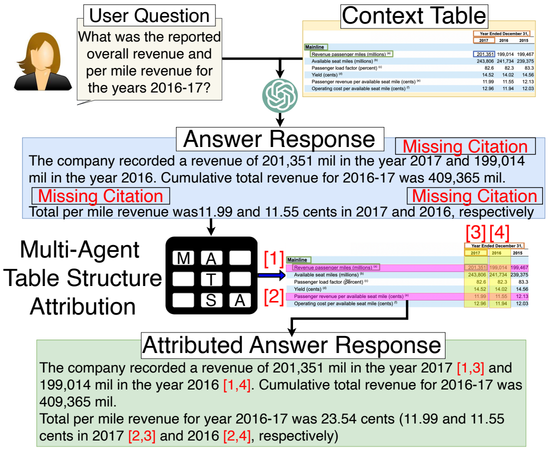
Demo Video: https://youtu.be/UFuNwvZFN18 **Demo Link**: matsa.ai

Abstract

Large Language Models (LLMs) have significantly advanced QA tasks through in-context learning but often suffer from hallucinations. Attributing supporting evidence grounded in source documents has been explored for unstructured text in the past. However, tabular data present unique challenges for attribution due to ambiguities (e.g., abbreviations, domain-specific terms), complex header hierarchies, and the difficulty in interpreting individual table cells without row and column context. We introduce a new task, Fine-grained Structured Table Attribution (FAST-Tab), to generate row and column-level attributions supporting LLM-generated answers. We present MATSA[1], a novel LLM-based Multi-Agent system capable of post-hoc Table Structure Attribution to help users visually interpret factual claims derived from tables. MATSA augments tabular entities with descriptive context about structure, metadata, and numerical trends to semantically retrieve relevant rows and columns corresponding to facts in an answer. Additionally, we propose TabCite, a diverse benchmark designed to evaluate the FAST-Tab task on tables with complex layouts sourced from Wikipedia and business PDF documents. Extensive experiments demonstrate that MATSA significantly outperforms SOTA baselines on TabCite, achieving an 8-13% improvement in F1 score. Qualitative user studies show that MATSA helps increase user trust in Generative AI by providing enhanced explainability for LLM-assisted table QA and enables professionals to be more productive by saving time on fact-checking LLM-generated answers. **Demo Website**: matsa.ai

1 Introduction

Recent advances in LLMs have enhanced question-answering capabilities (Brown et al., 2020; Achiam et al., 2023), but they are prone to hallucination, producing plausible-sounding yet non-factual information, which undermines user trust (Xu et al., 2024; Snyder et al., 2023). The absence of supporting evidence complicates the verification of LLM-generated outputs. Contemporary solutions address this by grounding claims in LLM-generated answers with citations from the document context (Ji et al., 2023). Previous works have explored instruction tuning (Kamalloo et al., 2023), in-context learning (Gao et al., 2023b), and NLI-based post-hoc attribution methods (Gao et al., 2023a) to link supporting passages to claims with varying levels of success in attributing free-form text.



What was the reported overall revenue and per mile revenue for the years 2016-17?

	Year Ended December 31,		
	2017	2016	2015
Mainline			
Revenue passenger miles (millions) (a)	201,351	199,014	199,467
Available seat miles (millions) (b)	243,806	241,734	239,375
Passenger load factor (percent) (c)	82.6	82.3	83.3
Yield (cents) (d)	14.52	14.02	14.56
Passenger revenue per available seat mile (cents) (e)	11.99	11.55	12.13
Operating cost per available seat mile (cents) (f)	12.96	11.94	12.03

The company recorded a revenue of 201,351 mil in the year 2017 and 199,014 mil in the year 2016. Cumulative total revenue for 2016-17 was 409,365 mil.
Total per mile revenue was11.99 and 11.55 cents in 2017 and 2016, respectively

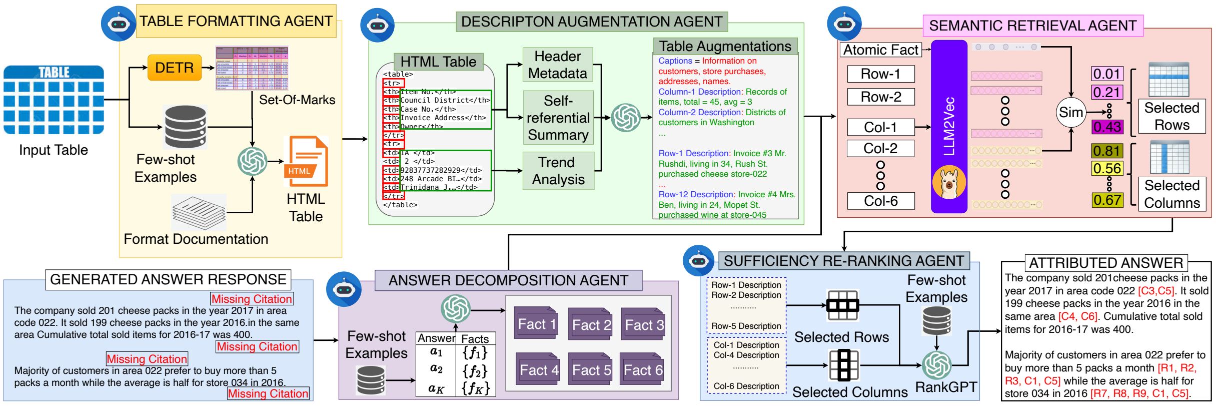
The company recorded a revenue of 201,351 mil in the year 2017 [1,3] and 199,014 mil in the year 2016 [1,4]. Cumulative total revenue for 2016-17 was 409,365 mil.
Total per mile revenue for year 2016-17 was 23.54 cents (11.99 and 11.55 cents in 2017 [2,3] and 2016 [2,4], respectively)

Figure 1: MATSA is a post-hoc table structure attribution approach that retrieves rows and columns supporting the factual claims in an LLM-generated answer in response to a question.

Tables are widely used for handling complex semi-structured data in various domains, including healthcare, finance, and education. Application of LLMs to tabular data presents unique challenges: hierarchical header structures, varying formats (e.g., JSON, HTML, CSV, Markdown), lack of straightforward serialization techniques, noisy content, and ambiguity in raw data (e.g., abbreviations, domain-specific terms) (Sui et al., 2023). Due to the high specificity of table data, attributing table structures at the row/column level in gener-

[1] **Demo Video**: https://youtu.be/UFuNwvZFN18

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing
System Demonstrations, pages 250–258
November 12-16, 2024 ©2024 Association for Computational Linguistics



```
<table>
<tr>
<th>Item No.</th>
<th>Council District</th>
<th>Case No.</th>
<th>Invoice Address</th>
<th>Owner</th>
</tr>
<tr>
<td>1A </td>
<td> 2 </td>
<td>92837737282929</td>
<td>248 Arcade BI…</td>
<td>Trinidana J.…</td>
</tr>
</table>
```

Captions = Information on customers, store purchases, addresses, names.
Column-1 Description: Records of items, total = 45, avg = 3
Column-2 Description: Districts of customers in Washington
...
Row-1 Description: Invoice #3 Mr. Rushdi, living in 34, Rush St. purchased cheese store-022
...
Row-12 Description: Invoice #4 Mrs. Ben, living in 24, Mopet St. purchased wine at store-045

The company sold 201 cheese packs in the year 2017 in area code 022. It sold 199 cheese packs in the year 2016.in the same area Cumulative total sold items for 2016-17 was 400.

Majority of customers in area 022 prefer to buy more than 5 packs a month while the average is half for store 034 in 2016.

The company sold 201cheese packs in the year 2017 in area code 022 [C3,C5]. It sold 199 cheese packs in the year 2016 in the same area [C4, C6]. Cumulative total sold items for 2016-17 was 400.

Majority of customers in area 022 prefer to buy more than 5 packs a month [R1, R2, R3, C1, C5] while the average is half for store 034 in 2016 [R7, R8, R9, C1, C5].

Figure 2: MATSA provides citations for generated answers grounded in table structures by orchestrating LLM agents: (1) Table Formatting Agent converts input table data into HTML format; (2) Description Augmentation Agent enriches raw tables with descriptions of row/column entities; (3) Answer Decomposition Agent decomposes the answer passage into atomic facts; (4) Semantic Retrieval Agent recalls relevant rows/columns based on semantic similarity; (5) Sufficiency Re-ranking Agent improves factual precision by retaining rows/columns required to collectively explain all factual claims in the answer statement.

ated answers remains under-explored. Prior methods for post-hoc answer attribution use embedding-based retrievers or LLM prompting and are limited to attributing entire tables rather than fine-grained structures (Huo et al., 2023). Hence, we introduce a novel task, Fast-Tab: Fine-grained Attribution over Structured Tables which involves identifying table rows and columns that support claims in an answer to a user's question.

We propose a novel multi-agentic system – MATSA: Multi-Agent Table Structure Attribution, (see Figure 1) that provides citations for generated answers based on table structures by utilizing multiple LLM agents: *(1) Table Formatting Agent* converts input table data into HTML format, which is crucial for linking data elements to their appropriate layout-specific fields. *(2) Description Augmentation Agent* enriches raw row/column entities with natural language descriptions to enhance the contextual understanding of table elements and reduce data misinterpretations. *(3) Answer Decomposition Agent* decomposes the answer passage into atomic facts, allowing each fact to be individually linked to specific table row/column citations. *(4) Semantic Retrieval Agent* extracts relevant rows/columns via embedding-based semantic similarity between row/column descriptions and answer facts, ensuring high recall for answer grounding.*(5) Sufficiency Re-ranking Agent* selects the minimal set of sourced rows and columns that collectively explain the answer, leveraging LLM reasoning to evaluate the utility of table structures beyond mere similarity.

Lastly, we propose a new benchmark - TabCite comprising of 8.5K table-QA pairs along with ground truth row/column-level attribution annotations, assembled by integrating three open-source datasets (ToTTo, FetaQA, AITQA) from diverse domains. The answer attributions may be derived from single or multiple table cells, and reflect a rich diversity of structure hierarchies. We conducted a user evaluation on diverse samples from TabCite to assess MATSA's utility in professional settings. Results show that participants find the fine-grained attributions to be accurate and useful in helping them more easily verify the accuracy of answers. Our **main technical contributions** are:

- **Fine-grained Table Structure Attribution (Fast-Tab)** task to generate row/column-level attributions to support factual claims in LLM-generated answers.
- **TabCite benchmark** of table QA and attributions sourced from Wikipedia and business PDF documents containing tables with complex header hierarchies.
- **MATSA - Multi-Agent Table Structure Attribution** framework that performs post-hoc table structure attribution via descriptive context augmentation of table entities to cite relevant rows/columns and outperforms SOTA baselines on TabCite by 8-13% F1 score.

Our **main system-level contributions** are:
(1) Interpretability: MATSA promotes interpretable answer attribution through *description augmentation agent* which provides logical rationales for the significance of each table entity in the LLM's reasoning process.

(2) Explainability MATSA is designed to explain the underlying reason to select various rows and columns to logically to support the answer text. To achieve this, it transcends simple textual similarity by introducing a *sufficiency re-ranking agent* that performs implicit multi-hop chain-of-thought reasoning to comprehensively extract all necessary evidence from the table.

(2) Reliability: By employing LLMs for table row/column-level citations, MATSA aims to assist professionals in domains such as business, education, and finance. This approach enables users to focus on more productive tasks by reducing time spent on fact-checking LLM-generated answers, thereby enhancing overall reliability.

2 Methodology

2.1 Fine-grained Structured Table Attribution

Let there be a table T with a distinct set of R rows and C columns. Given an input question q and its corresponding answer a, we propose a novel task of Fine-grained Structured Table Attribution (FAST-Tab) that aims to extract the set of top-n rows and top-m columns (collectively denoted by attribution set A_T), that is necessary and sufficient to explain how a is the correct and complete answer to q. Further, none of the artifacts in A_T should contradict the answer a.

2.2 MATSA

Figure 2 shows MATSA, an LLM-based multi-agent framework that provides citations for generated answers grounded in table structures by orchestrating the following LLM agents.

2.2.1 Table Formatting Agent

Tabular data frequently appears in PDF documents, necessitating conversion into LLM-friendly formats. Various table storage formats (e.g., CSV, JSON, XML, Markdown, HTML) exhibit different levels of information compression and present unique challenges for LLMs in comprehending table content. Given the extensive web data used in their training, LLMs often demonstrate superior proficiency in interpreting complex table layouts in HTML and XML formats. To convert input table data into HTML format, we employ a two-step process. First, we utilize the Detection Transformer (DETR) (Smock et al., 2022) to identify and mark row and column separators on table image renderings. Next, we leverage Large Multimodal Models (LMMs), such as GPT-4V, using few-shot set-of-mark prompting (Yang et al., 2023) to convert the marked table image into HTML format. This approach enables efficient transformation of diverse tabular data into a format that maximizes LLM comprehension and processing capabilities.

2.2.2 Table Description Augmentation Agent

Tabular data interpretation relies on accurately understanding the semantics of the cell-level information contextualized with structure metadata and underlying patterns across the table rows and columns. The raw content of a table may contain ambiguous information (e.g., abbreviations, domain-specific terms, signs, numbers with or without units, ill-defined row/column headers) that requires further clarification and may not have sufficient context for automated factual attribution. Towards this end, we utilize zero-shot LLM prompting to generate detailed descriptions for each row and column to explicitly augment raw table data. We consider the following information augmentation types:

(1) Header Metadata Augmentation: Headers are crucial for defining the meaning and context of row-column structured data, linking each cell item to its specific hierarchical fields. We prompt the LLM to supplement each cell item with multiple levels of associated row and column header information, ensuring comprehensive data categorization.

(2) Trend Analysis Augmentation: Statistical trend analysis of numerical data helps summarize key quantitative characteristics and tendencies across the table. We prompt the LLM to extract non-trivial quantitative comparisons, numerical ranges, and statistical data trends across all rows and columns.

(3) Self-Referential Summary Augmentation: Descriptions of data elements within a specific row or column help contextualize its categorical and numerical information in coherent natural language. We employ LLM prompting to generate descriptive narratives for each row and column, ensuring that the interrelationships between data items are thoroughly explained. The combined outputs from all three augmentation techniques act as a proxy for representing table rows and columns information in the attribution generation step.

2.2.3 Answer Decomposition Agent

Answer texts frequently contain multiple facts derived from various table rows and columns. To enhance interpretability and facilitate precise citations, it is crucial to distill attributable facts from an answer, such that each can be mapped to specific table elements. To address this challenge, we introduce an answer decomposition agent that extracts atomic facts, ensuring each statement is complete and independently verifiable without external dependencies. Inspired by (Min et al., 2023), we prompt LLM with few-shot examples to convert answer passages into a list of coherent and factual sentences. To prevent hallucinations, we use a pre-trained NLI model (RoBERTa (Wang et al., 2021)) to verify that each generated fact is entails the original answer passage.

2.2.4 Table Structure Attribution

We employ a two-pass retrieval strategy to identify the most relevant table rows and columns for attributions. We first generate a set of candidate rows/columns using embedding-based semantic matching to maximize recall, followed by a second-pass LLM-based re-ranking to dynamically retrieve rows and columns with high precision.

(1) Semantic Retrieval Agent: We use LLM-based embedding models, such as those from SentenceBert, BGE embeddings (Xiao et al., 2023), or LLM2Vec with a Llama-3 8B backbone (BehnamGhader et al., 2024), to obtain semantic embeddings for each row and column. Compared to previous encoder-only embeddings, decoder-only LLMs benefit from extensive large-scale pretraining. Instead of directly encoding table elements, we leverage the row/column descriptions generated by the Description Augmentation Agent to ensure that the fact sentences and table structure information are in-domain for the embedding model. For each fact sentence f_i, we select all rows/columns with an embedding similarity score between the fact embedding $e(f_i)$ and the table structure description embeddings ($e(r)$ or $e(c)$ $\forall r \in R, c \in C$) higher than a threshold η.

(2) Sufficiency Re-ranking Agent: While semantic retrieval identifies multiple supporting row/column citations based on semantic similarity to answer facts, it may lead to false positives. Attributions with unrelated supporting citations can reduce user trust in LLM-generated answers and may be perceived as a form of hallucination. To address this, we extend beyond mere textual similarity and focus on the collective utility of each extracted piece of evidence in forming a coherent chain of thoughts that logically supports the overall answer statement. Sufficiency Re-ranking Agent improves factual precision by retaining a minimal set of evidence required to sufficiently explain all factual claims in an answer. Inspired by the conceptualization of LLM function calling for fact verification (Katranidis and Barany, 2024), we repurpose LLM function calling to dynamically re-rank and retrieve relevant rows and columns, along with a "chain-of-thought" explanation that reasons about them in a multi-hop fashion. For a given answer passage a and a list of retrieved table rows/columns $d_1, d_2, \cdots, d_n$, we leverage the row/column descriptions as inputs and parse the output of the Sufficiency Re-ranking Agent to select the top-n rows and top-m columns as answer attributions. This approach promotes logical consistency in evidence and minimizes irrelevant citations. More details on prompt design in Supplementary Materials.

Dataset	**TottoQA**	**FetaQA**	**AITQA**
Size	7700	3004	513
Table Data Format	PDF	PDF	PDF
Table Domain	Wikipedia	Wikipedia	Financial Reports
Question Source	AI-generated	Human	Human
Answer Source	Human	Human	AI-generated
Contains Merged Cells	✗	✓	✗
Contains Column Hierarchy	✓	✗	✓
Contains Row Hierarchy	✗	✗	✓
Multiple Attribution Rows	✓	✓	✗
Multiple Attribution Columns	✓	✓	✗
# of Unique tables	7377	2876	112
Avg. Row Count	33	15	14
Max Row Count	2136	34	41
Avg. Column Count	5.2	5.6	5.2
Max Column Count	36	22	9
Avg. # of Words in Answer	14.9	19.8	12.2
Avg. # of Answer Sentence	2.3	2.4	2.2
Avg. # of Rows Attributed	1.5	3.5	1
Max # of Rows Attributed	436	32	1
Avg. # of Columns Attributed	2.4	3.4	1
Max # of Columns Attributed	15	15	1

Table 1: Data Statistics for `TabCite` Benchmark consisting of TottoQA, FetaQA, and AITQA corpus.

3 Experiments

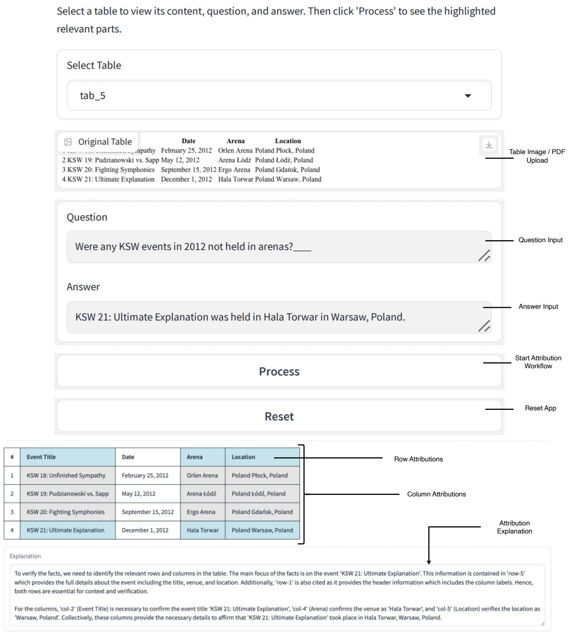
We evaluate the `MATSA` on our proposed `TabCite` benchmark. Tables in this benchmark are derived from Wikipedia pages and SEC filings, which are paired with questions, free-form answers, and ground truth row/column attributions. Table 1 gives data stastics about `TabCite` benchmark. `TabCite` is sourced by reformulating existing datasets:

(1) TOTTO (Parikh et al., 2020) is a Wikipedia-based open-domain table-to-text dataset containing short text descriptions of highlighted table cells. It lacks human-generated questions, hence we reformulated the content descriptions as answers and synthetically generated questions using GPT-4[2].

[2] https://openai.com/index/gpt-4/

Method	TabCite - FetaQA						TabCite - Totto						TabCite - AITQA					
	Row Attribution			Column Attribution			Row Attribution			Column Attribution			Row Attribution			Column Attribution		
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1
Post-hoc Retrieval (SentenceBert)	0.86	0.50	0.59	0.93	0.69	0.78	0.86	0.28	0.39	0.91	0.58	0.69	0.95	0.19	0.32	0.98	0.22	0.36
In-context Learning (GPT-4o)	0.76	0.77	0.73	0.93	0.88	0.89	0.95	0.65	0.74	0.94	0.51	0.66	0.96	0.64	0.74	0.95	0.39	0.55
MATSA (Ours)	0.74	0.92	**0.78**	0.95	0.90	**0.91**	0.82	0.78	**0.79**	0.87	0.70	**0.75**	0.94	0.85	**0.88**	0.92	0.47	**0.61**

Table 2: Performance comparison of MATSA with baselines for fine-grained table structure (rows and columns) attribution across FetaQA, Totto, and AITQA datasets in the TabCite benchmark. MATSA green achieves best F1 score across all settings.


Select a table to view its content, question, and answer. Then click 'Process' to see the highlighted relevant parts.

Select Table

tab_5

	Date	Arena	Location
mpathy	February 25, 2012	Orlen Arena	Poland Płock, Poland
2 KSW 19: Pudzianowski vs. Sapp	May 12, 2012	Arena Łódź	Poland Łódź, Poland
3 KSW 20: Fighting Symphonies	September 15, 2012	Ergo Arena	Poland Gdańsk, Poland
4 KSW 21: Ultimate Explanation	December 1, 2012	Hala Torwar	Poland Warsaw, Poland

Question

Were any KSW events in 2012 not held in arenas?___

Answer

KSW 21: Ultimate Explanation was held in Hala Torwar in Warsaw, Poland.

#	Event Title	Date	Arena	Location
1	KSW 18: Unfinished Sympathy	February 25, 2012	Orlen Arena	Poland Płock, Poland
2	KSW 19: Pudzianowski vs. Sapp	May 12, 2012	Arena Łódź	Poland Łódź, Poland
3	KSW 20: Fighting Symphonies	September 15, 2012	Ergo Arena	Poland Gdańsk, Poland
4	KSW 21: Ultimate Explanation	December 1, 2012	Hala Torwar	Poland Warsaw, Poland

Explanation

To verify the facts, we need to identify the relevant rows and columns in the table. The main focus of the facts is on the event 'KSW 21: Ultimate Explanation'. This information is contained in 'row-5' which provides the full details about the event including the title, venue, and location. Additionally, 'row-1' is also cited as it provides the header information which includes the column labels. Hence, both rows are essential for context and verification.

For the columns, 'col-2' (Event Title) is necessary to confirm the event title 'KSW 21: Ultimate Explanation', 'col-4' (Arena) confirms the venue as 'Hala Torwar', and 'col-5' (Location) verifies the location as 'Warsaw, Poland'. Collectively, these columns provide the necessary details to affirm that 'KSW 21: Ultimate Explanation' took place in Hala Torwar, Warsaw, Poland.

Figure 3: Demo App UI for MATSA

TOTTO includes tables with extreme size variations, merged cells, and complex column hierarchies, representative of real-world distributions.

(2) FetaQA (Nan et al., 2021) (Free-form Table Question Answering) is a dataset consisting of Table QA pairs from Wikipedia that mimic human-like multi-hop QA reasoning over evidence table cells to generate long-form coherent answers. While tables in FetaQA lack complex header hierarchies, the dataset is designed to require retrieving and reasoning over evidence cells from multiple rows for answer grounding.

(3) AITQA (Katsis et al., 2022) (Airline Industry Table QA) is a domain-specific dataset of tables gathered from US SEC 10-K annual reports of publicly traded airline companies that requires reasoning with complex column and row header hierarchies containing domain-specific vocabulary. Table distribution is similar to that found in scientific and business documents. Answers in AITQA are provided as singular table entities, which we converted into complete statements using GPT-4. We extracted the rows and columns corresponding to the supporting cells in above-listed datasets to get the set of ground truth row/column attributions

Baseline: We evaluate the effectiveness of MATSA with recent baselines: **(1) Few-shot In-Context Learning** (Gao et al., 2023b) prompts LLMs with few-shot examples to generate answers with in-line citations; **(2) Post-hoc Retrieval** (Gao et al., 2023b) using a dense retriever to retrieve top-k rows/columns for answer attribution.

Evaluation Metrics: As predictions output by MATSA are not ranked, we evaluate the attribution quality using Precision, Recall, and F1 score. Given a table with total D rows (or columns) , d' retrieved rows (or columns), and $\hat{d}$ ground truth rows (or columns) , we evaluate: (1) citation recall ($\sum_1^N \frac{d' \cap \hat{d}}{\hat{d}}$) to determine if the model captures all supporting rows/columns, and (2) citation precision ($\sum_1^N \frac{d' \cap \hat{d}}{d'}$), which identifies any irrelevant citations in the selected attribution set. Prioritizing citation recall helps emphasize answer credibility and verifiability while enhancing citation precision is crucial for better truthfulness and reduces the need for human review of extraneous attributions. For the simplicity of demo evaluation, we include randomly chosen 100 samples from each dataset split of our proposed benchmark.

LLM Archietctures: We use GPT-4o API through the Microsoft Azure platform for all our experiments. We also tried GPT-3.5 *(gpt3.5-turbo-16k-0613)* model but it performed consistently worse that GPT-4o.

Semantic Retriever architecture: We experimented with SentenceBert (Reimers and Gurevych, 2019), BGE embedding[3], and LLM2Vec with Llama-3 8B[4] as the embedding models. We use SentenceBert for final evaluations as it provided least latency. We use fused cosine similarity score to get top-k rows/columns, where $k = 5$ in each table.

Demo UI: We used Gradio for the demo UI hosted

[3] https://huggingface.co/BAAI/bge-base-en-v1.5

[4] https://huggingface.co/McGill-NLP/LLM2Vec-Meta-Llama-3-8B-Instruct-mntp

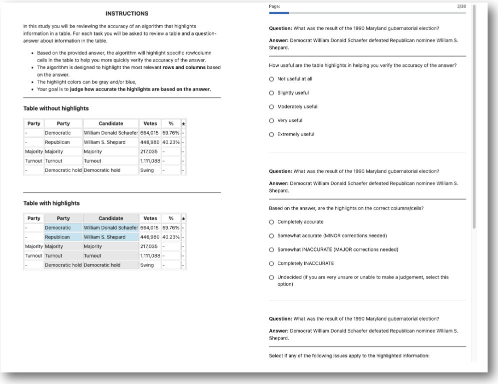
locally or on the AWS cloud platform.

4 MATSA Demo App

Figure 3 shows the MATSA demo app. The app was built using Gradio[5] and uses OpenAI GPT-4o and GPT-4V (vision) models. The interface includes an upload panel for table images and questions, option to type in the answer statement or let the LLM generate the answer based on table context. MATSA helps users visualize the cited rows and columns in different colors. The users also have the ability to read the LLM generated explanation for the row/column attributions, and can reset the interface to restart.

5 Results

Main Results: Table 2 compares the performance of MATSA with baseline methods on TabCite benchmark. We observe that MATSA significantly outperforms the baselines in terms of overall F1 scores for both row-wise and column-wise attribution settings. These results demonstrate that our multi-agent approach effectively captures the informative semantics of tabular entities, providing reliable answer citations. The post-hoc retriever baseline shows a severely degraded performance due to the inability of the retriever model to contextualize data in row and column cells. It suffers skewed recall as the lack of answer decomposition leads to many rows/columns being classified as relevant attributions, leading to high recall but low precision. Moreover, traditional retrieval models cannot dynamically adapt the value of k in their top-k selections based on attribution relevancy. The naive in-context learning baseline shows better performance compared to post-hoc retrieval, yet struggles to match high precision as in MATSA as instructing LLMs to retrieve relevant attributions at inference is challenging to simultaneously generate coherent answers and ground atomic facts in complex table structures. MATSA involves description augmentation that generates detailed natural language descriptions of rows and columns to improve cell-level entity contextualization and reduce noise in the retriever embedding. This contributes to its best performance among all models. The two-stage retrieve-and-rank pipeline in MATSA balances precision and recall, resulting in state-of-the-art F1 scores across all three datasets.



INSTRUCTIONS

In this study you will be reviewing the accuracy of an algorithm that highlights information in a table. For each task you will be asked to review a table and a question-answer about information in the table.

- Based on the provided answer, the algorithm will highlight specific row/column cells in the table to help you more quickly verify the accuracy of the answer.
- The algorithm is designed to highlight the most relevant **rows and columns** based on the answer.
- The highlight colors can be gray and/or blue,
- Your goal is to **judge how accurate the highlights are based on the answer.**

Table without highlights

Party	Party	Candidate	Votes	%	±
-	Democratic	William Donald Schaefer	664,015	59.76%	-
-	Republican	William S. Shepard	446,980	40.23%	-
Majority	Majority	Majority	217,035	-	-
Turnout	Turnout	Turnout	1,111,088	-	-
-	Democratic hold	Democratic hold	Swing	-	-

Table with highlights

Party	Party	Candidate	Votes	%	±
-	Democratic	William Donald Schaefer	664,015	59.76%	-
-	Republican	William S. Shepard	446,980	40.23%	-
Majority	Majority	Majority	217,035	-	-
Turnout	Turnout	Turnout	1,111,088	-	-
-	Democratic hold	Democratic hold	Swing	-	-

Page: 3/30

Question: What was the result of the 1990 Maryland gubernatorial election?

Answer: Democrat William Donald Schaefer defeated Republican nominee William S. Shepard.

How useful are the table highlights in helping you verify the accuracy of the answer?

○ Not useful at all

○ Slightly useful

○ Moderately useful

○ Very useful

○ Extremely useful

Question: What was the result of the 1990 Maryland gubernatorial election?

Answer: Democrat William Donald Schaefer defeated Republican nominee William S. Shepard.

Based on the answer, are the highlights on the correct columns/cells?

○ Completely accurate

○ Somewhat accurate (MINOR corrections needed)

○ Somewhat INACCURATE (MAJOR corrections needed)

○ Completely INACCURATE

○ Undecided (if you are very unsure or unable to make a judgement, select this option)

Question: What was the result of the 1990 Maryland gubernatorial election?

Answer: Democrat William Donald Schaefer defeated Republican nominee William S. Shepard.

Select if any of the following issues apply to the highlighted information:

Figure 4: Interface for user evaluation. Participants were presented with the question-answer and related table with and without attribution highlights. Participants rated the attribution accuracy and usefulness in helping verify the accuracy of the answer.

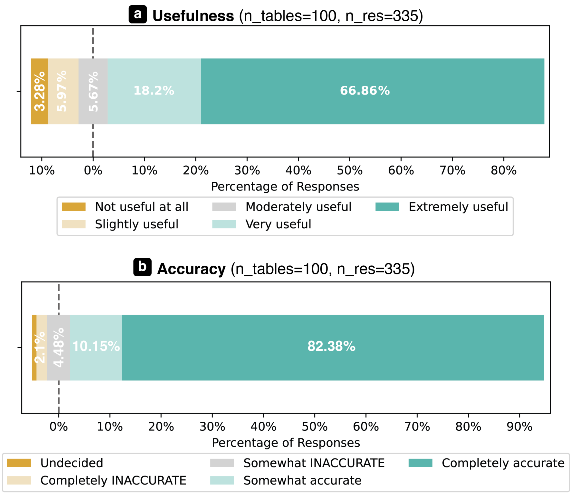


Figure 5: User evaluation ratings on attribution a) Usefulness and b) Accuracy.

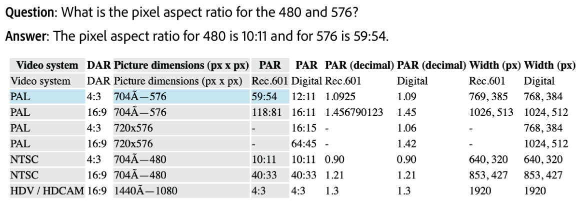
6 User Evaluation

We conducted a user evaluation to assess the attribution accuracy of MATSA and perceived usefulness of having fine-grained attribution on tables.

Recruitment & Methodology: Sixteen participants were recruited via Prolific[6]. Our evaluation dataset was comprised of 100 long-form Table QA pairs randomly sampled from our proposed TabCite corpus. Participants were asked to review the fine-grained attribution produced by MATSA shown as highlights obtained for the table QA . Participants were asked to rate (1) the usefulness of the attribution in helping them verify the accuracy of the answers, (2) the accuracy of the attribution, and (3) list any improvements on the attribution

[5] https://www.gradio.app

[6] https://www.prolific.com/

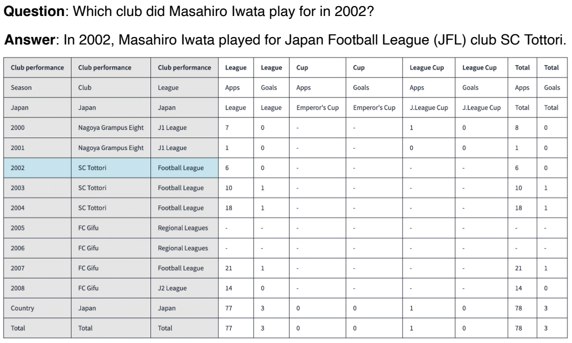


Question: What is the pixel aspect ratio for the 480 and 576?

Answer: The pixel aspect ratio for 480 is 10:11 and for 576 is 59:54.

Video system	DAR	Picture dimensions (px x px)	PAR	PAR	PAR (decimal)	PAR (decimal)	Width (px)	Width (px)
Video system	DAR	Picture dimensions (px x px)	Rec.601	Digital	Rec.601	Digital	Rec.601	Digital
PAL	4:3	704Ã—576	59:54	12:11	1.0925	1.09	769, 385	768, 384
PAL	16:9	704Ã—576	118:81	16:11	1.456790123	1.45	1026, 513	1024, 512
PAL	4:3	720x576	-	16:15	-	1.06	-	768, 384
PAL	16:9	720x576	-	64:45	-	1.42	-	1024, 512
NTSC	4:3	704Ã—480	10:11	10:11	0.90	0.90	640, 320	640, 320
NTSC	16:9	704Ã—480	40:33	40:33	1.21	1.21	853, 427	853, 427
HDV / HDCAM	16:9	1440Ã—1080	4:3	4:3	1.3	1.3	1920	1920

Figure 6: Example of Table QA pair from `TabCite` benchmark where question/answer are unclear as reported by evaluation participants.



Question: Which club did Masahiro Iwata play for in 2002?

Answer: In 2002, Masahiro Iwata played for Japan Football League (JFL) club SC Tottori.

Club performance	Club performance	Club performance	League	League	Cup	Cup	League Cup	League Cup	Total	Total
Season	Club	League	Apps	Goals	Apps	Goals	Apps	Goals	Apps	Goals
Japan	Japan	Japan	League	League	Emperor's Cup	Emperor's Cup	J.League Cup	J.League Cup	Total	Total
2000	Nagoya Grampus Eight	J1 League	7	0	-	-	1	0	8	0
2001	Nagoya Grampus Eight	J1 League	1	0	-	-	0	0	1	0
2002	SC Tottori	Football League	6	0	-	-	-	-	6	0
2003	SC Tottori	Football League	10	1	-	-	-	-	10	1
2004	SC Tottori	Football League	18	1	-	-	-	-	18	1
2005	FC Gifu	Regional Leagues	-	-	-	-	-	-	-	-
2006	FC Gifu	Regional Leagues	-	-	-	-	-	-	-	-
2007	FC Gifu	Football League	21	1	-	-	-	-	21	1
2008	FC Gifu	J2 League	14	0	-	-	-	-	14	0
Country	Japan	Japan	77	3	0	0	1	0	78	3
Total	Total	Total	77	3	0	0	1	0	78	3

Figure 7: Example of Table QA pair from `TabCite` benchmark where attribution is accurate as reported by evaluation participants.

or feedback. Figure 4 shows our hosted interface that was used for user study with participants recruited on Prolific. They were presented with the question-answer and the related table, with and without attribution highlights. Participants rated the attribution accuracy and usefulness in helping verify answer citations.

Usefulness & Accuracy: Overall, the participants had a positive feedback for the fine-grained table attributions produced by `MATSA`. Figure 5 shows the ratings for Usefulness and Accuracy. The majority of users found the attributions *Extremely useful* $(224/335, 66.86\%)$ and *Very useful* $(61/335, 18.5\%)$ for verifying the accuracy of table QA. Participants found the attributions to be Completely accurate $(276/335, 82.38\%)$ and Somewhat accurate requiring minor corrections $(34/335, 10.15\%)$. Through qualitative feedback, participants described the attributions as easy to understand, helpful in reducing reading time of the tables (*"I could sift through the table quickly"*) and making verification easier (*"...can help me to locate the answer quickly."*).

User Feedback: The participants also provided feedback for cases where attribution could be improved. In some cases participants reported additional row/columns could be included in the attribution to make them more helpful $(19/100)$. In other cases, some unnecessary row/columns could be removed $(15/100)$. Additionally, in our evaluation dataset a small portion of the QA pairs were found to have either an inaccurate answer or the question was unclear (Figure 6), which in turn impacted participant ratings of the usefulness and accuracy.

Qualitative Examples: Figure 7 shows an example table QA pair from the TabCite benchmark where attribution is accurate as reported by evaluation participants. Figure 6 shows an example table QA pair from the TabCite benchmark where question/answer are unclear as reported by evaluation participants. We found that a small portion of human generated question-answer pairs in FetaQA may be noisy leading to inconsistent attribution experience.

7 Target Audience

`MATSA` is targeted to help students, professionals, and other users of LLM-based chat systems interacting with PDFs or text document. Some of the common use cases that we envision for this system are: (1) enable users to fact check LLM-generated answers grounded in tabular data, (2) post-hoc text attribution for financial documents, product manuals, Wikipedia-style web pages, (3) generate annotation data for instruction-tuning LLM models to retrogressively generate inline citations with text.

System License: The MATSA system is a proprietary system developed for research experimentation and development. At this stage, we do not plan to publicly open-source it for any commercial or non-commercial purposes.

8 Conclusion

We introduce `FAST-Tab`, a novel task for fine-grained table structure attribution to provide citations from table rows and columns to support factual claims in LLM-generated answers to tabular questions. We present the `TabCite` benchmark, which includes table QA and row/column attributions from Wikipedia and business PDF documents with complex layouts. Our multi-agent LLM framework, `MATSA`, converts tables into HTML, augments raw table data with descriptive context, and retrieves semantically relevant rows/columns that support atomic facts in the answers. Future work may extend these methods to low-resource domains and other semi-structured documents, such as charts, info graphics, and diagrams.

9 Ethics Statement

We utilize the publicly available Table QA corpora—FetaQA (Nan et al., 2021), Totto (Parikh et al., 2020), and AITQA (Katsis et al., 2022)—for this research without introducing new human annotations. We preprocess the tables and PDF documents to obtain ground truth attribution annotations. Publicly accessible API-based LMMs and LLMs (e.g., GPT-4V, GPT-4, GPT-3.5) are employed in our experiments. All evaluations are conducted automatically without any human intervention. No Personally Identifiable Information (PII) is utilized at any stage of our experiments. The intended applications of our work are strictly for research purposes, and we do not endorse any commercial adaptation without adequate testing. Given the propensity of Large Language Models to hallucinate, we ensure that no LLM-generated text is used for training or fine-tuning downstream models in violation of commercial licenses. For a comprehensive understanding of LLM safety risks and mitigation strategies, we refer users to relevant works by (Kumar et al., 2024; Cui et al., 2024; Luu et al., 2024).

10 Limitations

1. **Limited to Table Structures in Documents**: Our work focuses on providing citations for LLM-generated answers using tabular information. All samples in our benchmark derive supporting citations exclusively from tables. While real-world applications involve complex documents that include unstructured text, charts, graphs, diagrams, and form fields, our task is a simplified approach to address a specific aspect of the broader issue of LLM hallucinations.

2. **English-only Evaluations**: Our study is confined to evaluating table structure attribution for table QA in English. Adapting to other low-resource languages will necessitate the collection of appropriate table QA and attribution datasets. Our proposed MATSA framework utilizes publicly available LLM APIs which have demonstrated reasonable language understanding capabilities across diverse languages. Hence, we encourage future work to adapt our task and framework for low-resource languages.

3. **LLM/LMM API Cost and Performance Fluctuations**: Our work leverages API-accessible Large Language Models and Large Multimodal Models. The cost associated with these model APIs varies based on the token count in the request and response, as well as image resolution and dimensions. Additionally, these API-based models are susceptible to performance fluctuations.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *ArXiv*, abs/2404.05961.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tianyu Cui, Yanling Wang, Chuanpu Fu, Yong Xiao, Sijia Li, Xinhao Deng, Yunpeng Liu, Qinglin Zhang, Ziyi Qiu, Peiyang Li, Zhixing Tan, Junwu Xiong, Xinyu Kong, Zujie Wen, Ke Xu, and Qi Li. 2024. Risk taxonomy, mitigation, and assessment benchmarks of large language model systems.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Siqing Huo, Negar Arabzadeh, and Charles Clarke. 2023. Retrieving supporting evidence for generative question answering. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 11–20.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. HAGRID: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv:2307.16883*.

Vasileios Katranidis and Gabor Barany. 2024. Faaf: Facts as a function for the evaluation of rag systems. *arXiv preprint arXiv:2403.03888*.

Yannis Katsis, Saneem Chemmengath, Vishwajeet Kumar, Samarth Bharadwaj, Mustafa Canim, Michael Glass, Alfio Gliozzo, Feifei Pan, Jaydeep Sen, Karthik Sankaranarayanan, and Soumen Chakrabarti. 2022. AIT-QA: Question answering dataset over complex tables in the airline industry. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 305–314, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Ashutosh Kumar, Sagarika Singh, Shiv Vignesh Murty, and Swathy Ragupathy. 2024. The ethics of interaction: Mitigating security threats in llms.

Quan Khanh Luu, Xiyu Deng, Anh Van Ho, and Yorie Nakahira. 2024. Context-aware llm-based safe control against latent risks.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Linyong Nan, Chia-Hsuan Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryscinski, Nick Schoelkopf, Riley Kong, Xiangru Tang, Murori Mutuma, Benjamin Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, and Dragomir R. Radev. 2021. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642.

Ben Snyder, Marius Moisescu, and Muhammad Bilal Zafar. 2023. On early detection of hallucinations in factual question answering. *ArXiv*, abs/2312.14183.

Yuan Sui, Jiaru Zou, Mengyu Zhou, Xinyi He, Lun Du, Shi Han, and Dongmei Zhang. 2023. Tap4llm: Table provider on sampling, augmenting, and packing semi-structured data for large language model reasoning. *ArXiv*, abs/2312.09039.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *ArXiv*, abs/2104.14690.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *ArXiv*, abs/2401.11817.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyue Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *ArXiv*, abs/2310.11441.