ChartCitor: Answer Citations for ChartQA via Multi-Agent LLM Retrieval

Kanika Goswami
IGDTUW, Delhi
India

Puneet Mathur
Adobe Research
USA

Ryan Rossi
Adobe Research
USA

Franck Dernoncourt
Adobe Research
USA

Abstract

Large Language Models (LLMs) can perform chart question-answering tasks but often generate unverified hallucinated responses. Existing answer attribution methods struggle to ground responses in source charts due to limited visual-semantic context, complex visual-text alignment requirements, and difficulties in bounding box prediction across complex layouts. We present ChartCitor, a multi-agent framework that provides fine-grained bounding box citations by identifying supporting evidence within chart images. The system orchestrates LLM agents to perform chart-to-table extraction, answer reformulation, table augmentation, evidence retrieval through pre-filtering and re-ranking, and table-to-chart mapping. ChartCitor outperforms existing baselines across different chart types. Qualitative user studies show that ChartCitor helps increase user trust in Generative AI by providing enhanced explainability for LLM-assisted chart QA and enables professionals to be more productive.

CCS Concepts

• **Information systems → Information extraction.**

Keywords

Visual Fact Checking, Information Extraction, Multimodal Retrieval, LLM Agents

ACM Reference Format:

Kanika Goswami, Puneet Mathur, Ryan Rossi, and Franck Dernoncourt. 2025. ChartCitor: Answer Citations for ChartQA via Multi-Agent LLM Retrieval. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Chart data finds extensive use across diverse domains such as healthcare, finance, and education. Recently, LLMs such as Llama-3.2 [16], Claude-3.5 Sonnet, and GPT-4V [1] have proven effective in utilizing in-context learning and visual prompting to interpret and reason over chart images. However, these models tend to hallucinate — generate answers with semantically plausible but factually incorrect information — which undermines their reliability and erodes user trust [14, 19]. While existing approaches attempt to address hallucination by grounding LLM-generated responses in source documents through citation mechanisms [7], charts present unique challenges: (i) complex mapping between visual elements and underlying data, (ii) limited contextual information due to compressed visual data representation, (iii) difficulty in localizing chart elements across diverse visualization types and layouts, and (iv) ambiguity in alignment between text descriptions and visual elements. Prior research has explored various approaches to address this challenge, including instruction tuning [8], in-context learning [5], and natural-language inference (NLI)-based post-hoc attribution methods [4]. However, these approaches have primarily focused on attributing entire charts rather than specific structural elements [6], limiting their practical utility. To address these limitations, we propose ChartCitor, a system that provides visual evidence for generated answers by identifying and highlighting relevant chart elements through bounding box annotations. ChartCitor works by orchestrating multiple specialized LLM agents to: (1) extract structured data table from charts, (2) break down answers into logical steps, (3) generate contextual descriptions for rows/columns, (4) identify supporting evidence through pre-filtering and re-ranking to connect specific table cells to claims, and (5) localize the selected cells in the chart image. ChartCitor helps professionals save time on fact-checking LLM-generated answers and enhances user trust by providing reliable and logically-explained citations sourced from charts.

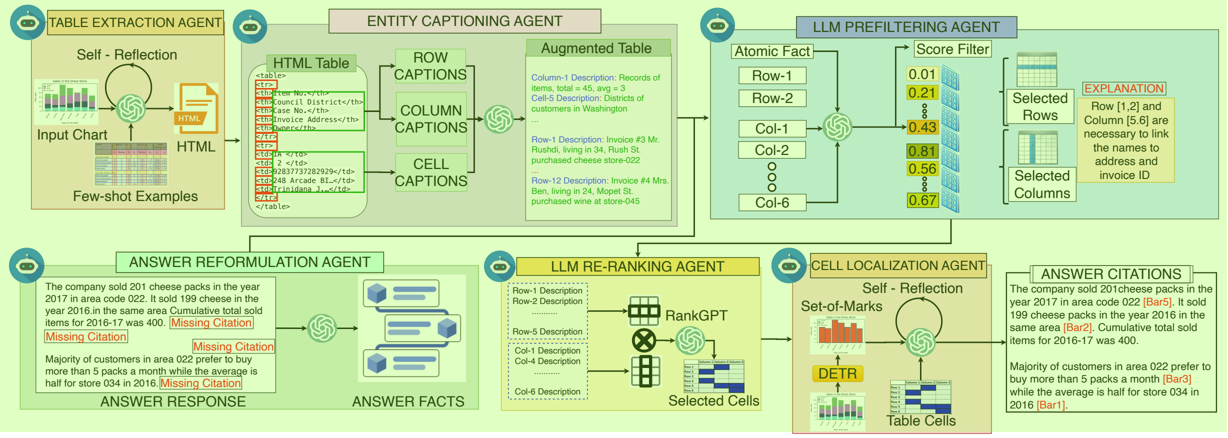
2 ChartCitor

We aim to solve the Fine-grained Structured Chart Attribution task which involves identifying graph elements (e.g bars, lines, pies in chart images) that support factual claims in a generated text response to a user's question. We propose ChartCitor (Fig. 1) – a multi-agent framework that provides fine-grained citations for generated answers grounded in chart image by orchestrating multiple LLM agents, which is explained as follows:

(1) Chart2Table Extraction Agent: Charts are predominantly present in PDFs or scanned documents that need to be converted into structured table formats (e.g., CSV, or HTML). We utilize GPT-4V to comprehend PDF images and output corresponding HTML without the need for external OCR using few shot prompting to identify cell data across each row/column. We use visual self-reflection [13] to

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn



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

Column-1 Description: Records of items, total = 45, avg = 3
Cell-5 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

Row [1,2] and Column [5.6] are necessary to link the names to address and invoice ID

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

The company sold 201cheese packs in the year 2017 in area code 022 [Bar5]. It sold 199 cheese packs in the year 2016 in the same area [Bar2]. Cumulative total sold items for 2016-17 was 400.

Majority of customers in area 022 prefer to buy more than 5 packs a month [Bar3] while the average is half for store 034 in 2016 [Bar1].

Figure 1: ChartCitor - a multi-agent framework that performs table extraction, answer reformulation, entity captioning, row/col retrieval, and cell localization in chart images to ground answers.

provide the GPT-4V with its own rendered HTML and data table output to check for consistency between the re-plotted LLM output and the original chart. In case of inconsistencies in the data extraction, the LLM refines its output until the extracted table data is error-free. **(2) Answer Reformulation Agent**: The answer to be attributed, which can be AI-generated or otherwise, may be composed of multiple facts, numerical formulations and multi-hop logic. Each fact may be sourced from a different row/column in the chart table. To facilitate precise citations, re-framing the answer statement into a chain of reasoning steps helps to better retrieve the correct citations from the table. We convert the answer statement into a hierarchy of reasoning thoughts/arguments via few-shot in-context prompting, ensuring the resultant answer arguments are independent sentences without any deviation in their collective meaning from the original answer statement.

(3) Entity Captioning Agent: Understanding tabular data extends beyond simple cell interpretation, requiring comprehension of how cell information relate to both the table's structure and its broader context. Tables often present analytical challenges through ambiguous content, including technical terminology, contextless numeric values, domain-specific symbols, and hierarchical row/column headers. These ambiguities impede reliable evidence extraction and citation validation through semantic matching. Our solution leverages LLMs in an unsupervised manner to generate rich, multi-layered contextual descriptions: **(i) Row Captioning**: Our system employs GPT-4o to generate comprehensive row-level descriptions that capture complex patterns across features, summarize temporal trends, highlight significant dates and provide comparative analysis with respect to outliers within each row. **(ii) Column Captioning**: We generate detailed captions for each column using GPT-4o to explain ambiguous measurement units, symbols, empty cell spaces, and technical relationship of it's contents with corresponding row headers. **(iii) Cell Captioning**: Row and column-level captions may highlight broader trends but the fine-grained cell level information needs to be contextualized in terms of its associated row and column headers. Captioning agent uses GPT-4o to describe the importance of each cell in the context of its associated row and columns.

Table Cell Retrieval: We use retrieve-then-rank approach to identify the most relevant table cells. Our two-step approach begins with LLM-based pre-filtering to reject irrelevant rows and columns. We then employ LLM re-ranking to retrieve the most precise cell-level matches, ensuring both comprehensive coverage and accuracy in the final selection.

(4) LLM Pre-filtering Agent: We hypothesize that some of the table rows/ columns are likely to be unrelated to the answer facts. Passing irrelevant and distracting table entities to the LLM-based re-ranker can mislead it, negatively impacting the ranking process. Inspired by [11], the LLM-based pre-filtering step uses chain-of-thought [18] followed by Plan and Solve [17] prompting techniques to generate a relevance score for each row/column based on the significance of its descriptive caption to the given answer statement (between 0 to 1). Additionally, we prompt the LLM to explain its rationale behind the score generation to enhance explainability and avoid hallucinations. We establish a specific threshold (usually $0.3 - 0.5$) for row/column filtering to retain potential citations that are sent to the re-ranker, discarding those falling below the threshold. This implementation significantly reduces the number of noisy of rows and columns that can misguide the re-ranker, leading to an improved citation retrieval performance.

(5) LLM Re-ranking Agent: LLMs providing citations that don't directly support their claims can be interpreted as a form of hallucination which may diminish user confidence. To solve this, we retrieve the set of table cells that are collectively both sufficient and directly relevant to the answer claims. We use RankGPT [15], a listwise LLM re-ranker to re-rank the table cells extracted from the intersection of rows and columns selected in the pre-filtering stage. Additionally, we prompt GPT-4o to provide a layer-of-thought [3] explanation of its rationale in ranking the cell items, enhancing the transparency in the citation mechanism by enforcing a logical coherence in the evidence chain.

(6) Cell Localization Agent: The final step maps cited table cells to their corresponding visual elements in the chart image. This agent leverages DETR trained [2] on ChartQA data to identify all possible data marks (bars, line segments, pie slices) using image processing

Method	IoU
Kosmos-2	3.89
LISA	4.34
GPT-4V (Direct Bbox Decoding)	12.5
Claude-3.5 (Sonnet Direct Bbox Decoding)	13.8
DETR [2]+ Set-of-Marks Prompting [21]	18.6
ChartCitor	**27.4**

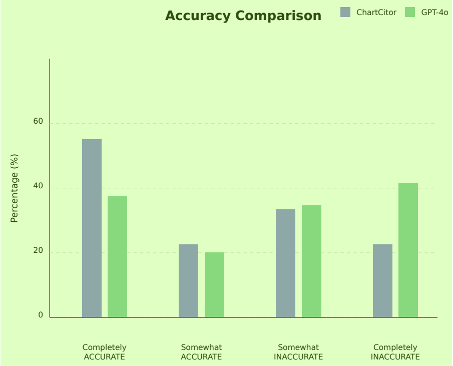


Figure 2: (a) Ablation analysis of multimodal feedback agents; (b) User Evaluation of ChartCitor

algorithms. GPT-4V with few-shot set-of-marks prompting [20] then identifies elements corresponding to the cited cells. The agent generates bounding box coordinates for the relevant visual elements, employing visual self-reflection to verify precise correspondence between highlighted regions and cited data points.

3 Implementation Details

All constituent agents utilize textual LLM APIs such as those provided by OpenAI (GPT-4o, GPT-4V) or Claude Sonnet-3.5. We use ChatGPT-4o as the base multimodal language model (MLLM) for ChartCitor. We convert data tables from TabCite benchmark [10] into bar/pie/line charts along with paired QA.

Evaluation: We adopt visual Intersection over Union (IoU) as principal metric for chart attribution tasks. Detected regions in the chart image are matched to ground truth regions (e.g., bars in barplot or pies in piechart) based on a threshold value of IoU $\geq$ 0.9. Unlike bar charts and pie charts, where detected regions can be matched to discrete ground truth regions, line charts involve discrete points. Since grounding models generate bounding boxes or regions, we compute the proportion of ground truth points covered within the detected region(s) over total points detected.

Baselines: (1) Zero-shot LLM Bounding Box Prompting – We prompt GPT-4o and Claude 3.5 Sonnet to predict normalized bounding box coordinates for chart components based on input text and the visual chart. (2) Kosmos-2 [12] is a multimodal LLM with text-to-visual grounding capabilities. It represents object locations as Markdown links for generating bounding boxes for visual grounding tasks. (3) LISA (Large Language Instructed Segmentation Assistant) [9] is a reasoning-based zero-shot segmentation model that generates masks from implicit and complex textual queries.

4 Results and Discussion

Table 2(a) shows quantitative results that demonstrate that ChartCitor consistently outperforms the baselines across all chart types, highlighting its robustness and effectiveness in visual chart understanding. ChartCitor achieves better performance compared to directly prompting LLMs to predict bounding boxes. Further, even using GPT-4V with set of marks prompting over detected chart elements show weak performance. Kosmos-2 and LISA perform poorly, with very low IoU scores, highlighting their inability to handle factual grounding in charts due to insufficient visual and numerical reasoning. Interestingly, all tested method including our proposed ChartCitor, zero-shot LMM prompting, LISA and KOSMOS2 struggle with interpreting complex geometrical proportions in pie charts due to their difficulty in handling non-rectangular bounding box segmentation task. Further, we conducted a user study (Fig. 2(b) to evaluate the citation accuracy and perceived utility of fine-grained chart attribution provided by ChartCitor. Five participants evaluated 250 randomly sampled question-answer pairs with associated chart images to study the usefulness and accuracy of the citations provided by ChartCitor compared to direct GPT-4o prompting. The evaluation results demonstrated strong positive reception, with participants rating the attributions as *Completely Accurate* (41% vs 28%) or *Somewhat Accurate* (17% vs 15%) for verifying chart-based question answering accuracy in ChartCitor and GPT-4o, respectively. Attributions were found to be more "Completely Inaccurate" ChartCitor than GPT-4o (17% vs 31%). Participants described the citations as a handy tool in making verification of LLM-generated answers easier (*"...can help me to quickly verify trends in charts, cutting down the time I spent on 10 documents from 5 hrs to 20 mins."*).

5 Conclusion

We introduced ChartCitor, which grounds LLM-generated QA responses to chart elements using agentic orchestration and set-of-marks prompting. The system outperforms baselines by 9-15% and shows promise for rich document QA over PDF collections. While currently effective for single-chart citations, future work can address multi-chart interactions, hallucination mitigation, and explicit citation-text mapping to enhance trustworthy multimodal content generation.

References

[1] 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* (2023).

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*. Springer, 213–229.

[3] Wachara Fungwacharakorn, Nguyen Ha Thanh, May Myo Zin, and Ken Satoh. 2024. Layer-of-Thoughts Prompting (LoT): Leveraging LLM-Based Retrieval with Constraint Hierarchies. *arXiv preprint arXiv:2410.12153* (2024).

[4] 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. 2023. 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)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 16477–16508. https://doi.org/10.18653/v1/2023.acl-long.910

[5] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6465–6488. https://doi.org/10.18653/v1/2023.emnlp-main.398

[6] 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*. 11–20.

[7] 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. *Comput. Surveys* 55, 12 (2023), 1–38.

[8] 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* (2023).

[9] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9579–9589.

[10] Puneet Mathur, Alexa Siu, Nedim Lipka, and Tong Sun. 2024. MATSA: Multi-Agent Table Structure Attribution. In *Conference on Empirical Methods in Natural Language Processing*. https://aclanthology.org/2024.emnlp-demo.26/

[11] Baharan Nouriinanloo and Maxime Lamothe. 2024. Re-Ranking Step by Step: Investigating Pre-Filtering for Re-Ranking with Large Language Models. *arXiv preprint arXiv:2406.18740* (2024).

[12] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824* (2023).

[13] Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *ArXiv* abs/2303.11366 (2023). https://api.semanticscholar.org/CorpusID:257636839

[14] Ben Snyder, Marius Moisescu, and Muhammad Bilal Zafar. 2023. On Early Detection of Hallucinations in Factual Question Answering. *ArXiv* abs/2312.14183 (2023). https://api.semanticscholar.org/CorpusID:266521062

[15] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14918–14937. https://doi.org/10.18653/v1/2023.emnlp-main.923

[16] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[17] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:258558102

[18] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv* abs/2201.11903 (2022). https://api.semanticscholar.org/CorpusID:246411621

[19] Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. 2024. Hallucination is Inevitable: An Innate Limitation of Large Language Models. *ArXiv* abs/2401.11817 (2024). https://api.semanticscholar.org/CorpusID:267069207

[20] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441* (2023).

[21] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chun yue Li, and Jianfeng Gao. 2023. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. *ArXiv* abs/2310.11441 (2023). https://api.semanticscholar.org/CorpusID:266149987