

Format Matters: The Robustness of Multimodal LLMs in Reviewing Evidence from Tables and Charts

Xanh Ho,¹ Yun-Ang Wu,^{2*} Sunisth Kumar,³
Florian Boudin,⁴ Atsuhiro Takasu,¹ Akiko Aizawa^{1,3}

¹National Institute of Informatics, Japan

³The University of Tokyo, Japan

{xanh, takasu, aizawa}@nii.ac.jp
sunisth@g.ecc.u-tokyo.ac.jp

²National Taiwan University

⁴Inria, LS2N, Nantes Université, France

r11944072@csie.ntu.edu.tw
florian.boudin@univ-nantes.fr

Abstract

With the growing number of submitted scientific papers, there is an increasing demand for systems that can assist reviewers in evaluating research claims. Experimental results are a core component of scientific work, often presented in varying formats such as tables or charts. Understanding how robust current multimodal large language models (multimodal LLMs) are at verifying scientific claims across different evidence formats remains an important and underexplored challenge. In this paper, we design and conduct a series of experiments to assess the ability of multimodal LLMs to verify scientific claims using both tables and charts as evidence. To enable this evaluation, we adapt two existing datasets of scientific papers by incorporating annotations and structures necessary for a multimodal claim verification task. Using this adapted dataset, we evaluate 12 multimodal LLMs and find that current models perform better with table-based evidence while struggling with chart-based evidence. We further conduct human evaluations and observe that humans maintain strong performance across both formats, unlike the models. Our analysis also reveals that smaller multimodal LLMs (under 8B) show weak correlation in performance between table-based and chart-based tasks, indicating limited cross-modal generalization. These findings highlight a critical gap in current models' multimodal reasoning capabilities. We suggest that future multimodal LLMs should place greater emphasis on improving chart understanding to better support scientific claim verification.

Code — <https://github.com/Alab-NII/tables-vs-charts>

1 Introduction

Scientific claim verification requires models to determine whether a given claim is supported or not, based on the provided evidence. This evidence can take the form of text, tables, or charts. In recent years, several datasets have been introduced for this task, including SciFact (Wadden et al. 2020), HealthVer (Sarrouti et al. 2021), SciTab (Lu et al. 2023), and MuSciClaims (Lal et al. 2025). Depending on the dataset, the evidence may be textual (e.g., SciFact), tabular (e.g., SciTab), or visual (e.g., MuSciClaims). However,

existing datasets typically represent evidence in only a single format (details in Section 2). While the recent SciVer dataset (Wang et al. 2025) incorporates multiple evidence modalities such as textual, tabular, and visual, it still presents tables in visual rather than text-based table format. Moreover, the different types of evidence in SciVer provide complementary information rather than expressing the same content across modalities.

In the era of generative AI and large language models (LLMs), researchers can now produce research papers more efficiently with support from AI agents (Press et al. 2024; Si, Yang, and Hashimoto 2025), leading to a significant increase in the number of submissions. As a result, automated peer review systems and tools that assist human reviewers are becoming increasingly important. A critical part of many research papers is the experimental results section, which presents key findings that support the main claims. These results are typically presented in tables or charts, depending on the author's preference. Understanding how multimodal LLMs handle different evidence formats is essential, especially as they are integrated into review systems. If LLMs perform well on one format but poorly on another, this could lead to biased or incomplete evaluations. Therefore, assessing their robustness across formats is vital for building reliable and generalizable AI-assisted review systems. However, it remains unclear whether current multimodal LLMs can consistently verify claims across different evidence formats. We argue that an effective review system must evaluate claims accurately, regardless of how the supporting evidence is presented.

From a different task perspective, Chart2Code and Data2Chart are tasks that focus on converting between charts and the code used to generate them, or on transforming tabular data into visual charts. Several datasets have been developed for these tasks, such as MatPlotBench (Yang et al. 2024), Plot2Code (Wu et al. 2025), and ChartMimic (Yang et al. 2025). These datasets typically provide paired examples of code and the corresponding chart, with an emphasis on the transformation between these two formats. While our research also involves both data formats, specifically charts and tables (assuming that tables can be derived from the underlying code), our focus is different. We treat these formats as distinct types of evidence for a downstream task, claim verification, rather than for format conversion.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

*Research conducted during internship at NII, Japan.

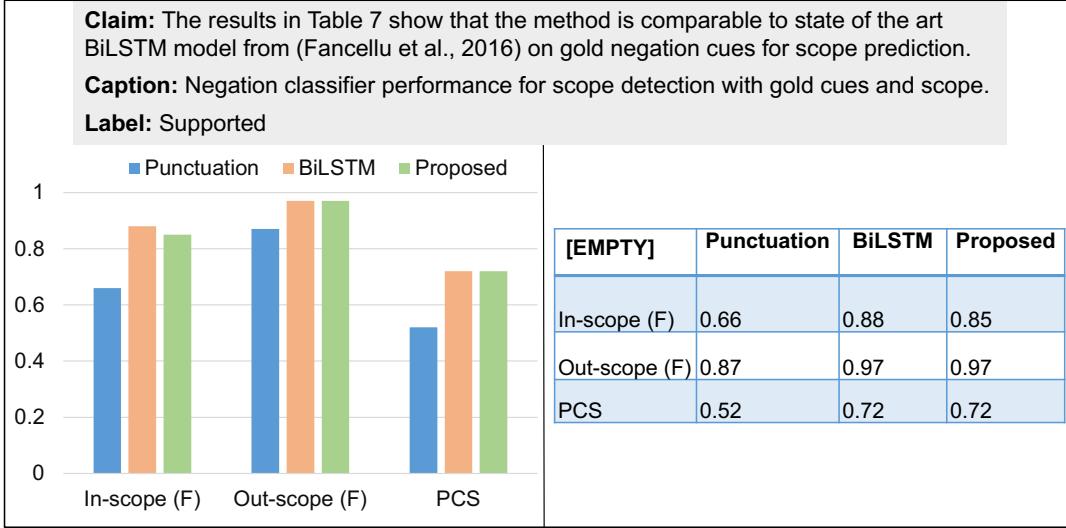


Figure 1: An example of the claim verification task in our experiment, featuring both types of evidence: table and chart formats that represent the same information. It is noted that the original example is from the SciTab dataset. To ensure a fair comparison between table- and chart-evidence formats, we modified claims by replacing references to tables with figures (e.g., “Table 7” → “Figure 7”) to make them compatible with chart-format evidence.

In this paper, we aim to investigate the robustness of current multimodal LLMs in performing claim verification across different evidence formats. Specifically, we consider both tables and charts as representations of the evidence data. We found that existing datasets do not fully meet the requirements of our task, as they typically lack aligned table and chart evidence that convey the same underlying information. To address this gap, we extend two existing datasets, namely SciTabAlign (Ho et al. 2025) and ChartMimic (Yang et al. 2025), to construct a new dataset tailored to our claim verification setting. In particular, we curate instances that consist of a scientific claim, a table serving as evidence, and a corresponding chart that represents the same data as the table. This alignment enables a systematic evaluation of multimodal LLM performance in verifying claims using different, yet equivalent, evidence formats. Figure 1 illustrates an example of the claim verification task used in our experiments.

Using the extended versions of these datasets, we conduct a series of experiments to assess model performance across various settings, including table-only input, chart-only input, and combined table–chart input. We also evaluate performance across different chart types: basic bar charts, symbol bar charts, line charts, and swapped charts. Our results from 12 multimodal LLMs show that current models struggle to verify claims when evidence is presented as charts but perform better when the same information is provided in tables, indicating a strong reliance on structured, text-like input. This discrepancy highlights a notable limitation in current multimodal LLMs’ ability to interpret visual data, even when the content is semantically equivalent. To better understand this gap, we asked human annotators to complete the same tasks. Unlike the models, humans performed consistently well across both formats, suggesting that the observed difficulty stems from model limitations rather than task am-

biguity. These findings underscore the need for improved multimodal reasoning and better alignment between visual and textual representations in future model development.

In summary, our main contributions are as follows. (1) We extend two existing datasets, SciTabAlign and ChartMimic, to create enhanced versions that support scientific claim verification using different evidence formats representing the same information. (2) We comprehensively evaluate 12 multimodal LLMs under three input settings: table-only, chart-only, and a combination of both. We also investigate performance across different chart types. (3) Our experimental results show that current multimodal LLMs struggle to process chart-based input, while performing better with text-based table input. Human performance on the same task confirms that people do not face similar difficulties with different evidence formats, highlighting a limitation in current multimodal LLMs.

2 Related Work

2.1 Claim Verification

Claim verification or fact-checking is a well-established research task, with numerous datasets proposed across various domains (Guo, Schlichtkrull, and Vlachos 2022), such as news articles (Wang 2017) and Wikipedia-based resources (Thorne et al. 2018). In this work, we focus specifically on datasets that involve tables or charts within scientific papers. Table 1 provides a summary of existing datasets relevant to this domain. Although datasets such as TabFact (Chen et al. 2020), InfoTabs (Gupta et al. 2020), and FEVEROUS (Aly et al. 2021) also include tables, they are based on Wikipedia and therefore fall outside the scope of our study. As a result, we do not include them in the table.

An important aspect of scientific claim verification is the

Year	Name	Main Task	Input	Output	Size	Domain	Dataset Creation
2021	SEM-TAB-FACTs (Wang et al. 2021)	Verdict	Single table + claim	Label	5,715	ScienceDirect	Crowdsourcing
2023	SciTab (Lu et al. 2023)	Verdict	Single table* + claim	Label	1,225	Computer science	Authentic claims + LLM with human verification
2025	SciAtomicBench (Zhang et al. 2025b)	Verdict	Single table + claim	Label	2,568	Multi	LLM with human verification
2025	SciVer (Wang et al. 2025)	Verdict	Multi paragraphs, tables, charts + claim	Label	3,000	Computer science	Experts
2025	MuSciClaims (Lal et al. 2025)	Verdict	Figures + claim	Label	918	Life sciences	Authentic claims + experts
2025	SciTabAlign (Ho et al. 2025)	Verdict	Single table* + claim	Label	372	Computer science	SciTab + Experts
2020	PlotQA (Methani et al. 2020)	QA	Single chart + question	Answer	28.9M	Online data sources (e.g., Open Government Data)	Semi-automatic
2022	ChartQA (Masry et al. 2022)	QA	Single chart + question	Answer	9,608	Various Websites (Statista, Pew, OWID, OECD)	Crowdsourcing
2024	ChartBench (Xu et al. 2024)	QA	Single chart + question	Answer	18.9K	Kaggle	LLM with human verification
2024	CharXiv (Wang et al. 2024)	QA	Single chart + question	Answer	2,323	arXiv	Humans
2024	MMC-benchmark (Liu et al. 2024)	QA	Chart(s) + question	Answer	2,000	Various sources (e.g., arXiv, Statista, VisText)	Humans + LLM
2025	DomainCQA (Zhong et al. 2025)	QA	Single chart + question	Answer	1,890	Astronomy	Semi-automatic
2025	ChartQAPro (Masry et al. 2025)	QA	Single chart + question	Answer	1,948	Various Websites (Pew, Tabelau, PPIC, OWID)	Humans + LLM
2024	SciTabQA (Ghosh et al. 2024)	QA	Single table* + question	Answer	822	Computer Science	Humans
2024	SciTaT (Zhang et al. 2025a)	QA	Single table* + text + question	Answer	953	Computer Science	LLM with human verification
2024	MatPlotBench (Yang et al. 2024)	Data2Chart	Raw data + Instruction	Chart	100	Matplotlib, Origin	Crawl
2025	Plot2Code (Wu et al. 2025)	Chart2Code	Chart(s) + Instruction	Code	132	Python’s matplotlib, Python’s plotly and R’s plotly	Crawl
2025	ChartMimic (Yang et al. 2025)	Chart2Code	Single chart + Instruction	Code	4,800	arXiv, Twitter, Reddit, Matplotlib, Stackoverflow	Humans
2025	ChartEdit (Zhao et al. 2025)	Chart2Code	Single chart + Instruction	Code	1,405	arXiv	Humans + LLM with human verification

Table 1: Existing related datasets and their information. It is noted that our list is not comprehensive, as we only include datasets related to experimental results or chart numbers. Also, we focus solely on tasks involving scientific papers; therefore, we may ignore datasets from the Wikipedia domain. Due to space constraints, we use the term *Verdict* to represent the *claim verification* task. * indicates that the table provided in the dataset is represented in a text-based table format (e.g., JSON), which can be directly used to generate a chart.

authenticity of the claims being evaluated. Among the existing datasets, only SciTab, MuSciClaims, and SciTabAlign reuse authentic claims extracted directly from original scientific papers. Furthermore, we observe that current datasets for claim verification do not include both table and chart formats representing the same underlying information. While some datasets do contain tables, these are often embedded as figures rather than provided in a structured, text-based format (e.g., JSON). This limits the ability to programmatically generate corresponding charts.

2.2 Other Scientific Tasks Involving Tables and Figures

In addition to claim verification, several related tasks have been explored in the literature, including question answering (QA), chart-to-code generation, and chart-to-text generation. For the QA task, the objective is to answer natural language questions based on structured data such as tables or visual data such as charts. Existing datasets can be broadly

categorized by their input modality: some focus solely on charts (e.g., (Methani et al. 2020; Masry et al. 2022; Xu et al. 2024; Wang et al. 2024; Zhong et al. 2025; Masry et al. 2025)), others on tables (e.g., (Ghosh et al. 2024; Zhang et al. 2025a)), and a few incorporate both charts and tables (e.g., (Foroutan et al. 2025)). These tasks share similarities with claim verification in that they require comprehension of structured or semi-structured data, but differ in terms of output type and the specific reasoning involved.

The chart-to-code generation task, sometimes referred to as derendering, involves generating the underlying code (e.g., Matplotlib) that would produce a given chart. Some studies focus solely on recovering the chart specification from the image (Wu et al. 2025; Yang et al. 2025), while others also address chart editing by predicting modifications or supporting interactive changes to the chart (Zhao et al. 2025). In contrast, the chart-to-text generation task aims to produce natural language descriptions or summaries of charts. While it shares visual understanding components

with chart-to-code generation, the output is unstructured text, which introduces distinct challenges in representation and language generation (Kantharaj et al. 2022). Our research differs from both of these directions. Rather than focusing on converting charts to code or to tables (which can be derived from the code), we investigate how different evidence formats such as tables and charts affect model performance on the claim verification task. This shifts the focus from generation to reasoning and evidence interpretation.

3 Datasets

In this section, we first introduce the existing datasets, SciTabAlign and ChartMimic, which are derived from scientific papers. We then describe the process we applied to create enhanced versions, SciTabAlign+ and ChartMimic+, which are used in our experiments.

3.1 Existing Datasets

SciTab and SciTabAlign. SciTab (Lu et al. 2023) is a scientific claim verification dataset that contains authentic scientific claims. Its tables are presented in text-based format, making them suitable for generating corresponding charts. SciTabAlign (Ho et al. 2025) is an extended version of SciTab, with explanations added for each claim label. Additionally, ambiguous cases have been removed, particularly those where the available information is insufficient to perform the task or where issues with the table or claim introduce ambiguity. Since our goal is to investigate the robustness of multimodal LLMs in handling different evidence formats, we chose to work exclusively with SciTabAlign, where all ambiguous cases have been removed. SciTabAlign includes 136 tables and 372 claims. Each sample consists of a text-based table format, a table caption, a claim, and a label for the claim. The label can be either Supported or Refuted.

ChartMimic. ChartMimic (Yang et al. 2025) is a specially designed dataset for the chart2code task. It consists of two sub-datasets: Direct Mimic and Customized Mimic. Direct Mimic is intended to evaluate a model’s ability to generate code based on a given chart image. In contrast, Customized Mimic tests a model’s ability to follow specific instructions to mimic the style of a given chart and use that style to generate code for new data. The resulting code is then used to produce a new chart. The dataset includes a wide variety of chart types, from popular ones like bar charts and line charts to less common ones like 3D charts. For the Direct Mimic task, the dataset contains 600 original samples. Each sample consists of a PNG image file and a corresponding Python code file that can be used to generate the chart.

3.2 Datasets Used in Our Experiments

Processed Version of SciTabAlign (SciTabAlign+). We construct SciTabAlign+ through a series of processing steps, as follows. (1) We first normalize the table data by removing HTML-like tags (e.g., `<bold>`, `<italic>`), bracket tags (e.g., [BOLD], [ITALIC]), and standardizing numeric values so that they can be easily converted to floats for chart generation. After this process, we retain 70 out of the original 136 tables. We manually inspected the discarded tables

and found several issues. Many contained empty cells, making them difficult to process automatically for accurate chart generation. Additionally, some tables included numerical values with ambiguous units (e.g., “M” for million), or textual information instead of score values, which are less relevant for our visualization purposes. As a result, we decided not to generate charts for these tables. (2) Starting with the 70 selected tables and 162 associated claims, we designed four types of charts: **Basic bar charts**, where different colors represent different bars; **Symbol bar charts**, where we replace colors with symbols such as “/” or “-” to form visual patterns; **Line charts**, using lines to connect data points; and **Swapped charts**, where we interchange the x-axis labels by transforming ‘methods’ into ‘metrics’ and vice versa. Figure 2 shows an example of the four chart types used in SciTabAlign+. We consider this extended dataset a *plus* version of SciTabAlign, called SciTabAlign+, where each claim can now be supported by both table and chart-based evidence.

In summary, our dataset includes 372 claims with table evidence and 648 claims (162 for each of the four chart types) with chart evidence. To ensure a fair comparison between the table-evidence and chart-evidence formats, if a claim contains references such as “the results in Table 4 show that...,” we replace the word table with figure when evaluating on the chart-only setting.

Processed Version of ChartMimic (ChartMimic+). We focus only on line charts and bar charts, excluding other types of charts, such as 3D charts, from the scope of this research. This choice is motivated by the fact that bar and line charts are not only the most common chart types in the ChartMimic dataset but also the most frequently used for presenting experimental results in academic papers. From the Direct Mimic subtask in ChartMimic, which contains 600 charts, we select only bar and line charts. Specifically, we obtain 70 line charts and 80 bar charts. As mentioned above, each sample in ChartMimic consists of a PNG image file and a corresponding Python file used to generate the chart. Using the provided Python code, we automatically extract the underlying table data. We then present the following information to the annotators (four NLP researchers): the extracted table, the corresponding bar or line chart, and the original Python code. Annotators are asked to verify and edit the table so that it accurately reflects the chart. If the chart and table are deemed suitable for the claim verification task, we ask the annotators to write one supported claim and one refuted claim for each table. We encourage annotators to write complex claims rather than simple comparative ones such as “A is better than B.” During the annotation process, we decided to exclude sub-charts, as our focus is on single-chart analysis. Multi-chart scenarios are left for future work. We also exclude charts generated using `np.random.normal` because they do not represent real data suitable for scientific content.

In total, we obtain 152 claims based on 52 bar charts and 24 line charts. To ensure consistency between each chart and its corresponding table, we include a caption field where annotators can add any information shown within the chart, such as embedded captions.

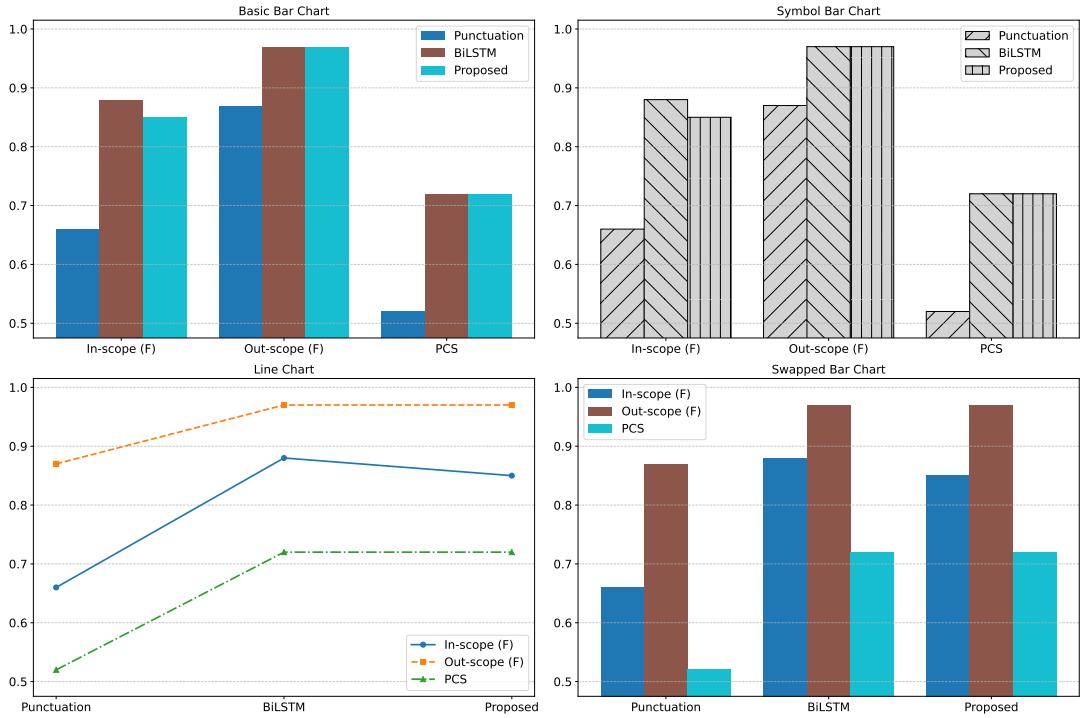


Figure 2: An example of four chart types used in SciTabAlign+.

4 Experimental Setup

Settings. We run our experiments under three settings: table-only input, chart-only input, and a combination of both table and chart as input. For SciTabAlign+, we use four different types of charts: bar charts, symbol charts, line charts, and swapped charts.

Models. We conduct our investigation using 12 open-source multimodal LLMs from four different families, all capable of processing both image and text inputs. Specifically, the models include InternVL3 (1B, 8B, 14B, and 38B) (Zhu et al. 2025), Qwen-VL 2.5 (3B, 7B, 32B, and 72B) (Bai et al. 2025), LLaVA-v1.6 (llava-v1.6-mistral-7b, llava-v1.6-vicuna-13b, and llava-v1.6-34b) (Li et al. 2024), and Llama-3.2 (11B-Vision) (Grattafiori et al. 2024). We use the instruct-tuned versions of models in our experiments.

Promptings. Following the previous work (Wang et al. 2025), we use zero-shot Chain-of-Thought prompting (Kojima et al. 2022) as the primary evaluation setting in our paper. Figure 3 shows the prompting used in experiments combining chart and table inputs. When no evidence type is used, we set it to *None* and simplify phrases like “Use the provided table and image” to “Use the provided table.”

Evaluation. Following the evaluation protocols used in the SciTab and SciTabAlign datasets, we adopt macro-F1 as our primary evaluation metric.

Adopted Zero-shot CoT Prompting

```
[{"type": "image", "image": "path/to/image"}, {"type": "text", "text": "table information\nUse the provided table and image, predict the label for this claim: {claim}; the label can be Supported or Refuted. Think step by step before answering. Please format your final answer within brackets as follows: <ans> YOUR ANSWER </ans>"}]
```

Figure 3: Input structure used for zero-shot CoT prompting with combined input from both the chart and the table.

5 Results

5.1 SciTabAlign+

Table 2 presents the macro-F1 scores of models on the SciTabAlign+ dataset under three input settings: table-only, chart-only, and the combination of both.

[Table vs. Chart] Comparing table-only and chart-only settings (Table 162 vs. Avg. columns), table-based in-

Model	Table (All)	Table (162)	Basic	Symbol	Line	Swapped	Avg.	Chart + Table
Qwen2.5-VL-3B	52.7	53.6	42.1	38.3	37.2	42.1	39.9	50.4
Qwen2.5-VL-7B	75.7	80.0	61.1	57.6	58.9	55.5	58.3	75.9
Qwen2.5-VL-32B	84.6	86.2	70.6	67.9	65.1	66.8	67.6	86.2
Qwen2.5-VL-72B	88.5	86.3	70.7	63.8	68.8	70.6	68.5	88.0
Llama-3.2-11B-Vision	68.9	68.8	52.3	51.2	56.3	50.3	52.5	59.8
LLaVA-v1.6-Mistral-7B	52.3	57.6	57.6	56.9	55.4	60.8	57.7	58.2
LLaVA-v1.6-Vicuna-13B	48.8	49.8	43.3	36.6	36.6	35.8	38.1	48.4
LLaVA-v1.6-34B	60.2	56.7	32.3	32.3	35.6	33.5	33.4	37.1
InternVL3-1B	31.1	32.6	21.7	28.1	26.1	17.1	23.3	34.1
InternVL3-8B	69.9	70.4	57.3	50.8	59.8	55.0	55.7	70.2
InternVL3-14B	81.5	81.1	62.1	59.1	61.9	66.3	62.4	84.9
InternVL3-38B	80.7	82.4	64.7	62.6	61.1	61.4	62.5	88.8

Table 2: Macro-F1 scores of the models on the **SciTabAlign+** dataset under three input settings: table-only, chart-only, and the combination of chart and table. For the combined setting, we use the basic chart type. **Table (All)** refers to results on the full dataset with 372 claims. **Table (162)** refers to results on 162 claims, matching the subset used for each chart type (basic, symbol, line, or swapped) in the chart-only setting. **Avg.** indicates the average score across the four chart types.

put consistently outperforms chart-based input across 11 models. The only exception is LLaVA-v1.6-Mistral-7B, where performance is nearly identical (57.6 vs. 57.7). The five largest performance gaps, 23.3, 21.7, 19.9, 18.7, and 18.6, are observed in LLaVA-v1.6-34B, Qwen2.5-VL-7B, InternVL3-38B, InternVL3-14B, and Qwen2.5-VL-32B, respectively. The remaining gaps range from 9.3 to 17.8. These substantial gaps indicate that most existing multimodal LLMs struggle with chart inputs while performing better with table inputs.

[Table vs. Combination] In some cases (e.g., Qwen2.5-VL-3B, Qwen2.5-VL-7B, Llama-3.2-11B-Vision, and LLaVA-v1.6-34B), using only table input yields better performance than the combined input. Conversely, some models such as InternVL3-1B, InternVL3-14B, and InternVL3-38B benefit from the combination. LLaVA-v1.6-34B performs poorly at handling chart information; even when provided with combined input, its performance remains low. It is worth noting that we used the same prompt and the same setup for all models in our experiments. We follow the instructions for handling image input as described on each model’s Hugging Face page. For example, Qwen recommends using the *process_vision_info* function from *qwen_vl_utils*, while LLaVA suggests using the *Image* class from *PIL*.

[Chart vs. Combination] The combination of chart and table input consistently outperforms chart-only input across all 12 models. The largest gap is 26.3 on InternVL3-38B, while the smallest is 0.5 on LLaVA-v1.6-Mistral-7B. For the remaining models, most gaps are larger than 10.0, except for Llama-3.2-11B-Vision (7.3) and LLaVA-v1.6-34B (3.7). These large gaps between chart-only input and the combination suggest that, although both types of evidence represent the same information, most models still struggle to effectively process chart data.

[Comparison Across Different Chart Types] Among

the four chart types presented in Section 3.2 (basic bar charts, symbol bar charts, line charts, and swapped charts), we observe that six out of twelve models achieved their best performance on basic bar charts. Line charts and swapped charts each had three models achieving the highest scores. Only one model, InternVL3-1B, performed best on the symbol bar chart; however, its score was relatively low, with an macro-F1 of just 28.1. Notably, one model, Qwen2.5-VL-3B, achieved similar top scores on two chart types, which brings the total count to 13 models instead of 12. The averages across all 12 models for the four chart types are 53.0, 50.4, 51.9, and 51.3 for basic bar charts, symbol bar charts, line charts, and swapped charts, respectively. In summary, symbol bar charts appear to be the most challenging, while the other chart types are generally easier for the models, especially basic bar charts, which achieve the highest scores.

5.2 ChartMimic+

Table 3 presents the Macro-F1 scores of the models on the ChartMimic+ dataset under three input settings: table-only, chart-only, and the combination of chart and table.

[Table vs. Chart] When comparing the table-only and chart-only inputs, we observe that 11 models perform better with table input than with chart input. The four largest performance gaps are 28.1, 24.8, 17.2, and 14.6 for LLaVA-v1.6-34B, LLaVA-v1.6-Vicuna-13B, InternVL3-1B, and Qwen2.5-VL-3B, respectively. We observe a similar pattern in the SciTabAlign+ dataset, where LLaVA-v1.6-34B also performs well with table input but worse with chart input. Qwen2.5-VL-7B is the only model that performs better with chart input than with table input.

[Table vs. Combination] For Qwen2.5-VL-3B, LLaVA-v1.6-Mistral-7B, LLaVA-v1.6-Vicuna-13B, LLaVA-v1.6-34B, InternVL3-1B, and InternVL3-8B, the table-only input outperforms the combination of chart and table inputs. For the remaining models, the combined input performs better

Model	Table	Chart	Chart + Table
Qwen2.5-VL-3B	73.6	59.0	71.6
Qwen2.5-VL-7B	83.3	87.0	87.1
Qwen2.5-VL-32B	93.0	86.7	95.4
Qwen2.5-VL-72B	92.1	90.1	94.7
Llama-3.2-11B-Vision	71.5	67.6	78.9
LLaVA-v1.6-Mistral-7B	59.8	48.1	49.7
LLaVA-v1.6-Vicuna-13B	60.0	35.2	46.1
LLaVA-v1.6-34B	61.4	33.3	42.8
InternVL3-1B	35.2	18.0	25.0
InternVL3-8B	90.4	80.0	82.7
InternVL3-14B	89.8	82.2	92.4
InternVL3-38B	90.1	86.2	94.7

Table 3: Macro-F1 scores of the models on the **ChartMimic+** dataset under three input settings: table-only, chart-only, and the combination of chart and table.

than the table-only input. The four largest gaps, 18.6, 13.9, 10.2, and 10.1, are observed in LLaVA-v1.6-34B, LLaVA-v1.6-Vicuna-13B, InternVL3-1B, and LLaVA-v1.6-Mistral-7B, respectively. These scores suggest that these models lack the ability to effectively integrate table and chart information in the combination setting.

[Chart vs. Combination] Similar to the SciTabAlign+ dataset, we observe that the combination of chart and table inputs outperforms the chart-only input across all 12 models. This observation further emphasizes that current multimodal LLMs still struggle when working with chart-only information. Since the chart and table convey the same underlying information, the models should be able to achieve performance comparable to either the table-only or the combined input setting.

5.3 Compared to Humans

To better understand the difficulty of the task for humans and to test how different evidence formats affect human performance, we randomly selected 50 samples from SciTabAlign+. These samples were split into two sub-tasks, similar to how they are used for models: one using table-only evidence and the other using chart-only evidence. We presented the two sub-tasks to two different annotators (both Master’s students in Computer Science) and asked them to perform the task.

The resulting Macro-F1 scores are 94.0 for the table-only evidence and 96.0 for the chart-only evidence. The Pearson correlation between the annotations of the two annotators is 0.887. These scores indicate that humans perform well with either tables or charts, as the format does not affect their performance when the input information is the same.

6 Correlation between Tables and Charts

To better understand the behavior of models when using table-only versus chart-only evidence, we analyze the correlation scores between these two settings. Table 4 presents

the Pearson correlation scores for the two datasets: SciTabAlign+ and ChartMimic+.

For SciTabAlign+, we observe that the correlation between the table-only and chart-only settings is generally low, with some small models such as InternVL3-1B even showing negative correlation. Overall, correlation scores tend to be higher for larger models and lower for smaller models. In contrast, the correlation scores for the ChartMimic+ dataset are higher than those for the SciTabAlign+ dataset. Large models such as Qwen2.5-VL-32B, Qwen2.5-VL-72B, and InternVL3-38B achieve scores above 0.7, indicating a strong correlation between the two settings.

Considering both datasets, we observe that models from the LLaVA and Llama families show low correlation between table-only and chart-only inputs when solving the task. For the Qwen and InternVL3 families, smaller models with fewer than 8 billion parameters do not show strong correlation between the two evidence formats. However, larger models (with more than 32 billion parameters) exhibit strong correlation on ChartMimic+, but not on SciTabAlign+.

Model	Basic	Symbol	Line	Swap	Mimic+
Qwen2.5-VL-3B	0.132	0.143	0.132	0.132	0.295
Qwen2.5-VL-7B	0.246	0.237	0.203	0.178	0.657
Qwen2.5-VL-32B	0.417	0.301	0.341	0.352	0.710
Qwen2.5-VL-72B	0.337	0.264	0.396	0.446	0.805
Llama-3.2-11B	0.144	0.113	0.121	0.065	0.400
LLaVA-Mistral-7B	0.029	0.175	0.051	0.145	0.146
LLaVA-Vicuna-13B	0.106	0.051	0.080	0.118	0.149
LLaVA-34B	0.059	0.059	0.252	0.125	NaN
InternVL3-1B	-0.055	-0.042	-0.100	0.176	-0.056
InternVL3-8B	0.163	0.143	0.182	-0.026	0.553
InternVL3-14B	0.264	0.194	0.151	0.302	0.593
InternVL3-38B	0.225	0.351	0.229	0.133	0.762

Table 4: Correlation between using table-only and chart-only evidence for the two datasets: SciTabAlign+ and ChartMimic+. For SciTabAlign+, we include four different chart types, which are presented as four columns: Basic Chart, Symbol Chart, Line Chart, and Swapped Chart.

7 Conclusion

In this paper, we extend two existing datasets, SciTabAlign and ChartMimic, to create enhanced versions that support scientific claim verification across different evidence formats conveying the same underlying information. Using these datasets, we conduct a comprehensive evaluation of 12 multimodal LLMs under three input settings: table-only, chart-only, and a combination of both. We also analyze model performance across different chart types. Our experimental results reveal that current multimodal LLMs perform better with table-based inputs but struggle with chart-based evidence. In contrast, human performance remains consistent across both formats, suggesting that the observed discrepancies reflect limitations in the models rather than in the task design. To improve the reliability of automated scientific review systems, future research should focus on enhancing multimodal LLMs’ ability to interpret and reason

over diverse visual formats, particularly charts, as a crucial step toward robust scientific claim verification.

Acknowledgments

We would like to thank the anonymous reviewers for their feedback and suggestions on the improvement of the paper. We also thank Tian Cheng Xia and Truc Hoang for their help with the human annotation tasks. We are grateful to the National Institute of Informatics (NII) for supporting this research. This work was supported by JSPS KAKENHI Grant Number 24K03231.

References

- Aly, R.; Guo, Z.; Schlichtkrull, M. S.; Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Cocarascu, O.; and Mittal, A. 2021. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.
- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv:2502.13923*.
- Chen, W.; Wang, H.; Chen, J.; Zhang, Y.; Wang, H.; Li, S.; Zhou, X.; and Wang, W. Y. 2020. TabFact: A Large-scale Dataset for Table-based Fact Verification. In *International Conference on Learning Representations*.
- Foroutan, N.; Romanou, A.; Ansaripour, M.; Eisenschlos, J. M.; Aberer, K.; and Lebret, R. 2025. WikiMixQA: A Multimodal Benchmark for Question Answering over Tables and Charts. In *Findings of the Association for Computational Linguistics: ACL 2025*, 24941–24958. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.
- Ghosh, A.; Bathini, V. S.; Ganguly, N.; Goyal, P.; and Singh, M. 2024. How Robust Are the QA Models for Hybrid Scientific Tabular Data? A Study Using Customized Dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 8258–8264. Torino, Italia: ELRA and ICCL.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; and et al. 2024. The Llama 3 Herd of Models. *arXiv:2407.21783*.
- Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10: 178–206.
- Gupta, V.; Mehta, M.; Nokhiz, P.; and Srikumar, V. 2020. INFOTABS: Inference on Tables as Semi-structured Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2309–2324. Online: Association for Computational Linguistics.
- Ho, X.; Kumar, S.; Wu, Y.-A.; Boudin, F.; Takasu, A.; and Aizawa, A. 2025. Table-Text Alignment: Explaining Claim Verification Against Tables in Scientific Papers. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2509–2517. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7.
- Kantharaj, S.; Leong, R. T.; Lin, X.; Masry, A.; Thakkar, M.; Hoque, E.; and Joty, S. 2022. Chart-to-Text: A Large-Scale Benchmark for Chart Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4005–4023. Dublin, Ireland: Association for Computational Linguistics.
- Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.
- Lal, Y. K.; Bandham, M.; Hasan, M. S.; Kashi, A.; Koupaee, M.; and Balasubramanian, N. 2025. MuSciClaims: Multi-modal Scientific Claim Verification. *arXiv:2506.04585*.
- Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv:2408.03326*.
- Liu, F.; Wang, X.; Yao, W.; Chen, J.; Song, K.; Cho, S.; Yacoob, Y.; and Yu, D. 2024. MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1287–1310. Mexico City, Mexico: Association for Computational Linguistics.
- Lu, X.; Pan, L.; Liu, Q.; Nakov, P.; and Kan, M.-Y. 2023. SCITAB: A Challenging Benchmark for Compositional Reasoning and Claim Verification on Scientific Tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7787–7813. Singapore: Association for Computational Linguistics.
- Masry, A.; Islam, M. S.; Ahmed, M.; Bajaj, A.; Kabir, F.; Kartha, A.; Laskar, M. T. R.; Rahman, M.; Rahman, S.; Shahmohammadi, M.; Thakkar, M.; Parvez, M. R.; Hoque, E.; and Joty, S. 2025. ChartQAPro: A More Diverse and Challenging Benchmark for Chart Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, 19123–19151. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.
- Masry, A.; Long, D. X.; Tan, J. Q.; Joty, S.; and Hoque, E. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2263–2279. Dublin, Ireland: Association for Computational Linguistics.
- Methani, N.; Ganguly, P.; Khapra, M. M.; and Kumar, P. 2020. PlotQA: Reasoning over Scientific Plots. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*.
- Press, O.; Hochlehnert, A.; Prabhu, A.; Uddandarao, V.; Press, O.; and Bethge, M. 2024. CiteME: Can Language Models Accurately Cite Scientific Claims? In *The Thirty-*

- eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track.*
- Sarrouti, M.; Ben Abacha, A.; Mrabet, Y.; and Demner-Fushman, D. 2021. Evidence-based Fact-Checking of Health-related Claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3499–3512. Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Si, C.; Yang, D.; and Hashimoto, T. 2025. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. In *The Thirteenth International Conference on Learning Representations*.
- Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 809–819. New Orleans, Louisiana: Association for Computational Linguistics.
- Wadden, D.; Lin, S.; Lo, K.; Wang, L. L.; van Zuylen, M.; Cohan, A.; and Hajishirzi, H. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7534–7550. Online: Association for Computational Linguistics.
- Wang, C.; Shen, Y.; Kuang, Z.; Cohan, A.; and Zhao, Y. 2025. SciVer: Evaluating Foundation Models for Multi-modal Scientific Claim Verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8562–8579. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.
- Wang, N. X. R.; Mahajan, D.; Danilevsky, M.; and Rosenthal, S. 2021. SemEval-2021 Task 9: Fact Verification and Evidence Finding for Tabular Data in Scientific Documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 317–326. Online: Association for Computational Linguistics.
- Wang, W. Y. 2017. “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 422–426. Vancouver, Canada: Association for Computational Linguistics.
- Wang, Z.; Xia, M.; He, L.; Chen, H.; Liu, Y.; Zhu, R.; Liang, K.; Wu, X.; Liu, H.; Malladi, S.; Chevalier, A.; Arora, S.; and Chen, D. 2024. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Wu, C.; Liang, Z.; Ge, Y.; Guo, Q.; Lu, Z.; Wang, J.; Shan, Y.; and Luo, P. 2025. Plot2Code: A Comprehensive Benchmark for Evaluating Multi-modal Large Language Models in Code Generation from Scientific Plots. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 3006–3028. Albuquerque, New Mexico: Association for Computational Linguistics.
- Xu, Z.; Du, S.; Qi, Y.; Xu, C.; Yuan, C.; and Guo, J. 2024. ChartBench: A Benchmark for Complex Visual Reasoning in Charts. *arXiv:2312.15915*.
- Yang, C.; Shi, C.; Liu, Y.; Shui, B.; Wang, J.; Jing, M.; XU, L.; Zhu, X.; Li, S.; Zhang, Y.; Liu, G.; Nie, X.; Cai, D.; and Yang, Y. 2025. ChartMimic: Evaluating LMM’s Cross-Modal Reasoning Capability via Chart-to-Code Generation. In *The Thirteenth International Conference on Learning Representations*.
- Yang, Z.; Zhou, Z.; Wang, S.; Cong, X.; Han, X.; Yan, Y.; Liu, Z.; Tan, Z.; Liu, P.; Yu, D.; Liu, Z.; Shi, X.; and Sun, M. 2024. MatPlotAgent: Method and Evaluation for LLM-Based Agentic Scientific Data Visualization. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 11789–11804. Bangkok, Thailand: Association for Computational Linguistics.
- Zhang, X.; Wang, D.; Wang, B.; Dou, L.; Lu, X.; Xu, K.; Wu, D.; and Zhu, Q. 2025a. SCITAT: A Question Answering Benchmark for Scientific Tables and Text Covering Diverse Reasoning Types. In *Findings of the Association for Computational Linguistics: ACL 2025*, 3859–3881. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.
- Zhang, Y.; Wang, Q.; Qian, C.; Liu, J.; Sun, C.; Zhang, D.; Abdelzaher, T.; Zhai, C.; Nakov, P.; and Ji, H. 2025b. Atomic Reasoning for Scientific Table Claim Verification. *arXiv:2506.06972*.
- Zhao, X.; Liu, X.; Haoyue, Y.; Luo, X.; Zeng, F.; Li, J.; Shi, Q.; and Chen, C. 2025. ChartEdit: How Far Are MLLMs From Automating Chart Analysis? Evaluating MLLMs’ Capability via Chart Editing. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 3616–3630. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.
- Zhong, L.; Lu, Y.; Yang, J.; Li, W.; Wei, P.; Wang, Y.; Duan, M.; and Zhang, Q. 2025. DomainCQA: Crafting Expert-Level QA from Domain-Specific Charts. *arXiv:2503.19498*.
- Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; Gao, Z.; Cui, E.; Wang, X.; Cao, Y.; Liu, Y.; Wei, X.; Zhang, H.; Wang, H.; Xu, W.; Li, H.; Wang, J.; Deng, N.; Li, S.; He, Y.; Jiang, T.; Luo, J.; Wang, Y.; He, C.; Shi, B.; Zhang, X.; Shao, W.; He, J.; Xiong, Y.; Qu, W.; Sun, P.; Jiao, P.; Lv, H.; Wu, L.; Zhang, K.; Deng, H.; Ge, J.; Chen, K.; Wang, L.; Dou, M.; Lu, L.; Zhu, X.; Lu, T.; Lin, D.; Qiao, Y.; Dai, J.; and Wang, W. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv:2504.10479*.

A Experimental Details

Running. We run all models on either a single NVIDIA A100 80 GB GPU or two of them. We use max_new_tokens=1024 for all models.

Evaluation. We adopt the precision_recall_fscore_support function from scikit-learn and use it in our evaluation.