Table-Text Alignment: Explaining Claim Verification Against Tables in Scientific Papers

Xanh Ho,[1] **Sunisth Kumar,**[2] **Yun-Ang Wu,**[3*]
Florian Boudin,[4] **Atsuhiro Takasu,**[1] and **Akiko Aizawa**[1,2]
[1]National Institute of Informatics, Japan [2]The University of Tokyo, Japan
[3]National Taiwan University [4]JFLI, CNRS, Nantes Université, France
{xanh, takasu, aizawa}@nii.ac.jp sunisth@g.ecc.u-tokyo.ac.jp
r11944072@csie.ntu.edu.tw florian.boudin@univ-nantes.fr

Abstract

Scientific claim verification against tables typically requires predicting whether a claim is supported or refuted given a table. However, we argue that predicting the final label alone is insufficient: it reveals little about the model's reasoning and offers limited interpretability. To address this, we reframe table–text alignment as an explanation task, requiring models to identify the table cells essential for claim verification. We build a new dataset by extending the SciTab benchmark with human-annotated cell-level rationales. Annotators verify the claim label and highlight the minimal set of cells needed to support their decision. After the annotation process, we utilize the collected information and propose a taxonomy for handling ambiguous cases. Our experiments show that (i) incorporating table alignment information improves claim verification performance, and (ii) most LLMs, while often predicting correct labels, fail to recover human-aligned rationales, suggesting that their predictions do not stem from faithful reasoning.[1]

1 Introduction

Claim verification against tables requires models to determine whether a natural language claim is supported or refuted based on structured tabular data. Several benchmarks have been proposed in the general domain, such as TabFact (Chen et al., 2020), INFOTABS (Gupta et al., 2020), and FEVEROUS (Aly et al., 2021), primarily focusing on Wikipedia tables. However, tables in scientific papers pose additional challenges: they are often denser, more structured, and require domain-specific reasoning.

Two datasets have recently addressed this task in the scientific domain: SEM-TAB-FACTS (SEM; Wang et al., 2021), which includes both claim verification and cell-level evidence selection, and SciTab (Lu et al., 2023), which focuses solely on claim verification. While SEM includes an alignment component, its claims are crowd-generated and simplified, limiting their representativeness. SciTab, in contrast, uses naturally occurring claims but lacks explicit annotations that explain why a given label is correct.

We argue that label prediction alone, as in SciTab, is not enough. It fails to reveal whether a model truly understands the table content, nor does it provide interpretable reasoning. For both evaluation and practical use, models need to go beyond classification and provide explanations grounded in tabular evidence.

From the perspective of scientific reading tools (Lo et al., 2023), table–text alignment is also crucial. It allows readers to quickly locate which parts of a table are referenced in the text, improving comprehension and accelerating the reading process. Such alignments could directly support scientific workflows by making tabular evidence more accessible and interpretable.

To address these limitations, we reframe table–text alignment as an explanation task for scientific claim verification. Specifically, we extend the SciTab dataset with human-annotated cell-level rationales. For each claim–table pair, annotators verify the claim label and highlight the minimal set of table cells needed to support the decision.

During annotation, we frequently encountered ambiguous cases in claim interpretation and evidence selection. To capture these edge cases systematically, we introduce a taxonomy of five ambiguity types in scientific table-based verification: (i) Table Conversion Errors, (ii) Additional Context Requirements, (iii) Unexpected Claim Types, (iv) Subjective Adjectives, and (v) Unclear Claims.

We use our dataset to evaluate various types of

[*]Research conducted during internship at NII, Japan.
[1]Our data and code are available at https://github.com/Alab-NII/SciTabAlign

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2509–2517
November 4-9, 2025

large language models (LLMs), including table-based models, open-source LLMs, and closed-source LLMs. Our experiments also incorporate three different prompting strategies. On average, our human-annotated cell-level rationales help improve the performance of the claim label prediction task. The results show that while models achieve high macro-F1 scores on the claim label prediction task, their performance on the cell selection task remains low—even for advanced models like GPT-4o. The highest score, 50.8, is achieved by Qwen 2.5 72B using CoT prompting. Further analysis of the correlation between the two tasks reveals that although LLMs often correctly predict the claim label, their ability to identify the corresponding explanation cells is still limited.

2 Related Work

Claim verification has been studied across multiple domains, including news (Wang, 2017), Wikipedia (Thorne et al., 2018; Jiang et al., 2020), scientific literature (Wadden et al., 2020; Ou et al., 2025), and medicine (Kotonya and Toni, 2020; Vladika et al., 2024). Beyond plain text, recent efforts have extended claim verification to structured or multimodal evidence, including tables (Chen et al., 2020; Lu et al., 2023), figures (Akhtar et al., 2024), knowledge graphs (Kim et al., 2023) and multimodal data (Yang et al., 2025b).

Among table-based datasets, SEM (Wang et al., 2021) and TabEvidence (Gupta et al., 2022) are most related to our work. However, SEM features simplified, crowd-generated claims, while TabEvidence is limited to two-column Wikipedia tables, lacking the complexity of scientific tables. Recent frameworks like Chain-of-Table (Wang et al., 2024) and Dater (Ye et al., 2023) include evidence selection steps, but only report label accuracy, without evaluating the relevance or quality of the selected evidence, limiting trust in their predictions.

In contrast, our work emphasizes explanation via alignment, explicitly evaluating whether the model selects the correct table cells needed for verification, providing a more faithful and interpretable assessment of reasoning.

3 Dataset Creation

In this section, we first briefly introduce the existing SciTab dataset, on which our work is based. We then describe the process of obtaining the extended version, SciTabAlign, with cell-level explanations. Finally, we propose a taxonomy of five common ambiguity types, which we hope future work considers to build more reliable claim verification datasets. We note that we remove all ambiguous cases from our dataset, reducing the number of claims from 868 to 372.

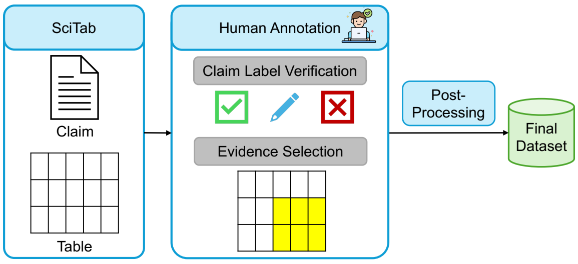

Figure 1: Overall dataset creation process.

3.1 Base Dataset: SciTab

We build on SciTab (Lu et al., 2023), the only available dataset for claim verification against scientific tables with naturally occurring claims. SciTab is derived from SciGen (Moosavi et al., 2021), a table-to-text generation dataset in which each sample consists of a scientific table and its corresponding textual description.

The dataset contains 1,224 claim–table pairs: 457 supported, 411 refuted, and 356 not enough information (NEI). Supported claims are sourced from original paper content, while refuted and NEI claims are generated by InstructGPT (Ouyang et al., 2022) and then manually verified. The original benchmark defines two settings: binary classification (supported vs. refuted) and three-class classification (supported, refuted, NEI), but focuses only on label prediction, without providing explanations.

3.2 Our Dataset: SciTabAlign

We extend SciTab by adding cell-level explanations, i.e. table regions required to support or refute each claim. We focus on the supported and refuted claims (868 total), leaving out NEI cases, which are typically under-specified. As shown in Figure 1, our annotation pipeline includes two tasks: claim label verification and evidence selection.

Human Annotation. Each annotator is given a claim, its label, the associated table, and the caption. Annotators first verify the correctness of the claim. If it is clearly supported or refuted, they mark it as *Good*; if the claim is unclear, malformed, or unsupported by the table, they choose *Do Noth-*

ing or optionally revise it (*Revised*). For *Good* or *Revised* claims, annotators highlight the minimal set of table cells required to determine the label. Annotation was performed by four NLP researchers (authors of this paper).

Post-Processing. After human annotation, we obtain 444 *Good* samples, 81 *Revised*, and 343 *Do Nothing*. We retain only the samples labeled as *Good*. Among these, we discard 66 samples in which all table cells are marked (non-informative) and 6 samples containing *NaN* values, resulting in a final dataset of 372 aligned samples (195 supported, 177 refuted).

Inter-Annotator Agreement. To assess annotation quality, we conducted a second round of labeling on 50 randomly selected tables (covering 137 claims) by a different annotator. Using the first annotation as ground truth, we obtained 75.2% precision, 89.1% recall, and 78.0% macro-F1 for cell-level overlap.

3.3 A Proposed Taxonomy

During annotation, we observed frequent edge cases where claim verification was hindered by poor table formatting, unclear language, or missing context. We propose a taxonomy of five ambiguity types based on annotator notes and discussion:

(i) Table Conversion Errors: artifacts introduced during table extraction (e.g. merged cells, missing entries, formatting loss).

(ii) Additional Context Requirements: Claims referencing abbreviations, statistical tests, or assumptions not recoverable from the table alone.

(iii) Unexpected Claim Types: Descriptive or meta-level claims (e.g., "*Table 4 lists the scores of different models.*") that require no reasoning.

(iv) Subjective Adjectives: Use of vague or non-quantifiable terms (e.g., "*poor*", "*substantial*", or "*a large margin*").

(v) Unclear Claims: Ambiguous references to table elements (e.g., "*this model*", "*these scores*").

We provide examples for each case in Appendix A. We hope this taxonomy will guide future dataset development and improve robustness in scientific claim verification tasks.

4 Experimental Settings

Models. We use three groups of models in our experiments: Table-based LLMs, Open-source LLMs, and Closed-source LLMs. For table-based LLMs, we use TAPAS-base and TAPAS-large (Herzig et al., 2020), pretrained for reasoning over tabular input. For open-source LLMs, we use the Instruction-tuned variants of Qwen 2.5 (7B and 72B, Yang et al., 2025a) and Llama 3.1 (8B and 70B, Grattafiori et al., 2024). For closed-source LLMs, we use GPT-4o (Hurst et al., 2024).

Prompting Strategies. Our dataset contains two subtasks: (1) claim label prediction and (2) cell-level evidence selection. We conduct experiments using three prompting strategies: zero-shot, few-shot, and Chain-of-Thought (CoT; Wei et al., 2022). For few-shot and CoT promptings, we use four demonstration examples selected from the revised subset of samples not included in the evaluation set, ensuring fair evaluation.

Tabular Representation. Following Wang et al. (2024), who found that the PIPE encoding format with explicit tags (e.g. Col and Row 1) outperformed HTML, TSV, and Markdown formats for tabular data, we adopt PIPE encoding for all of our experiments.

Evaluation Metrics. We use Macro-F1 to evaluate both tasks in our dataset: (1) claim label prediction and (2) cell-level evidence selection. For the claim label prediction task, we compare the predicted label with the ground-truth label. Our dataset contains two labels: *Supported* and *Refuted*. For the cell-level evidence selection task, both ground-truth and predicted evidence are represented as lists of (row, column) index tuples using PIPE encoding. For example, (1, 2) refers to the cell in row 1, column 2. We compare the two lists to obtain True Positives, False Positives, and False Negatives, and then calculate precision, recall, and F1 based on these values. The Macro-F1 score is computed as the average of the individual F1 scores.

5 Results

All results are shown in Table 1. It is noted that, due to cost constraints, we only run GPT-4o on 100 samples selected to match the label distribution of the entire dataset.

Claim Prediction Results. As expected, GPT-4o achieves the highest score. Larger models, such as Qwen 2.5 72B and Llama 3.1 70B, outperform their smaller 7B and 8B counterparts, and

Model	Claim Labeling			Cell Selection		
	Zero	Few	CoT	Zero	Few	CoT
TAPAS-base	48.1	-	-	-	-	-
TAPAS-large	51.6	-	-	-	-	-
Llama 3.1 8B	53.2	59.5	62.4	23.6	22.3	22.6
Llama 3.1 70B	75.2	75.0	73.9	31.8	28.8	36.8
Qwen 2.5 7B	66.3	68.1	67.9	20.7	16.6	17.0
Qwen 2.5 72B	83.5	84.7	81.5	32.8	46.7	**50.8**
GPT-4o	**88.4**	87.0	88.0	32.4	32.9	34.8

Table 1: Macro-F1 scores of the models on our dataset. 'Zero', 'Few', and 'CoT' denote zero-shot, few-shot, and CoT prompting, respectively.

all LLMs surpass the performance of the previous table-based model, TAPAS. We also observe that few-shot and CoT prompting are less effective for larger, well-instructed models like the 70B variants and GPT-4o on this familiar label classification task, but remain beneficial for smaller models.

Evidence Selection Results. Compared to claim label prediction, evidence cell selection is a more challenging task that most LLMs are unfamiliar with. The input consists of a claim and a table, and the output is a list of cell positions—each defined by a row and column index—required to determine the claim's label. This structured output format adds complexity, and overall, all models struggle to achieve high scores on this task. In the zero-shot setting, GPT-4o, Llama 3.1 70B, and Qwen 2.5 72B achieve comparable scores. Under few-shot and CoT prompting, GPT-4o's performance remains relatively stable, while Qwen 2.5 72B sees an 18.0 F1 improvement from zero-shot to CoT. CoT prompting also boosts Llama 3.1 70B's performance. In contrast, smaller models (7B–8B) show decreased performance under both few-shot and CoT prompting compared to zero-shot.

Overall, compared to the human agreement score (78.0 macro F1), the best model still falls short, indicating room for improvement on this task. Despite its difficulty and the possibility of multiple valid reasoning paths, our proposed evidence cells can be seen as a minimal, useful set for claim verification. In the era of black-box LLMs, focusing solely on the final label is insufficient—evidence selection is equally important for explainability. Our work takes a first step toward more interpretable evaluation and highlights the underlying reasoning abilities of models.

Model	Table	Exp.	Table + Exp.
Llama 3.1 8B	53.2	56.9	63.0
Llama 3.1 70B	75.2	80.1	80.9
Qwen 2.5 7B	66.3	67.5	69.8
Qwen 2.5 72B	83.5	80.6	81.9

Table 2: Macro-F1 scores of the models on our dataset using different types of input table contexts. "Exp." refers to our explanation table cells. For all experiments in this table, we use zero-shot prompting.

Divergent Effects of Few-Shot and CoT Prompting on Claim Labeling vs. Cell Selection. The claim labeling task is a binary classification problem (supported vs. refuted), which closely aligns with tasks that most LLMs are already exposed to during pretraining. In contrast, cell selection is a novel task with a different structure, likely unfamiliar to most models. We observe that for claim labeling, few-shot and CoT prompting benefit smaller models (7B–8B), while larger models (70B–72B) show little to no improvement, likely due to their stronger inherent reasoning capabilities. For cell selection, however, smaller models struggle with few-shot and CoT prompting, possibly because the demonstrations are not easily generalizable for this unfamiliar task. Larger models perform better in this setting, suggesting greater adaptability to task structure even when it deviates from pretraining distributions.

Effectiveness of Our Explanation Cells. To evaluate the effectiveness of our explanation cells, we assess models under two different settings: (1) using only our explanation table cells, and (2) using both the original table and our explanation table cells. The results are shown in Table 2. On average, we observe that using only our explanation cells or combining them with the original table leads to improved task performance.

6 Analyses

To better understand the correlation between the claim label prediction task and the cell evidence selection task, we categorize outcomes into four types: Correct–Correct, Correct–Incorrect, Incorrect–Correct, and Incorrect–Incorrect. For claim label prediction, correctness is easily determined based on whether the predicted label (Supported or Refuted) matches the ground truth. In contrast, cell evidence selection involves list-based predictions, making exact matches more challenging. Therefore, we consider two evaluation criteria: exact match (EM) and a relaxed case where an F1 score of 50.0 or higher is considered correct.

Claim	Cell	L 8B	L 70B	Q 7B	Q 72B	GPT
		Setting 1: Exact Match for Both Tasks				
C	C	0.0	0.0	0.0	4.6	0.0
C	I	63.4	73.9	68.0	73.4	88.0
I	C	0.0	0.0	0.0	0.0	0.0
I	I	36.6	26.1	32.0	22.0	12.0
		Setting 2: F1 >= 50 in Cell Selection				
C	C	10.5	26.1	4.3	44.1	30.0
C	I	53.0	47.8	63.7	33.9	58.0
I	C	5.6	10.5	2.7	8.9	7.0
I	I	30.9	15.6	29.3	13.2	5.0

Table 3: Categorical statistics (%) showing the correlation between the claim label prediction and cell evidence selection tasks. *C* and *I* denote Correct and Incorrect, respectively. *L* and *Q* denote Llama and Qwen, respectively. The results are from CoT prompting.

The percentage distribution of these cases is shown in Table 3. The case where both tasks are correct (C–C) is what we expect. However, as shown in the table, none of the models achieve a percentage of 50% for this case—even in the second setting, where an F1 score of 50.0 or higher is considered correct for the cell selection task. This suggests that while models often predict the claim label correctly, they lack the ability to select the minimal subset of table cells necessary to support that prediction.

7 Conclusion

In this work, we highlighted the limitations of scientific claim verification systems that focus solely on label prediction, arguing for the importance of interpretability through evidence selection. By reframing table-text alignment as an explanation task and introducing a new dataset with human-annotated cell-level rationales, we provide a more rigorous benchmark for evaluating model reasoning. Additionally, we proposed a taxonomy of ambiguous cases in claim verification against tables, which can support future work on dataset construction. Our findings demonstrate that while LLMs often predict the correct claim labels, they frequently fail to identify the minimal supporting evidence, revealing a gap between accuracy and faithful reasoning. This underscores the need for future work to prioritize not just correctness, but also alignment with human-understandable rationales in scientific fact verification tasks.

Limitations

Our work has several limitations. First, the annotation scale is modest, with 868 claims as input and only 372 claims in the final dataset, which may affect the statistical reliability and generalizability of the findings. Second, the dataset originates from a specific domain (computer science), which may limit its applicability to tables and claims from other domains. Third, the PIPE encoding method used may not be well-suited for handling complex table structures, suggesting the need for more robust encoding approaches.

Acknowledgments

We would like to thank the anonymous reviewers for their feedback and suggestions on the paper. This work was supported by JSPS KAKENHI Grant Number 24K03231.

Ethical Statement and Broader Impact

We built our dataset based on the publicly available SciTab dataset, which is released under the MIT License. We respect the terms of this license and provide appropriate attribution to the original authors. To extend the dataset, four NLP researchers manually annotated the data. We created and followed a detailed annotation guideline to ensure consistency, clarity, and fairness in the annotation process. The dataset does not include any personal or sensitive information.

Potential Risks

As this dataset consists of 372 samples for claim verification and is intended primarily for evaluation purposes, the risks are limited. The data does not include personal or sensitive information. However, potential risks include biases in the sample selection, which may affect the representativeness of the dataset and the generalizability of evaluation results. Additionally, the dataset could be misapplied outside its intended scope, leading to misleading conclusions if used as a training resource rather than for evaluation.

References

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl.

2024. ChartCheck: Explainable fact-checking over real-world chart images. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13921–13937, Bangkok, Thailand. Association for Computational Linguistics.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 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)*.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and et al. 2024. The llama 3 herd of models. *arXiv:2407.21783*.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.

Vivek Gupta, Shuo Zhang, Alakananda Vempala, Yujie He, Temma Choji, and Vivek Srikumar. 2022. Right for the right reason: Evidence extraction for trustworthy tabular reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3268–3283, Dublin, Ireland. Association for Computational Linguistics.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P. Goucher, and et al (OpenAI). 2024. Gpt-4o system card. *arXiv:2410.21276*.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. FactKG: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16190–16206, Toronto, Canada. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, Fangzhou Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Kinney, Aniket Kittur, Hyeonsu Kang, Egor Klevak, Bailey Kuehl, Michael Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. 2023. The semantic reader project: Augmenting scholarly documents through ai-powered interactive reading interfaces. *arXiv:2303.14334*.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 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*, pages 7787–7813, Singapore. Association for Computational Linguistics.

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

Jiefu Ou, William Gantt Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng, William Jurayj, Miriam Wanner, Shaobo Liang, Candice Morgan, Seunghoon Han, Weiqi Wang, Chandler May, Hannah Recknor, Daniel Khashabi, and Benjamin Van Durme. 2025. Claimcheck: How grounded are llm critiques of scientific papers? *arXiv:2503.21717*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 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)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 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)*, pages 317–326, Online. Association for Computational Linguistics.

William Yang Wang. 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)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024. Chain-of-table: Evolving tables in the reasoning chain for table understanding. In *The Twelfth International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

An Yang, Baosong Yang, and et al. 2025a. Qwen2.5 technical report. *arXiv:2412.15115*.

Bohao Yang, Yingji Zhang, Dong Liu, André Freitas, and Chenghua Lin. 2025b. Does table source matter? benchmarking and improving multimodal scientific table understanding and reasoning. *arXiv:2501.13042*.

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

A Dataset Creation

A.1 A Proposed Taxonomy

We present examples for our proposed taxonomy in Section 3.3 in Tables 4, 5, 6, 7, and 8, respectively.

Claim Comparing POS and SEM tagging (Table 5), we note that higher layer representations do not necessarily improve SEM tagging, while POS tagging does not peak at layer 1. We noticed no improvements in both translation (+0.9 BLEU) and POS and SEM tagging (up to +0.6% accuracy) when using features extracted from an NMT model trained with residual connections (Table 5).

Label Refuted

Table Caption Table 5: POS and SEM tagging accuracy with features from different layers of 4-layer Uni/Bidirectional/Residual NMT encoders, averaged over all non-English target languages.

Table

```
Uni | POS | 0 87.9 | 1 92.0 | 2 91.7 | 3 91.8 | 4 91.9
Uni | SEM | 81.8 | 87.8 | 87.4 | 87.6 | 88.2
Bi | POS | 87.9 | 93.3 | 92.9 | 93.2 | 92.8
Bi | SEM | 81.9 | 91.3 | 90.8 | 91.9 | 91.9
Res | POS | 87.9 | 92.5 | 91.9 | 92.0 | 92.4
Res | SEM | 81.9 | 88.2 | 87.5 | 87.6 | 88.5
```

Table 4: Example of (i) Table Conversion Errors. The column headers are merged with the data values. For example, "`0 87.9`" incorrectly combines the column name `0` and the value `87.9`.

Claim After removing the graph attention module, our model gives 24.9 BLEU points.

Label Supported

Table Caption Table 9: Ablation study for modules used in the graph encoder and the LSTM decoder

Table

```
[BOLD] Model | B | C
DCGCN4 | 25.5 | 55.4
Encoder Modules | [EMPTY] | [EMPTY]
-Linear Combination | 23.7 | 53.2
-Global Node | 24.2 | 54.6
-Direction Aggregation | 24.6 | 54.6
-Graph Attention | 24.9 | 54.7
-Global Node &Linear Combination | 22.9 | 52.4
Decoder Modules | [EMPTY] | [EMPTY]
-Coverage Mechanism | 23.8 | 53.0
```

Table 5: Example of (ii) Additional Context Requirements. `B` and `C` stand for BLEU and CHRF++, respectively, but this cannot be inferred from the claim, caption, or table alone. It requires additional context from the original paper.

Claim Table 4 lists the EM/F1 score of different models.

Label Supported

Table Caption Table 4: Exact match/F1-score on SQuad dataset. "#Params": the parameter number of Base. rnet*: results published by Wang et al. (2017).

Table

```
Model | #Params | Base | +Elmo
rnet* | - | 71.1/79.5 | -/-
LSTM | 2.67M | [BOLD] 70.46/78.98 | 75.17/82.79
GRU | 2.31M | 70.41/ [BOLD] 79.15 | 75.81/83.12
ATR | 1.59M | 69.73/78.70 | 75.06/82.76
SRU | 2.44M | 69.27/78.41 | 74.56/82.50
LRN | 2.14M | 70.11/78.83 | [BOLD] 76.14/ [BOLD] 83.83
```

Table 6: Example of (iii) Unexpected Claim Types. The claim simply describes what the table shows, similar to the caption, and does not require any reasoning or data to support it.

Claim [CONTINUE] RELIS significantly outperforms the other RL-based systems.

Label Supported

Caption Table 3: Results of non-RL (top), cross-input (DeepTD) and input-specific (REAPER) RL approaches (middle) compared with RELIS.

Table

[EMPTY]	DUC'01 R1	DUC'01 R2	DUC'02 R1	DUC'02 R2	DUC'04 R1	DUC'04 R2
ICSI	33.31	7.33	35.04	8.51	37.31	9.36
PriorSum	35.98	7.89	36.63	8.97	38.91	10.07
TCSum	<bold>36.45</bold>	7.66	36.90	8.61	38.27	9.66
TCSum-	33.45	6.07	34.02	7.39	35.66	8.66
SRSum	36.04	8.44	<bold>38.93</bold>	<bold>10.29</bold>	39.29	10.70
DeepTD	28.74	5.95	31.63	7.09	33.57	7.96
REAPER	32.43	6.84	35.03	8.11	37.22	8.64
RELIS	34.73	<bold>8.66</bold>	37.11	9.12	<bold>39.34</bold>	<bold>10.73</bold>

Table 7: Example of (iv) Subjective Adjectives. Whether the performance is considered "significant" depends on how the term is defined. Moreover, many argue that using the word "significant" requires the result to pass some form of statistical test. The original version of the first row is: `[EMPTY] | DUC'01 <italic>R</italic>1 | DUC'01 <italic>R</italic>2 | DUC'02 <italic>R</italic>1 | DUC'02 <italic>R</italic>2 | DUC'04 <italic>R</italic>1 | DUC'04 <italic>R</italic>2`.

Claim It closely matches the performance of ORACLE with only 0.40% absolute difference.

Label Supported

Caption Table 3: Accuracy of transferring between aspects. Models with † use labeled data from source aspects. Models with ‡ use human rationales on the target aspect.

Table

Source	Target	Svm	Ra-Svm‡	Ra-Cnn‡	Trans†	Ra-Trans‡†	Ours‡†	Oracle†
Beer aroma+palate	Beer look	74.41	74.83	74.94	72.75	76.41	[BOLD] 79.53	80.29
Beer look+palate	Beer aroma	68.57	69.23	67.55	69.92	76.45	[BOLD] 77.94	78.11
Beer look+aroma	Beer palate	63.88	67.82	65.72	74.66	73.4	[BOLD] 75.24	75.5

Table 8: Example of (v) Unclear Claims. It is unclear what entity the pronoun "it" refers to.