表格-文本对齐：解释针对科学论文中的表格的声明验证

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

针对表格的科学主张验证通常需要预测在给定表格的情况下主张是被支持还是被驳斥。然而，我们认为仅仅预测最终标签是不够的：它几乎无法揭示模型的推理过程，并且提供的解释能力有限。为了解决这一问题，我们将表格–文本对齐重新构建为一个解释任务，要求模型识别出验证主张所必需的表格单元。我们通过扩展 SciTab 基准数据集并加入人为标注的单元级理由来构建一个新的数据集。标注者验证主张标签并突出支持其决策所需的最小单元集合。在标注过程结束后，我们利用收集到的信息并提出处理模糊情况的分类法。我们的实验表明，(一) 结合表格对齐信息可以提高主张验证性能，而 (二) 大多数 LLM 在通常预测正确标签时，未能恢复人与模型对齐的理由，这表明它们的预测并非源于忠实的推理。[1]

1 介绍

针对表格的声明验证要求模型确定自然语言声明是否受到结构化表格数据的支持或驳斥。在一般领域中已提出了几个基准，如 TabFact (Chen et al., 2020) 、INFOTABS (Gupta et al., 2020) 和 FEVEROUS (Aly et al., 2021)，它们主要集中在维基百科的表格上。然而，科学论文中的表格提出了额外的挑战：它们通常更密集、更有结构，需要特定领域的推理。最近有两个数据集解决了科学领域中的这一任务：SEM-TAB-FACTS (SEM; **?**)，它包括声明验证和单元格级别的证据选择，以及专注于声明验证的 SciTab (Lu et al., 2023)。虽然 SEM 包含一个对齐组件，但其声明是由人群生成并简化的，限制了其代表性。相比之下，SciTab 使用自然发生的声明，但缺乏解释为何给定标签正确的明确注释。我们认为，仅仅像 SciTab 那样进行标签预测是不够的。它未能揭示模型是否真正理解表格内容，也未能提供可解释的推理。无论是为了评估还是实际应用，模型都需要超越分类，提供以表格证据为基础的解释。

从科学阅读工具的角度来看，(Lo et al., 2023)，表格与文本的对齐也至关重要。它使读者能够快速定位文本中引用了表格的哪些部分，提高理解力并加快阅读过程。这样的对齐可以通过使表格证据更易获取和解释，直接支持科学工作流程。为了解决这些限制，我们将表格与文本的对齐重新构想为一个科学论点验证的解释任务。具体而言，我们通过添加人工标注的单元级理由来扩展 SciTab 数据集。对于每一个论点–表格对，标注者验证论点标签并高亮支持决策所需的最少表格单元。

在注释过程中，我们经常遇到声明解释和证据选择中的模糊情况。为了系统地捕捉这些边缘情况，我们引入了一个关于科学表格验证中五种模糊类型的分类法：(i) 表格转换错误，(ii) 额外的上下文需求，(iii) 意外的声明类型，(iv) 主观形容词，和 (v) 不明确的声明。

我们使用我们的数据集来评估各种类型的大型语言模型 (LLMs)，包括基于表格的模型、开源 LLMs 和闭源 LLMs。我们的实验还包含了三种不同的提示策略。平均而言，我们人工标注的单元级推理有助于改善声明标签预测任务的性能。结果显示，虽然模型在声明标签预测任务上取得了较高的宏观 F1 分数，但它们在单元选择任务中的表现仍然较低，即使对于像 GPT-4o 这样的高级模型。最高得分 50.8 由 Qwen 2.5 72B 使用 CoT 提示策略实现。对这两个任务之间相关性的进一步分析揭示了虽然 LLMs 往往能正确预测声明标签，但它们识别对应解释单元的能力仍然有限。

2 相关工作

事实核查已经在多个领域进行研究，包括新闻 (Wang, 2017)，维基百科 (Thorne et al., 2018; Jiang et al., 2020)，科学文献 (Wadden et al., 2020; Ou et al., 2025)，以及医学 (Kotonya and Toni, 2020; Vladika et al., 2024)。除了纯文

[*]Research conducted during internship at NII, Japan.
[1]我们的数据和代码可在 https://github.com/Alab-NII/SciTabAlign 获取

本，近期的研究将事实核查扩展到结构化或多模态证据上，包括表格（Chen et al., 2020; Lu et al., 2023），图形（Akhtar et al., 2024），知识图谱（Kim et al., 2023）和多模态数据（Yang et al., 2025b）。在基于表格的数据集中，SEM（Wang et al., 2021）和 TabEvidence（Gupta et al., 2022）与我们的工作最为相关。然而，SEM 的特点是简化的、由大众生成的声明，而 TabEvidence 仅限于维基百科的双列表格，缺乏科学表格的复杂性。最新框架如 Chain-of-Table（Wang et al., 2024）和 Dater（Ye et al., 2023）包含了证据选择步骤，但仅报告标签准确性，没有评估所选证据的相关性或质量，从而限制了对其预测的信任。

相比之下，我们的工作强调通过对齐进行解释，明确评估模型是否选择了进行验证所需的正确表格单元，从而提供更真实且可解释的推理评估。

3 数据集创建

3.1 基础数据集：SciTab

我们建立在 SciTab（Lu et al., 2023）的基础上，这是唯一可用的用于对自然发生的科学表格声明进行验证的数据集。SciTab 来源于 SciGen（Moosavi et al., 2021），这是一个表格到文本生成数据集，其中每个样本都由一个科学表格及其相应的文本描述组成。

该数据集包含 1,224 个索引表对：457 个支持，411 个驳斥，和 356 个信息不足（NEI）。支持的索引来自原始论文内容，而驳斥和 NEI 索引则由 InstructGPT（Ouyang et al., 2022）生成，然后手动验证。原始基准定义了两种设置：二分类（支持与驳斥）和三分类（支持、驳斥、NEI），但仅关注标签预测，不提供解释。

3.2 我们的数据集：表格对齐的解释

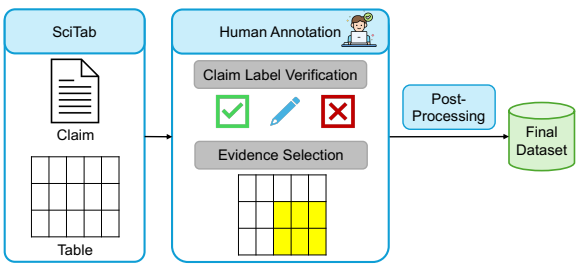

Figure 1: 整体数据集创建过程。

我们通过添加单元级别的解释来扩展 SciTab，即支持或反驳每个声明所需的表格区域。我们专注于已支持和已反驳的声明（共 868 个），忽略了通常没有明确指定的 NEI 案例。如图 1 所示，我们的标注流程包括两个任务：声明标签验证和证据选择。

每位标注者都会得到一个论断、其标签、相关的表格和标题。标注者首先验证论断的正确性。如果明显支持或反驳它，他们会将其标记为“好”；如果论断不明确、格式错误或未被表格支持，他们可以选择“不做任何事”或可选择地修订它（“修订”）。对于“好”或“修订”的论断，标注者会突出显示确定标签所需的表格单元格的最小集合。标注由四位自然语言处理研究人员（本文作者）进行。

我们舍弃了所有表格单元格被标记（非信息性）或包含 NaN 值的样本，最终得到 372 个对齐样本的数据集（195 个支持，177 个反驳）。

为了评估注释质量，我们安排另一位注释者对 50 个随机选取的表格（涵盖 137 个声明）进行了第二轮标注。使用第一次注释作为基准，我们获得了 75.2 % 的精确率，89.1 % 的召回率，以及 78.0 % 的宏观 F1 得分用于衡量单元格级别的重叠。

3.3 一个拟议的分类法

在注释过程中，我们观察到由于表格格式不佳、语言不清楚或缺少上下文而导致声明验证经常遇到特殊情况。我们基于注释者的记录和讨论提出了五种模糊类型的分类法：

(i) 表格转换错误：表格提取过程中引入的瑕疵（例如，合并单元格、丢失条目、格式丢失）。

(ii) 附加上下文要求：引用缩写、统计检验或不能仅从表格中恢复的假设的声明。

(iii) 意外的声明类型：描述性或元层次声明（例如，“表 4 列出了不同模型的分数。”）不需要推理。

(iv) 主观形容词：使用模糊或无法量化的术语（例如，“差劲的”、“大量的”或“大幅度的”）。

(v) 不明确的声明：对表格元素的模糊引用（例如，“这个模型”、“这些分数”）。

我们在附录 A 提供了每种情况的示例。我们希望这种分类法能够指导未来数据集的开发，并提高科学论点验证任务的稳健性。

4 实验设置

模型。 我们在实验中使用了三组模型：基于表格的 LLM、开源 LLM 和闭源 LLM。对于基于表格的 LLM，我们使用了 TAPAS-base 和 TAPAS-large（Herzig et al., 2020），预训练用于处理表格输入中的推理。对于开源 LLM，我们使用了 Qwen 2.5（7B 和 72B, ?）和 Llama 3.1（8B 和 70B, ?）的指令调优版本。对于闭源 LLM，我们使用了 GPT-4o（Hurst et al., 2024）。

提示策略。 我们的数据集包含两个子任务：（1）声明标签预测和（2）单元级证据选择。我们使用三种提示策略进行实验：零样本、少量样本和思维链 (CoT; **?**) 。对于少量样本和思维链的提示，我们使用四个示例，这些示例从未包含在评估集中的修订样本子集内选择，以确保公平评估。

表格表示 基于 Wang et al. (2024) 的研究结果，即使用显式标签（例如 Col 和 Row 1）的 PIPE 编码格式在处理表格数据时优于 HTML、TSV 和 Markdown 格式，我们在所有实验中采用了 PIPE 编码。

5 结果

遵循原始 SciTab 数据集，我们使用 Macro-F1 进行评估。所有结果如表 1 所示。值得注意的是，由于成本限制，我们仅在选取的与整个数据集标签分布相匹配的 100 个样本上运行 GPT-4o。

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: 模型在我们数据集上的 Macro-F1 分数。'Zero'、'Few' 和 'CoT' 分别表示零样本、少样本和 CoT 提示。

正如预期的那样，GPT-4o 达到了最高分。更大的模型，例如 Qwen 2.5 72B 和 Llama 3.1 70B，其表现超过了较小的 7B 和 8B 模型，并且所有大型语言模型都超越了先前基于表格的模型 TAPAS 的性能。我们还观察到，对于像 70B 变体和 GPT-4o 这样的较大且受过良好指示的模型，在这个熟悉的标签分类任务中，少样本和链式思维提示效率较低，但对于较小的模型仍然有益。

证据选择结果。 与声明标签预测相比，证据单元选择是一项更具挑战性的任务，大多数 LLMs 对此并不熟悉。输入由一个声明和一个表格组成，输出则是一个单元位置的列表——每个由行和列索引定义——用于确定声明的标签。这种结构化的输出格式增加了复杂性，总体而言，所有模型在此任务中都难以获得高分。在零样本设置中，GPT-4o、Llama 3.1 70B 和 Qwen 2.5 72B 取得了可比较的分数。在少样本和 CoT 提示下，GPT-4o 的性能相对稳定，而 Qwen 2.5 72B 从零样本到 CoT 的 F1 值提高了 18.0。CoT 提示也提升了 Llama 3.1 70B 的性能。相比之下，较小的模型（7B–8B）在少样本和 CoT 提示下的性能比零样本设置更差。

总体而言，与人类一致性评分（78.0 宏 F1）相比，最佳模型仍有不足之处，表明该任务仍有改进空间。尽管存在困难和多种合理推理路径的可能性，我们提出的证据单元可以视为声明验证的一个最小、有效的集合。在黑箱大模型的时代，仅关注最终标签是不够的——证据选择对可解释性同样重要。我们的工作首次迈出了更具可解释性评估的步伐，并突出显示了模型的潜在推理能力。

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: 在我们的数据集上，使用不同类型的输入表格上下文进行的模型宏平均 F1 分数。"Exp." 指的是我们的解释表单元格。对于本表中的所有实验，我们使用零样本提示。

我们解释单元的有效性 为了评估我们解释单元格的有效性，我们在两种不同的设置下对模型进行评估：(1) 仅使用我们的解释表单元格，(2) 同时使用原始表和我们的解释表单元格。结果显示在表 2 中。平均而言，我们观察到仅使用我们的解释单元格或将其与原始表结合使用都能提高任务性能。

6 分析

为了更好地理解声明标签预测任务与单元证据选择任务之间的相关性，我们将结果分为四种类型：正确–正确、正确–不正确、不正确–正确和不正确–不正确。对于声明标签预测，正确与否可以很容易地根据预测标签（支持或反驳）是否与真实值相符来确定。相比之下，单元证据选择涉及基于列表的预测，这使得精确匹配更具挑战性。因此，我们考虑了两种评估标准：精确匹配（EM）和一种放宽的情况，即 F1 分数达到或高于 50.0 被视为正确。

这些案例的百分比分布显示在表 3 中。我们期待的情况是两个任务都正确 (C–C)。然而，如表中所示，没有一个模型在这种情况下达到 50 % 的百分比——即使在第二种设置中，其中 F1 得分达到或超过 50.0 被视为单元选择任

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: 分类统计（%）显示了声明标签预测任务与单元证据选择任务之间的相关性。C 和 I 分别表示正确和错误，L 和 Q 分别表示 Llama 和 Qwen。这些结果来自 CoT 提示。

务的正确。这表明虽然模型通常能够正确预测声明标签，但它们缺乏选择支持该预测所需的最小表格单元子集的能力。

在这项工作中，我们强调了仅专注于标签预测的科学声明验证系统的局限性，主张通过证据选择来实现可解释性的重要性。通过将表格-文本对齐重新构想为解释任务并引入具有人工注释的细胞级理由的新数据集，我们为评估模型推理提供了更严格的基准。此外，我们提出了一个针对表格中声明验证的模糊案例分类法，这可以支持未来的数据集构建工作。我们的研究结果表明，尽管大型语言模型（LLMs）经常预测正确的声明标签，但它们常常无法识别出最小的支持证据，暴露了准确性与真实推理之间的差距。这强调了未来工作必须不仅关注正确性，还要注重与人类可理解的理由相一致的科学事实验证任务的重要性。

7

限制 我们的工作有几项限制。首先，注释规模较小，仅有 868 个断言作为输入，而最终数据集中只有 372 个断言，这可能影响结果的统计可靠性和广泛适用性。其次，数据集来自特定领域（计算机科学），这可能限制其对其他领域的表格和断言的适用性。第三，所使用的 PIPE 编码方法可能不适合处理复杂的表格结构，这表明需要更为稳健的编码方法。

8

伦理声明与更广泛的影响 我们根据公开可用的 SciTab 数据集构建了我们的数据集，该数据集是根据 MIT 许可证发布的。我们尊重该许可证的条款并对原始作者给予适当的归属。为了扩展数据集，四位 NLP 研究人员手动注释了数据。我们创建并遵循了详细的注释指南，以确保注释过程的一致性、清晰性和公正性。数据集不包括任何个人或敏感信息。

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 数据集创建

A.1 一个拟议的分类法

我们在第 3.3 节中的表格 4 、5 、6 、7 和 8 中分别展示了我们提出的分类法的示例。

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: (i) 表格转换错误的例子。列标题与数据值合并。例如，“`0 87.9`”错误地结合了列名称 `0` 和值 `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: （ii）额外上下文要求的示例。`B` 和 `C` 分别代表 BLEU 和 CHRF++，但这无法仅从声明、标题或表格中推断出来。需要引用原始论文中的额外上下文。

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: （iii）意外的论断类型的示例。该论断仅仅描述了表格所展示的内容，类似于标题，并不需要任何推理或数据来支持。

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: （iv）主观形容词的例子。将表现视为“显著”与否，取决于该词的定义方式。此外，许多人认为使用“显著”一词需要结果通过某种形式的统计检验。第一行的原始版本是：[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: （v）不明确的主张示例。代词“它”指代的实体不明确。