Overview of the Context24 Shared Task on Contextualizing Scientific Claims

Joel Chan[♢*] **Aakanksha Naik**[♣] **Matthew Akamatsu**[♡] **Hanna Bekele**[♡]
Erin Bransom[♣] **Ian Campbell**[♡] **Jenna Sparks**[♣]
[♢] University of Maryland
[♣] Allen Institute for AI
[♡] University of Washington

Abstract

To appropriately interpret and use scientific claims for sensemaking and decision-making, it is critical to contextualize them, not just with textual evidence that the claim was in fact asserted, but also with key supporting empirical evidence, such as a figure that describes a key result, and methodological details, such as the methods of data collection. Retrieving this contextual information when encountering claims in isolation, away from their source papers, is difficult and time-consuming for humans. Scholarly document processing models could help to contextualize scientific claims, but there is a lack of datasets designed for this task. Thus, we contribute a dataset of 585 scientific claims with gold annotations for supporting figures and tables, and gold text snippets of methodological details, that ground the key results behind each claim and run the Context24 shared task to encourage model development for this task. This report describes details of our dataset construction process, summarizes results from the shared task conducted at the 4th Workshop on Scholarly Document Processing (SDP), and discusses future research directions in this space. To support further research, we also publicly release the dataset on HuggingFace.

1 Introduction

People read and use scientific claims both within the scientific process (e.g., in literature reviews, problem formulation, making sense of conflicting data) and outside of science (e.g., evidence-informed deliberation). To appropriately interpret, appraise, and ultimately use claims for sensemaking and decision-making, it is critical to *contextualize* claims with key supporting empirical evidence (e.g., figures presenting key results) and methodological details (e.g., measures, sample).

The goal of contextualizing claims with *empirical* evidence differs slightly from other evidence identification tasks, in that it is aimed at retrieving very targeted (down to the sub-figure) multimodal evidence (i.e., figures or tables grounding a scientific claim), rather than simply a set of text snippets, as is common in claim verification tasks and datasets (Vladika and Matthes, 2023).

This distinction between empirical — vs. general textual — evidence for a claim is crucial for sensemaking and decision-making: knowing that a scientific claim has been asserted in a paper is significantly different from knowing that a scientific claim is associated with empirical evidence (or not) that can be further assessed. Without a clear distinction between these forms of evidence, downstream sensemaking and synthesis may be compromised by the spread of "ghost claims" that lack empirical evidence but are asserted as if they do have evidential weight. In a striking illustration of this, (Harzing, 1995) found that a widely circulated claim about high failure rates for expatriates who are sent abroad to work was based on misquotations from three articles, only one of which contained solid empirical evidence that was in fact contrary to the focal claim. Therefore, despite the increased difficulty, linking claims to empirical evidence is crucial for the scientific process.

However, scientific claims are often encountered and used in settings far removed from the source materials and data, such as in brief citation statements in research papers, or conversations in discussion threads or on social media. In such cases, retrieving needed contextual information in the moment can be difficult and time-consuming for humans. Scholarly document processing models trained for scientific claim verification and fact checking (Vladika and Matthes, 2023), figure captioning (Hsu et al., 2021), and data extraction for systematic reviews (Schmidt et al., 2023) could potentially assist with the task of contextualizing

[*]Corresponding author: joelchan@umd.edu

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 12–21
August 16, 2024 ©2024 Association for Computational Linguistics

scientific claims in these settings. However, the community currently lacks datasets specifically formulated for this task.

We bridge this gap by contributing a dataset of 585 scientific claims actually in use in lab notes and discussions for synthesis and research planning, across the domains of biology, computer science, and the social sciences. For each claim, the dataset includes gold annotations for figures/tables that ground the key empirical evidence for the claim (Track 1); for a subset of these claims, the dataset also includes gold examples of text snippets that describe the key methodological details that ground each claim (Track 2). As part of the 4th Workshop on Scholarly Document Processing (SDP) at ACL 2024, we ran a shared task with this dataset, with official submissions from a total of 6 teams for Track 1, and 2 teams for Track 2. In this report, we describe the details of the dataset construction, summarize the results from the shared task, and discuss implications and future research directions in this space. To support further research on this task, our full dataset (including gold test data) is available on HuggingFace [1].

2 Related Work

The task of contextualizing scientific claims can be formulated similarly to tasks like claim verification and fact checking. Additionally, modeling techniques used for this task have parallels in work on figure captioning and data extraction from scientific literature.

2.1 Claim Verification

Claim verification is typically formulated as the task of predicting support/refute relationships between claims and snippets of source texts, such as research abstracts (Vladika and Matthes, 2023). There are several datasets for scientific claim verification, but most widely used ones like SciFact (Wadden et al., 2020), SciFact-Open (Wadden et al., 2022), COVID-Fact (Saakyan et al., 2021), and HealthVER (Sarrouti et al., 2021) focus on pairing claims with textual evidence.

There are, however, a few more specific datasets that focus on verifying scientific claims against figures or tables. For instance, MuLMS-Img (Tarsi et al., 2024) includes 78 queries (and their corresponding figures) over materials science figures, sourced from domain experts, while SCITAB (Lu et al., 2023) includes 1.2k expert-annotated claims associated with *tables* from a corpus of computer science papers. Our work extends these datasets in terms of size (7x more queries than MulMS-Img), and breadth of scientific domains included (social sciences, HCI, and cell biology vs. materials science or computer science specifically).

2.2 Scientific Figure Captioning

There is a substantial body of work on figure captioning and alt-text generation, supported by many figure captioning datasets, such as SciCap (Hsu et al., 2021), SciOL (Tarsi et al., 2024), MedICaT (Subramanian et al., 2020), PMC-OA (Lin et al., 2023), and Multimodal ArXiv (Li et al., 2024).

The task of generating text describing a figure can be viewed as the inverse of the task of retrieving a figure that matches a scientific claim, thus these datasets and models built using them can be leveraged to build better techniques for our task. Some dataset, such as MedICaT (Subramanian et al., 2020) and PMC-OA (Lin et al., 2023), focus not just on overall figure captioning, but also sub-figure and sub-caption alignment. Thus, to the extent that these figures describe empirical evidence, subsets from these datasets can be as additional weak supervision for our task. However, figure captioning datasets at present do not focus solely on empirical evidence subsets, which can be hard to filter. Another complication to using these datasets directly for our task is that the quality and descriptiveness of figure captions in scientific texts can vary widely (Chintalapati et al., 2022).

2.3 Data Extraction from Scientific Literature

The subtask of retrieving methodological details relevant to a claim is analogous to the task of data extraction in systematic reviews, which has a long history in biomedical informatics and NLP (Schmidt et al., 2023). There are, however, limited public datasets for this task. EBM-NLP (Nye et al., 2018) is the primary public dataset available for use: it consists of 5k medical abstracts annotated with ***P****opulation*, ***I****ntervention* (collapsed with ***C****omparator*), and ***O****utcome* elements from the PICO framework (Richardson et al., 1995), which is widely used to structure systematic reviews in evidence-based medicine, often focusing specifically on randomized controlled trials (RCTs). PubMed PICO (Jin and Szolovits, 2018) is a much larger dataset with 24k RCT abstracts, but PIO

[1] https://huggingface.co/datasets/joelchan/contextualizing-scientific-claims

annotations are automatically assigned based on structured abstract section headings. There are also smaller, disease-specific datasets with PICO annotations (Hu et al., 2023), though these are not publicly available, and datasets with more fine-grained experimental findings annotated (Naik et al., 2024). Beyond medical abstracts, some datasets focus on information extraction and question answering over other scientific domains like computer science (Jain et al., 2020; Dasigi et al., 2021).

We extend this body of work by focusing on scientific claims from a wider range of scientific domains, and with a more specific focus on extracting methodological details.

3 Shared Task Tracks

3.1 Track 1: Evidence Identification

Task formulation. Given a scientific claim c, and a set of figures ($F = \{f_1, ..., f_n\}$) and tables ($T = \{t_1, ..., t_n\}$) from the paper that contains the claim, models must produce a ranking of all items in $F \cup T$, based on whether they provide supporting evidence for the claim c. Therefore, this task can be modeled as an image-text retrieval task. Figure 1 shows an example supporting figure for a claim in our dataset.

Evaluation. We use standard retrieval metrics to assess model performance — normalized discounted cumulative gain (nDCG) at 5 and 10. An issue that complicates relevance computation in our case is that many figures are compound figures, with one or more sub-figures providing supporting evidence for a claim in place of the entire figure. This was especially true for biology-related claims in our dataset (see Table 2). For instance, in Figure 1, the claim is supported by sub-figure (C) but a model might produce a ranking like Figure 1, Figure 1(C). In this case, the top-ranked prediction is a parent of the gold figure, and could be considered partially correct. We tackle this by assigning a partial relevance score of 0.5 to the top-ranked figures/tables that were parent or sub-figures of a gold figure/table.

3.2 Track 2: Grounding Context Identification

Task formulation. Given a scientific claim c, and the full-text of the research paper that contains the claim, models must return a set of text snippets $S = \{s_1, ..., s_n\}$ that describe key methodological details for the experiments that provide empirical evidence supporting the claim. Therefore, this task can be modeled as query-based extractive summarization.

Evaluation. We use automated summarization evaluation metrics to assess model performance — ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020). Specifically, predicted text snippets are mapped to their best-match gold snippets based on highest ROUGE/BERT scores, which are then averaged to produce the final score.

4 Dataset Collection and Preprocessing

Our dataset curation process consists of three stages: (i) claim construction, (ii) collecting annotations for evidence and grounding context identification, and (iii) sourcing supplementary data (e.g., extracted figures/tables, paper full-texts, etc.) to ease model development for participants.

4.1 Claim Construction

The claims used in this shared task were created in the course of naturalistic synthesis work across a range of real-world research contexts listed below:

- `akamatsulab`: Literature review and research planning in a cell biology lab.
- `megacoglab`: Literature review and research planning in a human-computer interaction lab.
- `BIOL403`: Synthesis activities in a microbiology seminar.
- `social-media`: Focused interdisciplinary systematic review of effects of social media on polarization.

A total of 585 claims were created across all contexts, split into sets of 474 and 111 claims, sourced from 229 and 46 unique research papers, for training and testing respectively. The mean number of figures and tables in each paper was approximately 6 and 2, though with a considerable range (up to 24 and 18 figures per paper in train and test, and 19 and 14 tables in train and test).

4.2 Annotation Process

Track 1. The annotation process for track 1 requires identifying key figures and tables that provide supporting evidence for a claim. These figures and tables were captured as screenshots from paper PDFs, alongside figure/table numbers, using a note-taking tool called Roam Research.[2] Because all claims were created in the course of ongoing

[2] https://roamresearch.com/

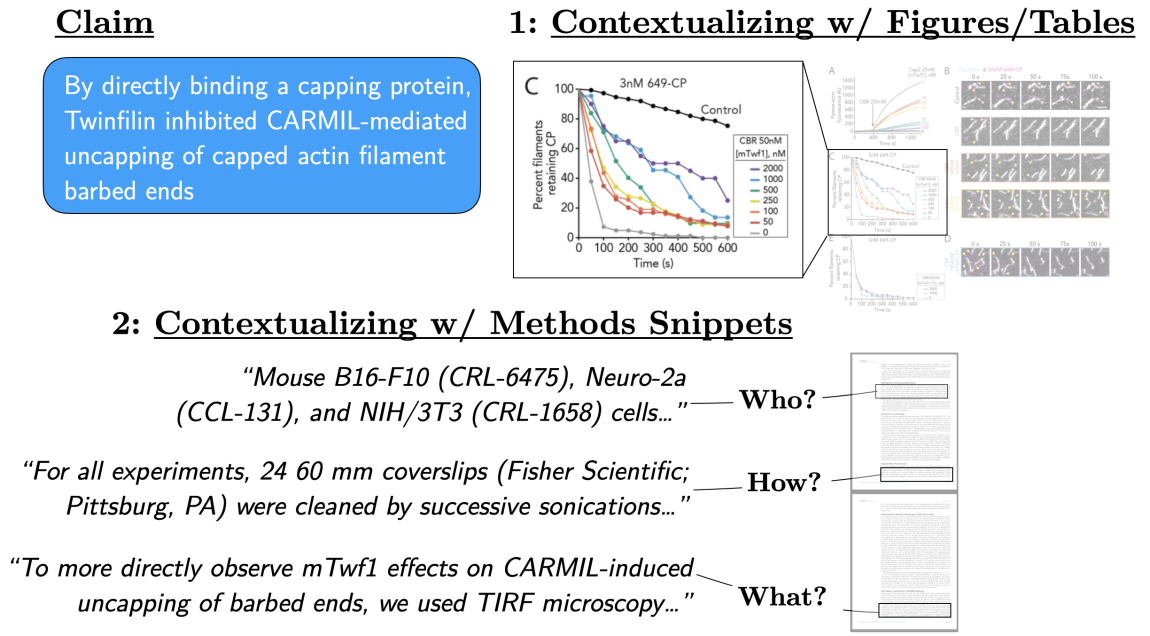

Figure 1: Example claim with figures/tables (Track 1) and methods snippets (Track 2) that contextualize the empirical evidence behind the claim. Relevant figures are often subsets of compound figures, and methods snippets are often distributed throughout source texts.

Dataset	Domain(s)	Train	Test
`akamatsulab`	Cell biology	213	51
`BIOL403`	Cell biology	60	
`social-media`	Social sciences (political science, economics, HCI)	78	
`megacoglab`	Various (HCI, psychology, economics, public health)	123	60
Total		**474**	**111**

Table 1: Number and source of scientific claims for training and test splits in Track 1.

Dataset	Figure	Table	Subfigure
Training split			
`akamatsulab`	205	4	165
`BIOL403`	60	0	49
`social-media`	57	26	2
`megacoglab`	51	89	5
Test split			
`akamastulab`	51	0	50
`megacoglab`	37	41	2

Table 2: Number of claims contextualized by figures, tables, and/or subfigures, by dataset.

synthesis work, all screenshots and labels were captured by researchers with relevant domain knowledge. To further ensure accuracy, the first and third authors, who are PIs of each research group (Chan for `social-media` and `megacoglab`; Akamatsu for `akamatsulab` and `BIOL403`) either personally captured the relevant screenshots and labels, or verified screenshots and labels captured by students or research assistants. Notes from Roam were exported to markdown and figure/table labels were parsed out using a Python script. The first author again manually verified each figure label for accuracy. These labels were collected for all 474 training claims and 111 test claims. The breakdown of the number of claims in training and test splits for track 1 across various research contexts is shown in Table 1. Also, the breakdown of type of context (figure, table, and/or subfigure) is shown in Table 2). Note that subfigures are very common n the biology-focused subset of the data, while tables as context are more common in the social-science/HCI subset o the data.

Track 2. The annotation process for track 2 requires identifying all methodological details for experiments that support a claim. Since our claims covered a range of disciplines and methodologies, including experimental studies, biological compu-

Dataset	Domain(s)	Dev	Test
akamatsulab	Cell biology	28	49
megacoglab	Various (HCI, psychology, economics, public health)	14	60
Total		**42**	**109**

Table 3: Number and source of scientific claims for dev and test splits in Track 2.

tational simulations, observational studies, and so on, we conceptualized a generalized schema for defining three aspects of methodological context that are needed to interpret and appraise the empirical evidence for a claim (see Figure 1 for examples from our data):

- **What** observable measures/data were collected
- **How** (with what methods, analyses, etc.) from
- **Who(m)** (which participants, what dataset, what population, etc.)

This annotation schema was primarily shaped by the expressed information needs of the researchers who wrote and use the claims in this dataset, and informed in part by other schemas for synthesis, such as the PICO framework for medical systematic reviews (Richardson et al., 1995), which also focus on the who (population), how (intervention and comparator), and what (outcome). While this schema enabled us to reliably identify all methodological details researchers needed across the range of disciplines and methodologies in our data, in practice it was often difficult to cleanly distinguish between some of the schema categories. For instance, in observational studies, data collection and preparation details could either be classified as *what* data were used in analysis or *how* data was obtained; this separation between data collection and analysis is somewhat cleaner in experimental studies, where there are substantially more details on procedure, materials, and design. Thus, we did not include snippet categorization into who/how/what, and only assessed model performance on snippet identification. However, we described our annotation schema to participants to clarify task requirements and provide additional motivation for modeling approaches.

All methodological details were captured as text quotes from PDFs copied into the Roam Research note-taking tool. As with figure/table snippets, the first and third authors either personally captured relevant text quotes, or verified quotes captured by students or research assistants. Due to the high labor cost of obtaining text quotes for track 2, which requires exhaustive annotation over paper full-texts, we did not annotate the entire training set of claims. Instead, we collected and released annotations for a small development set of 42 claims (out of 474), and posed this as a few-shot task. For testing, we collected annotations for 109 claims. The research context breakdown of development and test splits for track 2 are shown in Table 3.

4.3 Supplementary Data Sourcing

Extracting figures, tables and captions. Track 1, which focuses on ranking all figures and tables from a paper based on whether they provide supporting evidence for a claim, requires models to accurately extract figures and tables from paper PDFs. Our preliminary exploration showed that even state-of-the-art toolkits such as PaperMage (Lo et al., 2023) perform poorly on figure and table extraction, particularly for cell biology ($\sim$20% accuracy). This motivated us to get all figures, tables and captions manually extracted from paper PDFs and release them as supplementary data.

For this annotation process, all paper PDFs were rasterized and uploaded to the Label Studio data labeling platform. [3] Annotators then labeled all figures, tables and captions by drawing bounding boxes, annotating figures at the smallest sub-figure level. We recruited three annotators with experience in reading and writing scientific text from Upwork.[4] All annotators were trained to do the task on a small pilot set of PDFs annotated by two authors, and paid $20-$24 an hour. Bounding box coordinates collected via this process were used then to crop out figures, tables and captions. Sub-figures were concatenated into a compound figure since the entire figure might be labeled as supporting evidence in some instances. For example, if a Figure 1 had subfigures 1A, 1B, and 1C, we provided a compound Figure 1 besides Figure 1A, Figure 1B and Figure 1C as candidates for ranking. Lastly, we also ran optical character recognition (OCR) using the Nougat library (Blecher et al., 2023) on caption images and released caption texts.

Extracting full-texts. Track 2 requires participants to extract methodological details from paper full-texts. To make participation easier, we released

[3] https://labelstud.io/
[4] https://www.upwork.com/

parsed full-texts for all PDFs extracted using PaperMage (Lo et al., 2023). Because we could not guarantee the accuracy of section labels, the full-text parses did not have section information.

Silver data generation. Since track 2 has limited training data, we release additional unlabeled text resources to encourage exploration of weak supervision-style approaches. Specifically, we also provided full text parses for 17,007 papers from 1-2 hop in-bound and out-bound citations of the original papers present in our training data.

The training, development, and blind test data (without gold annotation) for both tracks are available on HuggingFace[5].

5 Official Results

The training data for this shared task was released on Github, alongside evaluation scripts, on April 11, 2024. The test phase was released as an open competition on Eval.Ai[6] on May 31, 2024. Participants were allowed to upload a maximum of five submissions per day during the seven-day testing phase. We received official submissions from a total of 6 teams for track 1, and 2 teams for track 2. The official results from the leaderboard, as well as the performance of some baseline systems, are shown in Tables 4 and 5 for tracks 1 and 2, respectively.

Team	nDCG@5	nDCG@10
CSIRO-LT	**0.73**	**1.06**
OSX	0.64	0.69
UW24	0.26	0.30
larch24	0.25	0.29
KISTI	0.17	0.22
SLTD	0.14	0.14
Baselines		
CC-Sim	0.28	0.36
CC-Sim (Sci)	0.29	0.36
CC-LLM-Rank	0.33	0.36
LLaVA	0.20	0.29
GPT-4o	0.64	0.68

Table 4: Official results for Track 1. For each team, we report the performance of the highest-scoring run submitted to the leaderboard

[5] https://huggingface.co/datasets/joelchan/contextualizing-scientific-claims

[6] https://eval.ai/web/challenges/challenge-page/2306/overview

5.1 Track 1

Baseline systems. For track 1, we report the performance of the following baselines:

- **CC-Sim:** Ranking figures/tables in decreasing order of cosine similarity between the claim and figure/table caption texts embedded using MPNet (Song et al., 2020). We use the all-mpnet-base-v2 version from Sentence Transformers (Reimers and Gurevych, 2019).[7]
- **CC-Sim (Sci):** Same as above, except claim and caption texts are embedded using SPECTER (Cohan et al., 2020).
- **CC-LLM-Rank:** Prompting an LLM to rank figures/tables in decreasing order of support given a claim and figure/table caption texts. We use GPT-4-Turbo.
- **LLaVA:** Prompting a multimodal LLM to produce relevance scores given a claim and figure/table pair, which are then used for ranking. We use the llava-1.5-7b-hf version from Huggingface.
- **GPT-4o:** Same as above, except using GPT-4o.

Participant submissions. Though six teams participated, we only received reports from the top three teams: CSIRO-LT, OSX, and UW24. Interestingly, all three adopted different approaches to the task. Team OSX opted for the most straightforward approach of prompting multimodal LLMs, but explored sophisticated prompting strategies such as chain-of-thought prompting, multiple retries, etc., which significantly boosted their performance. Team UW24 focused on data augmentation and leveraged datasets for related tasks like scientific image captioning to train a better scientific image-text similarity model, but this technique was less successful. Lastly, CSIRO-LT had the most complicated system design consisting of an ensemble of retrieval approaches, which differed in retrieval strategy used (BM25, embeddings, etc.) and how text *documents* representing figures/tables are constructed (LLM-based descriptions, captions, etc.). Consistent with the findings of other teams, they also observed that their text-based systems outperformed all approaches that tried to train an image-text similarity model.

The top-performing team (CSIRO-LT) achieved nDCG@5 = 0.73 and nDCG@10 = 1.06. Note that

[7] https://sbert.net/docs/sentence_transformer/pretrained_models.html

Team	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L
CSIRO-LT	**0.87**	**0.35**	**0.18**	**0.27**
OSX	0.86	0.32	0.17	0.26
Baselines				
GPT-4	0.70	0.28	0.16	0.22
Mixtral-8x22B	0.80	0.29	0.14	0.23

Table 5: Official results for Track 2. For each team, we report the performance of the highest-scoring run submitted to the leaderboard

our partial relevance score modification can sometimes lead to nDCG scores > 1 if both the correct sub-figures as well as parent figures are present in the top ranked predictions. With the exception of the top two teams, all others had worse results than our weakest baseline, indicating the difficulty of this task. On the other hand, GPT-4o achieves very strong performance with our simple zero-shot baseline prompting strategy, almost matching the performance of the team ranked second, indicating that closed LLMs are now improving the tractability of this task.

5.2 Track 2

Baseline systems. For track 2, we report the performance of the following baselines:

- **GPT-4:** Prompting an LLM to return all relevant quotes about methodological details given a claim and the paper full-text. We use GPT-4-Turbo.
- **Mixtral-8x22B:** Same as above with an open LLM, we use Mixtral-8x22B-Instruct-v0.1 owing to its large context length which can ingest full-texts.

Participant submissions. Only CSIRO-LT and OSX participated in track 2, with OSX continuing to rely on prompting LLMs. CSIRO-LT again leveraged retrieval approaches, additionally exploring rule-based and LLM-based postprocessing of retrieved snippets to further push performance.

For track 2, both participating teams had very close scores, improving moderately over our baselines. We note here that the BERTScores were in general substantially higher than the ROUGE scores. Interestingly, we observed that Mixtral achieved slightly better performance than GPT-4 indicating that this task is tractable for SOTA open LLMs too (unlike track 1 where LLaVA significantly underperformed GPT-4o).

6 Discussion and Future Directions

Overall, our baseline experiments and participant submissions for both shared task tracks demonstrate that the problem of contextualizing scientific claims is still challenging but starting to become tractable. For instance, baseline performance was quite low, but two teams were able to substantially outperform these baselines while still leaving some room for improvement. Our experiences constructing the dataset, and reflections on participants' technical reports, also suggest fruitful future directions.

Better image-text similarity models. Contrastively trained image-text similarity models like CLIP could be beneficial for our task since they can adapt to other scientific domains with little training data (necessary for real-world use given high costs of collecting annotated data). However, a common theme across participants' reports was that using image-text similarity models did not perform well compared to relying entirely on text or prompting multimodal LLMs. Exploring why models like CLIP do not work well on our task and how to improve them could lead to stronger models for our task, as well as advancements in image-text representation learning. One potential direction may be to explore augmentation of figures with textual mentions (Yang et al., 2023).

PDF pre-processing. An ideal system for contextualizing scientific claims would be able to operate end-to-end from research paper PDFs. However, state-of-the-art PDF preprocessing toolkits (e.g., PaperMage (Lo et al., 2023)) were not able to extract figures/tables with sufficient accuracy, particularly for some domains like biology papers; thus, to make this shared task tractable, we provided manually extracted figures/tables/captions to participants. This suggests that further work is needed to improve PDF preprocessing tasks such as

structured content extraction (Shen et al., 2022; Lo et al., 2023), or converting PDFs to more accessible HTML (Wang et al., 2021), which could enable development of better end-to-end contextualization methods, especially across scientific domains.

Evaluation metrics for context identification. Finally, from an evaluation perspective, we suggest that better evaluation metrics may be needed for assessing the extraction of methodological details as context for claims. Similar to other summarization tasks, in our setting, methodological details can be presented at multiple points in a paper, and models might choose different snippet(s) expressing the same information. Our choice of BERTScore as a complement to ROUGE partially mitigated this issue: however, this approach may overestimate performance since it relies on overall text similarity rather than assessing whether all relevant facts are retrieved. Thus, better automated evaluation for this subtask is still an open problem. Future work could explore adapting factuality-style evaluations, such as extracting atomic facts from gold snippets and measuring whether they are covered/supported by model-retrieved snippets (Min et al., 2023) or explore model-based evaluation using LLMs as judges (Bubeck et al., 2023).

7 Conclusion

In this paper, we described a new dataset and results from a shared task for contextualizing scientific claims with empirical evidence and methodological details. Our dataset consists of 585 scientific claims with gold annotations for figures and tables, and gold text snippets of methodological details, that ground the key results behind each claim. Experiments with text embedding and (multimodal) LLM-based baselines demonstrate the challenging nature of the task. Performance gains over these baselines from official submissions to the Shared Task for this dataset at SDP 2024 show the tractability of the task, and reveal promising directions for future work, such as improving image-text similarity models, PDF-processing, and automated evaluation metrics for context identification.

Acknowledgements

The authors would like to thank William Ammon, Anthea Weng, and Nessma Hassan for their help with figure, table and caption annotation, Kyle Lo and Joseph Chee Chang for their assistance with PaperMage and Gunjan Chhablani for supporting our Eval.Ai platform setup during the testing phase.

References

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural Optical Understanding for Academic Documents. *arXiv preprint*. ArXiv:2308.13418 [cs].

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Sanjana Shivani Chintalapati, Jonathan Bragg, and Lucy Lu Wang. 2022. A Dataset of Alt Texts from HCI Publications: Analyses and Uses Towards Producing More Descriptive Alt Texts of Data Visualizations in Scientific Papers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '22, pages 1–12, New York, NY, USA. Association for Computing Machinery.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics. 00002.

Anne-Wil K. Harzing. 1995. The persistent myth of high expatriate failure rates. *The International Journal of Human Resource Management*, 6(2):457–474. Publisher: Routledge _eprint: https://doi.org/10.1080/09585199500000028.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. 2021. SciCap: Generating Captions for Scientific Figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yan Hu, Vipina K Keloth, Kalpana Raja, Yong Chen, and Hua Xu. 2023. Towards precise PICO extraction from abstracts of randomized controlled trials using a section-specific learning approach. *Bioinformatics*, 39(9):btad542.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Di Jin and Peter Szolovits. 2018. PICO Element Detection in Medical Text via Long Short-Term Memory Neural Networks. In *Proceedings of the BioNLP 2018 workshop*, pages 67–75, Melbourne, Australia. Association for Computational Linguistics.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal ArXiv: A Dataset for Improving Scientific Comprehension of Large Vision-Language Models. *arXiv preprint*. ArXiv:2403.00231 [cs].

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-CLIP: Contrastive Language-Image Pre-training using Biomedical Documents. *arXiv preprint*. ArXiv:2303.07240 [cs] version: 1.

Kyle Lo, Zejiang Shen, Benjamin Newman, Joseph Chang, Russell Authur, Erin Bransom, Stefan Candra, Yoganand Chandrasekhar, Regan Huff, Bailey Kuehl, Amanpreet Singh, Chris Wilhelm, Angele Zamarron, Marti A. Hearst, Daniel Weld, Doug Downey, and Luca Soldaini. 2023. PaperMage: A Unified Toolkit for Processing, Representing, and Manipulating Visually-Rich Scientific Documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 495–507, Singapore. Association for Computational Linguistics.

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.

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

Aakanksha Naik, Bailey Kuehl, Erin Bransom, Doug Downey, and Tom Hope. 2024. CARE: Extracting experimental findings from clinical literature. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4580–4596, Mexico City, Mexico. Association for Computational Linguistics.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

W Scott Richardson, Mark C Wilson, Jim Nishikawa, and Robert S Hayward. 1995. The well-built clinical question: a key to evidence-based decisions. *ACP journal club*, 123(3):A12–3.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics. 00015.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based Fact-Checking of Health-related Claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lena Schmidt, Ailbhe N. Finnerty Mutlu, Rebecca Elmore, Babatunde K. Olorisade, James Thomas, and Julian P. T. Higgins. 2023. Data extraction methods for systematic review (semi)automation: Update of a living systematic review. *F1000Research*, 10:401.

Zejiang Shen, Kyle Lo, Lucy Lu Wang, Bailey Kuehl, Daniel S. Weld, and Doug Downey. 2022. VILA: Improving Structured Content Extraction from Scientific PDFs Using Visual Layout Groups. *Transactions of the Association for Computational Linguistics*, 10:376–392.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Sanjay Subramanian, Lucy Lu Wang, Ben Bogin, Sachin Mehta, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. MedICaT: A Dataset of Medical

Images, Captions, and Textual References. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2112–2120, Online. Association for Computational Linguistics.

Tim Tarsi, Heike Adel, Jan Hendrik Metzen, Dan Zhang, Matteo Finco, and Annemarie Friedrich. 2024. SciOL and MuLMS-Img: Introducing A Large-Scale Multimodal Scientific Dataset and Models for Image-Text Tasks in the Scientific Domain. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4548–4559, Waikoloa, HI, USA. IEEE.

Juraj Vladika and Florian Matthes. 2023. Scientific Fact-Checking: A Survey of Resources and Approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

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. 00092.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-Open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lucy Lu Wang, Isabel Cachola, Jonathan Bragg, Evie Yu-Yen Cheng, Chelsea Haupt, Matt Latzke, Bailey Kuehl, Madeleine N van Zuylen, Linda Wagner, and Daniel Weld. 2021. Scia11y: Converting scientific papers to accessible html. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–4.

Zhishen Yang, Raj Dabre, Hideki Tanaka, and Naoaki Okazaki. 2023. SciCap+: A Knowledge Augmented Dataset to Study the Challenges of Scientific Figure Captioning. *arXiv preprint*. ArXiv:2306.03491 [cs].

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.