Overview of the Fifth Workshop on Scholarly Document Processing

Tirthankar Ghosal[a] **Philipp Mayr**[b]
Anita de Waard[c] **Aakanksha Naik**[d] **Amanpreet Singh**[d]
Dayne Freitag[e] **Georg Rehm**[f,g] **Sonja Schimmler**[h,i] **Dan Li**[c]

Abstract

The workshop on Scholarly Document Processing (SDP) started in 2020 to accelerate research, inform policy, and educate the public on natural language processing for scientific text. The fifth iteration of the workshop, SDP 2025 was held at the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025) in Vienna as a hybrid event. The workshop saw a great increase in interest, with 26 submissions, of which 11 were accepted for the research track. The program consisted of a research track, invited talks and four shared tasks: (1) SciHal25: Hallucination Detection for Scientific Content, (2) SciVQA: Scientific Visual Question Answering, (3) ClimateCheck: Scientific Fact-checking of Social Media Posts on Climate Change, and (4) Software Mention Detection in Scholarly Publications (SOMD 25). In addition to the four shared task overview papers, 18 shared task reports were accepted. The program was geared towards NLP, information extraction, information retrieval, and data mining for scholarly documents, with an emphasis on identifying and providing solutions to open challenges.

1 Workshop description

Scholarly literature serves as the primary vehicle for scientists and academics to record and disseminate their findings, playing a vital role in driving knowledge forward and enhancing human well-being.

As the volume of scholarly literature continues to grow, automated methods in NLP, information retrieval, text mining, and document understanding are increasingly essential to address challenges such as information overload, disinformation, and reproducibility (Hołyst et al., 2024). While notable progress has been made, scholarly texts present unique characteristics that demand dedicated research efforts. This workshop aims to serve as a venue for tackling these challenges and to foster the development of tasks and resources specific to scientific document processing. Our long-term goal is to establish scholarly and scientific texts as a core domain within NLP research, complementing ongoing work on web and news content.

The first Scholarly Document Processing (SDP) workshop was co-located online with the EMNLP 2020 conference (Chandrasekaran et al., 2020), and provided a dedicated venue for those working on SDP to submit and discuss their research. Following this success and the demonstrated need for venues to foster discussions around scholarly NLP, SDP 2021 co-located with NAACL (Beltagy et al., 2021), SDP 2022 with COLING (Cohan et al., 2022), SDP 2024 (Ghosal et al., 2024) with ACL again aimed to connect researchers and practitioners from different communities working with scientific literature and data and created a premier meeting point to facilitate discussions on open problems in SDP.

SDP 2025 invited submissions from all communities that explore both the applications and challenges of processing scholarly and scientific documents. Relevant topics included, but were not limited to, large language models (LLMs) for science, representation learning, information

[a]Oak Ridge National Laboratory, USA
[b]GESIS – Leibniz Institute for the Social Sciences, Germany
[c]Elsevier, Netherlands
[d]Allen Institute for AI, USA
[e]SRI International, USA
[f]Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany
[g]Humboldt-Universität zu Berlin, Germany
[h]Fraunhofer FOKUS, Germany
[i]Technische Universität Berlin, Germany

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 1–6
July 31, 2025 ©2025 Association for Computational Linguistics

extraction, document understanding, summarization, and question-answering. We also welcomed work on discourse modeling, argumentation mining, network analysis, bibliometrics and scientometrics, as well as research integrity and reproducibility – including new challenges introduced by generative AI. Additional areas of interest included peer review technologies, metadata and indexing, dataset availability, research infrastructure, and digital libraries. We further encouraged contributions on improving inclusion and representation in scholarly work, designing LLM-based interfaces for interacting with scientific documents, and examining the broader societal impact of scholarly communication.

2 Program

The SDP 2025 workshop consisted of keynote talks, a research track and a shared task track. SDP 2025 received 26 submissions for the research track, of which 11 were accepted (42% acceptance rate). Since the workshop will be hybrid, there will be both in-person and virtual presentations at the conference venue and online. Topics of the presentations run the gamut, and include: Scientific Misconduct Detection, Scholarly Impact Prediction, Novelty Assessment in Scientific Literature, Information Extraction from Scientific PDFs, Dataset Reference Extraction, Citation and Document Attribution, Literature Discovery via Natural Language Queries, Mathematical Term Disambiguation, LaTeX Code Generation Evaluation, Clinical Trial Translation Prediction, Climate Misinformation Mapping, Abstract Screening in Systematic Reviews. As expected, we see a sharp increase in papers that employ large language models for downstream SDP tasks. The full program with links to papers, videos and posters is available at https://sdproc.org/2025/program.html.

3 Shared Task Track

SDP 2025 hosted four shared tasks. All four shared tasks had their own organizing committees consisting of several members of the SDP 2025 organizers and/or other collaborators. Detailed overview papers of the shared tasks are referred to and followed in the proceedings.

3.1 Hallucination Detection for Scientific Content (SciHal25)

Organizers: Dan Li, Bogdan Palfi, Colin Kehang Zhang

Generative AI-powered academic research assistants are transforming how research is conducted. These systems enable users to pose research-related questions in natural language and receive structured, concise summaries supported by relevant references. However, hallucinations – unsupported claims introduced by large language models – pose a significant challenge to fully trusting these automatically generated scientific answers.

The SciHal25 task (Li et al., 2025) invites participants to detect hallucinated claims in answers to research-oriented questions. This task is formulated as a multi-label classification problem, each instance consists of a question, an answer, an extracted claim, and supporting reference abstracts. Participants are asked to label claims under two subtasks: (1) coarse-grained detection with labels Entailment, Contradiction, or Unverifiable; and (2) fine-grained detection with a more detailed taxonomy including 8 types. The dataset consists of claim-level annotations designed to evaluate the factual consistency between claims in generated answers and their cited references within scientific retrieval-augmented generation (RAG) systems. The data are primarily derived from Scopus AI, an in-house research assistant tool powered by a RAG system indexing millions of scientific abstracts. The dataset is divided into 3,592 training, 500 validation, and 500 test instances. Subtask 1 saw 83 submissions across 9 teams while subtask 2 saw 38 submissions across 6 teams, resulting in a total of 5 published technical reports. System reports from top three participating teams as well as an overview paper summarizing future directions are included in the workshop proceedings.

3.2 SciVQA: Scientific Visual Question Answering

Organizers: Ekaterina Borisova and Georg Rehm

Data visualisations such as figures (i. e., charts and diagrams) are ubiquitous in scholarly publications. Researchers use *scientific figures* to present and compare results with prior works as well as to enhance the understanding of their findings (Clark and Divvala, 2016). Hence, extracting and interpreting information from figures is beneficial

for a wide array of tasks in scholarly document processing, including visual question answering (VQA). However, reasoning over scientific figures is challenging as they are inherently multimodal, diverse in types, and contain domain-specific concepts (Meng et al., 2024; Zhou et al., 2023; Huang et al., 2024).

The SciVQA shared task (Borisova et al., 2025) aims to shed light on the capabilities of current multimodal large language models to recognise and link visual elements (i. e., colour, shape, size, height, direction, position) of scientific figures with textual content (e. g., captions, legends, axis labels) for the VQA task. Participants were invited to develop VQA systems based on the novel SciVQA dataset containing 3,000 images of scientific figures from the ACL Anthology[1] and arXiv[2], and a total of 21,000 QA pairs.[3] The key focus of the SciVQA challenge is on closed-ended QA pairs, both *visual*, i. e., addressing visual attributes of a figure, and *non-visual*, i. e., not targeting visual elements of a figure. SciVQA was hosted on the Codabench platform (Xu et al., 2022), and the submitted systems were evaluated using precision, recall, and F1 scores of ROUGE-1, ROUGE-L (Lin, 2004), and BERTScore (Zhang* et al., 2020).[4] The competition attracted 20 registered participants, with seven submissions to the leaderboard and five papers reporting the solutions.

3.3 ClimateCheck: Scientific Fact-checking of Social Media Posts on Climate Change

Organizers: Raia Abu Ahmad, Aida Usmanova, and Georg Rehm

The rapid spread of climate-related discourse on social media has created new opportunities for public engagement, but it has also amplified the spread of mis- and disinformation (Fownes et al., 2018; Al-Rawi et al., 2021). As online platforms increasingly shape public understanding of scientific issues, it becomes essential to develop tools that can link everyday claims to trustworthy sources. While NLP has made significant strides in tasks such as misinformation detection (Aldwairi and Alwahedi, 2018; Aïmeur et al., 2023), scientific entity extraction (Hafid et al., 2022; Hughes and Song, 2024), and scientific document understanding (Dagdelen et al., 2024), the challenge of grounding social media claims about climate change in scientific literature remains largely underexplored.

To bridge this gap, we organised ClimateCheck (Abu Ahmad et al., 2025b), a shared task aimed at automating the verification of climate-related claims from social media using scholarly publications as evidence. Hosted on Codabench (Xu et al., 2022) during April/May 2025, the task included two subtasks: (1) Retrieving relevant scientific abstracts for a given claim, and (2) Classifying the claim's veracity based on the retrieved evidence. The competition drew 27 registered users and 13 active teams, 10 of which submitted to the leaderboazed. Participants worked with a curated dataset of 435 climate-related claims written in lay language and a corpus of 394,269 scientific abstracts (Abu Ahmad et al., 2025a). In Subtask I, abstracts retrieval, systems were evaluated using Recall@K ($K = 2, 5, 10$) and Binary Preference to account for incomplete annotations. In Subtask II, claim verification, classification performance was measured using the weighted F1-score and Recall@10 from the previous subtask to encourage both accuracy and evidence coverage. The ClimateCheck dataset and evaluation suite are publicly available, providing a resource for further research on bridging scientific knowledge and public discourse.[5,6]

3.4 Software Mention Detection in Scholarly Publications (SOMD 25)

Organizers: Sharmila Upadhyaya, Wolfgang Otto, Frank Krüger, Stefan Dietze

Scientific research is increasingly data-centric, and software plays a vital role across disciplines by enabling the collection, analysis, and interpretation of research data. As such, software has emerged as a critical scholarly artifact whose identification is essential for ensuring the transparency, reproducibility, and collaborative nature of scientific inquiry. However, the heterogeneous and informal nature of software mentioned in scholarly publications presents ongoing challenges for accurate detection and disambiguation. To ad-

[1] https://aclanthology.org
[2] https://arxiv.org
[3] https://huggingface.co/datasets/katebor/SciVQA
[4] https://www.codabench.org/competitions/5904/
[5] https://huggingface.co/datasets/rabuahmad/climatecheck
[6] https://huggingface.co/datasets/rabuahmad/climatecheck_publications_corpus

dress this, we organized the second iteration of the Software Mention Detection (SOMD2025[7]) shared task as part of the Scholarly Document Processing (SDP) workshop at ACL 2025. The objective of SOMD2025 is to foster community-driven development of joint frameworks for Named Entity Recognition (NER) and Relation Extraction (RE) targeting software mentions and their associated attributes. This edition builds on the previous SOMD2024 task but emphasizes a joint evaluation setting, better reflecting real-world information extraction pipelines. The shared task consists of two phases: Phase I focuses on model development using a gold-standard dataset. At the same time, Phase II introduces an out-of-distribution (OOD) test set to evaluate generalizability. Despite 18 registered participants, only six teams completed two phases and submitted system descriptions. Participants applied diverse strategies, including joint and pipeline architectures, leveraging pre-trained language models and data augmentation using LLM-generated samples. The evaluation was based on a macro-averaged F1 score for NER and RE components, reported as the SOMD score. The top-performing systems achieved a SOMD score of 0.89 in Phase I and 0.63 in Phase II, underscoring the difficulty of generalization in OOD scenarios. These results show clear improvements over the baselines and show that while current methods perform well in in-distribution data, generalization remains a significant challenge.

4 Workshop Review and Outlook

SDP is evolving along with other fields of AI. The increasing maturity of generative LLMs provides new opportunities and poses new challenges. The way in which tasks traditionally associated with literature mining are addressed has changed dramatically over the life of the workshop series. Generative AI has not obviated tasks such as retrieval, extraction, and summarization, but has enabled researchers to explore interesting variants of these tasks and to shift focus from understanding to prediction. The same burgeoning of research, attributable to generative AI's democratizing effect, has created new problems for the conduct of science, raising interest in automated support for peer review and the enforcement of scholarly integrity.

As we consider future iterations of the workshop, we are discussing ways to respond to these trends. With SDP 2025 we have begun to present a more varied set of shared tasks, each highlighting challenges unique to the automated processing of the scholarly literature. As we proceed with planning and advertising, a key objective will be to elicit high-quality submissions from researchers interested in the use and meta-linguistic aspects of scholarly communication.

5 Conclusion

The Workshop on Scholarly Document Processing is part of a virtual cycle. Advances in SDP have given rise to powerful new tools, such as Google's Co-Scientist or Elsevier's ScienceDirect AI, that derive value from the communications of scholars and return value to scholars through sophisticated new forms of research facilitation. To the extent that these tools succeed, both the pace of scholarly discovery and the volume of scholarly communication will increase.

But SDP research is not just an amplifier. We believe and hope that the research fostered at our workshop will open new lines of inquiry across a range of disciplines and relieve scientists of tedious or rote aspects of their labor. We hope that our work will ultimately increase the number and diversity of people that can make meaningful scholarly contributions.

6 Program Committee

1. Aida Usmanova, Leuphana Universität Lüneburg
2. Akiko Aizawa, National Institute of Informatics
3. Allan Hanbury, Complexity Science Hub & Technische Universität Wien
4. Allen G Roush, Oracle
5. Anita De Waard, Utrecht University
6. Antonio Pieri, Elsevier
7. Biswadip Mandal, University of Texas at Dallas
8. Boris Veytsman, Chan Zuckerberg Initiative & George Mason University
9. Buse Sibel Korkmaz, Imperial College London
10. Dan Li, Elsevier
11. Daniel Acuna, University of Colorado at Boulder

[7] https://www.codabench.org/competitions/5840/

12. Dayne Freitag, SRI International
13. Ekaterina Borisova, Technische Universität Berlin & Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)
14. Fabio Barth, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)
15. Hamed Alhoori, Northern Illinois University
16. Hiroki Teranishi, Nara Institute of Science and Technology & RIKEN
17. Ibrahim Al Azher, Northern Illinois University
18. Ioana Buhnila, University of Lorraine
19. James Dunham, Georgetown University
20. Jay DeYoung, Allen Institute for Artificial Intelligence
21. Miftahul Jannat Mokarrama, Northern Illinois University
22. Neil R. Smalheiser, University of Illinois at Chicago
23. Nicolau Duran-Silva, Universitat Pompeu Fabra
24. Petr Knoth, Open University
25. Philipp Mayr, GESIS – Leibniz Institute for the Social Sciences
26. Pierre Senellart, Ecole Normale Supérieure
27. Raia Abu Ahmad, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)
28. Roman Kern, Know Center GmbH & Technische Universität Graz
29. Sameera Horawalavithana, Pacific Northwest National Laboratory
30. Sebastian Schellhammer, GESIS – Leibniz Institute for the Social Sciences
31. Sharmila Upadhyaya, GESIS – Leibniz Institute for the Social Sciences
32. Shiyuan Zhang, University of Illinois at Urbana-Champaign
33. Soham Chitnis, New York University
34. Sotaro Takeshita, University of Mannheim
35. Tamjid Azad, Northern Illinois University
36. Taro Watanabe, Nara Institute of Science and Technology
37. Tohida Rehman, Jadavpur University
38. Tosho Hirasawa, Tokyo Metropolitan University
39. Wojtek Sylwestrzak, University of Warsaw
40. Wolfgang Otto, GESIS – Leibniz Institute for the Social Sciences
41. Xinyuan Lu, National University of Singapore
42. Yoshitomo Matsubara, Yahoo!
43. Yupeng Cao, Stevens Institute of Technology
44. Wuhe Zou, Netease Group

Acknowledgements

The organizers wish to thank all those who contributed to this workshop series: The researchers who submitted papers, the keynote speakers, the many reviewers who generously offered their time and expertise, and the participants of the workshop.

Philipp Mayr received funding from Deutsche Forschungsgemeinschaft under grant: MA 3964/8-2; the OUTCITE project (Backes et al., 2024) and the European Union under the Horizon Europe grant OMINO - Overcoming Multilevel INformation Overload (Hołyst et al., 2024) under grant number 101086321.

Georg Rehm was supported through the project NFDI for Data Science and Artificial Intelligence (NFDI4DS) as part of the non-profit association National Research Data Infrastructure (NFDI e. V.). The NFDI is funded by the Federal Republic of Germany and its states.

References

Raia Abu Ahmad, Aida Usmanova, and Georg Rehm. 2025a. The ClimateCheck dataset: Mapping social media claims about climate change to corresponding scholarly articles. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria.

Raia Abu Ahmad, Aida Usmanova, and Georg Rehm. 2025b. The ClimateCheck shared task: Scientific fact-checking of social media claims about climate change. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria.

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Ahmed Al-Rawi, Derrick O'Keefe, Oumar Kane, and Aimé-Jules Bizimana. 2021. Twitter's fake news discourses around climate change and global warming. *Frontiers in Communication*, 6:729818.

Monther Aldwairi and Ali Alwahedi. 2018. Detecting fake news in social media networks. *Procedia Computer Science*, 141:215–222.

Tobias Backes, Anastasiia Iurshina, Muhammad Ahsan Shahid, and Philipp Mayr. 2024. Comparing Free Reference Extraction Pipelines. *International Journal on Digital Libraries*.

Iz Beltagy, Arman Cohan, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Keith Hall, Drahomira Herrmannova, Petr Knoth, Kyle Lo, Philipp Mayr, Robert Patton, Michal Shmueli-Scheuer, Anita de Waard, Kuansan Wang, and Lucy Lu Wang. 2021. Overview of the second workshop on scholarly document processing. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 159–165, Online. Association for Computational Linguistics.

Ekaterina Borisova, Nikolas Rauscher, and Georg Rehm. 2025. SciVQA 2025: Overview of the first scientific visual question answering shared task. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria. Association for Computational Linguistics.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard Hovy, Philipp Mayr, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview of the First Workshop on Scholarly Document Processing (SDP). In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 1–6, Online. Association for Computational Linguistics.

Christopher Clark and Santosh Divvala. 2016. Pdffigures 2.0: Mining figures from research papers. In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, pages 143–152.

Arman Cohan, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Drahomira Herrmannova, Petr Knoth, Kyle Lo, Philipp Mayr, Michal Shmueli-Scheuer, Anita de Waard, and Lucy Lu Wang. 2022. Overview of the third workshop on scholarly document processing. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 1–6, Gyeongju, Republic of Korea. Association for Computational Linguistics.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Jennifer R Fownes, Chao Yu, and Drew B Margolin. 2018. Twitter and climate change. *Sociology Compass*, 12(6):e12587.

Tirthankar Ghosal, Amanpreet Singh, Anita De Waard, Philipp Mayr, Aakanksha Naik, Orion Weller, Yoonjoo Lee, Zejiang Shen, and Yanxia Qin. 2024. Overview of the fourth workshop on scholarly document processing. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 1–6.

Salim Hafid, Sebastian Schellhammer, Sandra Bringay, Konstantin Todorov, and Stefan Dietze. 2022. Scitweets-a dataset and annotation framework for detecting scientific online discourse. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3988–3992.

Janusz A. Hołyst, Philipp Mayr, Michael Thelwall, Ingo Frommholz, Shlomo Havlin, Alon Sela, Yoed N. Kenett, Denis Helic, Aljoša Rehar, Sebastijan R. Maček, Przemysław Kazienko, Tomasz Kajdanowicz, Przemysław Biecek, Boleslaw K. Szymanski, and Julian Sienkiewicz. 2024. Protect our environment from information overload. *Nature Human Behaviour*, 8:402–403.

Kung-Hsiang Huang, Hou Pong Chan, Yi R. Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models.

Anthony James Hughes and Xingyi Song. 2024. Identifying and aligning medical claims made on social media with medical evidence. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8580–8593.

Dan Li, Bogdan Palfi, Colin Kehang Zhang, Jaiganesh Subramanian, Adrian Raudaschl, Yoshiko Kakita, Anita Dewaard, Zubair Afzal, and Georgios Tsatsaronis. 2025. Overview of the scihal25 shared task on hallucination detection for scientific content. In *The 5th Workshop on Scholarly Document Processing @ ACL 2025*.

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.

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. ChartAssistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7775–7803, Bangkok, Thailand. Association for Computational Linguistics.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

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*.

Mingyang Zhou, Yi Fung, Long Chen, Christopher Thomas, Heng Ji, and Shih-Fu Chang. 2023. Enhanced chart understanding via visual language pretraining on plot table pairs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1314–1326, Toronto, Canada. Association for Computational Linguistics.