Overview of the Fourth Workshop on Artificial Intelligence for Scientific Publications (WASP 2026)

Atilla Kaan Alkan
Center for Astrophysics
Harvard & Smithsonian
atilla.alkan@cfa.harvard.edu

Felix Grezes
Center for Astrophysics
Harvard & Smithsonian
felix.grezes@cfa.harvard.edu

Kelly Lockhart
Center for Astrophysics
Harvard & Smithsonian
kelly.lockhart@cfa.harvard.edu

Alberto Accomazzi
Center for Astrophysics
Harvard & Smithsonian
aaccomazzi@cfa.harvard.edu

Tirthankar Ghosal
Oak Ridge National Laboratory
ghosalt@ornl.gov

Abstract

The Workshop on Artificial Intelligence for Scientific Publications (WASP) brings together research on methods that read, organize, connect, and evaluate the scientific literature. Its fourth edition was held online as part of AACL-IJCNLP 2026 on November 10, 2026. It received 43 submissions and accepted 25: 15 main-track papers and 10 system papers for the AstroCLIMB shared task on linking astronomy figures and captions to the literature. This overview aims to give the community, including those who did not attend, an accessible account of what the workshop presented and found. For each main-track paper, it summarizes the approach, evaluation, main results, and the authors' interpretation and next steps. We then draw out shared themes and future directions. Many contributions are deployed systems built with archives and curators. They combine language models with deterministic validation, keep source evidence attached to extracted values, and design evaluations around scientific use. Several papers find that lean, focused systems perform as well as more elaborate ones. The authors highlight directions such as automatic verification of grounding, shared gold standards, normalization of quantities and entities, figure understanding, and checking that citations are relevant and real.

1 Introduction

Scientific results are communicated through prose, tables, figures, references, and descriptions of experimental conditions. Making this material computationally usable requires more than finding a relevant document. Systems must recover what was measured and under which conditions, connect claims to evidence and to cited work, and keep enough context for a result to be compared or reused. Prior work frames these problems as scientific entity and relation extraction (Luan et al., 2018), generative structured extraction (Dagdelen et al., 2024), citation-informed representations (Cohan et al., 2020), full-paper question answering (Dasigi et al., 2021), claim verification (Wadden et al., 2020) and figure question answering (Borisova et al., 2025). WASP is one venue for this research, alongside the Scholarly Document Processing workshop (Ghosal et al., 2025). It began as the Workshop on Information Extraction from Scientific Publications (WIESP; (Ghosal et al., 2022)), whose 2022 and 2023 editions hosted the DEAL (Grezes et al., 2022) and FOCAL (Grezes et al., 2023) shared tasks. The 2025 edition (Lockhart et al., 2025), the first under the WASP name, hosted TRACS (Grezes et al., 2025). The fourth edition, co-located with AACL-IJCNLP 2026, introduced the multimodal AstroCLIMB shared task (Grezes et al., 2026a). This paper is written for readers who did not attend. For each main-track paper, we summarize what the authors did, how they evaluated it, what they found, how they interpret it, and the next steps they identify (Section 3). Sections 5 and 6 give our synthesis as organizers, with every point tied to specific papers. Overall, the 2026 program moves from *can a model extract this?* toward *can the extracted record be trusted, traced to its source, and used for science?*

2 The Workshop

Submissions and review WASP 2026 received 43 submissions through OpenReview, each reviewed by program committee members (Appendix A), typically by two reviewers. In the main track, we accepted 15 papers (10 long and 5 short);

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 1–9
November 10, 2026 ©2026 Association for Computational Linguistics

in the shared task, we accepted all 10 submissions.

Program The one-day online program, held on November 10, 2026, combined oral and poster presentations, two keynotes, a panel discussion, and a shared-task session. In her keynote, *Salience and Inquiry in Scientific Literature*, Jessy Li (The University of Texas at Austin) framed the reading of scientific literature as an implicitly interactive discourse process, in which readers identify salient information while continually asking questions. Using summarization as a behavioral probe, she showed that the questions models answer remain strikingly consistent across summary lengths. She also discussed question-driven discourse models linking passages and figures, and AI review pilots at large conferences. In the second keynote, *From General-Purpose LLMs to Grounded Tools for Scientific Discovery*, Pavlos Protopapas (Harvard University) drew on two projects built on NASA's Astrophysics Data System (ADS) to discuss how retrieval, tool use, verification, and explicit workflows can make LLM outputs about the literature more transparent and assessable, and which limitations remain. His central question was how to design AI tools that support scientific judgment and let researchers inspect and challenge their results. The panel discussion, *Easier to Write than to Verify? Generative AI and the Crisis in Scholarly Publishing*, brought together researchers and representatives of learned societies to discuss how the scholarly record can remain trustworthy now that generative models are used at every stage of publishing, by authors, reviewers, publishers, and indexing services alike. The starting point was an asymmetry: the cost of producing a plausible paper has fallen much faster than the cost of checking one, while submission volumes outpace the pool of willing referees and AI-assisted paper mills produce work that is hard to detect. The panelists debated whether AI-assisted review should be banned, tolerated, or required as a first pass, and how its use should be disclosed; whether detecting AI-generated text is a solvable problem or the wrong one to pursue; what a machine-verifiable paper, with claims explicitly linked to their underlying data and code, would look like and who would bear its cost; and how citation as a measure of impact changes when discovery increasingly happens through AI-generated answers. A recurring question was whether AI writing tools should raise the bar for what a paper contains rather than simply increase how many papers are published. Rather than settle these questions, the discussion aimed to identify where NLP and AI research can make the most difference.

Domains and Contributions Table 1 summarizes the 15 main-track papers. Six concern astronomy literature or archives. The others cover agronomy, metabolomics, medicine, biomedical research funding, NLP research, and a six-domain benchmark. Several papers come from, or were built with, institutions that maintain scientific records such as Astro-COLIBRI (Reichherzer et al., 2021) or the NASA/IPAC archives. Three describe systems already in production or used by other projects (Schüssler et al., 2026; Susemiehl et al., 2026; Koblischke and Bovy, 2026).

3 Main-Track Findings

For each paper, we describe the problem and approach, the evaluation and main results, and the authors' own reading of their findings and next steps.

3.1 Extraction Pipelines in Operation

Schüssler et al. (2026) converts free-text GCN Circulars in time-domain astronomy into auditable, event-linked observation records. Regular expressions supply hints, an LLM (`gpt-5.4-mini`) fills a closed schema in which every observation quotes its supporting excerpt, and deterministic code normalizes times, filters, and exposures. In an audit by seven domain reviewers, 223 of 231 reports were fully confirmed and the remaining 8 were partly correct, with 99.80% of definite field decisions confirmed. A cross-check against three redshift catalogs found no misread values. Since 2016, the pipeline has produced 68,393 observations, now served in real time by Astro-COLIBRI (Reichherzer et al., 2021). The authors summarize the design as "the LLM handles linguistic variation, while code enforces numerical and interface invariants," and identify an independently annotated corpus as the next step to enable recall measurement.

Susemiehl et al. (2026) describe a pipeline that drafts database load files from new papers. Local open-weight LLMs extract object names and parameters with a source snippet for every value, guided by a similar previously processed paper. A second model then checks the result, tool calls handle unit conversion, and tables are parsed program-

Paper	Domain	Contribution	Data and scale	How it is evaluated	Headline result (as reported)
Extraction pipelines in operation (§3.1)					
Schüssler et al. (2026)	Time-domain astronomy	Deployed LLM + rule pipeline	26,811 GCN Circulars → 68,393 observations	Expert audit of 231 outputs; redshift catalog cross-check	99.80% of definite field decisions confirmed; 8/231 reports partly correct
Susemiehl et al. (2026)	Astronomy archives (NED, NEA)	Deployed multi-stage LLM pipeline	100 articles per archive vs. human load files	Strict match to historical load files	Macro recall 0.89 (NED) / 0.83 (NEA); in daily use by curators
Koblischke and Bovy (2026)	Extragalactic astronomy	LLM-agent-built resource (GalaxyWiki)	156,006 astro-ph papers → 44,332 discussions of 13,616 galaxies	Quote matching; manual review of 120 records	96.5% of quotes match source after normalization; 118/120 records correct
Kumar et al. (2026)	Legacy LENR literature	Multimodal extraction architecture	60-paper unannotated pilot	Pilot: throughput and validation behaviour	58.3% of papers triaged out before extraction; 48.5 s per paper
Extraction measured against ground truth and downstream use (§3.2)					
Lu et al. (2026)	Agronomy (intercropping)	Comparison of 3 LLM strategies × 6 open models	90 papers, 42 fields	Field-level vs. curated data; re-run of published meta-regression	Direct zero-shot best (similarity-adjusted F1 0.577); published effect direction recovered with all six models
Rovai (2026)	Metabolomics (GC)	Task, rule-based extractor, comparability census	355 MetaboLights studies; 167,331 Workbench pairs	Two blind LLM raters validated on planted errors	No fabricated values in audit; comparable pairs 120 → 171
Bölücü et al. (2026)	Six domains	Task formalization and benchmark	20 repurposed datasets	Record-aligned cell-level metrics; zero-shot LLM/VLM baselines	Challenging for zero-shot baselines; normalization gives large gains
Pranjić et al. (2026)	CS abstracts, news	Context-fusion keyphrase model	kp20k, Inspec, KPTimes	F1@5/@10 vs. SEKE	Comparable on scientific benchmarks with 170.9M vs. 274M parameters
Grounded answers and citations (§3.3)					
Recski et al. (2026)	NLP literature (ACL Anthology)	Extractive QA system and gold set	∼114K papers; 100 annotated query–chunk pairs	Word-level span F1	150M student 53.6 vs. best LLM extractor 48.7
Goyal (2026)	Medical QA	RL with a citation-existence verifier	Train: 1,479 COVID-QA questions; test: 40 MedAESQA questions	Rate of non-resolving references	Hallucinated references 85.5% → 76.7%; 71% more verified references
Basseri et al. (2026)	Astro-, earth and planetary science	Dataset (LCRC) and retrieval study	14,735 citing sentences; 52,618-paper corpus	Hit rate, Recall, NDCG@k	+8.0% relative Recall@25 over dense retrieval, no fine-tuning
Bibliographies and knowledge flows (§3.4)					
Grezes et al. (2026b)	Telescope bibliographies	Dataset (TRACS) and zero-shot baselines	89,579 papers; Chandra, HST, JWST	Per-telescope F1	Telescope F1 up to 82.4 zero-shot; usage classification is harder
Toney-Wails et al. (2026)	NIH-funded research	Citation-aware summarization (CARDS)	4,826 diffusion pathways; 81 evaluated	Human annotation and LLM judge	CARDS judged more specific in 70.7% of paired comparisons
Figures and claims (§3.5)					
Naiman (2026)	Astronomical sky images	VQA dataset and model probe	2,001 figures, 47,357 questions; 255 figures × 3 LMMs	Accuracy; log-ratio error	Domain conventions (e.g., sky coordinates) challenge low-cost LMMs
Centonze (2026)	Biomedical claims (SciFact)	Claim-preservation framework	40 reformulated claims (illustrative)	Exploratory qualitative analysis	Shows how subtle epistemic shifts can be detected

Table 1: The 15 main-track papers of WASP 2026, grouped as in Section 3. Results are those reported by the authors in their own settings; they are not comparable across rows.

matically. Against 100 human-made load files per archive, macro recall reaches 0.89 (NED) and 0.83 (NEA), with median times of 7.4 and 33.4 minutes per paper, compared with 15-45 minutes for a curator. The authors explain that development deliberately prioritized recall, since curators remove extra values more easily than they find missing ones, and that NEA is harder because its records carry about 27 parameters per object. Curators use the tool daily, and the next goal is to improve precision and parsing.

In astronomy, databases link galaxies to papers but do not show whether a paper actually discusses a galaxy. GalaxyWiki (Koblischke and Bovy, 2026) aims to fill this gap. GPT-5.5 agents read 87,899 astro-ph papers and, for every galaxy discussed in more than two sentences, copied verbatim quotes and wrote a summary; discussions were then grouped by sky position. The result is 44,332 discussions from 18,641 papers, covering 13,616 galaxies. 96.5% of quotes match the source text, and 118 of 120 manually reviewed records are correct. The authors see this as an example of how language models make previously impractical projects feasible through parallelizable, verifiable tasks. Other projects already use the resource, and the authors plan to extend the corpus and evaluate coverage.

Kumar et al. (2026) targets historical experimental literature on Low Energy Nuclear Reactions,

which buries data in multi-column layouts and plots. It combines layout-aware parsing, a low-cost triage step that sets aside non-empirical papers and decorative images, and multimodal extraction into a schema that records failed experiments and time series. Typed validators reject physically impossible values. A 60-paper pilot shows that triage sets aside 58.3% of papers before the expensive model runs, at 48.5 s per paper. The authors' central methodological message is to separate clearly what code enforces from what the prompt requests. Automated verification of source quotes and evaluation against a gold standard are their stated next steps.

3.2 Extraction Against Ground Truth and Downstream Use

Data extraction for meta-analysis (Lu et al., 2026) Can LLMs automate the data-extraction step of an agricultural meta-analysis? The authors re-extract 42 fields from 90 intercropping papers and compare direct zero-shot prompting, a two-step workflow, and a multi-agent system across six open-weight models. Beyond field-level scores, they refit the published regression of land-equivalent ratio on temporal niche differentiation on each extracted dataset. Direct prompting performs best across all six models (best similarity-adjusted F1: 0.577) and recovers the positive effect reported in the original study for every model, though the estimated magnitudes vary. The authors describe a "complexity trap": extra stages lost context or carried early errors forward. They conclude that LLMs are currently best used to assist researchers, with domain-specific training as a promising direction.

Method parameters from protocol prose (Rovai, 2026) This paper proposes measuring metadata quality by whether two records can be *compared*, rather than by how many fields are filled. A census of all 167,331 pairs of gas-chromatography analyses in Metabolomics Workbench shows that a single field, the retention-index reference series, determines most comparability. The author therefore defines recovering the oven program and reference series from the protocol text as an extraction task with explicit abstention, and builds a precise rule-based extractor. Two blind LLM raters, first validated on planted errors, audit the outputs: 65% of programs are fully correct, none contains a fabricated value, and the series extractor is correct on 57 of 66 recoveries. Recovered series raise the number of comparable pairs from 120 to 171. The author argues that the extraction pipeline should be credited with the extracted values and releases all values with their source sentences as a seed gold standard.

Scientific Table Generation (Bölücü et al., 2026) Structured extraction from papers has been studied under many names. This paper unifies it as Scientific Table Generation: filling a schema-defined table of records from PDFs. The benchmark brings together 20 datasets from six domains, with an evaluation protocol that aligns records before scoring cells and adds metrics for numerical values. Zero-shot baselines with text LLMs and a page-level VLM show that the task remains challenging (best average F1 13.78, o1-mini). In pilot experiments, normalizing units and values improves scores markedly (e.g., from 3.59 to 18.56 F1). The authors identify numerical representation, unit and entity normalization, entity linking, and table and figure understanding as key open challenges, and invite dataset contributions from new domains.

Keyphrase extraction with context fusion (Pranjić et al., 2026) This paper keeps token-level representations of the target sentence while compressing every surrounding sentence into a single vector, so that document context is available without processing the whole document token by token. With 170.9M parameters, the model performs comparably to the larger SEKE system (274M) on the scientific benchmarks kp20k and Inspec, and slightly better on kp20k with full training (37.5 vs. 36.3 F1@5). The authors' error analysis shows that many predictions counted as wrong partially overlap gold keyphrases. They point to candidate selection, phrase boundaries, and ablations of the context component as next steps.

3.3 Grounded Answers and Citations

ACL-Verbatim (Recski et al., 2026) To make literature question answering free of hallucination by design, ACL-Verbatim returns only verbatim spans from about 114K ACL Anthology papers, which are released in Markdown. The authors build a manually annotated evaluation set and train a compact 150M ModernBERT extractor on LLM-generated labels. It reaches the best word-level F1 (53.6, vs. 48.7 for the strongest LLM extractor) at a fraction of the cost. The authors highlight its precision: it abstains on most irrelevant chunks, so it also filters out irrelevant search results. They see

the approach as a blueprint for efficient extractive QA in other domains, and plan larger annotation efforts.

Reinforcement learning against a citation verifier (Goyal, 2026) Since we can objectively check whether a reference exists, this paper uses it as a reward. A deterministic verifier resolves generated references in CrossRef and OpenAlex, and Llama models are trained against it with GRPO. On medical questions from a different dataset, the share of hallucinated references drops from 85.5% to 76.7%, and the model produces 71% more verified references, with no sign of reward hacking in a six-part audit. The trained model favors canonical sources, which helps on common questions. The author notes that verifying relevance, and not only existence, is the natural next step toward fully verifiable answers.

Local citation recommendation (Basseri et al., 2026) This paper releases LCRC, a corpus of 14,735 sentences from astrophysics review papers with their citations, and studies recommendation from content alone, without citation counts or the citation graph. Simple, general-purpose choices work best: a general embedding model, plain text chunks, and a mixed query that combines the sentence with and without its preceding context. Together with BM25 fusion, this improves Recall@25 by 8.0% (relative) without any fine-tuning. The authors explain that cited sentences often resemble each other more than the cited source, so citation captures more than semantic similarity. Testing other domains is planned.

3.4 Bibliographies and Knowledge Flows

TRACS (Grezes et al., 2026b) This paper documents the dataset behind the WASP 2025 shared task: 89,579 papers from the curated bibliographies of Chandra, HST, and JWST, labeled with a taxonomy developed with bibliographers (*science*, *instrumentation*, *mention*, *not_telescope*). Zero-shot baselines show a clear gap between a frontier model (telescope F1 up to 82.4) and smaller open models, and show that classifying *how* a telescope is used is harder than identifying it. JWST, the youngest telescope, is the hardest to recognize. The authors propose workflows in which models triage easy cases for bibliographers, and plan to add ground-based telescopes.

CARDS (Toney-Wails et al., 2026) How does knowledge flow from one research area to another? CARDS links NIH-funded publications to the clusters of a map of science and follows citations across clusters to build 4,826 research diffusion pathways. Each pathway is then summarized from the actual cited and citing papers rather than from cluster descriptions. Human annotators judged the CARDS summaries more specific in 70.7% of paired comparisons, and an LLM judge found no hallucinations. The case studies trace both basic-to-applied progressions (e.g., toward CAR-T therapies) and diffusion across fields. The authors plan to apply the method to studying the impact of research funding.

3.5 Figures and Claims

Sky-image visual question answering (Naiman, 2026) This paper introduces the first VQA dataset for astronomical sky images: 2,001 figures split between contour plots, real telescope images, and synthetic sky images that share the same axes, each also in a version that mimics scanned historical journals, with 47,357 questions. Low-cost multimodal models struggle with domain conventions: they often confuse contour plots with sky images, rarely recognize the synthetic images, and misread sexagesimal axis ticks when estimating angular sizes. A review of 104 published figures yields similar results. The author concludes that domain-specific VQA resources are needed and plans to cover additional figure types, such as spectra.

Claim preservation under AI reformulation (Centonze, 2026) When an AI system rewrites a scientific sentence, the result can be fluent and semantically similar, yet the claim can change. This paper proposes a Claim-Preservation Framework that describes a claim by its propositional content, epistemic force, evidential relation, argumentative function, and qualification and attribution. An exploratory application to SciFact claims shows how the framework catches subtle shifts, for example, “decreases risk” becoming “is associated with a decreased risk.” The author positions claim preservation as an evaluation dimension distinct from similarity and factuality, and outlines a benchmark with expert annotation as the next step.

4 Shared Task: AstroCLIMB

AstroCLIMB (Grezes et al., 2026a), Astronomy Citation Linking from Illustrations – a Multimodal Benchmark, asks systems to classify pairs of figures and captions from astronomy papers as same_figure, same_paper, related_papers, or unrelated_papers. The benchmark is built from open-access astronomy papers with more than 94,000 figures. The Kaggle competition used 10,000 training and 10,000 test pairs, scored with macro-F_1, and attracted 23 participants in 15 teams (136 submissions); ten system papers were accepted. The dedicated overview paper analyzes the task, results, and systems.

5 Cross-Cutting Themes

Looking across the papers, we see several shared directions.

Combining language models with deterministic code Deployed systems let language models interpret variable prose. At the same time, deterministic code enforces schemas, units, physical bounds, and identifiers (Schüssler et al., 2026; Susemiehl et al., 2026; Koblischke and Bovy, 2026; Kumar et al., 2026). Rovai (2026) takes this further, using a rule-based extractor and LLMs as validated auditors.

Keeping evidence attached to every value Extraction systems increasingly return the source text behind each value: excerpts, snippets, quotes, verbatim spans or source sentences with explicit attribution (Schüssler et al., 2026; Susemiehl et al., 2026; Koblischke and Bovy, 2026; Recski et al., 2026; Rovai, 2026). The authors see the next step in verifying this link automatically (Koblischke and Bovy, 2026; Kumar et al., 2026) and in checking that a source actually supports the claim it is attached to (Goyal, 2026; Centonze, 2026).

Lean configurations perform strongly Several papers find that focused, simpler systems perform as well as or better than more elaborate ones: direct prompting for meta-analysis extraction (Lu et al., 2026), a compact extractor for verbatim QA (Recski et al., 2026), plain chunks and general-purpose embedders for citation recommendation (Basseri et al., 2026), and a smaller keyphrase model (Pranjić et al., 2026).

Evaluation designed around scientific use Papers propose evaluations that reflect how outputs will be used: refitting a published meta-regression (Lu et al., 2026), counting the comparisons that records allow (Rovai, 2026), cross-checking against independent catalogs (Schüssler et al., 2026), record-level alignment (Bölücü et al., 2026) and word-level span overlap (Recski et al., 2026). Several authors also show that exact matching can undercount useful output (Pranjić et al., 2026; Naiman, 2026; Susemiehl et al., 2026).

Designing for the end user Systems are tuned for their users. Some favor precision and treat abstention as a valid answer because downstream systems find fabricated values hard to detect (Rovai, 2026; Recski et al., 2026). Others favor recall because curators review every output (Susemiehl et al., 2026). Several propose keeping experts in the loop through curator interfaces, triage, or assistance (Susemiehl et al., 2026; Grezes et al., 2026b; Lu et al., 2026).

What models know, and what remains hard Model behavior reflects the literature they were trained on: recent facilities are harder to recognize (Grezes et al., 2026b), and models favor frequently cited sources (Goyal, 2026; Basseri et al., 2026). Figures and numerical quantities remain the most challenging content (Naiman, 2026; Bölücü et al., 2026; Lu et al., 2026).

6 Outlook

The authors' future work points to several directions. **Verified grounding:** automatic checks that quotes are literal (Kumar et al., 2026), verifiers for the relevance and support of citations (Goyal, 2026), and claim-preservation benchmarks (Centonze, 2026). **Shared gold standards:** independently annotated corpora to measure coverage (Schüssler et al., 2026; Koblischke and Bovy, 2026) and larger expert annotations (Recski et al., 2026; Centonze, 2026), building on released resources such as STG, TRACS, LCRC, and audited extractions; Rovai (2026) proposes a new shared task. **Normalization and linking** of units and entities (Bölücü et al., 2026; Schüssler et al., 2026). **New domains**, telescopes and object types (Basseri et al., 2026; Grezes et al., 2026b; Koblischke and Bovy, 2026; Naiman, 2026).

These papers increasingly involve the people who maintain scientific records and focus on whether a record can be audited, traced, and used. WASP aims to keep this exchange between NLP

and scientific practice going.

7 Conclusion

WASP 2026 gathered 15 main-track papers, spanning deployed archive pipelines, benchmarks, retrieval and QA systems, and evaluation frameworks, and the `AstroCLIMB` shared task. Together they show a community building practical, evidence-keeping systems, designing evaluations around scientific use, and working closely with the experts who curate the literature.

Ethical Considerations

Some workshop organizers co-authored accepted main-track papers (Grezes et al., 2026b; Schüssler et al., 2026; Basseri et al., 2026). We did not assign these submissions to any co-authors for review; other program committee members reviewed them. In this overview, we summarize them the same way we summarize all other papers.

Acknowledgments

We thank all authors, reviewers, speakers, panelists, and shared-task participants.

References

Benjamin Basseri, Pavlos Protopapas, Kemeng Zhang, Alberto Accomazzi, Stephanie Jarmak, and Anna Kelbert. 2026. Local citation recommendation without graph signals. In *Proceedings of the Fourth Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Necva Bölücü, Huichen Yang, Roelien C. Timmer, Stephen Wan, Brian Jin, Changhyun Lee, Andreas Duenser, and Jessica Irons. 2026. Scientific table generation: Structured information extraction from scientific literature. In *Proceedings of the Fourth Workshop on Artificial Intelligence for Scientific Publications*, 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 Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 182–210, Vienna, Austria. Association for Computational Linguistics.

Laura Centonze. 2026. Beyond fluency: A framework for analyzing epistemic and argumentative shifts in AI-assisted scientific text reformulation. In *Proceedings of the Fourth Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

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.

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*, 15:1418.

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.

Tirthankar Ghosal, Sergi Blanco-Cuaresma, Alberto Accomazzi, Robert M. Patton, Felix Grezes, and Thomas Allen, editors. 2022. *Proceedings of the first Workshop on Information Extraction from Scientific Publications*. Association for Computational Linguistics, Online.

Tirthankar Ghosal, Philipp Mayr, Anita De Waard, Aakanksha Naik, Amanpreet Singh, Dayne Freitag, Georg Rehm, Sonja Schimmler, and Dan Li. 2025. Overview of the fifth workshop on scholarly document processing. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 1–6, Vienna, Austria. Association for Computational Linguistics.

Rushank Goyal. 2026. Towards verifiable answers: Reinforcement learning with a verifier reduces hallucinated references in literature-based question answering. In *Proceedings of the Fourth Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Felix Grezes, Thomas Allen, Tirthankar Ghosal, and Sergi Blanco-Cuaresma. 2023. Function of citation in astrophysics literature (FOCAL): Findings of the shared task. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 143–147, Bali, Indonesia. Association for Computational Linguistics.

Felix Grezes, Jennifer Lynn Bartlett, Kelly Lockhart, Alberto Accomazzi, Ethan Seefried, and Tirthankar Ghosal. 2025. Overview of TRACS: the telescope reference and astronomy categorization dataset & shared task. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Felix Grezes, Sergi Blanco-Cuaresma, Thomas Allen, and Tirthankar Ghosal. 2022. Overview of the first

shared task on detecting entities in the astrophysics literature (DEAL). In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 1–7, Online. Association for Computational Linguistics.

Felix Grezes, Kelly Lockhart, Alberto Accomazzi, Atilla Alkan, Ethan Seefried, and Tirthankar Ghosal. 2026a. Astroclimb: Astronomy citation linking from illustrations - a multimodal benchmark (overview of the wasp 2026 shared task). In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Felix Grezes, Kelly Lockhart, Alberto Accomazzi, Jennifer Lynn Bartlett, Ethan Seefried, Anjali Pandiri, and Tirthankar Ghosal. 2026b. TRACS: Telescope reference & astronomy categorization dataset and shared task. In *Proceedings of the Fourth Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Nolan Koblischke and Jo Bovy. 2026. Extracting published discussions of 13,616 galaxies with language models. In *Proceedings of the Fourth Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Gaurav Kumar, Anasse Bari, and David J. Nagel. 2026. Bridging the data bottleneck: A multi-modal generative extraction pipeline for legacy scientific literature. In *Proceedings of the Fourth Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Kelly Lockhart, Alberto Accomazzi, Felix Grezes, and Tirthankar Ghosal. 2025. Overview of the third workshop for artificial intelligence for scientific publications. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 1–4, Mumbai, India and virtual. Association for Computational Linguistics.

Zehao Lu, Xingguo Xiong, Wopke van der Werf, Thijs L. van der Plas, and Ioannis N. Athanasiadis. 2026. Can generative AI automate data extraction for meta-analysis? a case study on intercropping research. In *Proceedings of the Fourth Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

J. P. Naiman. 2026. Star light, star bright: Can multimodal models parse an astronomy visual question answering dataset right? In *Proceedings of the Fourth Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Marko Pranjić, Kristijan Armeni, and Senja Pollak. 2026. Fusing sentence-level and document-level representations for keyphrase extraction. In *Proceedings of the Fourth Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Gábor Recski, Szilveszter Tóth, Nadia Verdha, István Boros, and Ádám Kovács. 2026. ACL-Verbatim: Hallucination-free question answering for research. In *Proceedings of the Fourth Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

P. Reichherzer, F. Schüssler, V. Lefranc, A. Yusafzai, A. K. Alkan, H. Ashkar, and J. Becker Tjus. 2021. Astro-colibri—the coincidence library for real-time inquiry for multimessenger astrophysics. *The Astrophysical Journal Supplement Series*, 256(1):5.

Fabio Rovai. 2026. Recovering method parameters from protocol prose: Task definition, audited extraction, and the comparability it licenses. In *Proceedings of the Fourth Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Fabian Schüssler, S. Bisero, M. Cellier, J. Cinosi, A. Ciric, B. Cornejo, I. Jaroschewski, Atilla Kaan Alkan, W. Kiendrébéogo, and A. Saint-Paul. 2026. From astronomical circulars to auditable follow-up records: Schema-constrained LLM extraction with human verification. In *Proceedings of the Fourth Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Nicholas Susemiehl, Nima Chartab, Meca Lynn, and Ricky Nilsson. 2026. Accelerating literature data extraction using AI at the IPAC archives. In *Proceedings of the Fourth Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Autumn Toney-Wails, Katherine N. Quinn, and Steph Batalis. 2026. CARDS: Citation-aware research diffusion summarization. In *Proceedings of the Fourth Workshop on Artificial Intelligence for Scientific Publications*, Online. 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.

A Program Committee

We thank the members of the WASP 2026 Program Committee for their contributions to the review process and for their valuable feedback to the authors. The Program Committee members are listed below.

Antoine Gauquier — École Normale Supérieure, Paris

Alberto Accomazzi — Center for Astrophysics, Harvard & Smithsonian

Arne Binder — German Research Center for AI

Daniel E. Acuna — University of Colorado Boulder

Joyita Chakraborty — National Institute of Technology Durgapur

Kelly Lockhart — Center for Astrophysics, Harvard & Smithsonian

Elyon Eyimife — Stevens Institute of Technology

Ibrahim Al Azher — Northern Illinois University

Mengyang Liu — Shanghai Jiao Tong University

Harsh Sakhrani — University of Southern California

Shuntaro Yada — Tsukuba University; Tokyo Institute of Technology

Bhargab Choudhury — Gauhati University

Sharmila Upadhyaya — GESIS – Leibniz Institute

Tosho Hirasawa — LayerX

Sridevi Wagle — Pacific Northwest National Laboratory

Sujit Pal — Elsevier

Atilla Kaan Alkan — Center for Astrophysics, Harvard & Smithsonian

Rahul Nagurtha — Google

Tirthankar Ghosal — Oak Ridge National Laboratory

Sergey Feldman — Allen Institute for Artificial Intelligence

Neil R. Smalheiser — University of Illinois at Chicago

Pawin Taechoyotin — University of Colorado Boulder

Yagmur Ozturk — Université Grenoble Alpes

Wojtek Sylwestrzak — University of Warsaw

Taro Watanabe — Nara Institute of Science and Technology

Tim Schopf — National Institute of Informatics

Felix Grezes — Center for Astrophysics, Harvard & Smithsonian

Biswadip Mandal — Amazon

Ethan Seefried — Oak Ridge National Laboratory

Thomas S. Allen — Center for Astrophysics, Harvard & Smithsonian

Sharmin Afrose — Oak Ridge National Laboratory

Ran Elgedawy — University of Tennessee, Knoxville

Swati Rajwal — Emory University

Sandeep Kumar — Mohamed bin Zayed University of Artificial Intelligence

Sanjay Das — Oak Ridge National Laboratory