WASP 2026

The 4th Workshop on Artificial Intelligence for Scientific Publications

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications

November 10, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-456-9

Preface

We are excited to welcome you to WASP at AACL-IJCNLP 2026, the Fourth Workshop for Artificial intelligence for Scientific Publications. This year the workshop is being held online on November 10, 2026.

Building on the success of the previous Workshops for Artificial Intelligence for Scientific Publications at IJCNLP-AACL 2022, 2023, and 2025, the Fourth Workshop on Artificial intelligence for Scientific Publications (WASP) at AACL-IJCNLP 2026 aims to establish itself as a pivotal platform for promoting discussions and research in the field of Natural Language Processing (NLP) and Artificial Intelligence (AI). This gathering brings together esteemed experts and renowned organizations with students and early-career researchers who are interested and invested in efforts to extract and mine the world's scientific knowledge from research papers. Their collaboration focuses on developing advanced algorithms, models, and tools that lay the foundation for future machine comprehension of scientific literature. The fourth iteration of WASP specifically concentrates on various topics related to Artificial Intelligence research for/with scientific publications.

We especially welcome participation from academic and research institutions, government and industry labs, publishers, and information service providers. Projects and organizations using NLP/ML techniques in their text mining and enrichment efforts are also welcome to participate. We strongly encourage the participation of students, researchers, and science practitioners from diverse backgrounds.

WASP 2026 includes one shared task where we invite teams (individuals and groups) to tackle the problem of recreating the partial citation graph from scientific figures, with AstroCLIMB: Astronomy Citation Linking from Illustrations - a Multimodal Benchmark.

WASP 2026 received 43 submissions of which 25 were accepted (10 long papers, 5 short papers, and 10 shared task system papers). We are thankful to our program committee members for helping us curate a strong WASP 2026 program.

On behalf of the program co-chairs:
Tirthankar Ghosal, Alberto Accomazzi, Kelly Lockhart, Felix Grezes, and Atilla Kaan Alkan.

Organizing Committee

Program Chairs

Alberto Accomazzi, Center for Astrophysics | Harvard & Smithsonian
Tirthankar Ghosal, Oak Ridge National Laboratory
Felix Grezes, Center for Astrophysics | Harvard & Smithsonian
Kelly Lockhart, Center for Astrophysics | Harvard & Smithsonian
Atilla Kaan Alkan, Center for Astrophysics | Harvard & Smithsonian

Program Committee

Reviewers

Antoine Gauquier, Ecole Normale Supérieure, Ecole Normale Supérieure de Paris
Alberto Accomazzi, Center for Astrophysics | Harvard & Smithsonian
Arne Binder, German Research Center for AI
Daniel E. Acuna, University of Colorado at 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, Institute of Library, Information and Media Science, Tsukuba University, Tokyo Institute of Technology
Bhargab Choudhury, Department of Information Technology, 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 at Boulder
Yagmur Ozturk, Université Grenoble Alpes
Wojtek Sylwestrzak, University of Warsaw
Taro Watanabe, Nara Institute of Science and Technology, Japan
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

Invited Talk

Salience and Inquiry in Scientific Literature

Jessy Li
The University of Texas at Austin



Tuesday, November 10, 2026 –

Abstract: In this talk, we view the processing of scientific literature as an (implicitly) interactive discourse process: the reader is constantly abstracting out the most salient information, while staying inquisitive about the content. I will discuss the role of questions, first in deriving information salience from research documents. Using summarization as a behavioral probe, we reveal that across different lengths of summaries, the questions that models tend to answer remain intriguingly consistent. Next, we show how questions play important roles in connecting different pieces of a scientific document via generative discourse models, and extend existing frameworks to multimodal settings, modeling the connections across passages and figures. Finally, I will end the talk with brief discussions on AI review pilots at recent large conferences.

Bio: Jessy Li (https://jessyli.com/) is an associate professor in the Linguistics Department and (by courtesy) the Computer Science Department at the University of Texas at Austin. She received her Ph.D. from the Department of Computer and Information Science at the University of Pennsylvania in 2017. Her research interests are in computational linguistics and natural language processing, specifically discourse and pragmatics, generation and evaluation, and language & code. She is a recipient of an NSF CAREER Award, ACL and EMNLP Outstanding Paper Awards, an ACM SIGSOFT Distinguished Paper Award, among others. Jessy is an AI research lead at the NSF-Simons CosmicAI Institute. She is also the Secretary of NAACL, and a Program Chair of the EMNLP 2026 conference.

Invited Talk

From General-Purpose LLMs to Grounded Tools for Scientific Discovery

Pavlos Protopapas
Harvard University



Tuesday, November 10, 2026 –

Abstract: Large language models offer new ways to search, interpret, and synthesize scientific literature. Yet their responses can reflect biases in their training data, uneven coverage of research, and claims that are difficult to verify. How can we make these models useful while retaining control over the evidence and reasoning behind their outputs? Drawing on two recent research projects, I will discuss building applications on top of NASA's Astrophysics Data System (ADS) that connect language models to scholarly sources and guide their use through explicit workflows. The talk will explore how retrieval, tool use, and verification can help make outputs more transparent and assessable, while examining the limitations that remain. The central question is how to design AI tools that support scientific judgment and allow researchers to inspect and challenge their results.

Bio: Pavlos Protopapas is the Scientific Program Director for the master's programs in data science and computational science engineering at the Harvard John A. Paulson School of Engineering and Applied Sciences. As an educator, he currently teaches the CS109 course series for basic and advanced data science and a course in Machine Learning Operations at Harvard University. He previously taught the capstone courses for the master's programs, and also a course in deep reinforcement learning. His courses are also available through the Harvard Extension School and on HarvardX. In addition he has participated in data science schools around the world in Chile, Colombia, India, and Rwanda. Pavlos has a PhD in theoretical physics from the University of Pennsylvania and his research interests are at the intersection of astronomy, machine learning, and statistics.

Table of Contents

Overview of the Fourth Workshop on Artificial Intelligence for Scientific Publications (WASP 2026)
Atilla Kaan Alkan, Felix Grezes, Kelly Lockhart, Alberto Accomazzi and Tirthankar Ghosal . . . 1

AstroCLIMB: Astronomy Citation Linking from Illustrations - a Multimodal Benchmark and Overview of the WASP 2026 Shared Task
Felix Grezes, Atilla Kaan Alkan, Kelly Lockhart, Alberto Accomazzi, Ethan Seefried and Tirthankar Ghosal . . . 10

From Astronomical Circulars to Auditable Follow-up Records: Schema-Constrained LLM Extraction with Human Verification
Fabian Schüssler, Sofia Bisero, M. Cellier, J. Cinosi, A. Ciric, Bernardo Cornejo, Ilja Jaroschewski, Atilla Kaan Alkan, R. Weizmann Kiendrebeogo and A. Saint-Paul . . . 17

ACL-Verbatim: hallucination-free question answering for research
Gábor Recski, Szilveszter Tóth, Nadia Verdha, István Boros and Ádám Kovács . . . 23

Auditable Multi-View Hosted Multimodal Judging for AstroCLIMB
Kewen Zhu and Junyuan Tan . . . 34

Can Generative AI Automate Data Extraction for Meta-Analysis? A Case Study on Intercropping Research
Zehao Lu, Xingguo Xiong, Wopke van der Werf, Thijs L. van der Plas and Ioannis N. Athanasiadis 38

HCMUS_Beyonders at AstroCLIMB: Same Paper or Citing Paper? Linking Astronomy Figures with OCR, Citation Graphs, and LLMs
My-Linh Bang, Phu-Quy Nguyen-Lam, Duy-Minh Dao-Sy, Trung-Kiet Huynh, Phu-Hoa Pham and Chi-Nguyen Tran . . . 52

Local Citation Recommendation Without Graph Signals
Benjamin Basseri, Pavlos Protopapas, Kemeng Zhang, Alberto Accomazzi, Stephanie Jarmak and Anna Kelbert . . . 61

CARDS: Citation-Aware Research Diffusion Summarization
Autumn Toney-Wails, Katherine N. Quinn and Steph Batalis . . . 77

Bridging the Data Bottleneck: A Multi-Modal Generative Extraction Pipeline for Legacy Scientific Literature
Gaurav Kumar, Anasse Bari and David J. Nagel . . . 91

GRAIL: Compact Multimodal Similarities for Astronomy Citation Linking
Egor Fedosov . . . 99

Parameter-Efficient Adaptation of Qwen3-VL for Astronomy Citation Linking at AstroCLIMB 2026
Syed Mohaiminul Hoque . . . 104

Modality-Conditioned Ensembles and Zero-Shot Cross-Encoder Reranking for Relation Classification Between Astronomy-Paper Objects
M. Tahsan Zahid, Most. Atkia Farzana and Tonmoy Sarker Sourav . . . 112

Fusing Sentence-Level and Document-Level Representations for Keyphrase Extraction
Marko Pranjić, Kristijan Armeni and Senja Pollak . . . 122

Beyond Fluency: A Framework for Analyzing Epistemic and Argumentative Shifts in AI-Assisted Scientific Text Reformulation
Laura Centonze 133

TRACS: Telescope Reference & Astronomy Categorization Dataset and Shared Task
Felix Grezes, Kelly Lockhart, Alberto Accomazzi, Jennifer Lynn Bartlett, Ethan Seefried, Anjali Pandiri and Tirthankar Ghosal 142

Public Scientific-Figure Adaptation for AstroCLIMB with a 4B Vision–Language Model
Irving Ernesto Quezada Ramírez 154

Accelerating Literature Data Extraction using AI at the IPAC Archives
Nicholas Susemiehl, Nima Chartab, Meca Lynn and Ricky Nilsson 166

Extracting Published Discussions of 13,616 Galaxies with Language Models
Nolan Koblischke and Jo Bovy 174

Star Light, Star Bright: Can Multimodal Models Parse an Astronomy Visual Question Answering Dataset Right?
J. P. Naiman 180

PSGS@2026: From Figures to Citation Graphs: QLoRA Based Multimodal Linking for AstroCLIMB
Prajwal Ghimire, Shiv Ram Saud, Aashish Mahato, Sundeep Dawadi, Ramesh Saud, Sunil Regmi and Manoj Shakya 192

Towards Verifiable Answers: Reinforcement Learning with a Verifier Reduces Hallucinated References in Literature-Based Question Answering
Rushank Goyal 199

Beyond Visual Similarity: Parameter-Efficient Vision–Language Models for Scientific Citation Linking on AstroCLIMB
Ashmani Kumar, Shubham Sharma, K S Ramalakshmi and Ankush Chopra 210

Multimodal Feature Integration for Pairwise Relationship Classification in Astronomy Literatures
Raj Mahesh Mane, Ananya Kalia, Aashna Shetty, Rutal Mahajan and Preksha Pareek 216

Scientific Table Generation: Structured Information Extraction from Scientific Literature
Necva Bölücü, Huichen Yang, Roelien C. Timmer, Stephen Wan, Brian Jin, Changhyun Lee, Andreas Duenser and Jessica Irons 224

Recovering Method Parameters from Protocol Prose: Task Definition, Audited Extraction, and the Comparability It Licenses
Fabio Rovai 247

Modality-Specific and Relation-Aware Modeling for Scientific Figure Linking
Bodhisatta Maiti and Debshree Chowdhury 262

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

AstroCLIMB: Astronomy Citation Linking from Illustrations - a Multimodal Benchmark and Overview of the WASP 2026 Shared Task

Felix Grezes[1], **Atilla Kaan Alkan**[1], **Kelly Lockhart**[1], **Alberto Accomazzi**[1], **Ethan Seefried**[2] **and Tirthankar Ghosal**[2]
[1]Center for Astrophysics, Harvard & Smithsonian, USA
[2]Oak Ridge National Laboratory, USA
[1]{felix.grezes,atilla.alkan, kelly.lockhart,aaccomazzi}@sao.si.edu
[2]{seefriedej, ghosalt}@ornl.gov

Abstract

Scientists rely on figures to share their discoveries, but this makes the information contained in the figures difficult to parse, archive, and search. Multimodal neural-network models, designed to handle multiple types of data, promise to extract this information, but have not yet been widely tested and adopted by the astronomy community. In partnership with astroexplorer.org, we propose a novel dataset with over 100K figures and captions, and an associated task as a benchmark to evaluate a model's multimodal capabilities. The task is to reconstruct the citation graph of astronomy papers solely from their figures and captions. AstroCLIMB was the shared task for WASP 2026. It attracted 15 participants on the open competition platform Kaggle, of which 10 submitted system papers. Results show that zero-shot methods and general purpose metrics underperform compared to finetuned methods, indicating that scientific figures are still hard to understand for currently available multimodal models. The dataset, benchmark task, and associated code are publicly available online.

1 Introduction

The Science Explorer (SciX) (Accomazzi and SciX Team, 2026), formerly the Astrophysics Data System (Kurtz et al., 2000), is a digital library portal for researchers in astronomy, Earth science, heliophysics, physics, and planetary science, with the mission of facilitating and increasing research efficiency. Currently, SciX processes the textual information present in the papers it ingests. Author names, keywords, title and abstract, references, and many more are harvested, cleaned, categorized, and finally formatted into a rich data schema, allowing SciX users to search and filter for papers most useful to them.

However, missing from that harvest is the information present in the various figures in the papers. Unlike text which is linear in nature, these figures allow researchers to parse and absorb new scientific findings holistically. In astronomy, the image itself is often the valuable information, and any text attempting to describe it would be incomplete; for example a notable star formation in a specific location in the sky. Ideally, scientists could use natural language to describe figures to search and find papers of interest, even if the query doesn't strictly relate to the textual information in the caption or body of the papers. An example of query that is not currently possible is:

```
Find JWST papers with a plot showing a red to blue
heatmap and black contour lines.
```

Harvesting the information from figures is non-trivial: classical OCR systems can fail when the text is rotated at different angles over different color shades in the same figure; multiple plots are often combined into one figure; relevant information might be missing from the direct caption. Image 1 shows an example of these difficulties.

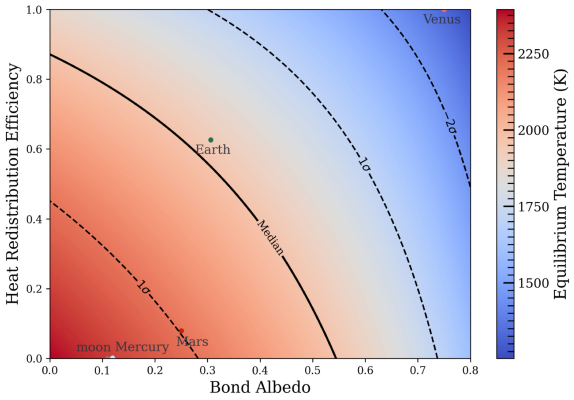


Figure 1: Sample astronomy image. To extract relevant information, a system needs to be able to read text at 5 different angles over continuous shades of blue and red, translate the XY and color coordinates of the Earth position to numbers, and understand that this is likely a JWST paper based on the context of the image and caption. Image courtesy of astroexplorer.org, original by Smith et al. (2026)

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 10–16
November 10, 2026 ©2026 Association for Computational Linguistics

Since the breakout of CLIP (Radford et al., 2021), neural network models that harness the technology of transformers to process multiple modes of information (text, image, audio, etc...) have proliferated. Nearly all frontier models offer multimodal variants, both closed (Gemini Team, 2025; Anthropic Team, 2026; OpenAI Team, 2025) and open source (Qwen Team, 2026; DeepSeek Team, 2026; Gemma Team, 2026) just to name a few. While these models are capable of embedding images and text in a shared vector space, and showcase that these embeddings preserve some of the shared meaning between images and associated text, they have not been rigorously tested on astronomy scientific figures.

An potential approximation for semantic understanding comes from the scientific citation graph, in which authors connect papers by their references. Similarly to Word2Vec (Mikolov et al., 2013) or more modern semantic preserving embedding methods, a thesis of this paper is that a model that can approximate the citation graph of the astronomy papers citation graph from its figures is a model that has some understanding of these.

To test this thesis, we contribute a novel dataset and benchmark: the Astronomy Citation Linking from Illustrations - a Multimodal Benchmark, or AstroCLIMB.

Contributions

- An open dataset consisting of over 100K figures from over 10K recent open access astronomy papers, along with their associated caption, links to citing and referring papers, and metadata, hosted publicly on Huggingface[1].
- A shared task platform to benchmark multimodal models, hosted publicly on Kaggle[2].
- A baseline analysis, evaluating off-the-shelf multimodal LLMs on the task.
- A description of the systems developed by participants to WASP@IJCNLP-AACL 2026[3]

2 Dataset Description

Since its launch in 2013, the Astronomy Image Explorer (AIE, astroexplorer.org) has archived hundreds of thousands of high-resolution images published in the open-access journals of the American Astronomical Society. Alongside the image, metadata containing information about author, journal, citation, format, and keywords is also stored. In partnership with the AIE, we harvested images and associated information for years ranging from 2022 to 2025. After filtering for images that may not be open-access, the final full dataset size is 102629 pairs of image and captions, coming from 12671 unique papers, of which 10843 have citation or reference links to other papers in the dataset. Figure 2 shows the distributions of reference and citation count per paper.

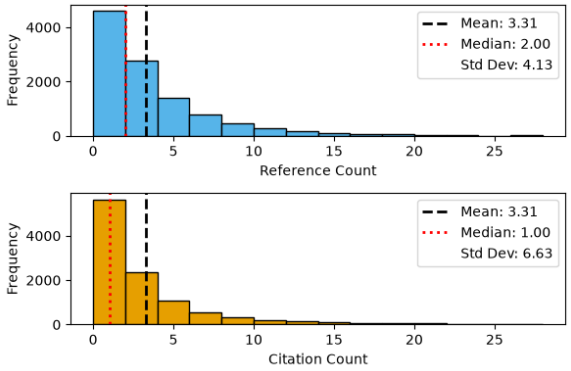


Figure 2: Distributions of reference and citation count for the papers within AstroCLIMB.

The dataset was split into 94233 rows for training and 8396 for validation. It should be noted that because AstroCLIMB is a benchmark rather than an inherently useful task, the 'training' mentioned here is about simple statistics computations rather than fully finetuning model weights.

2.1 Technical Details

The full dataset splits are Huggingface `Dataset` objects. The training split look as follows:

```
Dataset({
    features: ['image',
               'UUID',
               'Image ID',
               'Paper DOI',
               'Paper Title',
               'Image Caption',
               'Image Authors',
               'References DOIs',
               'Citing DOIs'],
    num_rows: 94233
})
```

The first row provides an example:

```
train_dataset[0] == {
'image':    <PIL.PngImagePlugin.PngImageFile
            image mode=RGB size=1000x1066>,
'UUID':     '10.3847.1538-4357.ac983d_apj[...]',
'Image ID': 'apjac983df5',
'Paper DOI': '10.3847/1538-4357/ac983d',
```

[1] huggingface.co/datasets/adsabs/AstroCLIMB
[2] kaggle.com/competitions/astroclimb
[3] ui.adsabs.harvard.edu/WIESP/2026/

```
'Paper Title': 'The Radius of PSR J0740+6[...]',
'Image Caption': 'Effect of the SW backgr[...]',
'Image Authors': 'Authors: Tuomo Salmi, S[...]',
'References DOIs': ['10.3847/1538-3881/ac4ae6'],
'Citing DOIs': ['10.3847/2041-8213/ace5b2',
    '10.3847/2041-8213/ad5beb',
    '10.3847/1538-4357/ae145d',
    '10.3847/1538-4357/ad81d2',
    '10.3847/1538-4357/acf9a0',
    '10.3847/1538-4357/ad7255',
    '10.3847/2041-8213/ad9f3c',
    '10.3847/2041-8213/ad5f02',
    '10.3847/1538-4357/acf49d',
    '10.3847/2041-8213/ad5a6f']}
```


Going through each of these:

- `'image': <PIL.PngImagePlugin.PngImageFile image mode=RGB size=1000x1066` is the image data for this figure.
- `Image Caption` is the other important feature of this row. It is the text caption associated with this figure.
- `References DOIs` and `Citing DOIs` are lists that link to other DOIs which are either referenced by or cite the current row's DOI. Note that are only DOIs listed elsewhere in the dataset, not every reference and citation in the full paper.
- `'UUID': '10.3847.1538-4357.ac983d_apjac983df5_8'` serves as a unique identifier for the row.
- `'Paper DOI': '10.3847/1538-4357/ac983d'` is the Digital Object Identifier (DOI) of the paper from which the figure was extracted.
- `'Image ID': 'apjac983df5'` is an identifier to distinguish from other figures in the same paper.
- `Paper`, `Title`, and `Image Authors` provide potentially useful metadata.

3 Shared Task Description

The goal of the shared task for WASP 2026 is multi-class, single-label classification. The inputs are pairs of objects. It could be two figures, a figure and a caption, or two captions. The possible outputs are:

- `same_paper` if the two objects originated from the same scientific paper (same DOI).
- `related_papers` if the two objects originated from related papers (one object's DOI cites the DOIs of the second object).
- `same_figure` if the two objects are actually from the same scientific figure (only for figure+caption pairs).
- `unrelated_papers` if the two objects originated from unrelated papers (none of the previous cases).

Input pairs only have one label and the relations between objects is symmetrical. This dataset is intended to be more of a benchmark for multimodal models, rather than a meaningful task in itself. As such, participants were not necessarily expected to fully train large models solely for AstroCLIMB. Basic statistics or incorporation into a larger multimodal training process seemed more appropriate.

From the `References DOIs` and `Citing DOIs` in the Huggingface dataset columns we can build the version of the dataset used on Kaggle for automatic benchmarking. However with over 100K rows in the full dataset, there are over 5B possible pairs of objects. The version of the dataset on Kaggle contains a small, balanced sample of those pairs. Both `train.csv` and `test.csv` contain 10K rows with 1K sample for each combination possible of image+image, image+caption, caption+caption input features with `same_figure`, `same_paper`, `related_papers` and `unrelated_papers` output labels. Submissions were evaluated using macro-F1 score over the four target classes.

4 Baseline Experiments

To establish reference performance for AstroCLIMB, we evaluated two open-weight vision-language models (VLMs): Llama-3.2-11B-Vision-Instruct (11B parameters) (Grattafiori et al., 2024) and Gemma-4-31B-it (31B parameters) (Gemma Team, 2026). Both models were evaluated zero-shot, without task-specific fine-tuning or in-context examples. These experiments provide a reference for assessing how general-purpose multimodal models perform on scientific figure and citation relationship classification.

4.1 Experimental Setup

Each pair was presented through a chat-formatted prompt defining the four relationship classes: `same_figure`, `same_paper`, `related_papers`, and `unrelated_papers`. Objects were provided in their original modalities, either as images or text,

Metric	Llama 3.2 11B	Gemma 4 31B
Overall macro-F1	0.2145	**0.4890**
Per-class F1		
Same figure	0.11	**0.87**
Same paper	0.42	**0.63**
Related papers	0.32	**0.46**
Unrelated papers	0.01	0.00
Modality-specific macro-F1		
Text + image	0.1662	**0.4212**
Image + image	0.1463	**0.3765**
Text + text	0.3426	**0.3891**

Table 1: Zero-shot VLM performance on AstroCLIMB. Overall macro-F1 is reported from the official Kaggle public leaderboard. Modality-specific macro-F1 is computed over structurally valid classes for each pair type: four classes for text-image pairs and three for image-image and text-text pairs.

followed by a `Label:` cue. The prompt explicitly specified that `same_figure` is only applicable to an image and its corresponding caption.

Both models were loaded using Hugging Face Transformers and evaluated in bfloat16 precision on NVIDIA A100 80GB GPUs. Inference was performed independently for each pair using greedy decoding with a maximum of 10 generated tokens. The first recognized class label was taken as the prediction, defaulting to `unrelated_papers` if no valid label was generated. No calibration, class balancing, or task-specific post-processing was applied.

We report macro-averaged F1 using the official Kaggle public leaderboard, along with locally computed class-level and modality-specific metrics.

4.2 Baseline Performance

Table 1 presents the results of both zero-shot baselines. Gemma substantially outperformed Llama, achieving a macro-F1 of 0.4890 compared with 0.2145. However, performance varied considerably across relationship categories and input modalities.

Gemma performed particularly well on direct figure-caption correspondence, achieving an F1 score of 0.87 for `same_figure`, with recall of 1.00 and precision of 0.77. In contrast, Llama achieved an F1 of only 0.11 for this category. Both models performed less effectively when distinguishing relationships between different scientific papers.

Modality-specific results reveal additional differences. Llama performed substantially better on text-text pairs (macro-F1 0.3426) than image-image pairs (0.1463). Gemma exhibited a smaller difference between these configurations, achieving 0.3891 and 0.3765, respectively. These results suggest that the models differ not only in overall classification performance but also in their handling of visual inputs.

Model	Free generation	Forced choice
Llama 3.2 11B	0.5%	100.0%
Gemma 4 31B	0.0%	68.9%

Table 2: Percentage of predictions assigned to `unrelated_papers` under greedy generation (full test set, $n = 10{,}000$) and forced-choice label scoring (stratified sample, $n = 450$). The two columns are descriptive and are not directly matched-sample comparisons.

We additionally examined predictions that violated the task's structural constraints. Although `same_figure` is only valid for image-caption pairs, Llama assigned this label to 1,064 of 3,000 image-image pairs (35.5%) and 151 of 3,000 text-text pairs (5.0%). Gemma made the same error on only 2 image-image pairs (0.07%) and 10 text-text pairs (0.3%). This substantial difference indicates that Gemma more reliably follows the modality-dependent definition of direct figure-caption correspondence.

4.3 Analysis of Prediction Behavior

Despite their differences in overall performance, both models exhibited a striking tendency to avoid predicting `unrelated_papers` during greedy generation. Although this category accounts for 30% of the evaluation set, Llama predicted it only 48 times out of 10,000 pairs, while Gemma never predicted it. Instead, Llama predominantly predicted `same_paper` (5,971 predictions), whereas Gemma favored `related_papers` (5,879 predictions).

To investigate whether this behavior was specific to the generation procedure, we conducted an additional forced-choice diagnostic on a stratified sample of 450 test pairs, containing 150 examples of each input configuration. Rather than generating labels freely, we computed the teacher-forced, length-normalized log-likelihood of each candidate label as a continuation of the original prompt. The label with the highest average token log-likelihood was selected.

The forced-choice results differ dramatically from free generation. Llama selected `unrelated_papers` for all 450 sampled pairs, while Gemma selected it for 310 pairs (68.9%). Neither model selected `related_papers` under

forced-choice scoring in this sample, despite its frequent occurrence in free-generation predictions.

This reversal indicates that the observed absence of `unrelated_papers` predictions is strongly dependent on the output selection procedure, rather than demonstrating that the models cannot assign likelihood to this relationship. However, forced-choice scoring introduces its own biases, and a high likelihood for a particular label does not necessarily indicate correct relational reasoning. Further experiments would be needed to determine whether calibrated or constrained prediction improves classification accuracy.

Overall, these results highlight two challenges for general-purpose VLMs on AstroCLIMB. First, the models differ substantially in their ability to respect modality-dependent relationship definitions. Second, their predicted relationship distributions are highly sensitive to how candidate labels are selected, suggesting that evaluation protocols and output calibration are important considerations when using generative VLMs for scientific relationship classification.

5 Participant Systems

The AstroCLIMB shared task attracted 15 participating teams on Kaggle, of which 10 submitted system papers to WASP 2026.

- Zhu and Tan (2026) test a hosted multimodal judge under a fixed, auditable multi-view protocol. The experiments show that this constrained approach substantially outperforms a content-only lexical and image-statistics baseline on local validation.

- Bang et al. (2026) combine text-centered OCR representations, citation-graph pretraining, and fine-tuned large language models. Their approach demonstrates that ensembling complementary architectures overcomes fine-grained confusions between same-paper and citation-linked pairs.

- Fedosov (2026) evaluates a factorized gradient-boosted classifier that merges frozen semantic embeddings with explicit visual and textual features. The results reveal that rich symmetric pair features provide significant performance gains over raw cosine similarities, whereas nested model selection yields no aggregate benefit.

- Hoque (2026) applies parameter-efficient adapters and probability ensembling to a vision-language model. The author finds that model averaging and validity constraints improve test performance, while broader visual adaptation and auxiliary training provide no further improvements.

- Sourav et al. (2026) evaluate modality-conditioned ensembles and zero-shot cross-encoder reranking. The experiments show that representation quality serves as the primary performance driver, whereas extensive feature engineering and distant supervision yield minimal gains.

- Quezada Ramírez (2026) evaluates public scientific-figure adaptation for a vision-language model. The author shows that pretraining on unlabeled public pairs improves downstream validation performance, whereas alternative generative and reinforcement-learning setups yield no additional benefit.

- Ghimire et al. (2026) evaluate parameter-efficient multimodal fine-tuning for vision-language models. The experiments show that low-rank adaptation substantially improves performance on challenging citation-relation classes, while demonstrating that unadapted few-shot reasoning yields no additional gains.

- Ramalakshmi et al. (2026) test parameter-efficient vision-language models across diverse modalities. Results show that fine-tuned generative backbone models achieve strong performance, while highlighting that textual semantics provide more reliable citation signals than visual features alone.

- Mane et al. (2026) build a multimodal feature-based pipeline combining textual, visual, semantic, and domain-oriented signals with a gradient-boosted classifier. The experiments show that complementary representations provide useful predictive features, while demonstrating that distinguishing related papers from other relationship types remains a persistent challenge.

- Maiti and Chowdhury (2026) evaluate a modality-specialized and relation-aware pipeline combining custom encoders with decomposed correspondence gates. They show

Team / Model	Macro-F1
Quezada Ramírez (2026)	0.75
Bang et al. (2026)	0.74
Ramalakshmi et al. (2026)	0.74
Hoque (2026)	0.74
Sourav et al. (2026)	0.69
Ghimire et al. (2026)	0.64
Maiti and Chowdhury (2026)	0.61
Gemma-4-31B-it (Zero-shot)	0.49
Fedosov (2026)	0.48
Mane et al. (2026)	0.48
Zhu and Tan (2026)	0.40
Random Baseline	0.25
Llama-3.2-11B-Vision (Zero-shot)	0.21

Table 3: AstroCLIMB shared-task leaderboard results and zero-shot VLM baselines, reported using macro-F1. Scores are rounded to two decimal places.

that separating exact match prediction from residual paper relations substantially improves performance, while identifying cross-paper citation links remains a persistent challenge.

We report the results of the participating teams in table 3.

6 Conclusion and Future Directions

In this work, we introduced AstroCLIMB, a novel multimodal dataset and benchmark task designed to evaluate the capability of neural network models to parse and understand scientific figures within astronomy literature. By framing the task around reconstructing citation and reference relationships from figures and captions, we provided a rigorous proxy for semantic comprehension of visual data in digital library portals like the Science Explorer (SciX).

Our baseline evaluations with off-the-shelf vision-language models—such as Llama-3.2-11B-Vision-Instruct and Gemma-4-31B-it—demonstrated that zero-shot general-purpose models struggle significantly with scientific illustrations, exhibiting sensitivity to output selection procedures and limitations in respecting modality-dependent relationship definitions. However, the response of the research community through the WASP 2026 shared task highlights the viability of addressing these challenges. With 15 participating teams on Kaggle and 10 submitted system papers, participants showed that strategies such as domain-adapted fine-tuning, parameter-efficient adapters (e.g., QLoRA), modality-conditioned ensembling, and factorized gradient-boosted classifiers can substantially boost performance, raising top macro-F1 scores to 0.75.

Despite these advances, several challenges remain. Participant systems and baseline analyses consistently identified that distinguishing cross-paper citation links and fine-grained relational categories (such as related papers versus same paper) remains a persistent hurdle. We hope that future works will use AstroCLIMB to build systems that enable natural language visual search and intelligent artifact discovery for digital libraries.

7 Ethics Statement

The authors of this paper follow principles of transparency and reproducibility. The dataset and code described are publicly available and open source, ensuring accessibility for verification and future research. Large language models were employed solely as baseline comparisons in our experiments, in a non-generative mode only, as classifiers. We acknowledge that LLMs may carry inherent biases present in their training data, and we have taken care to document these limitations in our analysis. The use of LLMs as baselines does not constitute endorsement of their outputs, but rather provides a standardized benchmark for evaluating our proposed methods. We are committed to responsible research practices and have considered the potential societal impacts of this work throughout the research process.

References

Alberto Accomazzi and SciX Team. 2026. The Science Explorer: An Open Science Discovery Platform in the Earth and Space Sciences. In *46th COSPAR Scientific Assembly. Held 1-9 August*, volume 46, page 3448.

Anthropic Team. 2026. The claude model family: Opus, sonnet, haiku. anthropic.com/system-cards. Model Card and Technical Specifications, archived URL: archive.ph/sVkGb.

My-Linh Bang, Phu-Quy Nguyen-Lam, Duy-Minh Dao-Sy, Trung-Kiet Huynh, Phu-Hoa Pham, and Chi-Nguyen Tran. 2026. HCMUS_beyonders at AstroCLIMB: Same paper or citing paper? linking astronomy figures with OCR, citation graphs, and LLMs. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

DeepSeek Team. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv e-prints*, arXiv:2606.19348. huggingface.co/deepseek-ai.

Egor Fedosov. 2026. GRAIL: Compact multimodal similarities for astronomy citation linking. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Gemini Team. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv e-prints*, arXiv:2507.06261.

Gemma Team. 2026. Gemma 4 technical report. *arXiv e-prints*, arXiv:2607.02770. huggingface.co/google.

Prajwal Ghimire, Shiv Ram Saud, Aashish Mahato, Sundeep Dawadi, Sunil Regmi, and Manoj Shakya. 2026. PSGS@2026: From figures to citation graphs: QLoRA based multimodal linking for AstroCLIMB. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Syed Mohaiminul Hoque. 2026. Parameter-efficient adaptation of Qwen3-VL for astronomy citation linking. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Michael J. Kurtz, Guenther Eichhorn, Alberto Accomazzi, Carolyn S. Grant, Stephen S. Murray, and Joyce M. Watson. 2000. The NASA Astrophysics Data System: Overview. , 143:41–59.

Bodhisatta Maiti and Debshree Chowdhury. 2026. Modality-specific and relation-aware modeling for scientific figure linking. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Raj Mane, Ananya Kalia, Aashna Shetty, Rutal Mahajan, and Preksha Pareek. 2026. Multimodal relationship classification of astronomy figures and captions. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv e-prints*, arXiv:1301.3781.

OpenAI Team. 2025. OpenAI GPT-5 System Card. *arXiv e-prints*, arXiv:2601.03267.

Irving Ernesto Quezada Ramírez. 2026. Public scientific-figure adaptation for AstroCLIMB with a 4B vision-language model. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Qwen Team. 2026. Qwen-image-2.0-rl technical report. *arXiv e-prints*, arXiv:2606.27608. huggingface.co/Qwen.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv e-prints*, arXiv:2103.00020.

K S Ramalakshmi, Shubham Sharma, Ashmani Kumar, and Ankush Chopra. 2026. Beyond visual similarity: Parameter-efficient vision–language models for scientific citation linking. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Cole Smith, Brandon Park Coy, Megan Weiner Mansfield, Rafael Luque, Anjali A. A. Piette, Qiao Xue, Michael Zhang, Jacob L. Bean, Madison Brady, Peter Gao, and et al. 2026. JWST MIRI reveals a potential atmosphere on the ultra-hot rocky planet TOI-431b. *arXiv e-prints*, arXiv:2609.03029.

Tonmoy Sarker Sourav, M. Tahsan Zahid, and Most. Atkia Farzana. 2026. Modality-conditioned ensembles and zero-shot cross-encoder reranking for relation classification between astronomy-paper objects. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Kewen Zhu and Junyuan Tan. 2026. Auditable multi-view hosted multimodal judging for AstroCLIMB. In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

From Astronomical Circulars to Auditable Follow-up Records: Schema-Constrained LLM Extraction with Human Verification

Fabian Schüssler[1] **S. Bisero**[1] **M. Cellier**[1,2] **J. Cinosi**[1] **A. Ciric**[1] **B. Cornejo**[1]
I. Jaroschewski[1] **A. Kaan Alkan**[3] **W. Kiendrébéogo**[1] **A. Saint-Paul**[1,2]

[1]Université Paris-Saclay / CEA / IRFU, F-91191 Gif-sur-Yvette, France
[2]École d'ingénieurs aéronautique et spatiale (IPSA), Paris, France
[3]Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA
astro.colibri@gmail.com

Abstract

GCN Circulars distribute crucial information on astronomical observations of transient phenomena, but their flexible prose mixes observations, limits, metadata, contacts, and interpretation. We present an auditable pipeline that converts these reports into event-linked, structured records. Regular expressions provide domain hints; a large language model returns strict schema-constrained JSON; deterministic stages normalize values and emit review flags. Seven domain reviewers assessed 231 outputs from 210 distinct Circulars, fully confirming 223 reports and marking eight partly correct. They confirmed 25,827 of 25,880 definite field decisions (99.80%); errors concentrated in timing and facility attribution. Comparison with independent redshift datasets found semantic catalog disagreements but no values misread by our pipeline. Applied to the GCN Circular archive back to 2016, the pipeline has structured 26,811 reports into 68,393 observations, now served in real time by the Astro-COLIBRI platform. We release the extraction pipeline as an open-source package and the derived records as a continuously updated resource.

1 Introduction

Gamma-ray bursts (GRBs) and other transients fade within minutes to days, requiring rapid coordination across worldwide observatories. Many results appear in GCN Circulars, a public rapid-publication channel for time-domain astronomy (Barthelmy, 1998; NASA Goddard Space Flight Center, 2026). A Circular is a natural language report that can mix detections, limits, times, exposures, filters, facilities, redshifts, and contacts in prose or compact tables, making them difficult to aggregate automatically.

Extraction requires more than entity recognition: a number may be an afterglow magnitude, a limit, or calibration-star photometry; a time may be absolute, trigger-relative, or an exposure midpoint. Planned observations and candidate-host redshifts must also be distinguished from completed measurements and spectroscopic redshift determinations.

This paper studies the Circular-to-record component deployed in Astro-COLIBRI (Reichherzer et al., 2021, 2023), whose main flow is shown in Figure 1. Our contributions are:

1. a typed, observation-level extraction task for time-domain reports, each measurement emitted with its time, exposure, band, value, detection status, facility, and a mandatory supporting excerpt (Section 3);
2. a hybrid architecture surrounding a schema-constrained model call with deterministic pre-analysis and post-validation, running in production (Section 4);
3. a field-level audit of the deployed system and an external redshift cross-check against three catalogs, reported with their limits (Section 5); and
4. an open-source release of the pipeline and the derived archive of >68k observations (Section 6).

2 Related Work

Astronomical text. TDAC introduced an annotated time-domain corpus and named-entity-recognition baselines (Alkan et al., 2022); astroECR added coreference, astrophysical relations, and object normalization (Alkan et al., 2024). General-purpose LLM studies found strong prompt sensitivity in multimessenger extraction (Sotnikov and Chaikova, 2023). Closest to this work, Sharma et al. (2026) apply LLMs to the GCN archive for topic classification and redshift recovery. We differ in three ways: the extraction runs on new Circulars

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 17–22
November 10, 2026 ©2026 Association for Computational Linguistics

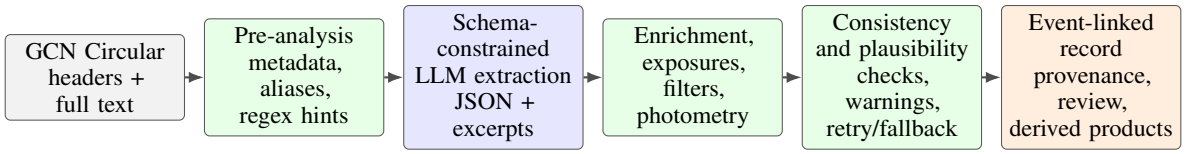


Figure 1: Main architecture. Deterministic stages surround the LLM call, preserve the source and extracted excerpts, and surface ambiguity as warnings. Schema or plausibility failures trigger retry or fallback before an event-linked record is emitted.

in real time and its output is linked to a live event record rather than analyzed offline; the schema is closed and every observation carries a supporting excerpt and full provenance metadata, so each value is traceable to the text that produced it; and we report a field-level human audit of the deployed system. Redshift is where the two efforts touch, and Section 5.3 evaluates ours against three external catalogs.

Scientific information extraction. Document-level scientific IE has long targeted typed records rather than isolated mentions, as in SciERC (Luan et al., 2018) and SciREX (Jain et al., 2020). Dagdelen et al. (2024) showed that LLMs can emit structured JSON records of scientific knowledge directly, which is the template we follow. Our setting differs in that records are consumed within minutes of publication, so an unparseable or implausible output is an ingestion failure rather than a scoring loss.

Constrained generation. PICARD constrains auto-regressive decoding for semantic parsing (Scholak et al., 2021), and grammar-constrained decoding generalizes this to structured tasks without finetuning (Geng et al., 2023). Constraint is not free: Tam et al. (2024) report that format restrictions can degrade reasoning. We accept that trade-off and offset it by keeping the model's job narrow: deterministic code, not the model, performs arithmetic, unit conversion, and cross-field checking (Section 4.3). The mandatory verbatim excerpt makes every field checkable against its source, in the spirit of attribution evaluation for generated text (Rashkin et al., 2023).

3 Task Definition

Given the full text and headers of a Circular, the task is to emit a single closed-schema record. The record contains astronomical source names, a report classification, contacts, an array of observations, and warnings. Report fields cover authorship, facility, instrument, detection, and redshift evidence. Each observation is one measurement or limit and carries time, exposure, band, magnitude or flux, uncertainty, limit status, calibration, notes, and a mandatory source excerpt.

Two properties distinguish this from span annotation. First, the schema is *closed*: additional properties are rejected, so a value that does not fit the record cannot be smuggled into a free-text field, and any output that validates is directly ingestible. Second, the unit of extraction is the observation rather than the mention: a report of six filters yields six rows, each independently timed, calibrated and attributed, and a row is only usable if every one of those fields is correct. Figure 2 shows a Circular fragment and one row it produces.

4 Hybrid Extraction Pipeline

4.1 Advisory pre-analysis

Regular expressions over the subject, headers, and body propose event aliases and locate optical/NIR terms, magnitude-like expressions, redshifts, and contacts. These are advisory prompt hints, never accepted as observations without checking the full text. The stage also normalizes contacts and preserves the unmodified source.

4.2 Schema-constrained model call

One request contains the identifier, subject, hints, and full text; the response must satisfy the schema. The deployment uses `gpt-5.4-mini` with low reasoning effort and `gpt-5.5` as fallback. Transient failures receive two additional attempts per model, while schema or plausibility failures can trigger fallback directly. Model settings are recorded. Fallback was used for 94 of 26,811 analysed reports (0.35%).

Domain rules exclude planned observations and comparison-, calibration-, or host-galaxy photometry. Multi-filter results become separate rows; tables are read by column first, then by row. Repeated exposures retain their notation while count, per-frame duration, and total are normalized. Non-

Input: GCN Circular 40471 (excerpt)

```
The Swift/UVOT began settled observations of
the field of GRB 250516A 108 s after the BAT
trigger (Beardmore et al., GCN Circ. 40467).

Filter    T_start(s) T_stop(s) Exp(s)   Mag
white          108      1715     411  20.5+/-0.3
v              650      1765     136      >18.5
b              576      1691     117      >19.2
u              320      1838     362      >19.7
w1             699      1814     117      >18.8
w2             626      1741      97      >19.1
```

Output: one element of `observations`

```
{"observation_type": "detection",
 "filter_raw": "white",
 "mag_raw": 20.5, "mag_err": 0.3,
 "is_upper_limit": false,
 "observatory": "Swift", "instrument": "UVOT",
 "observatory_short_name": "UVOT",     <-- wrong
 "telescope": "Swift",                 <-- wrong
 "time_raw": "108 s after the BAT trigger",
 "exposure_s": 411,
 "calibration": "Breeveld et al. 2011, ...",
 "source_excerpt": "white  108  1715  411
                    20.5 +/- 0.3"}
```

Figure 2: A Circular fragment containing the complete six-row table and one record it produces. Each table row becomes an independently attributed observation. `source_excerpt` carries the verbatim text supporting the shown row, so every value can be checked against its origin. This example is also one of the eight partly correct reports of Section 5.2: the model identified the facility correctly in `observatory` and `instrument`, but filled the redundant `observatory_short_name` and `telescope` fields with the two values transposed, an error that repeats across all six observations.

optical observations are flagged so their fluxes cannot populate magnitude fields, and every observation quotes its evidence.

4.3 Deterministic enrichment and review flags

Post-processing parses absolute, Modified Julian Date, and trigger-relative times; normalizes filters, facilities, and exposures; and derives observation midpoints from explicit values, intervals, durations, or plausible exposures in that order. Approximations remain flagged. Optical measurements are mapped, when possible, to an observer-frame R_C-equivalent AB display scale while raw values are retained.

Checks flag missing values or times, conflicting representations, implausible dates, pre-trigger observations, and unknown filters. Unresolved candidates and warnings are returned rather than dropped: the LLM handles linguistic variation, while code enforces numerical and interface invariants.

5 Evaluation

5.1 Corpora and protocol

The pipeline was developed on 4,537 Circulars from 2025, with manual inspection of roughly 30 random reports. All 1,775 Circulars published from 2026 January 1 through June 30 completed the held-out operational workflow after configured retries. This measures completion, not accuracy.

Seven reviewers with astronomy and development expertise completed 231 assessments of 210 distinct Circulars from 2020–2026: 115 recent and 95 randomly selected earlier reports, with twenty independently re-reviewed. They compared all extracted fields with the original text, assigned report- and field-level verdicts, and recorded corrections and omissions. Review neither reran the parser nor changed production data.

Field group	Rejected	Reports	Uncertain
Observation time	28	5	7
Facility/instrument	18	4	0
Contact information	4	1	0
Classification	2	1	0
Photometry/limits	1	1	0
Total	53	8	7

Table 1: Rejected and uncertain field verdicts. Content/provenance, filter/calibration, exposure, and redshift had none. "Reports" counts assessments containing a rejection.

5.2 Human-verification results

Reviewers confirmed 223 reports (96.5%), marked eight partly correct, and rejected none. Of 25,887 field verdicts, 25,827 were confirmed, 53 rejected, and seven uncertain: 99.80% of definite decisions were confirmed. One observation was missing.

Timing and facility attribution dominate (Table 1). All 53 rejections occur in the eight partly correct reports, and a single report accounts for 24 of them. Field verdicts are therefore non-independent, making eight of 231 reports the conservative error summary. The twenty re-reviewed reports had unanimous labels and 2,105/2,105 agreeing overlapping field decisions, but contained no rejected fields and thus do not test difficult cases.

That worst case is the example of Figure 2, and it is instructive: its 24 rejections are not 24 independent mistakes but two, each repeated down six rows. The model resolved the facility correctly in `observatory` and `instrument`, then

filled the redundant `observatory_short_name` and `telescope` with the values transposed; and it offset the reported start time instead of taking the midpoint of the reported start–stop interval. A closed schema carrying several near-synonymous fields for one concept invites the first, and our validation checks never compared those fields against each other. Both are cheap to defend against (cross-validate redundant fields, and prefer the interval midpoint when both bounds are present).

5.3 External redshift cross-check

A crucial parameter of astrophysical observations is the redshift (or distance) of the source. Of 468 events with extracted redshifts, 249 appear in GRBweb (Coppin, 2026); 228 (91.6%) agree within $\Delta z = 0.01$, with median absolute difference zero. All 21 residuals were read correctly: nine were limits or conditional associations, seven candidate-host photometric estimates, and five preliminary values. The three reference catalogs themselves agree pairwise for only 96–97% of shared events (Greiner, 2026; NASA/GSFC Swift Science Center, 2026), and our extraction agrees with Greiner for 201/205 events and with *Swift* for 138/139. The comparison supports faithful reading, but shows the schema should distinguish measurements, limits, candidate hosts, conditional associations, and superseded values.

6 Deployment, Scale and Released Artifacts

The pipeline runs on arriving Circulars in real time and has additionally been applied to the archive back to 2016, associating 26,811 reports with 5,787 events, extracting 68,393 observations from 618 facilities, and recovering 84,833 contact records (snapshot: 2026 July 30). Upper limits outnumber detections roughly two to one (42,926 against 20,990), a property of the resource rather than a defect: the non-detections that constrain a fading afterglow are recorded on the same footing as the detections. Of the optical and NIR measurements, 47,515/50,015 (95.0%) reached the normalized plotting scale.

Availability. We release `astro-colibri-circular-parser` v1.0.0 under the BSD 3-Clause license, with a versioned Zenodo archive.[1]

[1] `https://github.com/astro-transients/astro_colibri_circular_parser`; `https://doi.org/10.5281/zenodo.21887735`.

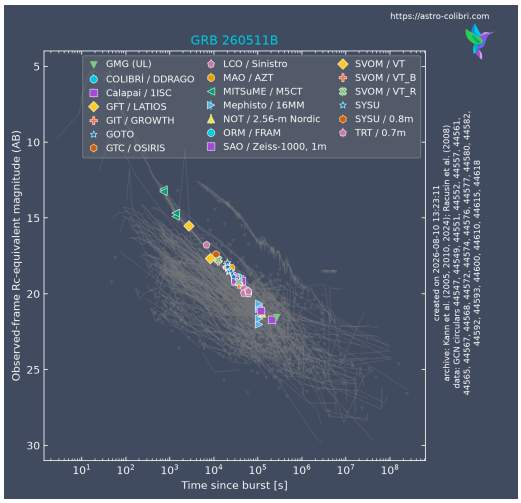


Figure 3: One downstream product: extracted photometry for GRB 260511B (colored markers) placed against archival afterglow measurements (gray). This context figure is intended for rapid inspection, not publication-level photometric analysis.

The package provides Python and command-line interfaces, a cached example for inspecting deterministic stages without an LLM request, and a credential-free test suite. Event association supports the public API, a caller-supplied backend, or offline use without association. Derived summaries and context figures are publicly viewable in Astro-COLIBRI; CSV and IVOA VOTable exports require a user account. The source Circulars are public documents. These products support triage and coordination rather than calibrated catalog analysis; Section 5.2 bounds their reliability.

Downstream, the records drive the optical-afterglow context figure of Figure 3, which plots a live event's extracted photometry against an archival afterglow population on a common observer-frame scale, while retaining links back to the source records and excerpts. The broader deployment and observer-facing products are described by Schüssler et al. (2026).

7 Conclusion

The hybrid pipeline converts free-text Circulars into typed, auditable records by surrounding schema-constrained generation with deterministic guidance and validation. Reviewers confirmed 96.5% of assessed outputs and localized errors mainly to timing and facility attribution. Redshift comparison found minor semantic disagreements but no misread values. Provenance and human verification remain essential.

Limitations

The human verification is an internal precision audit, not an independent benchmark: all reviewers contributed to the project, one supplied approximately 80% of assessments, and sampling emphasized recent reports without stratification by report type or difficulty. Reviewers primarily judged produced fields. Although they could record omissions, recall requires an independently annotated corpus enumerating every target before system output is seen. We therefore report no recall.

Field decisions repeat schema instances and are dominated by larger reports, so they are not 25,880 independent trials. The multiply reviewed subset contained no errors. Redshift validation covers one scalar field, catalogs disagree, and the schema loses whether a value is a limit, candidate-host estimate, or preliminary result.

The proprietary models can be retired or silently updated; recorded identifiers, inputs, and outputs improve traceability but not exact reproduction. Performance is untested for other languages, private reports, other genres, and genuinely multi-source Circulars. The normalized plot uses astrophysical assumptions and is only a context aid.

Ethical Considerations

GCN Circulars are public, but contact details are personal information supplied for scientific communication; deployments should preserve provenance and prevent bulk messaging. LLM output is not authoritative metadata without checks or review. Proprietary inference creates cost, access, and reproducibility concerns, and human users remain responsible for scientific conclusions.

Acknowledgments

The authors acknowledge ANR (French National Research Agency) for its support of the project "Multi-messenger Observations of the Transient Sky (MOTS)" no. ANR-22-CE31-0012. The authors also acknowledge support from the ACME project funded through the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No. 101131928. The authors acknowledge the use of AI-assisted writing and language editing tools during manuscript preparation; all content, interpretations, and conclusions were reviewed and approved by the authors. We thank Pierre Zweigenbaum, Cyril Grouin, and Felix Grèzes for collaboration and discussions on various aspects of the use of AI/NLP for time-domain astrophysics.

References

Atilla Kaan Alkan, Felix Grezes, Cyril Grouin, Fabian Schussler, and Pierre Zweigenbaum. 2024. Enriching a time-domain astrophysics corpus with named entity, coreference and astrophysical relationship annotations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 6177–6188. ELRA and ICCL.

Atilla Kaan Alkan, Cyril Grouin, Fabian Schussler, and Pierre Zweigenbaum. 2022. TDAC, the first corpus in time-domain astrophysics: Analysis and first experiments on named entity recognition. In *Proceedings of the First Workshop on Information Extraction from Scientific Publications*, pages 131–139. Association for Computational Linguistics.

Scott D. Barthelmy. 1998. The GRB coordinates network. In *Gamma-Ray Bursts, 4th Huntsville Symposium*, volume 428 of *AIP Conference Proceedings*, pages 99–103.

Paul Coppin. 2026. GRBweb: A comprehensive catalog of gamma-ray bursts. Accessed 2026-07-30.

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.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952, Singapore. Association for Computational Linguistics.

Jochen Greiner. 2026. Table of all well-localized gamma-ray bursts. Accessed 2026-08-07.

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. 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.

NASA Goddard Space Flight Center. 2026. GCN Circulars. Accessed 2026-06-27.

NASA/GSFC Swift Science Center. 2026. Swift GRB table. Accessed 2026-08-07.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Patrick Reichherzer, Fabian Schüssler, Valentin Lefranc, Julia Becker Tjus, Jayson Mourier, and Atilla Kaan Alkan. 2023. Astro-COLIBRI 2: An advanced platform for real-time multi-messenger discoveries. *Galaxies*, 11(1):22.

Patrick Reichherzer, Fabian Schüssler, Valentin Lefranc, Ali Yusafzai, Atilla Kaan Alkan, Halim Ashkar, and Julia Becker Tjus. 2021. Astro-COLIBRI: The COincidence LIBrary for real-time inquiry for multimessenger astrophysics. *The Astrophysical Journal Supplement Series*, 256(1):5.

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. 2021. PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9895–9901. Association for Computational Linguistics.

Fabian Schüssler, S. Bisero, M. Cellier, A. Ciric, B. Cornejo, I. Jaroschewski, A. Kaan Alkan, W. Kiendrébéogo, and A. Saint-Paul. 2026. AI-assisted extraction of follow-up observations from GCN circulars in Astro-COLIBRI. Preprint, arXiv:2608.23270.

Vidushi Sharma, Ronit Agarwala, Judith L. Racusin, Leo P. Singer, Tyler Barna, Eric Burns, Michael W. Coughlin, Dakota Dutko, Courey Elliott, Rahul Gupta, Ashish Mahabal, and Nikhil Mukund. 2026. Large language model-driven analysis of general coordinates network (GCN) circulars. *The Astrophysical Journal Supplement Series*, 283(1):30.

Vladimir Sotnikov and Anastasiia Chaikova. 2023. Language models for multimessenger astronomy. *Galaxies*, 11(3):63.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US. Association for Computational Linguistics.

ACL-Verbatim: hallucination-free question answering for research

Gábor Recski[1,2], Szilveszter Tóth[2], Nadia Verdha[1], István Boros[2], Ádám Kovács[2]
[1]TU Wien, [2]KR Labs
Correspondence: recski@krlabs.eu

Abstract

Researchers need efficient and reliable methods for collecting high-quality information from trusted sources, but modern tools for AI-assisted research still suffer from the tendency of Large Language Models (LLMs) to produce factually inaccurate or nonsensical output, commonly referred to as hallucinations. We apply the extractive question answering system VerbatimRAG (Kovacs et al., 2025) to research papers in the ACL Anthology[1], directly mapping user queries to verbatim text spans in retrieved documents. We contribute a novel ground truth dataset for the task of mapping user queries to relevant text spans in research papers, and use it to train and evaluate a variety of extractive models. Human annotation is performed by NLP researchers and is based on synthetic user queries generated using a custom pipeline based on the ScIRGen methodology (Lin et al., 2025), paired with chunks of research papers retrieved by VerbatimRAG. On this benchmark, a 150M-parameter ModernBERT token classifier trained on silver supervision from our pipeline achieves the best word-level F1 (53.6), ahead of the strongest evaluated LLM extractor (48.7).

1 Introduction

Researchers rely on scientific literature as a trusted source of information, but finding the relevant evidence in large paper collections remains difficult. Modern AI tools, especially those based on large language models (LLMs), offer a substantial increase in the efficiency of information search but introduce major risks to both individual users and organizations. Question Answering using LLMs lacks transparency and reliability. Even retrieval-augmented generation (RAG) systems (Gupta et al., 2024), which use a combination of document retrieval and generative AI, are prone to most major issues of LLMs, including a tendency to produce factually inaccurate, irrelevant, or nonsensical output, commonly referred to as hallucinations (Huang et al., 2025). Answers provided by LLMs cannot be trusted unless they are independently fact-checked, yet such verification remains a tedious process and is often omitted due to "algorithm appreciation, where people tend to prefer algorithmic judgment over human judgment, even when the algorithm's processes are not fully transparent" (Logg et al., 2019)

We present *ACL-Verbatim*, an application of the VerbatimRAG framework for extractive question answering (Kovacs et al., 2025) to the task of question answering from research papers in the ACL Anthology. While we contribute an end-to-end RAG system, complete with document preprocessing, indexing, and retrieval, our main focus is *extraction*, the task of identifying spans in retrieved text chunks that are most useful for satisfying the information need conveyed by the user query. It is this step that differentiates VerbatimRAG from other RAG frameworks and which enables question answering without hallucinations. In order to train and evaluate models on this task, we also create a pipeline for the automatic generation of search queries and perform manual annotation of text chunks to create a small ground truth dataset of 100 query-chunk pairs. Finally, we show that fine-tuning a compact extraction model on silver data generated by a strong LLM yields the best word-level F1 on our benchmark while using far fewer parameters than the evaluated LLM extractors.

2 Related work

All question answering systems that allow an LLM to generate the final answer are prone to producing output that is factually incorrect, inconsistent with the provided evidence, or nonsensical, a phenomenon that is commonly referred to as hallucination (Kaddour et al., 2023; Huang et al., 2025;

[1] https://aclanthology.org/

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 23–33
November 10, 2026 ©2026 Association for Computational Linguistics

Ji et al., 2023). Retrieval Augmented Generation (RAG) systems reduce intrinsic hallucinations by grounding the model in external sources, but extrinsic hallucinations can still occur due to LLMs' tendency to override retrieved information with their own prior "knowledge". RAG models continue to hallucinate (Niu et al., 2024), limiting their use in complex and high-risk domains such as medical, legal, or financial question answering (Lozano et al., 2023; Magesh et al., 2024). A range of methods have recently been proposed for hallucination detection. Frameworks such as RAGAS (Es et al., 2024) and ARES (Saad-Falcon et al., 2024) rely on specialized LLMs for large-scale hallucination detection but are not suitable for real-time prediction. Other LLM-based methods include approaches that use stochastic sampling (Manakul et al., 2023) or multi-step verification (Friel and Sanyal, 2023). Classifier models trained on hallucination datasets such as RAGTruth (Niu et al., 2024) include RAG-HAT (Song et al., 2024), RAGHalu (Zimmerman et al., 2024), and LettuceDetect (Ádám Kovács and Recski, 2025). Approaches to hallucination detection that investigate connections between responses and source documents include Luna (Belyi et al., 2025) and FACTOID (Rawte et al., 2024). Despite these recent efforts, hallucinations in RAG systems continue to limit the applicability of LLMs in real-world question answering tasks.

LLMs also suffer from lack of explainability, reducing both accountability and user trust. Mechanisms for generating post-hoc explanations of neural networks' predictions are notoriously unreliable, and intuitive but wrong self-explanations offered by LLMs create additional risk by inflating users' perception of their trustworthiness (Madsen et al., 2024; Chen et al., 2024b). These risks are inherent to systems that allow neural language models to generate the final output presented to users, even when these models have been specialized for the domain of academic research (Beltagy et al., 2019; Taylor et al., 2022; Viswanathan et al., 2023).

VerbatimRAG (Kovacs et al., 2025) is an open-source RAG framework that tackles the issue of hallucinations by taking an extractive approach to retrieval-augmented question answering that only returns text spans that are taken verbatim from source documents. VerbatimRAG combines standard retrieval with extraction, the task of highlighting the parts of some input text that are relevant for answering some user query, for which the framework offers multiple approaches, including LLMs prompted for the extraction tasks as well as smaller models fine-tuned for the extraction/highlighting task, such as Provence (Chirkova et al., 2025) or the Zilliz Semantic Highlighter (Zhang and Chen, 2026). Our experiments as well as our newly contributed dataset rely on the ACL Anthology[2], a public resource that has served as the basis of dozens of research datasets over the past decades (Bollmann et al., 2023), including large-scale corpora such as NLP Scholar (Mohammad, 2020), NLP-Explorer(Parmar et al., 2020), and the most recent ACL-OCL corpus (Rohatgi et al., 2023), each of which provides valuable additional metadata for publications and enables advanced analysis of NLP research.

3 Corpus creation

In this section we describe the corpus creation process, including data collection, preprocessing, segmentation, the generation of synthetic queries, as well as the human annotation process. Our pipeline is designed to allow for incremental updates of our dataset based on updates of the ACL anthology, detailed instructions are provided in the `acl-verbatim` repository[3]. The version of the dataset that served as the basis for the annotation and evaluation described in this paper is based on the state of the ACL Anthology in February 2026.

3.1 Data collection and preprocessing

The ACL Anthology (Gildea et al., 2018; Bollmann et al., 2023) is a public library of over 120,000 research papers from the domains of computational linguistics and natural language processing. Metadata as well as full-text PDFs of papers are distributed under permissive licenses (CC BY 4.0 as of 2016) and programmatic access is provided via a GitHub repository[4] and a Python library. We use these utilities to process all PDF files and to extract all paper metadata. Downloading, filtering, and preprocessing of papers was based on metadata extracted from the ACL Anthology on February 26, 2026. Entries for 120 034 papers were processed, out of which 114 567 were mapped to PDFs for further processing. The remaining 5k papers are hosted by third-party publishers and not covered by the permissive license of the ACL Anthology, these entries were discarded to maintain the flexible terms of the final dataset.

[2] https://aclanthology.org/
[3] anonymized
[4] https://github.com/acl-org/acl-anthology/

PDFs are converted to markdown format using the open-source `docling` library[5], resulting in 114 475 markdown files (a total of fewer than 100 papers were skipped due to a variety of docling errors that were not further investigated). Docling's DocumentConverter is invoked using default settings. All text-based content including headers, lists, tables, figure captions etc. are rendered in markdown, while other figures and some formulas are replaced by placeholder text indicating that some content has been discarded. We release this dataset of markdown files under a CC-BY 4.0 license on Huggingface[6].

Markdown documents are indexed using the VerbatimRAG library described in Section 2. For segmenting papers we implement a custom chunking strategy developed specifically for markdown-formatted research papers. This involves parsing section structure, segmenting papers along section boundaries, and prefixing each text chunk section and subsection titles to improve retrieval performance. The markdown chunker also prevents tables and code blocks from being split, and controls the minimum and maximum size of chunks, which we set to 500 and 5000 characters, respectively. Chunks are then indexed both for full text search and for dense vector search using the `granite-embedding-english-r2` embedding from IBM[7] (Awasthy et al., 2025).

3.2 Query generation and human annotation

This section describes the steps of creating a ground truth dataset mapping user queries to text spans in retrieved papers relevant for answering these queries. As a first step we create a sample of 333 papers in the ACL Anthology, randomly choosing from all English-language papers. Then we retrieve indexed chunks for these papers from the ACL-Verbatim index and randomly choose a single chunk for each paper. We then generate 3 synthetic queries for each chunk, following the ScIRGen methodology (Lin et al., 2025). This two-step process involves prompting an LLM to generate a list of question types that could be answered by a given paragraph, then using in-context learning for each question type to generate questions. We extend this pipeline by a third step that converts long and linguistically sophisticated questions to shorter and more fragmented queries that are more characteristic of real-world user queries. An end-to-end example is presented in Appendix A, full prompts are reproduced in Appendix D.

Generated queries are used to retrieve chunks from the VerbatimRAG index (see Section 3.1), and the top 5 chunks per query are used as input for the annotation task. Annotators then perform tasks for pairs of query and chunk. First, each chunk is annotated for relevance using a binary label. Chunks must be marked `relevant` if and only if the annotator considers the chunk to be relevant for satisfying the information need conveyed by the search query. If a chunk is labeled `irrelevant`, no further annotation takes place[8]. For chunks considered relevant, annotators must also indicate (highlight) the span or spans of text within the chunk that are most relevant for answering the query. If a table or figure is considered relevant, annotators are instructed to highlight its caption. This way a sequential labeling task is defined, mapping each relevant text chunk to one or more continuous sequences of its tokens. Annotation is performed via an excel sheet, created programmatically from the output of the ACL-Verbatim system and postprocessed to create the JSON-formatted gold dataset. All components of the pipeline for query generation and annotation are published as open-source software on GitHub[9].

3.3 Annotation challenges

The core annotation task of determining the text fragments within a section of a research paper that are best suited to fulfill the information need conveyed by a search query is quite complex and raises several methodological issues. While the sampling process described above yielded a total of 906 queries, and retrieving the top 5 chunks for each produced 4530 query-chunk pairs, the human-annotated portion of our dataset contains only the first 20 queries and a total of 100 chunks. Annotation as well as the adjudication of differences among annotators was performed by authors of this paper, NLP researchers with some variety in their fields and level of experience. In this section we describe some of the key challenges we have encountered during the manual annotation process and discuss some implications for the extraction

[5] https://docling-project.github.io/docling/

[6] anonymized

[7] https://huggingface.co/ibm-granite/granite-embedding-english-r2

[8] Additionally, a small number of chunks were marked with a question mark to indicate that the relevance judgement doesn't make sense, this was the case in particular for chunks containing bibliography sections of papers. Such chunks were also not annotated further.

[9] anonymized

task underlying the `verbatim` approach to question answering. We illustrate issues with examples cherry-picked from the 20 queries in the manually annotated dataset.

A key challenge that is specific to the domain of academic research is that for many queries the annotation of the most relevant spans within a section of a paper requires considerable domain-specific expertise and careful consideration. For example, the (synthetic) query *parsing merge predicate sequence equivalence conditions*, which is the simplified version of the more elaborate (synthetic) question *What are the three conditions under which two instantiated sequences are considered equivalent by the parsing merge predicate?*, which in turn was generated based on a short section describing an algorithm in the paper *LR Recursive Transition Networks for Earley and Tomita Parsing* (Perlin, 1991). Retrieving the top 5 chunks from our index yields sections from 4 different papers. Perlin 1991 is not among them, but all of them are concerned with algorithms for syntactic parsing and all of them would appear to be potentially relevant based on their vocabulary. It is only by reading through and developing a basic understanding for each section that the annotators could make the first of two judgements, the binary decisions on whether these chunks are relevant at all and whether one should proceed with the extraction task to identify relevant spans. The final dataset categorizes two of the chunks as relevant, the introductory section of the 1989 EACL paper on *Parsing and Derivational Equivalence* (Hepple and Morrill, 1989) and a paragraph describing a core algorithm in the 2010 ACL paper *Dynamic Programming for Linear-Time Incremental Parsing* (Huang and Sagae, 2010). Further annotation mapped both of these chunks to spans that should be extracted (highlighted) in response to the search query, which in one case reduced the 4700 character algorithm description to a single sentence of 92 characters (1.96%) *"The key observation for dynamic programming is to merge 'equivalent states' in the same beam"*, while the introductory section of the other paper was mostly relevant and only a few sentences were omitted from the extraction, reducing 1902 characters to 1447 (76.08%). This example illustrates the diversity of extracted spans, the difficulty and subjectivity of both annotation steps, and the meticulous effort required to create even a small high-quality dataset.

4 Extraction experiments

We evaluate extractive models on the manually annotated benchmark introduced in the previous sections. The benchmark contains 20 synthetic queries paired with the top-5 retrieved chunks per query, yielding 100 query–chunk pairs in total. Of these, 47 chunks are annotated as relevant and contain 78 gold evidence spans, while the remaining 53 chunks are irrelevant and have no gold spans. We report extraction metrics on all 100 rows. Three families of extractive systems are compared. First, we evaluate LLM-based span extractors. These models receive a question together with the retrieved chunk and must return verbatim evidence spans. We evaluate Mistral Small 2603, Nemotron-120B-A12B, GLM-5, and Qwen 3.6 35B. For Mistral, Nemotron, and Qwen, we compare a default extraction prompt against a paragraph-oriented prompt designed to encourage broader evidence selection. Second, we evaluate extractive pruning and highlighting baselines. Zilliz Semantic Highlight (Zhang and Chen, 2026) selects relevant sentences or token spans from the chunk, while Provence (Chirkova et al., 2025) prunes irrelevant sentences from the context using a DeBERTa-v3 reranker-style architecture. Zilliz follows the same general token-scoring formulation as Provence, but is trained as a bilingual semantic-highlighting model on top of a BGE-M3 reranker backbone (Chen et al., 2024a). Provence operates with a native context budget of 512 tokens and internally splits longer chunks.

Finally, we also train a compact self-contained student model, for which we generate silver supervision from the ACL Anthology corpus using synthetic queries and retrieved chunks, following the steps described in Section 3.2. The current release is based on 2000 sampled papers, which yielded 5892 synthetic queries, 32480 raw silver query–chunk rows, and 23235 retained rows after filtering. The final split contains 20916 silver training rows and 2319 development rows in canonical form, and 20920 / 2319 tokenized training / development windows at the full 8192-token ModernBERT context. At this context length roughly every silver row fits in a single window; the total amount of silver supervision the student sees is approximately 10k positive rows with spans and 11k negative rows. Evidence density (span characters over chunk characters) is 11.7%, i.e. a roughly 1:8 token-level class imbalance.

Our student architecture is a query-conditioned token classifier over an 8192-token ModernBERT backbone, with binary token labels and sliding-window inference. The input is the concatenation of question and chunk, and the output is a binary evidence label per token, decoded into character spans. We compare two backbones: the vanilla `answerdotai/ModernBERT-base` MLM checkpoint and the `Alibaba-NLP/gte-reranker-modernbert-base` cross-encoder, which has been post-trained on query–passage relevance. The silver teacher is Qwen 3.6 35B with the paragraph-oriented extraction prompt. We train for 5 epochs with batch size 8 at learning rate 2×10^{-5}; the best checkpoint is selected by silver-dev token F1. Two post-processing steps are applied at inference: spans shorter than 10 characters are dropped, and neighboring spans separated by at most 20 characters are merged. These two rules remove token-level fragmentation in which the model emits a "shotgun" of short pseudo-spans around genuine evidence tokens. The student model based on the reranker backbone will be released as `acl-verbatim-modernbert`.

We propose several overlap-based metrics for comparing extracted text spans against the ground truth on all 100 query–chunk rows in the benchmark. Our primary method of evaluation and comparison is word-level precision and recall, which compares the sets of words covered by gold and extracted spans. We prefer this metric because it isolates a model's ability to highlight the right words in a piece of text and is not sensitive to whether span boundaries are correctly predicted and whether models (and annotators) prefer fewer and longer spans or many shorter ones. The best configurations for each model type were selected by comparing word-level F1 scores. Additionally, *Containment* measures whether a large enough part of predicted spans are contained by gold spans (e.g. *Containment @ 0.8* is the ratio of spans that are at least 80% covered by a gold span). Analogously, *Coverage @ 0.8* is the number of gold spans that are at least 80% covered by predicted spans, divided by the total number of gold spans. Overlaps between spans are measured at the character level.

The right choice of metrics for evaluating span extraction depends heavily on the nature of the extraction task. One of the most common span extraction tasks in natural language processing is Named Entity Recognition (NER), where system outputs are commonly evaluated using span-level precision and recall, which requires exact matches between span boundaries and will penalize even the smallest mismatch by considering it as both a false positive and a false negative. We argue for our choice of metrics with a simple example. Consider a sequence of 100 characters in a retrieved chunk of text that is annotated as containing two relevant spans of 45 characters each, with a 10 character long break between them. Then consider a system that predicts this entire span of 100 characters as relevant, i.e. its mistake is merging the 10 irrelevant characters with the 90 relevant ones. This prediction would achieve span-level precision and recall scores of zero, since it has not made any correct predictions. In contrast, the word-level precision and recall of the system are 0.9 and 1.0, respectively. Finally, the *Containment* @ t ratio is 1 for all $t < 0.9$ and 0 for $t \geq 0.9$, while *Coverage* @ t is 1 for any value of t, expressing alternative preferences in evaluating extraction.

In addition to the ACL-specialized model evaluated here, we release a multi-domain sibling model, `verbatim-rag-modern-bert-v2`[10], trained on `verbatim-spans`[11]. This dataset combines our ACL silver data with RAGBench (Friel et al., 2025), a large-scale benchmark of retrieval-augmented question answering examples across industry-oriented domains, and Squeez (Ádám Kovács, 2026), a task-conditioned tool-output pruning dataset built from coding-agent tool observations. We evaluate this generic model separately in its model card, including QASPER (Dasigi et al., 2021), a scientific QA benchmark over NLP papers, as an out-of-training-domain test set. On ACL gold the generic model reaches 46.3 word-level F1, compared to 30.1 for Zilliz Semantic Highlight and 34.4 for Provence, and it also outperforms both baselines on RAGBench, Squeez, and QASPER.

5 Results

Table 1 compares extractor models on the 100 rows of the manually annotated benchmark. The best Word-F1 is achieved by our reranker-initialized ModernBERT student (53.63), ahead of the strongest LLM extractors, GLM-5 (48.71), Mistral Small (46.94), and Qwen with the paragraph prompt (46.73), while using 3 to 4 orders of magnitude fewer parameters. The generic multi-domain

[10] anonymized
[11] anonymized

Model	*Word*			*Containment @ t*			*Coverage @ t*			Latency
	Prec.	Rec.	F1	1.0	0.8	0.5	1.0	0.8	0.5	
LLM extractors										
Mistral Small 2603[a]	40.41	55.99	46.94	43.86	46.24	49.73	17.95	39.74	57.69	0.78
Mistral Small 2603[b]	34.22	**73.03**	46.61	27.70	33.61	39.04	48.72	**65.38**	**75.64**	1.07
Nemotron-120B-A12B[a]	34.14	50.92	40.88	36.62	36.62	42.88	43.59	43.59	52.56	0.38
Nemotron-120B-A12B[b]	30.15	62.72	40.73	29.93	31.29	44.90	**58.97**	62.82	62.82	0.54
GLM-5[a]	44.50	53.80	48.71	**52.87**	**54.25**	**58.36**	20.51	34.62	53.85	1.04
Qwen 3.6 35B[a]	44.74	40.88	42.72	49.05	50.95	53.33	15.38	25.64	41.03	1.63
Qwen 3.6 35B[b]	39.43	57.35	46.73	34.78	36.96	47.83	44.87	51.28	61.54	1.20
Pruning / highlighting baselines										
Zilliz Semantic Highlight[c]	46.97	22.11	30.07	33.33	35.46	39.53	2.56	8.97	19.23	1.04
Provence[d]	27.58	45.70	34.40	21.56	22.84	29.23	16.67	21.79	41.03	2.40
Our students (150M params)										
acl-verbatim-modernbert [e]	**65.43**	45.43	**53.63**	27.27	31.47	48.22	35.90	37.18	42.31	0.47
verbatim-rag-modern-bert-v2 [f]	63.00	36.58	46.29	30.30	31.82	46.97	26.92	32.05	38.46	0.40

Table 1: Extractor results on the gold benchmark (100 query–chunk pairs: 47 relevant and 53 irrelevant). See Section 4 for metric definitions. Values are percentages, except latency, which is seconds per row. Latencies for Zilliz, Provence, and ModernBERT were measured on CPU.

[a] Default extraction prompt. [b] Paragraph-oriented extraction prompt. [c] Zilliz token-span output at threshold 0.3. [d] Native Provence setting with internal splitting for contexts longer than 512 tokens. [e] Threshold 0.2 with min-span length 10 and merge gap 20. [f] `verbatim-rag-modern-bert-v2`, threshold 0.2 with min-span length 30 and merge gap 20.

ModernBERT model also remains competitive on ACL gold (46.29 Word-F1), outperforming the public extractive baselines despite not being specialized only for the ACL Anthology. Our ACL-specialized model also achieves the highest word-level precision. Unlike the LLM extractors, it often abstains on irrelevant chunks. LLMs, in particular those used with paragraph-oriented prompts, achieve higher recall and higher span-level coverage, but achieve substantially lower precision, extracting evidence from many chunks that are irrelevant for the query. This trade-off is of particular importance in our context of retrieval-augmented question answering, where high-precision extraction models are effective filters of irrelevant search results. This difference is directly observable if we compare the results of our best model with an LLM-based extractor that achieves much higher recall. On the 100 chunks in the evaluation dataset, 53 of which had no gold spans annotated, our model chose not to predict any spans for 60 chunks while the paragraph-based Mistral model abstained only 35 times. We also illustrate this behavior with a cherry-picked example. For the query *hate speech detection downsampled training examples number*, one of the top retrieved chunks is a subsection describing the experimental dataset in the paper *EDAL: Entropy based Dynamic Attention Loss for HateSpeech Classification* (Fahim et al., 2023). This text provides lots of detail about topics closely related to the query, including statistics on class labels, but it is not at all concerned with downsampling. Our model correctly chose not to extract any spans, and so did the Zilliz highlighting model and some of the LLM-based extractors, including the default Mistral model, both Nemotron models, and the paragraph-based Qwen model. However, the four remaining models each choose some false positive spans, including texts on merging labels and texts as well as tables on dataset sizes.

6 Conclusion

We described an application of the VerbatimRAG architecture to over 100K research papers in the ACL Anthology, contributed a manually annotated dataset for the core extraction task, and presented a set of experiments showing that small customized encoder-decoder architecture trained with synthetic data outperforms zero-shot LLM-based extraction on this task, at a fraction of the cost. We release all components of our pipeline as open-source software. We believe that combining the VerbatimRAG approach with the task-oriented training of extractive models provides a blueprint for the efficient deployment of high-performing hallucination-free question answering systems across a variety of domains.

Limitations

The validity of our conclusions is limited by the size of the manually annotated dataset that was the basis of both quantitative and qualitative evaluation. The high complexity of the annotation task, described in detail in Section 3.2, also limited our ability to measure agreement between multiple annotators. to implement a rigorous adjudication process for resolving differences among annotators, or to develop detailed and objective annotation guidelines. We believe that all these steps will be possible if our approach is applied to more narrowly defined question answering use-cases that in turn lead to more objective extraction tasks. Furthermore, the extraction models trained using synthetic training data may reproduce unintended bias present in the output of LLMs, which may lead to such bias being reinforced and propagated by our models.

References

Parul Awasthy, Aashka Trivedi, Yulong Li, Meet Doshi, Riyaz Bhat, Vignesh P, Vishwajeet Kumar, Yushu Yang, Bhavani Iyer, Abraham Daniels, Rudra Murthy, Ken Barker, Martin Franz, Madison Lee, Todd Ward, Salim Roukos, David Cox, Luis Lastras, Jaydeep Sen, and Radu Florian. 2025. Granite embedding r2 models. *Preprint*, arXiv:2508.21085.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Masha Belyi, Robert Friel, Shuai Shao, and Atindriyo Sanyal. 2025. Luna: A lightweight evaluation model to catch language model hallucinations with high accuracy and low cost. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 398–409, Abu Dhabi, UAE. Association for Computational Linguistics.

Marcel Bollmann, Nathan Schneider, Arne Köhn, and Matt Post. 2023. Two decades of the ACL Anthology: Development, impact, and open challenges. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 83–94, Singapore. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen Mckeown. 2024b. Do models explain themselves? Counterfactual simulatability of natural language explanations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 7880–7904. PMLR.

Nadezhda Chirkova, Thibault Formal, Vassilina Nikoulina, and Stéphane Clinchant. 2025. Provence: efficient and robust context pruning for retrieval-augmented generation. *Preprint*, arXiv:2501.16214.

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.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Md Fahim, Dr. Amin Ahsan Ali, Md Ashraful Amin, and Akm Mahbubur Rahman. 2023. EDAL: Entropy based dynamic attention loss for HateSpeech classification. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 775–785, Hong Kong, China. Association for Computational Linguistics.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2025. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *Preprint*, arXiv:2407.11005.

Robert Friel and Atindriyo Sanyal. 2023. Chainpoll: A high efficacy method for llm hallucination detection. *Preprint*, arXiv:2310.18344.

Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann, and Martín Villalba. 2018. The ACL Anthology: Current state and future directions. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 23–28, Melbourne, Australia. Association for Computational Linguistics.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *Preprint*, arXiv:2410.12837.

Mark Hepple and Glyn Morrill. 1989. Parsing and derivational equivalence. In *Fourth Conference of the European Chapter of the Association for Computational Linguistics*, Manchester, England. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Liang Huang and Kenji Sagae. 2010. Dynamic programming for linear-time incremental parsing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1077–1086, Uppsala, Sweden. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *Preprint*, arXiv:2307.10169.

Adam Kovacs, Paul Schmitt, and Gabor Recski. 2025. KR labs at ArchEHR-QA 2025: A verbatim approach for evidence-based question answering. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 69–74, Vienna, Austria. Association for Computational Linguistics.

Junyong Lin, Lu Dai, Ruiqian Han, Yijie Sui, Ruilin Wang, Xingliang Sun, Qinglin Wu, Min Feng, Hao Liu, and Hui Xiong. 2025. Scirgen: Synthesize realistic and large-scale rag dataset for scientific research. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 5619–5630, New York, NY, USA. Association for Computing Machinery.

Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

Alejandro Lozano, Scott L Fleming, Chia-Chun Chiang, and Nigam Shah. 2023. Clinfo.ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. *Preprint*, arXiv:2310.16146.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *Preprint*, arXiv:2405.20362.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Saif M. Mohammad. 2020. NLP scholar: A dataset for examining the state of NLP research. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 868–877, Marseille, France. European Language Resources Association.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Monarch Parmar, Naman Jain, Pranjali Jain, P. Jayakrishna Sahit, Soham Pachpande, Shruti Singh, and Mayank Singh. 2020. Nlpexplorer: Exploring the universe of nlp papers. In *Advances in Information Retrieval*, pages 476–480, Cham. Springer International Publishing.

Mark Perlin. 1991. LR recursive transition networks for Earley and Tomita parsing. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 98–105, Berkeley, California, USA. Association for Computational Linguistics.

Vipula Rawte, S. M Towhidul Islam Tonmoy, Krishnav Rajbangshi, Shravani Nag, Aman Chadha, Amit P. Sheth, and Amitava Das. 2024. Factoid: Factual entailment for hallucination detection. *Preprint*, arXiv:2403.19113.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: Advancing open science in computational linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361, Singapore. Association for Computational Linguistics.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. Ares: An automated evaluation framework for retrieval-augmented generation systems. *Preprint*, arXiv:2311.09476.

Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, and Cheng Niu. 2024. RAG-HAT: A hallucination-aware tuning pipeline for LLM in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1548–1558, Miami, Florida, US. Association for Computational Linguistics.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *Preprint*, arXiv:2211.09085.

Felipe Viegas, Washington Cunha, Christian Gomes, Antônio Pereira, Leonardo Rocha, and Marcos Goncalves. 2020. CluHTM - semantic hierarchical topic modeling based on CluWords. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8138–8150, Online. Association for Computational Linguistics.

Vijay Viswanathan, Luyu Gao, Tongshuang Wu, Pengfei Liu, and Graham Neubig. 2023. DataFinder: Scientific dataset recommendation from natural language descriptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10288–10303, Toronto, Canada. Association for Computational Linguistics.

Cheney Zhang and Jiang Chen. 2026. How we built a semantic highlight model to save token cost for rag. https://huggingface.co/blog/zilliz/zilliz-semantic-highlight-model. Hugging Face community article, published January 15, 2026.

Ilana Zimmerman, Jadin Tredup, Ethan Selfridge, and Joseph Bradley. 2024. Two-tiered encoder-based hallucination detection for retrieval-augmented generation in the wild. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 8–22, Miami, Florida, US. Association for Computational Linguistics.

Ádám Kovács. 2026. Squeez: Task-conditioned tool-output pruning for coding agents. *Preprint*, arXiv:2604.04979.

Ádám Kovács and Gábor Recski. 2025. Lettucedetect: A hallucination detection framework for rag applications. *Preprint*, arXiv:2502.17125.

A Query generation example

B Detailed model comparison

Table 2 reports containment and coverage metrics for the extractor configurations evaluated in Table 1. All metrics are computed on the full 100-row benchmark, including the 53 irrelevant query–chunk pairs as negative examples.

C Threshold selection for the student model

The student is a binary token classifier, so span decisions depend on a probability threshold at inference. Table 3 reports the all-row gold benchmark scores for the GTE-reranker student at $t \in \{0.2, 0.3, 0.4, 0.5\}$ with the same post-processing held constant: spans shorter than 10 characters are dropped, and neighbouring spans separated by at most 20 characters are merged.

The best Word-F1 is obtained at t=0.2. Increasing the threshold improves precision but reduces recall, which lowers F1 on this benchmark.

D Prompts for query generation

D.1 Question-type classification prompt

```
You are a researcher generating questions and
   answers to find relevant
information within a specific domain. Below
   are the potential question types.
Choose the type that best fits the field
   information and the user's purpose.

1. Verification: questions seeking a simple
   yes/no confirmation.
2. Disjunctive: questions presenting multiple
   alternatives.
3. Concept Completion: questions starting with
    Who/What/When/Where.
4. Example: questions asking for instances of
   a concept.
5. Feature Specification: questions about
   properties or characteristics.
6. Quantification: questions seeking numerical
    or measurable information.
7. Definition: questions asking for the
   meaning of a term or concept.
8. Comparison: questions asking for
   similarities or differences.
9. Interpretation: questions asking for
   inference over observed patterns.
10. Causal Antecedent: questions about causes
   or reasons.
11. Causal Consequence: questions about
   outcomes or results.
12. Goal Orientation: questions about
   objectives or intentions.
13. Instrumental/Procedural: questions asking
   how to achieve a goal.
14. Enablement: questions about conditions
   enabling an action.
15. Expectation: questions about anticipated
   or missing outcomes.
16. Judgmental: questions asking for
   evaluation or opinion.
17. Assertion: statements indicating lack of
   knowledge.
18. Request/Directive: requests to summarize,
   analyze, or search.

Task: Based on the following text from a
   research paper, return the most
appropriate 3 question types that could be
   answered by this text. Give me the
name of each type and not other information.
   Return ONLY valid JSON -- an
array of objects, no markdown or explanations.

Text: {chunk}
```

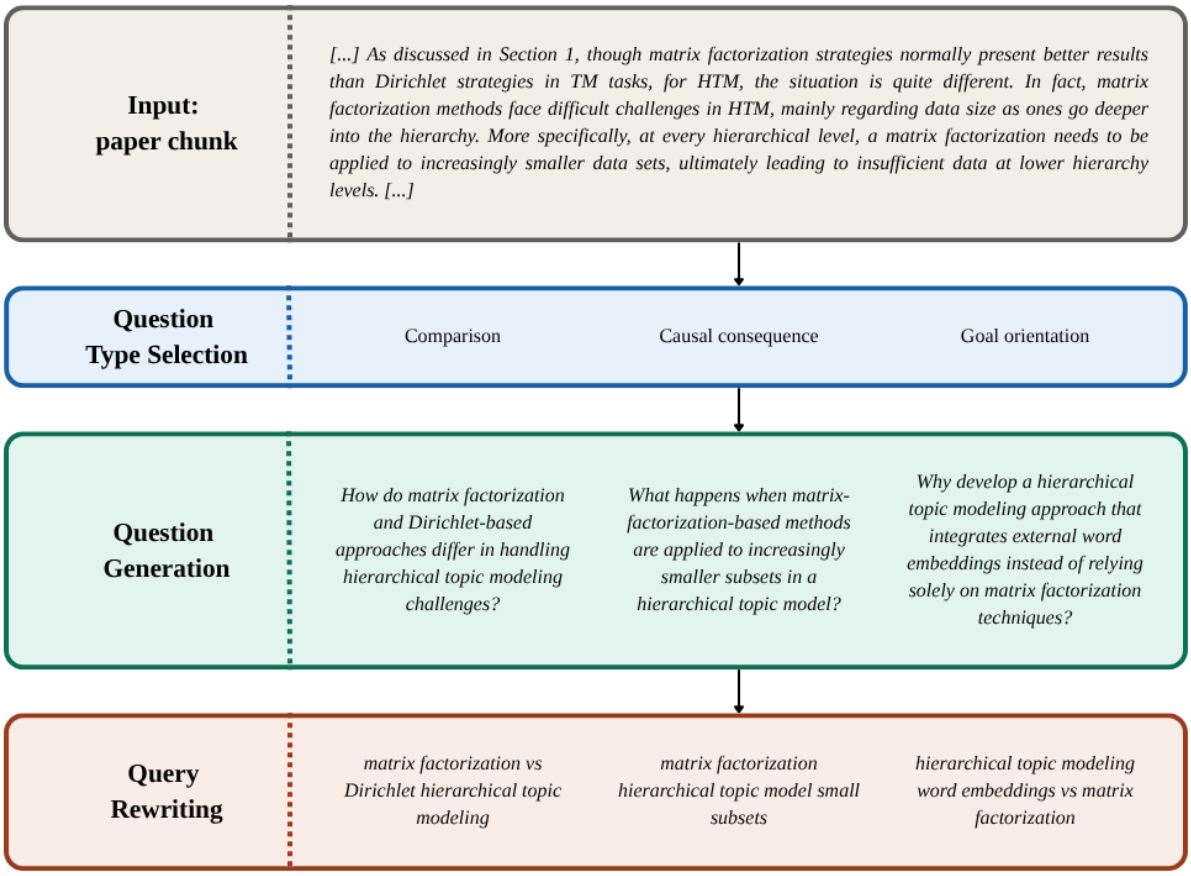


Figure 1: Generation of synthetic user queries, based on the ScIRGen methodology. The example shows a chunk from the paper *CluHTM - Semantic Hierarchical Topic Modeling based on CluWords* (Viegas et al., 2020)

D.2 Question generation prompt

```
You are a researcher asking questions aiming
   to find information in research
papers.

Content of paper: {chunk}

Please generate one question that can be
   answered by the above text and which
belongs to the question type below.

- Question Type: {q_type}
- Question Description: {q_def}
- Question Example: {q_ex}

Instructions:
1. Only return a question without any other
   information.
2. Use neutral terms like "a dataset", "data
   collection method", or
   "research approach", instead of references
   like "the study" or
   "this dataset".
3. The question should be short and simple,
   resembling what a user might
   type into a search engine.
4. The question should be answerable based on
   the text above.
```

D.3 Query rewriting prompt

```
You are a researcher using a search engine to
   find information.

Your question: {question}

Please generate a search query that you would
   use to find the answer to this
question.

Instructions:
1. Only return a search query without any
   other information.
2. The query should be short and simple,
   resembling what a user might type
   into a search engine.
3. The query does not need to be grammatical.
```

E Prompts for extraction

E.1 Default VerbatimRAG extraction prompt

```
Extract EXACT verbatim text spans from
   multiple documents that answer the
question.

Rules
```

	Containment @ t			Coverage @ t		
Model	0.5	0.8	1.0	0.5	0.8	1.0
Mistral Small 2603	0.497	0.462	0.439	0.577	0.397	0.179
Mistral Small 2603 + paragraph	0.390	0.336	0.277	0.756	0.654	0.487
Nemotron-120B-A12B	0.429	0.366	0.366	0.526	0.436	0.436
Nemotron-120B-A12B + paragraph	0.449	0.313	0.299	0.628	0.628	0.590
GLM-5	0.584	0.542	0.529	0.538	0.346	0.205
Qwen 3.6 35B	0.533	0.510	0.490	0.410	0.256	0.154
Qwen 3.6 35B + paragraph	0.478	0.370	0.348	0.615	0.513	0.449
Zilliz Semantic Highlight	0.395	0.355	0.333	0.192	0.090	0.026
Provence reranker-pruner	0.292	0.228	0.216	0.410	0.218	0.167
ACL-Verbatim GTE-reranker (t=0.2 + merge)	0.482	0.315	0.273	0.423	0.372	0.359

Table 2: Detailed extractor metrics on the full 100-row gold benchmark. Containment measures how much of a predicted span lies inside a gold span; coverage measures how much of a gold span is covered by a prediction.

Threshold	Word-P	Word-R	Word-F1	IoU-F1@0.5
0.2	0.654	0.454	**0.536**	**0.389**
0.3	0.667	0.421	0.516	0.338
0.4	0.678	0.403	0.506	0.365
0.5	**0.701**	0.380	0.493	0.336

Table 3: Threshold sweep for the GTE-reranker student on the full 100-row gold benchmark.

```
1. Extract only text that explicitly addresses
    the question.
2. Never paraphrase, modify, or add to the
   original text.
3. Preserve original wording, capitalization,
   and punctuation.
4. Order spans within each document by
   relevance, most relevant first.
5. Include complete sentences or paragraphs
   for context.

Output format
Return a JSON object mapping document IDs to
   span arrays ordered by relevance:
{
  "doc_0": ["most relevant span", "next most
   relevant span"],
  "doc_1": ["most relevant from doc 1"],
  "doc_2": []
}

If no relevant information exists in a
   document, use an empty array.

Your task
Question: {{ question }}

Documents:
{{ documents }}

Extract verbatim spans from each document:
```

E.2 Paragraph-style extraction prompt

```
Extract verbatim supporting passages from each
    document that answer the
question.

What to extract
A supporting passage is the complete portion
   of the document a researcher
would highlight to justify the answer,
   including:
- the sentence(s) that directly address the
   question;
- preceding setup sentence(s) that introduce
   the topic, methodology, or
  figure being referenced;
- concluding interpretation sentence(s) that
   summarize implications;
- table captions when the table itself is
   relevant.

Prefer a single continuous paragraph over
   multiple fragments of the same
paragraph. Only split into multiple spans when
    relevant content is in
non-adjacent parts of the document.

Rules
1. Use EXACT text from the document; no
   paraphrasing or edits.
2. Preserve original wording, capitalization,
   and punctuation.
3. If no passage in the document supports the
   answer, return an empty array.
4. Order spans within each document by
   relevance, most relevant first.

Output format
Return JSON mapping document IDs to arrays:
{
  "doc_0": ["first supporting passage", "
   second supporting passage"],
  "doc_1": ["passage from doc 1"],
  "doc_2": []
}

Your task
Question: {{ question }}

Documents:
{{ documents }}

Extract supporting passages from each document:
```

Auditable Multi-View Hosted Multimodal Judging for AstroCLIMB

Kewen Zhu
Independent

Junyuan Tan
University of California, Berkeley

Abstract

We study AstroCLIMB, a four-way relationship-classification task over pairs of astronomy figure/caption endpoints. Starting from a leakage-audited lexical and low-level image-statistics baseline (M7), we evaluate a hosted multimodal judge under a fixed, auditable protocol. GPT-5.6 Luna receives only the two endpoint contents and produces integer class probabilities for three views in both endpoint directions. On 2,000 frozen validation pairs, all 12,000 calls succeeded. Equal-weight fusion reaches macro-F1 0.67, versus M7's 0.40 (paired component-bootstrap delta $+0.27$, 95% CI $[+0.24, +0.29]$); Luna-only reaches 0.68. These are conditional local-validation results, not an official Luna leaderboard score. We therefore report an auditable system study without claiming paper-level generalization.

1 Introduction

AstroCLIMB (Grezes et al., 2026; AstroCLIMB Organizers, 2026) asks a system to classify a pair of astronomy figure/caption objects as *same figure*, *same paper*, *related papers*, or *unrelated papers*. Pairs may be image–image, image–caption, or caption–caption. The task tests whether systems can combine heterogeneous scientific endpoints while avoiding shortcuts from identifiers and metadata. The benchmark's *related papers* target is citation-based, while the original judging prompt describes substantive relatedness; throughout, we treat the latter as a semantic proxy rather than citation-edge ground truth.

Our M7 system uses symmetric word-level TF–IDF summaries, endpoint availability, and deterministic low-level image statistics. It obtains 0.40 macro-F1 on the frozen observable-content component split and 0.40 on the official public run (submission 56179217). Our research question is: under a fixed, auditable protocol and the same frozen split, how much does a hosted multimodal judge add to this content-only baseline? The resulting system, AstroLex-Luna, emphasizes provenance and failure visibility: anonymous cache keys, digest-bearing receipts, no stored payloads, and an all-six-or-fallback contract.

Our contributions are a documented content-only baseline and request protocol; an auditable report of one cached hosted-judge run, with conditional uncertainty and explicit failure handling; and an error analysis that separates benchmark semantics from the prompt's relatedness proxy. The contribution is an evaluation procedure and empirical analysis, not a new model architecture or a new aggregation algorithm.

This work is a systems evaluation, not a claim that a proprietary endpoint is universally superior. Related scientific multimodal resources include SciMMIR (Wu et al., 2024), MISTI (Taechoyotin and Acuna, 2024), and Multimodal ArXiv (Li et al., 2024); graph-based representations such as SPECTER (Cohan et al., 2020) use information excluded here.

2 Data and leakage audit

The official training and test files each contain 10,000 pair rows. Separate DOI, paper-title, author, citation-metadata, and image-identifier fields are excluded from prediction features. Identifiers and content hashes support splitting and traceability. Class labels supervise M7 training and score validation predictions; they are never supplied to the hosted judge. Contaminated sample/solution values are likewise excluded from modeling and selection.

We project permitted content to decoded captions, image bytes, modality, and endpoint SHA-256 digests. The fixed component-greedy-v2 split (seed 20260911) has 8,000 training and 2,000 validation rows, 8,671 observable-

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 34–37
November 10, 2026 ©2026 Association for Computational Linguistics

content components, exact validation class counts 200/600/600/600, and zero endpoint or pair overlap. The term "leakage-audited" below means observable-content component disjoint only: paper_disjoint=false because paper identities are unavailable for every endpoint. One endpoint is shared between distributed train and test files, so we do not claim a fully independent official test set. This content-overlap audit cannot detect whether the hosted model encountered the papers during pretraining.

3 Method

M7 is our frozen lexical and image-statistics baseline. Its 328 features combine symmetric word TF–IDF pair summaries, availability, and 64 low-level image statistics (color, brightness, gradient, and spatial summaries). A class-balanced multinomial logistic regression is fit using training-only TF–IDF, standardization, $C = 1$, and at most 2,000 iterations.

Hosted judge. For each pair, GPT-5.6 Luna (medium reasoning effort) is queried for direct, counterfactual, and boundary/ambiguity views, each in A/B and B/A directions. Inputs contain only raw captions and preprocessed images (EXIF transpose, RGB, longest edge at most 1536, JPEG quality 90); no tools are enabled, temperature is unset, and API storage is disabled (`store=false`). The six responses are aggregated as specified below.

Response and aggregation protocol. We use model alias `gpt-5.6-luna`. Each response must contain exactly four integer fields, in class order SF/SP/RP/UP, each in $[0, 100]$ and summing to 100; malformed responses count as failures. The view instructions are "View: direct. Classify the relation.", "View: counterfactual. Ignore superficial wording/layout and classify.", and "View: boundary. Classify conservatively at the nearest class boundary." Each follows the same label definitions and is issued with both endpoint orders. Appendix A reproduces the common prompt; the accompanying `PROTOCOL.md` also records the response schema and settings.

For class c, we compute

$$g_c = \left[\prod_{k=1}^{6} \frac{\max(p_{kc}, 10^{-9})}{100} \right]^{1/6}, \qquad (1)$$

System	Macro	SF	SP	RP	UP
M7 baseline	0.40	0.39	0.49	0.29	0.45
Luna-only	0.68	0.82	0.70	0.39	0.82
M7+Luna fusion	0.67	0.80	0.71	0.37	0.81

Table 1: Frozen validation results from one cached API run. SF, SP, RP, and UP denote same figure, same paper, related papers, and unrelated papers.

where p_{kc} is the returned integer. We normalize g, set SF to zero for image–image and text–text pairs, and renormalize. The same mask is applied to M7 before equal-weight probability fusion; prediction is by argmax, with ties resolved in the stated class order. Table 1 reports the original unmasked M7 baseline; its hard predictions and F1 are unchanged by masking in this run. Pooling both orders symmetrizes the cached ensemble; it does not guarantee identical responses on a new API run.

Fusion, failure handling, and uncertainty. The primary prediction is a fixed 0.5/0.5 fusion of masked M7 and Luna. A pair is aggregated only if all six calls succeed; any failure discards the partial Luna result and falls back to masked M7. This all-six-or-fallback behavior is implemented and tested; the reported run had no failures. We use 2,000 component-bootstrap replicates. The confidence interval is a fixed-run conditional component-bootstrap interval: it describes sampling variation conditional on these cached responses and does not include API randomness, model-alias changes, or serving drift. The live validation run required an explicit external-API flag and used an append-only response cache. The documented split and code specify how to rerun M7 with authorized data. Cached Luna responses permit reaggregation of this run; neither the protocol nor access to the service guarantees identical fresh responses.

4 Results

On the frozen validation set, M7 scores 0.40 macro-F1, Luna-only 0.68, and the fixed fusion 0.67. The fusion improves over M7 by $+0.27$; its paired component-bootstrap 95% CI is $[+0.24, +0.29]$. Luna-only has conditional bootstrap CI $[0.66, 0.70]$. The fused per-class F1 values are 0.80 (same figure), 0.71 (same paper), 0.37 (related papers), and 0.81 (unrelated papers); related papers remains the hardest class. Values shown here and in Table 1 are two-significant-digit

renderings; exact values remain in the evidence artifacts.

Fusion as a negative result. Fixed fusion is 0.011 macro-F1 below Luna-only. It raises same-paper F1 from 0.70 to 0.71, but lowers the other three class scores (Table 1); deltas are computed from unrounded values. We retain it as the prespecified comparison, not as the best-performing predictor. These results support the value of the hosted judge relative to M7, but do not support an overall macro-F1 benefit from adding M7 to Luna. No fusion weight was retuned on these outcomes, and we do not claim statistical significance for the Luna-only versus fusion difference.

Descriptive modality slices show fused macro-F1 of 0.52 for image–image ($n = 571$), 0.58 for image–caption ($n = 838$), and 0.51 for text–text ($n = 591$). Each slice averages over all four labels; SF has no support and contributes zero F1 in image–image and text–text. These availability-confounded scores should not be directly ranked as evidence of modality-specific understanding.

Error analysis. Table 2 shows the fused confusion matrix. Of 600 related-paper pairs, 263 (43.8%) are predicted as same paper and 165 (27.5%) as unrelated; only 142 (23.7%) are recovered. The reverse same-paper-to-related error occurs in 16 of 600 pairs (2.7%), showing an asymmetric boundary error. Luna-only has the same dominant failure: 276 related pairs become same paper, while 153 are correctly recovered. Fusion reduces the former count to 263 but increases related-to-unrelated errors from 147 to 165, leaving related-paper F1 lower overall.

The original prompt describes related papers as "objects from different but substantively related papers", whereas the task label requires a citation link. Content similarity is therefore an imperfect proxy for the target relation. This mismatch is a plausible contributor to the errors, but the confusion counts alone do not establish its cause. We preserve the original prompt and cached predictions rather than retroactively relabeling this experiment.

Exploratory call-count comparison. Using immutable cached responses, direct A/B (one call) scores 0.65, direct bidirectional (two) 0.64, all three A/B views (three) 0.66, and the full bidirectional three-view protocol (six) 0.67 under the same fusion. The six-versus-three difference (0.0081) is post-hoc and exploratory; it does not establish a causal contribution of any prompt view. Counts and cost do not by themselves demonstrate a benefit. Establishing a visual reasoning mechanism would require controlled image/caption interventions, which we did not run. We also cannot prove absence of pretraining contamination from this audit.

Gold \ Pred.	SF	SP	RP	UP
SF	200	0	0	0
SP	60	489	16	35
RP	30	263	142	165
UP	12	23	17	548

Table 2: Fused confusion counts on the 2,000 validation pairs. Rows are gold labels; columns are predictions. Abbreviations follow Table 1.

Auditability. The receipt records model, protocol, cache and prediction digests, SDK version, external-API status, and 12,000 successful calls. Summing the cache usage records gives 20,232,306 input and 3,783,736 output tokens; at the rates recorded for this run, USD 0.20/1.20 per million input/output tokens, estimated cost is USD 8.59. The accompanying protocol and `EVIDENCE.json` map the reported aggregates to hashed result and prediction artifacts. These records bind aggregates to local artifacts; hashes alone do not provide independent access to those artifacts. Raw endpoint payloads and per-pair predictions are not redistributed in this bundle. A fresh evaluation requires authorized competition data and access to the hosted endpoint, and remains sensitive to service drift.

5 Conclusion

The official system result in this paper is the M7 public run (score 0.40, submission 56179217). Luna is a locally evaluated hosted component, not an official system run. Under the observable-content component split and this fixed cached run, the constrained hosted judge substantially improves local validation over M7. This conclusion is conditional on one fixed split and cached service run; it does not establish paper-level generalization or a causal mechanism.

6 Limitations

The split is not paper-disjoint, so the results do not establish paper-level generalization. Luna is

proprietary, its serving behavior may drift, and the conditional bootstrap intervals do not capture API randomness. We cannot prove that the hosted model was free of pretraining contamination. Raw endpoint content can also contain visible titles or citations even though separate metadata fields are excluded. The six-view protocol does not establish causal visual or cross-modal understanding; controlled image/caption interventions would be needed for mechanism attribution. The fixed fusion was specified before reading the reported validation outcome, whereas Luna-only and call-count comparisons are exploratory. The result is neither a new architecture nor an official Luna leaderboard score. The prompt uses substantive relatedness as a proxy for the citation-edge definition of related papers; the cached results therefore also reflect this semantic mismatch, rather than an exact operationalization of citation linking.

Acknowledgements

Generative AI use. Generative AI tools, including OpenAI Codex, assisted with code implementation, experimental analysis, and manuscript drafting and revision. The authors are responsible for the scientific claims, references, and final content. The evaluated GPT-5.6 Luna predictions were produced under the protocol described in Section 3.

References

AstroCLIMB Organizers. 2026. AstroCLIMB: Astronomy citation linking from illustrations. WASP 2026 shared task and dataset documentation. Accessed 2026-09-11.

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. Association for Computational Linguistics.

Felix Grezes, Kelly Lockhart, Alberto Accomazzi, Atilla Alkan, Ethan Seefried, and Tirthankar Ghosal. 2026. 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.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387. Association for Computational Linguistics.

Pawin Taechoyotin and Daniel Acuna. 2024. Misti: Metadata-informed scientific text and image representation through contrastive learning. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 155–164. Association for Computational Linguistics.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, Noura Al Moubayed, Jie Fu, and Chenghua Lin. 2024. SciMMIR: Benchmarking scientific multi-modal information retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12560–12574. Association for Computational Linguistics.

A Frozen judging instructions

The following common prefix is reproduced from the evaluated protocol, with line wrapping only:

> Labels: same_figure=exactly an image and its own caption; same_paper=distinct objects from one paper; related_papers=objects from different but substantively related papers; unrelated_papers=no substantive relation. Return integer probabilities summing to 100.

Each request appends one of the three view instructions in Section 3 and the two endpoint content blocks. Each view is called once with A/B and once with B/A. The four JSON keys are the label names above; values must be integers between 0 and 100, with no additional fields and a total of 100. Application-level validation enforces the total as well as schema compliance.

The original prompt equates relatedness with substantive content similarity, including in its description of the unrelated class. A pair can be topically similar without a citation link, and a citation can connect objects with dissimilar content. Accordingly, this prompt does not exactly operationalize the benchmark's citation-based target. A citation-aligned prompt comparison would require new predictions and evaluation; no such comparison was performed for this paper. We retain the original prompt and scores to keep that distinction inspectable.

Can Generative AI Automate Data Extraction for Meta-Analysis? A Case Study on Intercropping Research

Zehao Lu[1,*], **Xingguo Xiong**[2,*], **Wopke van der Werf**[1],
Thijs L. van der Plas[1], **Ioannis N. Athanasiadis**[1]
[1]Wageningen University & Research, Wageningen, the Netherlands
[2]Zhejiang Academy of Agricultural Sciences
{zehao.lu,wopke.vanderwerf,thijs.vanderplas,ioannis.athanasiadis}@wur.nl, xiongxg@zaas.ac.cn
[*]Equal contribution.

Abstract

Meta-analysis is the synthesis of information from multiple sources to arrive at an overarching conclusion. There is a large need for meta-analysis in agricultural research to synthesize what is known and analyze overarching patterns. Extracting data from published literature is, however, labor-intensive, time-consuming, and tedious, and is impeded by a lack of standardization in research design, units of measurement, and terminology. These challenges are particularly evident in the domain of crop species mixtures, also called intercropping. With the growing capabilities of LLMs, many recent attempts have focused on building systems and tools to automate data collection, yet rigorous assessment against human-labeled ground truth is often missing. In this research, we evaluate three LLM-based approaches—direct zero-shot prompting, a staged workflow, and a multi-agent system—with six open-weight models to extract data from the intercropping literature. The results are evaluated against the manually curated ground truth and through a downstream statistical analysis. Overall, direct zero-shot prompting is the strongest and most consistent approach, achieving the highest mean similarity-adjusted F1 of 0.577, although none of the approaches is close to fully accurate. In the downstream analysis, most model–approach combinations recover the direction of the relationship between the predictor and outcome variables, but do not estimate its magnitude accurately. [1]

1 Introduction

Modern agriculture faces the dual challenge of increasing food production for a growing global population while adapting to climate change (Godfray et al., 2010). Meeting these demands requires a transition from conventional monoculture toward more diversified cropping systems (Tamburini et al., 2020). Intercropping is one such system, in which two or more crop species are grown together to improve resource-use efficiency and ecological resilience (Brooker et al., 2014). However, its benefits vary across crop combinations, pedoclimatic conditions, and management practices such as sowing dates, planting densities, and fertilizer inputs (Antle et al., 2017; Grassini et al., 2015). Evidence from different environments must therefore be synthesized to identify suitable crop combinations and management practices.

Meta-analysis systematically combines quantitative evidence from published studies to assess outcomes across diverse conditions (Gurevitch et al., 2018). This generally involves identifying eligible studies, extracting comparable outcomes and their uncertainty, and synthesizing the evidence statistically. In agriculture, this evidence can help evaluate management strategies and provide empirical data for models exploring production and climate-adaptation scenarios (Antle et al., 2017; Holzworth et al., 2014). Meta-regression extends this approach by examining whether study-level factors explain variation in outcomes across studies. This makes it necessary to extract not only the reported outcomes, but also the context in which they were observed. Because crop responses depend strongly on local soil, weather, and management conditions, a reported yield value is meaningful only when its experimental context is retained (Antle et al., 2017; Grassini et al., 2015). Meta-analysis in agricultural sciences therefore requires structured datasets that preserve not only numerical measurements, but also experimental designs, treatment comparisons, environmental conditions, and measurement units.

Constructing these datasets is particularly difficult for intercropping research, due to the lack of standardization and widely-accepted protocols, variations in equipment, and variability of metrics reported in each local study. A single trial may

[1]Code, prompts, and evaluation scripts are available at https://github.com/com3dian/meta_analysis_agents.

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 38–51
November 10, 2026 ©2026 Association for Computational Linguistics

report dozens of related measurements across intercropped plots and their corresponding sole-crop controls. Studies also differ widely in their objectives, terminology, management conditions, outcomes, and units (Connolly et al., 2001). The information required for one experimental record is often scattered throughout a paper: trial designs may be described in the text, numerical results reported in multi-level tables, and units or exceptions provided only in captions or footnotes. Although human experts can interpret and connect this information, manual curation requires substantial time and effort, creating a major bottleneck for large-scale evidence synthesis.

Earlier attempts to automate evidence synthesis relied mainly on task-specific NLP pipelines, with separate components for document filtering, entity recognition, and relation extraction (Callaghan et al., 2021; Sietsma et al., 2024; G. et al., 2023; Chebbi et al., 2024; Rezayi et al., 2022). Although effective for well-defined subtasks, these pipelines can propagate errors between stages and often struggle to assemble information into complete experimental records. Many relevant relationships extend beyond individual sentences or paragraphs and require information to be connected across the full document (Jain et al., 2020).

Large language models provide a more general approach: a single model can follow different extraction schemas, integrate broader context, and perform multiple extraction tasks (Kojima et al., 2023; Dagdelen et al., 2024). This flexibility has led to growing interest in LLM-based scientific information extraction (Dagdelen et al., 2024; Gupta et al., 2024) and agentic designs such as ReAct, plan-and-execute, and AFlow (Yao et al., 2023; Erdogan et al., 2025; Zhang et al., 2025). Such designs may help decompose complex tasks, but additional stages can also introduce context loss and error propagation. Whether they improve extraction from complex agricultural studies remains an open question.

In this work, we investigate whether generative AI can reliably extract complex intercropping trial data. We compare a direct zero-shot approach, a staged workflow, and a multi-agent system across six LLMs. We evaluate their outputs against human-curated records and test their ability to reproduce a published meta-analysis. Our main contributions are:

1. We formulate the structuring of complex agricultural trial data as an information extraction task and establish an annotated evaluation benchmark.
2. We compare three LLM-based approaches on multi-level trial structures and cross-table attribute binding.
3. We assess how extraction errors affect the reproduction of published meta-analytic findings.

2 Problem Statement

2.1 Structured Data Extraction

Given the textual content of a research paper extracted from a PDF, our task is to identify information relevant to a predefined schema and organize it into a set of structured records. Information available only in images is outside the scope of the task. A paper may contain zero, one, or multiple records, and the information needed for one record may appear in different parts of the paper.

Each record represents one experimental comparison and contains a fixed set of fields defined by the schema. These fields may describe the study conditions, treatments, outcomes, and observed effects. Their standardization varies across disciplines, from protocol-based clinical trials to more heterogeneous agricultural field studies (Chan et al., 2013; Eagle et al., 2017). The task requires extracting the correct values while preserving the connections among values from the same comparison.

2.2 Challenges Specific to Agricultural Studies

Intercropping experiments often span multiple sites, seasons, and treatments, requiring each mixture to be matched with two sole-crop comparators under comparable management. This is difficult when fertilizer inputs or planting densities differ among treatments (Yu et al., 2015; Liu et al., 2024). The production syndromes identified by Li et al. (2020) further show that crop composition, spatial design, management, timing, location, and yields must remain correctly linked. Outcomes may include crop-level yields, system-level yields, or indices such as the land equivalent ratio (Brooker et al., 2014), often reported under nested table headers and in different units. Breaking these relationships can produce incorrectly linked or omitted records and distort downstream analysis.

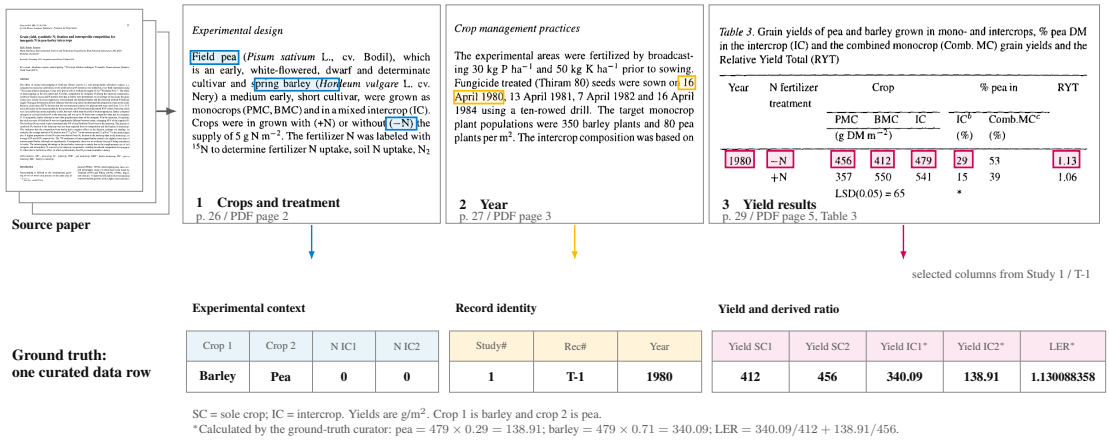


Figure 1: Example of structured data extraction from Jensen's pea–barley intercropping paper (Jensen, 1996), one of the 90 papers in our annotated corpus. Highlighted evidence from the experimental design, crop management, and yield table is linked to one ground-truth record. Asterisks mark values calculated from the reported table data.

3 Data & Experiment

This section introduces the data used and the experimental setup of this study. We use three generative-AI-based approaches to repeat the data-extraction stage of Yu et al. (2015), hereafter referred to as the reference study. We then reproduce the subsequent analysis by fitting the same linear mixed-effects model specified in the original R analysis to the LLM-extracted outputs.

3.1 Data

The reference study manually collected 746 records from 100 intercropping papers: the 50 most highly cited papers in the analysed corpus and 50 selected at random. Only 90 of the 100 paper folders could be reliably mapped to the corresponding ground-truth studies, so we conduct the extraction experiment on these 90 papers.

The source papers were preprocessed using MinerU (Wang et al., 2026) and converted from PDF into Markdown documents, which serve as the textual input to the three extraction approaches.

The complete ground-truth dataset contains 135 variables, including bibliographic and administrative information outside the extraction task. We evaluate the predefined subset of 42 extraction variables without modifying their definitions or ground-truth values.

3.2 Extraction Schema and Key Variables

We use the field definitions provided by the author of the reference study to clarify what information is contained in each intercropping record. Definitions of all 42 extracted fields are provided in Appendix Table 2. Among these fields, Temporal niche differentiation (TND) and land-equivalent ratio (LER) are central to the reproduced analysis.

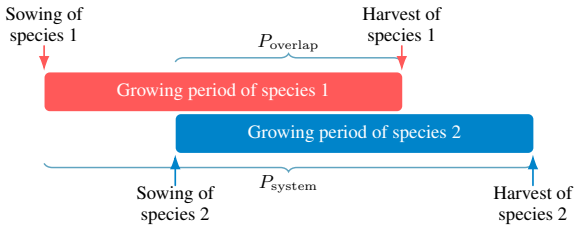


Figure 2: Illustration of temporal niche differentiation. Adapted from Yu et al. (2015).

Temporal niche differentiation measures the difference between the growing periods of the two crop species in the case of relay intercropping. Relay intercropping is the combination with two species with differences in growing period, which overlap only partially.

$$TND = 1 - \frac{P_{\text{overlap}}}{P_{\text{system}}}, \tag{1}$$

where P_{overlap} is the simultaneous growth period and P_{system} spans the period from the sowing of the first-sown crop till the harvest of the last-harvested crop. A TND of 0 indicates complete overlap, while higher values indicate greater temporal separation (Figure 2). The maximum value of TND is 1, indicating complete separation of growing periods.

The land-equivalent ratio measures how efficiently an intercropping system uses land relative to growing the two crop species separately. It is

calculated as

$$LER = \frac{Y_{\mathrm{IC1}}}{Y_{\mathrm{SC1}}} + \frac{Y_{\mathrm{IC2}}}{Y_{\mathrm{SC2}}}, \tag{2}$$

where Y_{IC1} and Y_{IC2} are species yields under intercropping, and Y_{SC1} and Y_{SC2} are their corresponding sole-crop yields. An LER above 1 indicates greater land-use efficiency under intercropping.

3.2.1 Analysis in the Reference Study

We reproduce the main analysis conducted in the reference study. This subsection describes that analysis and identifies the information required to reproduce it. The authors investigated the following research question:

> *Do crop combinations have a higher LER when their TND is larger, i.e. when they overlap (and compete) less?*

To answer this question, the authors modelled the relationship between temporal niche differentiation (TND) and the land-equivalent ratio (LER). They used a linear mixed-effects regression with random intercepts for experiments nested within studies:

$$\begin{aligned} LER_{ijk} = \beta_0 + \beta_1 TND_{ijk} + u_{\mathrm{Study},i} \\ + u_{\mathrm{Experiment},ij} + \epsilon_{ijk}. \end{aligned} \tag{3}$$

Here, β_0 is the expected LER at $\mathrm{TND} = 0$, and β_1 is the change in LER per unit increase in TND. The terms $u_{\mathrm{Study},i}$ and $u_{\mathrm{Experiment},ij}$ capture variation between studies and nested experiments, respectively, while ϵ_{ijk} denotes residual variation.

3.3 Experiment

3.3.1 Extraction Approaches

We designed three LLM-based approaches: direct zero-shot prompting, a staged workflow, and a multi-agent system (MAS).

The prompts were developed based on referenced research and expert knowledge of intercropping meta-analysis. We tested different prompt versions during pilot runs and selected those that showed the best performance and stability. For the staged workflow, we followed previous work on separating evidence identification from record construction. In the first step, a labeller identifies relevant values and tags the supporting text. In the second step, an extractor uses the labeled text to construct the final records.

During the prototyping of the agentic approach, we tested several designs, including ReAct (Yao et al., 2023) and Reflexion (Shinn et al., 2023). However, these repeated reasoning steps did not consistently improve the output and often accumulated errors. Reflexion showed a similar problem: even when previous error patterns were provided, the system did not reliably avoid them. We therefore selected a plan-and-execute design in which agents use tools to identify and label relevant evidence before producing and refining the records (Erdogan et al., 2025). Further details about the agents, prompts, and tools are provided in the Appendix B.

3.3.2 Language Model Choice

Regarding the LLM choice, we used open-weight models available through the SURF research infrastructure. This choice was motivated by concerns about sharing unpublished scientific work with commercial LLM providers and the possibility that related results could be pursued or released before the original researchers are ready to publish (Buckmaster, 2026). The six LLMs included QWEN3.6-27B-FP8, QWEN3.6-35B-A3B-FP8, QWEN3.5-122B-A10B-NVFP4, GEMMA-4-31B-IT-NVFP4, LLAMA-70B, and GPT-OSS-120B.

4 Evaluation

We evaluate the extraction approaches in two steps: comparing the extraction outputs against the ground truth using a rigorous evaluation pipeline, and reproducing the experiment conducted by human researchers in the original meta-analysis study using the extraction outputs. The first step aims to assess how well the LLM-based approaches agree with the human-annotated ground truth and how performance differs across models and methods. The second step examines whether the experiment can be reproduced using imperfect extraction outputs and how much the resulting findings differ from those reported by the human researchers.

4.1 Evaluation Against Ground Truth

We evaluate the outputs at the field level using comparison rules tailored to different data types. These rules were reviewed and agreed upon with the senior author of the reference study who was deeply involved in curation of the ground-truth dataset. We first introduce the types of extraction error and their corresponding metrics, and then describe the

evaluation pipeline used to align records, apply the comparison rules, and aggregate the results.

4.1.1 Error Types

The output of the three approach (direct, workflow, MAS) are saved in `csv` format, as is the ground truth. We classify the mismatches between the predictions and ground truth into three types, each measured by a corresponding metric:

- **Incorrectness**: Extracted values disagree with the corresponding ground-truth values in matched field values. Agreement is measured by the *similarity score*, which ranges from 0 to 1 and is averaged over fields where both records contain a value.
- **Hallucination**: A predicted field contains a value, but its ground-truth counterpart is missing. Populated fields in unmatched predicted records also count as unsupported extractions. The *factual rate* measures the proportion of populated predicted fields that have a populated ground-truth counterpart
- **Incompleteness**: A ground-truth field contains a value, but its predicted counterpart is missing. Populated fields in unmatched ground-truth records also count as omissions. The *completeness rate* measures the proportion of populated ground-truth fields that have a populated predicted counterpart.

4.1.2 Evaluation Metrics

Similarity Score We define a field-level similarity function that compares an extracted value with its corresponding ground-truth value and returns a score between 0 and 1. A score of 1 indicates full agreement under the applicable comparison rules, while 0 indicates no match. Intermediate scores represent partial agreement. The comparison depends on the field type, as follows:

- **Text and categorical values**: After lowercasing and removing whitespace, we compute character-level ROUGE-L F1 (Lin, 2004). Identical normalized strings receive 1, while partial overlap receives proportional credit.
- **Numerical values**: Values receive 1 when their relative difference is at most 2%; an exact match is required when the ground truth is zero. Values that match only after removing decimal points receive 0.5, reflecting a possible scale or unit-conversion error. All other mismatches receive 0.
- **Numerical values with units**: Compatible units are normalized and converted to the ground-truth unit before applying the 2% tolerance. If either unit is missing, we use the unit-free numerical comparison; incompatible or invalid quantity comparisons receive 0.
- **Dates**: We normalize calendar and Excel serial dates and sequentially compare full dates, partial month–day or month–year dates, relative days or ranges, and years. Applicable rules return 1 or 0; otherwise, we use character-level ROUGE-L F1.
- **Boolean and missing values**: Equivalent Boolean forms, such as *Yes*, *true*, and 1, are normalized before exact comparison. A comparison receives 0 if either value is missing, including when both are missing.

Factual Rate To assess hallucination error, we define the factual rate as the proportion of non-empty predicted fields that have a non-empty ground-truth counterpart after record alignment. This metric measures the presence of a corresponding value, regardless of whether the predicted and ground-truth values agree.

Completeness rate To assess omissions, we define the completeness rate as the proportion of non-empty ground-truth fields that have a non-empty predicted counterpart after record alignment. This metric measures how much of the ground-truth information is covered by the extraction, regardless of whether the values agree.

Combined Evaluation Score The three metrics capture complementary aspects of extraction quality. The similarity score S measures the correctness of aligned values, the factual rate F penalizes unsupported predictions, and the completeness rate C penalizes omitted information. To summarize these dimensions in a single score, we first calculate a similarity-adjusted factual rate:

$$F_{\text{adj}} = S \times F, \tag{4}$$

where S is the similarity score and F is the factual rate. We then combine F_{adj} with the completeness rate C using their harmonic mean:

$$E_{\text{combined}} = \frac{2F_{\text{adj}}C}{F_{\text{adj}} + C} = \frac{2SFC}{SF + C}. \tag{5}$$

The resulting score ranges from 0 to 1, with higher values indicating better overall extraction quality. Multiplying similarity by the factual rate

ensures that a populated predicted field contributes fully only when its value is also correct. The harmonic mean follows the principle of the F-measure and produces a high combined score only when both the similarity-adjusted factual rate and completeness rate are high (Schütze et al., 2008). We refer to this metric as the *similarity-adjusted F1 score*.

4.1.3 Evaluation pipeline

For each approach, we compare the extracted records with the ground-truth records from the same paper. The evaluation includes papers with an output CSV from at least one approach; missing outputs from other approaches are treated as empty sets of records.

We first compare every predicted record with every ground-truth record from the same paper. For each pair, we calculate record-level similarity as the arithmetic mean of the evaluated field scores. To account for interchangeable crop ordering, we retain the higher-scoring original or swapped crop-1/crop-2 assignment.

We sort all candidate record pairs by descending similarity and perform greedy one-to-one matching. A pair is accepted only if neither record has already been matched. No minimum similarity score is required.

After matching, each non-empty field in an unmatched predicted record is counted as an unsupported extraction, while each non-empty field in an unmatched ground-truth record is counted as an omission. For matched records, we evaluate field presence and agreement using the retained crop orientation. We calculate paper-level similarity, factual rate, and completeness rate, then average these metrics across papers for each approach.

4.2 Reproducing the Reference Analysis

Because the extracted datasets are imperfect, we investigate whether they nevertheless support the same or a similar scientific conclusion as the reference study. Automating meta-analysis should support not only accurate extraction but also the complete process from the literature to a final conclusion. We therefore examine how extraction errors propagate into the downstream analysis and whether they alter its findings.

For each model and extraction approach, we first construct the variables required for the regression. When a usable LER value is reported in an extracted record, we use that value directly. When LER is unavailable, we calculate it from the four required sole-crop and intercrop yields:

$$LER = \frac{Y_{\mathrm{IC1}}}{Y_{\mathrm{SC1}}} + \frac{Y_{\mathrm{IC2}}}{Y_{\mathrm{SC2}}}. \tag{6}$$

Records without either a reported LER or all four required yields are excluded. We similarly use the extracted TND when available or calculate it from the crops' sowing and harvest dates. Only records for which both LER and TND can be obtained are retained.

Within each study, the retained records are matched one-to-one with the reference records using experimental-design information only. LER, TND, yields, and dates are not used during this matching. The matched ground-truth Experimental ID is used solely to recover the nested grouping structure required by the statistical model; no ground-truth outcome values are introduced into the reproduced analysis.

We apply the linear mixed-effects model described in the previous section, with LER as the response, TND as the fixed-effect predictor, and random intercepts for experiments nested within studies, fitting it separately for each model and extraction approach.

LLM	Direct	Workflow	MAS
QWEN3.6-27B	48/0/42	41/5/44	50/12/28
QWEN3.6-35B-A3B	86/0/4	76/0/14	57/29/4
GEMMA-4-31B	69/1/20	65/0/25	74/6/10
QWEN3.5-122B	71/0/19	68/0/22	46/34/10
LLAMA-70B	74/1/15	73/3/14	55/22/13
GPT-OSS-120B	75/1/14	73/2/15	68/13/9

Table 1: Run outcomes across the 90 target papers. Each cell reports non-empty/empty/missing outputs. Direct denotes the direct LLM approach, Workflow the staged workflow, and MAS the multi-agent system.

5 Results

We applied all three extraction approaches with each LLM to the same 90 papers. We counted runs that produced a parsable table with at least one record. Individual LLM requests had a three-minute timeout, although multi-stage runs could take longer. Failures included API or parsing errors, empty tables, and missing outputs. Table 1 summarizes the resulting coverage.

For each model, the evaluation includes papers for which at least one approach produced an output CSV. Within this set, a missing or empty output

from a particular approach is treated as an empty prediction. Consequently, all non-empty ground-truth fields for that approach are counted as omissions.

5.1 Field-Level Extraction Results

We applied the evaluation pipeline described above to the outputs of each LLM and extraction approach. Figure 3 presents the mean similarity-adjusted F1 score for all papers included in the evaluation of each model. A paper is included when at least one of the three approaches produces an output CSV. For the included papers, a missing or empty output from an approach is evaluated as an empty prediction.

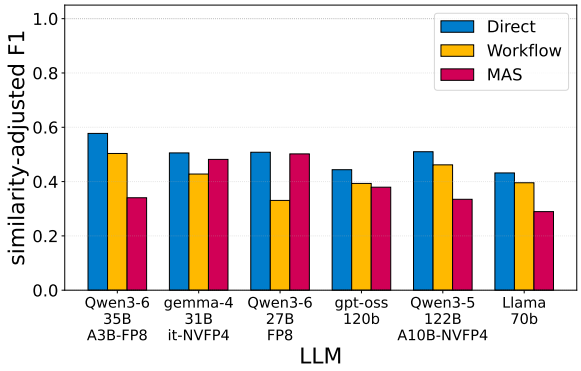


Figure 3: Mean similarity-adjusted F1 by LLM and extraction approach. Papers with output from at least one approach are included; missing or empty outputs are treated as empty predictions.

Figure 3 shows that the direct LLM approach achieves the highest score for all six models. Performance also varies across LLMs, but the most consistent pattern is the difference between extraction approaches: increasing method complexity does not improve extraction quality. Contrary to our initial expectation, the staged workflow and multi-agent system generally perform worse than the direct approach.

This comparison is partly influenced by differences in run coverage. As shown in Table 1, the direct approach produces more non-empty outputs overall. We therefore conduct an additional comparison restricted, for each model, to papers for which all three approaches produce at least one structured record. Figure 4 presents the results on this common-output subset.

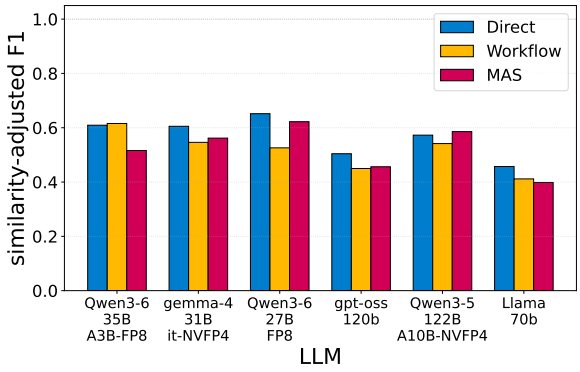


Figure 4: Mean similarity-adjusted F1 scores restricted, for each LLM, to papers for which all three extraction approaches produced at least one structured record.

The restricted comparison leads to a similar overall conclusion. The direct approach obtains the highest score for four of the six models. The workflow is slightly better for Qwen3.6-35B, while the multi-agent system is slightly better for Qwen3.5-122B. Thus, although the ranking is not identical for every model, the results provide no consistent evidence that increasing method complexity improves extraction quality.

5.2 Downstream Reproduction Results

We reproduce the reference analysis by fitting the same linear mixed-effects model specification separately to the data produced by each LLM and extraction method, as described in Section 4.2. Because extraction was conducted on 90 papers, we use the model fitted to the corresponding 90 ground-truth studies as the primary reference. This linear regression model estimates a positive TND effect of $\beta = 0.290$ with a 95% confidence interval of $[0.166, 0.413]$. We also show the effect reported for the complete reference dataset of 100 studies ($\beta \approx 0.211$).

Figure 5 shows the reproduced relationships. The direct approach produces a positive TND coefficient for all six LLMs, and five of these six estimates are significantly greater than zero. The Qwen3.6-35B model provides the most consistent qualitative reproduction: all three extraction methods produce significant positive coefficients. However, their estimated magnitudes are larger than the 90-study ground-truth coefficient. In contrast, the direct Qwen3.6-27B estimate ($\beta = 0.295$) is numerically closest to the ground truth, but its confidence interval includes zero. The GPT-OSS multi-agent fit produces a coefficient estimate but no valid standard error or confidence interval.

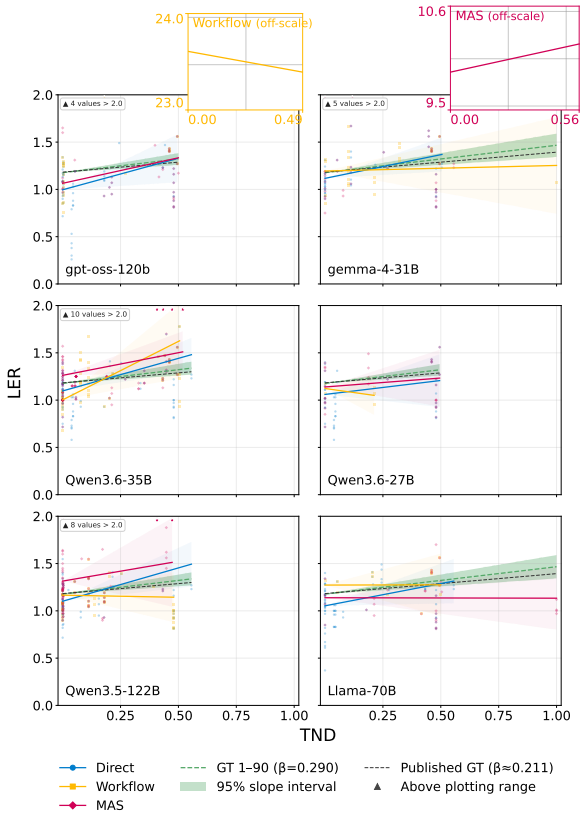


Figure 5: Estimated relationship between TND and LER for each LLM and extraction approach, compared with the ground-truth baselines. Insets show fits outside the main plotting range.

Overall, the automatically extracted data often recover the direction of the relationship, particularly when using the direct approach, but they do not reliably reproduce its magnitude or statistical certainty. The results therefore support the use of generative-AI extraction for qualitative exploration, but not as a substitute for human-curated data when precise quantitative conclusions are required.

Taken together, the field-level evaluation and reproduction experiment provide no consistent evidence that increasing method complexity improves extraction quality or downstream analytical reliability. We discuss possible explanations for this result in Section 6.1.

6 Discussion & Conclusion

Compared with the ground truth, direct zero-shot extraction was generally the strongest approach, although none of the approaches was fully accurate. It consistently recovered the positive relationship between TND and LER, but not the magnitude of the reference effect.

6.1 Complexity Trap

We manually examined the outputs for a fixed-seed sample of eight papers to better understand why the structured pipelines (workflow and MAS) performed less well on average. The two-step workflow sometimes lost context when papers were reduced to tagged evidence blocks, while MAS could carry mistakes from its initial direct extraction into later steps. In this sample, MAS produced eight records for a paper containing three ground-truth records, fifteen for a paper containing two, and no records for two other papers. Additional stages may help in some cases, but they also create more opportunities to omit, duplicate, or incorrectly associate evidence. The stronger overall performance of direct extraction may therefore be partly explained by its shorter processing path.

These observations align with reports of errors at agent handoffs (Lin et al., 2026), error amplification in agent systems (Kim et al., 2026), and mixed benefits of multi-agent decomposition on geometry benchmarks (E Sobhani et al., 2026). Complex pipelines therefore depend on a good fit with the task and careful checking of intermediate outputs.

6.2 Toward More Efficient Evidence Synthesis

Extracting data for an intercropping meta-analysis is time-consuming. Papers may report several experiments across different sites or years, with multiple treatments in each experiment. Researchers must therefore identify the appropriate intercrop and sole-crop treatments and decide which comparisons are scientifically meaningful.

The process is also iterative because intercropping meta-analysis is not yet fully standardized and new response metrics continue to be developed and evaluated. Researchers often move repeatedly between data retrieval, analysis, and conceptualization before arriving at an extraction and analysis strategy that fits the available evidence. Experience from conducting the reference meta-analysis suggests that a small meta-analysis, such as an MSc thesis project, may take about six months. A larger meta-analysis covering 50–150 studies may require two to three years of a PhD project (Yu et al., 2015). As experience in synthesizing intercropping research grows, there is an opportunity to standardize and streamline this process and make greater use of automated extraction.

Our results suggest that LLMs could reduce some of this manual work, but their outputs still re-

quire careful checking. Direct zero-shot extraction was generally the most reliable approach, although many records and fields remained missing, incorrect, or wrongly associated. This provides cautious support for earlier findings on LLM-based information extraction for evidence synthesis (Polak and Morgan, 2024). In the downstream analysis, direct extraction recovered a positive relationship between TND and LER across all six models, but the estimated effects were generally larger than the reference effect and varied in precision. These findings suggest that LLMs are currently more suitable for assisting researchers with data extraction than for fully automating the process without human verification.

LLMs may also reduce the time required for evidence synthesis by supporting study screening and full-text eligibility assessment, although their outputs must still be checked before drawing scientific conclusions (Delgado-Chaves et al., 2025; Taherinezhad et al., 2026; Wang et al., 2025).

6.3 Relevance Beyond Intercropping

Intercropping studies are particularly difficult to conduct meta-analysis on because of their nested experimental structure and the complex relationships among treatments, crops, and values reported across tables. Similar challenges occur in other domains, such as clinical trials (Nye et al., 2018) and materials-science datasets (Zheng et al., 2023).

The results from direct zero-shot prompting suggest that this approach may be transferred to other domains. However, its precision remains limited, and recovering the direction of a relationship appears easier than estimating its magnitude accurately. Domain-specific post-training may therefore be needed to improve performance.

Generative AI usage

AI coding tools (Claude, Cursor, Codex) were used to co-write parts of the code base and for minor copy-editing of the manuscript. All code has been manually verified and the authors take full responsibility of all presented methods and results.

Acknowledgements

This work was partially supported by the Horizon Europe project PHENET – Tools and Methods for Extended Plant Phenotyping and Envirotyping Services of European Research Infrastructures (Grant Agreement No. 101094587) and the Dutch Research Council (NWO) Large-Scale Research Infrastructures (LSRI) programme through the LTER-LIFE infrastructure (Grant No. 184.036.014).

References

John M. Antle, James W. Jones, and Cynthia E. Rosenzweig. 2017. Next generation agricultural system data, models and knowledge products: Introduction. *Agricultural Systems*, 155:186–190.

Rob W. Brooker, Alison E. Bennett, Wen-Feng Cong, Tim J. Daniell, Timothy S. George, Paul D. Hallett, Cathy Hawes, Pietro P. M. Iannetta, Hamlyn G. Jones, Alison J. Karley, Long Li, Blair M. McKenzie, Robin J. Pakeman, Eric Paterson, Christian Schöb, Jianbo Shen, Geoff Squire, Christine A. Watson, Chaochun Zhang, and 3 others. 2014. Improving intercropping: a synthesis of research in agronomy, plant physiology and ecology. *New Phytologist*, 206(1):107–117.

Tristan Buckmaster. 2026. Public statement. Accessed: 2026-09-11.

Max Callaghan, Carl-Friedrich Schleussner, Shruti Nath, Quentin Lejeune, Thomas R. Knutson, Markus Reichstein, Gerrit Hansen, Emily Theokritoff, Marina Andrijevic, Robert J. Brecha, Michael Hegarty, Chelsea Jones, Kaylin Lee, Agathe Lucas, Nicole van Maanen, Inga Menke, Peter Pfleiderer, Burcu Yesil, and Jan C. Minx. 2021. Machine-learning-based evidence and attribution mapping of 100, 000 climate impact studies. *Nature Climate Change*, 11(11):966–972.

An-Wen Chan, Jennifer M Tetzlaff, Douglas G Altman, Andreas Laupacis, Peter C Gøtzsche, Karmela Krleža-Jerić, Asbjørn Hróbjartsson, Howard Mann, Kay Dickersin, Jesse A Berlin, et al. 2013. Spirit 2013 statement: defining standard protocol items for clinical trials. *Annals of internal medicine*, 158(3):200–207.

Abir Chebbi, Guido Kniesel, Nabil Abdennadher, and Giovanna Dimarzo. 2024. Enhancing named entity recognition for agricultural commodity monitoring with large language models. In *Proceedings of the 4th Workshop on Machine Learning and Systems*, EuroSys '24, page 208–213. ACM.

J Connolly, HC Goma, and K Rahim. 2001. The information content of indicators in intercropping research. *Agriculture, ecosystems & environment*, 87(2):191–207.

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).

Fernando M. Delgado-Chaves, Matthew J. Jennings, Antonio Atalaia, Justus Wolff, Rita Horvath, Zeinab M. Mamdouh, Jan Baumbach, and Linda Baumbach. 2025. Transforming literature screening: The emerging role of large language models in systematic reviews. *Proceedings of the National Academy of Sciences*, 122(2).

Mahbub E Sobhani, Md. Faiyaz Abdullah Sayeedi, Mohammad Nehad Alam, Proma Hossain Progga, and Swakkhar Shatabda. 2026. Do multi-agents solve better than single? evaluating agentic frameworks for diagram-grounded geometry problem solving and reasoning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 27–47, Rabat, Morocco. Association for Computational Linguistics.

Alison J. Eagle, Laura E. Christianson, Rachel L. Cook, R. Daren Harmel, Fernando E. Miguez, Song S. Qian, and Dorivar A. Ruiz Diaz. 2017. Meta-analysis constrained by data: Recommendations to improve relevance of nutrient management research. *Agronomy Journal*, 109(6):2441–2449.

Lutfi Eren Erdogan, Nicholas Lee, Sehoon Kim, Suhong Moon, Hiroki Furuta, Gopala Anumanchipalli, Kurt Keutzer, and Amir Gholami. 2025. Plan-and-act: Improving planning of agents for long-horizon tasks. *Preprint*, arXiv:2503.09572.

Veena G., Vani Kanjirangat, and Deepa Gupta. 2023. Agroner: An unsupervised agriculture named entity recognition using weighted distributional semantic model. *Expert Systems with Applications*, 229:120440.

H. Charles J. Godfray, John R. Beddington, Ian R. Crute, Lawrence Haddad, David Lawrence, James F. Muir, Jules Pretty, Sherman Robinson, Sandy M. Thomas, and Camilla Toulmin. 2010. Food security: The challenge of feeding 9 billion people. *Science*, 327(5967):812–818.

Patricio Grassini, Lenny G.J. van Bussel, Justin Van Wart, Joost Wolf, Lieven Claessens, Haishun Yang, Hendrik Boogaard, Hugo de Groot, Martin K. van Ittersum, and Kenneth G. Cassman. 2015. How good is good enough? data requirements for reliable crop yield simulations and yield-gap analysis. *Field Crops Research*, 177:49–63.

Sonakshi Gupta, Akhlak Mahmood, Pranav Shetty, Aishat Adeboye, and Rampi Ramprasad. 2024. Data extraction from polymer literature using large language models. *Communications Materials*, 5(1).

Jessica Gurevitch, Julia Koricheva, Shinichi Nakagawa, and Gavin Stewart. 2018. Meta-analysis and the science of research synthesis. *Nature*, 555(7695):175–182.

Dean P. Holzworth, Neil I. Huth, Peter G. deVoil, Eric J. Zurcher, Neville I. Herrmann, Greg McLean, Karine Chenu, Erik J. van Oosterom, Val Snow, Chris Murphy, Andrew D. Moore, Hamish Brown, Jeremy P.M. Whish, Shaun Verrall, Justin Fainges, Lindsay W. Bell, Allan S. Peake, Perry L. Poulton, Zvi Hochman, and 27 others. 2014. Apsim – evolution towards a new generation of agricultural systems simulation. *Environmental Modelling & Software*, 62:327–350.

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.

Erik Steen Jensen. 1996. Grain yield, symbiotic n2 fixation and interspecific competition for inorganic n in pea-barley intercrops. *Plant and soil*, 182(1):25–38.

Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. 2026. Towards a science of scaling agent systems. *Preprint*, arXiv:2512.08296.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Chunjie Li, Ellis Hoffland, Thomas W. Kuyper, Yang Yu, Chaochun Zhang, Haigang Li, Fusuo Zhang, and Wopke van der Werf. 2020. Syndromes of production in intercropping impact yield gains. *Nature Plants*, 6(6):653–660.

Bohan Lin, Kuo Yang, Zelin Tan, Yingchuan Lai, Chen Zhang, Guibin Zhang, Xinlei Yu, Miao Yu, Xu Wang, Yudong Zhang, and Yang Wang. 2026. Agentask: Multi-agent systems need to ask. *Preprint*, arXiv:2510.07593.

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.

Yalin Liu, TjeerdJan Stomph, Fusuo Zhang, Chunjie Li, and Wopke van der Werf. 2024. Nitrogen input strategies impact fertilizer nitrogen saving by intercropping: a global meta-analysis. *Field Crops Research*, 318:109607.

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)*, page 197–207. Association for Computational Linguistics.

Maciej P. Polak and Dane Morgan. 2024. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1).

Saed Rezayi, Zhengliang Liu, Zihao Wu, Chandra Dhakal, Bao Ge, Chen Zhen, Tianming Liu, and Sheng Li. 2022. Agribert: Knowledge-infused agricultural language models for matching food and nutrition. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, IJCAI-2022, page 5150–5156. International Joint Conferences on Artificial Intelligence Organization.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Anne J. Sietsma, Emily Theokritoff, Robbert Biesbroek, Iván Villaverde Canosa, Adelle Thomas, Max Callaghan, Jan C. Minx, and James D. Ford. 2024. Machine learning evidence map reveals global differences in adaptation action. *One Earth*, 7(2):280–292.

Moein Taherinezhad, Sebastian Maier, Gerardo Vitagliano, Francesco Pierri, and Stefan Feuerriegel. 2026. Autosynthesis: An agentic system for automated meta-analysis. *arXiv preprint*.

Giovanni Tamburini, Riccardo Bommarco, Thomas Cherico Wanger, Claire Kremen, Marcel GA Van Der Heijden, Matt Liebman, and Sara Hallin. 2020. Agricultural diversification promotes multiple ecosystem services without compromising yield. *Science advances*, 6(45):eaba1715.

Bin Wang, Tianyao He, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Tao Chu, Yuan Qu, Zhenjiang Jin, Weijun Zeng, Ziyang Miao, et al. 2026. Mineru2. 5-pro: Pushing the limits of data-centric document parsing at scale. *arXiv preprint arXiv:2604.04771*.

Zifeng Wang, Lang Cao, Qiao Jin, Joey Chan, Nicholas Wan, Behdad Afzali, Hyun-Jin Cho, Chang-In Choi, Mehdi Emamverdi, Manjot K. Gill, Sun-Hyung Kim, Yijia Li, Yi Liu, Yiming Luo, Hanley Ong, Justin F. Rousseau, Irfan Sheikh, Jenny J. Wei, Ziyang Xu, and 5 others. 2025. A foundation model for human-ai collaboration in medical literature mining. *Nature Communications*, 16(1).

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Yang Yu, T. J. Stomph, David Makowski, and Wopke van der Werf. 2015. Temporal niche differentiation increases the land equivalent ratio of annual intercrops: A meta-analysis. *Field Crops Research*, 184:133–144.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. 2025. Aflow: Automating agentic workflow generation. *Preprint*, arXiv:2410.10762.

Zhiling Zheng, Oufan Zhang, Christian Borgs, Jennifer T. Chayes, and Omar M. Yaghi. 2023. Chatgpt chemistry assistant for text mining and the prediction of mof synthesis. *Journal of the American Chemical Society*, 145(32):18048–18062.

A Schema Definition

Table 2: Definitions of the 42 fields in an extracted intercropping record.

Field	Definition
Year of data	Year or years in which the experimental data were collected.
Duration of experiment	Total duration of the experiment, expressed in days, years, or growing seasons.
Experimental design	Experimental layout used in the study, such as a randomized complete block design.
Sowing date 1	Date on which the first crop species in the intercropping system was sown.
Sowing date 2	Date on which the second crop species in the intercropping system was sown.
Harvest date 1	Date on which the first crop species was harvested.
Harvest date 2	Date on which the second crop species was harvested.
TND	Temporal niche differentiation index explicitly reported in the paper.
Lat	Latitude of the experimental site in decimal degrees.
Lon	Longitude of the experimental site in decimal degrees.
Continent	Continent in which the experimental site is located.
Crop species 1	Common or scientific name of the first crop species in the intercropping system.
Crop species 2	Common or scientific name of the second crop species in the intercropping system.
Crop type 1	Agronomic functional type of the first crop, such as cereal, legume, oilseed, or root crop.
Crop type 2	Agronomic functional type of the second crop, such as cereal, legume, oilseed, or root crop.
Fodder crop 1	Boolean indicator of whether the first crop is grown for fodder, forage, or biomass rather than for grain or seed.
Fodder crop 2	Boolean indicator of whether the second crop is grown for fodder, forage, or biomass rather than for grain or seed.
Intercropping pattern	Spatial arrangement of the crops, classified as row, strip, or mixed intercropping.
Density ic 1	Plant density of the first crop species in the intercropping treatment.
Density ic 2	Plant density of the second crop species in the intercropping treatment.
Density sc 1	Plant density of the first crop species when grown as a sole crop.
Density sc 2	Plant density of the second crop species when grown as a sole crop.
RDT	Relative density total of the intercropping system when explicitly reported.
N input SC1	Amount of nitrogen fertilizer applied to the sole-crop treatment of the first crop species.
N input SC2	Amount of nitrogen fertilizer applied to the sole-crop treatment of the second crop species.
N input IC1	Amount of nitrogen fertilizer attributed to the first crop species in the intercropping treatment.
N input IC2	Amount of nitrogen fertilizer attributed to the second crop species in the intercropping treatment.
N total in IC	Total amount of nitrogen fertilizer applied to the intercropping treatment.
N Unit	Unit used to report nitrogen fertilizer inputs, such as kg N ha^{-1}.
Replications SC1	Number of replicates for the sole-crop treatment of the first crop species.
Replications SC2	Number of replicates for the sole-crop treatment of the second crop species.
Replications IC1	Number of replicates for the first crop species in the intercropping treatment.
Replications IC2	Number of replicates for the second crop species in the intercropping treatment.
Data source	Location of the extracted values in the publication, such as a table, figure, or supplementary material.
unified yield sc 1	Yield of the first crop species when grown as a sole crop. Grain yield is used for grain crops, while total dry-matter or biomass yield is used for fodder crops.
unified yield sc 2	Yield of the second crop species when grown as a sole crop. Grain yield is used for grain crops, while total dry-matter or biomass yield is used for fodder crops.
unified yield ic 1	Intercrop yield of the first crop species. For multi-cut forage crops, this is the cumulative seasonal yield across cuts.
unified yield ic 2	Intercrop yield of the second crop species. For multi-cut forage crops, this is the cumulative seasonal yield across cuts.
Yield unit	Common unit used for the four sole-crop and intercrop yield fields, such as t ha^{-1}, kg ha^{-1}, or g m^{-2}.
PLER 1	Partial land-equivalent ratio of the first crop species when explicitly reported.
PLER 2	Partial land-equivalent ratio of the second crop species when explicitly reported.
LER	Total land-equivalent ratio of the intercropping treatment when explicitly reported.

B Prompts and Agent Configurations

B.1 Direct Zero-Shot Prompt

The direct approach used the following prompt template. At runtime, `[SCHEMA]` was replaced by the complete 42-field schema and `[PAPER CONTENT]` by the paper text. The model returned structured output conforming to the same schema.

> You are an expert agricultural data extraction specialist. Your task is to extract measured crop yield information and related agronomic and contextual variables from scientific research papers. Maximize recall: your output list should include every distinct experiment-level or table-level data row supported by the paper, rather than a summary.
>
> **Step 1: Anchor identification**
>
> Scan the entire content, including paragraphs, all tables row by row, captions, footnotes, and supplements. Locate every sentence or cell reporting measured yields, densities, inputs, or other schema fields. Ignore model evaluation metrics, yield gaps, and simulated outputs.
>
> **Step 2: Contextual reasoning**
>
> For each anchor, gather the associated context, including year, treatment, fertilization level, cropping system, species, and block or plot when reported.
>
> **Step 3: Completeness, evidence, and confidence**
>
> If a field is missing from a record, search the following locations before leaving it null:
>
> - Year, season, or duration: abstract, methods, table-row labels, and column headers.
> - Location or coordinates: site description, methods, and tables.
> - Crops, species, or intercropping pattern: abstract, methods, and table or figure labels.
> - Treatments: experimental design, methods, and table columns.
> - Densities, inputs, and nutrients: methods and numerical table cells.
> - Yield columns: map sole-crop and intercrop yields to the corresponding schema fields for each row.
>
> Assign confidence as high, medium, or low, but still output the row when its values are supported by the paper.
>
> **Step 4: Record construction**
>
> Build one schema record for each distinct experimental observation:
>
> - Treat each row of a main results table, such as a distinct year $\times$ treatment $\times$ system combination, as a separate record.
> - When a table contains separate columns for the two species and their intercrop yields, fill all applicable yield fields for that row.
> - For factorial or split-plot experiments, create a separate record for each factor combination with its own numerical result.
> - Keep the original units and do not convert them. Use `null` only when a field is not reported for that row.
>
> Remove duplicates only when two records have the same year, treatment level, cropping system, and species context. When uncertain, keep both records.
>
> **Target data**
>
> Include field-measured yields, dry matter, densities, nutrient applications, and other values defined by the schema. Exclude model evaluation metrics, predictions, and correlation-only statistics.
>
> Extract all rows supported by the main results tables. The number of records should generally be similar to, or greater than, the number of distinct table rows and relevant treatment levels. For intercropping tables, produce separate records or fill the sole-crop and intercrop yield columns for each row. Dry-matter measurements reported in units such as $\mathrm{g\,m^{-2}}$ and $\mathrm{kg\,ha^{-1}}$ are valid.
>
> **Meta-analytic schema**
>
> `[SCHEMA]`
>
> Analyze the following paper and extract all records according to the schema:
>
> `[PAPER CONTENT]`
>
> Extract all records from the paper following the schema. Prefer too many distinct records over too few.

B.2 Staged Workflow Prompts

The staged workflow consisted of two steps. The first step selected and tagged schema-relevant evidence from the paper. The second step used the tagged text to produce the structured records.

B.2.1 Step 1: Evidence Labelling

The following prompt was used for the labelling step. `[SCHEMA FIELDS]` was replaced by the names and descriptions of all schema fields, and `[PAPER CONTENT]` was replaced by the paper text.

> You are a document labeller for downstream structured extraction.
>
> **Task:** Produce a single field `labeled_document` containing only the paragraphs, table blocks, captions, and footnotes that hold schema-relevant evidence. Drop all other text. Within each retained block, add inline XML tags so that the next step can produce one output record per experimental row.
>
> **Schema fields:** Tag names must match these field names exactly:
>
> `[SCHEMA FIELDS]`
>
> **Keep:**
>
> - Yield, density, nutrient, experimental-design, crop, site, year, and treatment values.
> - Results tables together with their captions and footnotes. Retain the complete table rather than a summary.

- Methods sentences reporting densities, fertilizer inputs, experimental design, sowing or harvest dates, or geographical coordinates.
- Any sentence containing a schema-relevant number or label.

Drop:

- Sections without schema-relevant values, such as general background, acknowledgements, references, and unrelated discussion.
- Do not rewrite or paraphrase retained text. Copy the source wording and only add XML tags.

Tagging rules:

- Wrap every schema-relevant number or label using `<FieldName>exact source text</FieldName>`.
- In tables, tag every cell that corresponds to a schema field, especially cells in individual yield, density, and input rows.
- Tag distinct years, treatments, and nitrogen levels so that records can be separated in the next step.
- Do not invent text; only wrap spans found in the paper.
- Preserve the table structure and separate retained blocks with a blank line.

Input document:

`[PAPER CONTENT]`

Return only one structured object containing `labeled_document`.

B.2.2 Step 2: Record Extraction

The second step reused the direct zero-shot prompt in Appendix B.1, with the labelled output from the first step supplied as the paper content. The following instruction was added before that prompt:

> The paper below includes XML tags in the form `<FieldName>verbatim text</FieldName>`, where the tag name matches a schema field. Use these tags as evidence, but also examine every table row and caption in the tagged text to maximize the number of complete records. The absence of a tag does not necessarily mean that the corresponding information is absent.

The following completeness instructions were appended:

> **Output completeness:**
>
> - Produce as many records as the evidence supports, typically one record for each distinct row in the main results tables.
> - For intercropping tables, populate all available yield columns for each row rather than combining the table into a small number of summary records.
> - Do not merge rows that differ by year, treatment, or nitrogen level unless the schema explicitly requires an aggregate record.
> - Examine the complete tagged document, including the methods, results, and all retained tables.
> - Use `null` only when a field is genuinely unavailable after examining all relevant evidence.
> - When uncertain between one combined record and several records, prefer separate records that preserve the structure of the source table.

B.3 Agents and Tools

The MAS used a planning agent followed by four execution agents. Table 3 summarizes their roles and tool access. Only the labeller used a callable tool; the other agents operated on the document and intermediate outputs provided in their context.

Table 3: Agents used in the MAS pipeline.

Agent	Description
Planner	Creates the four-step execution plan and assigns each step to the corresponding agent.
Value identifier	Scans the complete paper and identifies values associated with fields in the extraction schema.
Labeller	Uses the `xml_tag_from_field_values` tool to insert field-specific XML tags around the identified values while preserving the original document.
Direct extractor	Applies the direct zero-shot extraction procedure to the original paper and produces an initial set of structured records.
Record extractor	Refines the initial records using the XML-labelled document, filling missing fields and correcting values when supported by the tagged evidence.

HCMUS_Beyonders at AstroCLIMB: Same Paper or Citing Paper? Linking Astronomy Figures with OCR, Citation Graphs, and LLMs

My-Linh Bang[1,2,†] **and** **Phu-Quy Nguyen-Lam**[1,2,†] **and** **Duy-Minh Dao-Sy**[1,2]
Trung-Kiet Huynh[1,2] and **Phu-Hoa Pham**[1,2] and **Chi-Nguyen Tran**[1,2]
[1]Faculty of Information Technology, University of Science (HCMUS), Vietnam
[2]Vietnam National University - Ho Chi Minh City (VNU-HCM), Vietnam
{23122009, 23122048}@student.hcmus.edu.vn
{23122041, 23122039, 23122030, 23122044}@student.hcmus.edu.vn
[†]Equal contribution

Abstract

AstroCLIMB asks whether two objects taken from astronomy papers, each a figure or a caption, show the same figure, come from the same paper, come from papers linked by a citation, or are unrelated. We describe the two-stage system of team HCMUS_Beyonders, which reached a macro-F1 of 0.7421 on the full test set. In the first stage, cross-encoders read both objects of a pair jointly. Each figure is represented by the text that optical character recognition (OCR) extracts from it, an input that scored higher than every image-embedding and image-patch input we tried. Because test papers are unseen, the cross-encoders are pretrained on 600K pairs generated from the citation graph of the public AstroCLIMB corpus. An error analysis shows that this stage most often confuses pairs from the same paper with pairs from citing papers, and that none of the cues we tested separates the two. The second stage therefore adds a different kind of model rather than more features: a Qwen3-8B large language model (LLM) fine-tuned with low-rank adaptation (LoRA). On its own it scores lower than the first stage, but over a third of each stage's errors are not made by the other. Averaging the two raises held-out macro-F1 by 0.016 (95% confidence interval $[+0.002, +0.030]$) and test macro-F1 from 0.7072 to 0.7329. Adding a second LLM, Gemma-4-12B, without held-out validation brings the test score to 0.7421, a gain within test noise.

1 Introduction

AstroCLIMB (Grezes et al., 2026), which follows earlier shared tasks of this workshop series (Grezes et al., 2022, 2023, 2025), asks a system to relate two objects taken from astronomy papers, where an object is a figure image or a figure caption. The system must decide whether the two objects are the same figure (*same_figure*), come from the same paper (*same_paper*), come from two papers connected by a citation (*related_papers*), or are unrelated (*unrelated_papers*). Such links support figure-level search in the NASA Astrophysics Data System, which released the data.

The task is difficult for three reasons. First, most astronomy figures are plots, so encoders trained on photographs capture their style but not their content, which lives in axis labels, object names, and legends. Second, only 10,000 labelled pairs exist, and the test pairs come from papers absent from both the training set and the public AstroCLIMB corpus (§2), so a model must generalise to unseen papers. Finally, a paper and the papers it cites usually share a topic, and a pair carries no title, author, or identifier that would separate them.

We address the first two difficulties in a first stage built on text-centred cross-encoders, models that read the two objects of a pair as one input sequence. Instead of embedding figure pixels, we read the text inside each figure with optical character recognition (OCR), and a BERT-style model (Devlin et al., 2019) reads it together with the other object. Before training on the task data, we pretrain these models on 600K labelled pairs generated from the citation graph of the public corpus.

An error analysis of this first stage shows that the remaining gap is the third difficulty: it most often confuses *same_paper* with *related_papers*, failing on *same_paper* pairs that look different and *related_papers* pairs that look alike. None of the further cues we tested separates them, including object names, cross-references, and a contrastive objective that removes topic. The second stage therefore adds a different kind of model instead of another feature: large language models (LLMs) fine-tuned with low-rank adaptation (LoRA). Their errors only partly coincide with those of the first stage; we do not attribute this to their size (Appendix G).

Our contributions are: (1) evidence that OCR text represents astronomy figures better than the image embeddings and image patches we tried,

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 52–60
November 10, 2026 ©2026 Association for Computational Linguistics

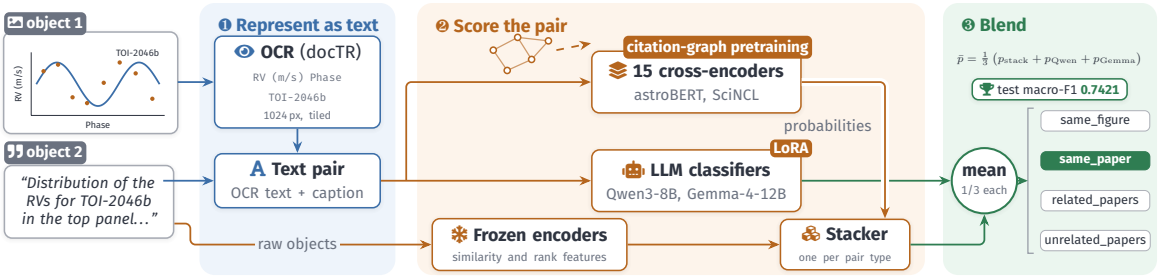


Figure 1: System overview. Images enter the text models as OCR text. In the first stage, cross-encoder probabilities and similarities from frozen (not fine-tuned) encoders feed a stacker; the second stage averages its output with two LLMs. The label is the argmax, with *same_figure* allowed only for image–caption pairs.

which did not include a vision–language LLM; (2) citation-graph pretraining, which raised test macro-F1 by 0.022; (3) an error analysis that locates the dominant error, and a blend with two LoRA-tuned LLMs that raises test macro-F1 from 0.7072 to 0.7421. Our code is available at `https://github.com/NLPQuy/WASP_2026`. Related work is discussed in Appendix I.

2 Task and Data

Each of the 10,000 training and 10,000 test pairs has one of three *pair types*: image–image, image–caption, or caption–caption. *same_figure* can only hold for image–caption pairs, which leaves ten allowed combinations of pair type and label; each has exactly 1,000 training pairs. Systems are ranked by macro-F1 over the four classes on the full test set. The leaderboard shows this *test score* after every submission, with no hidden split (Appendix D).

Test papers appear neither in the training set nor in the *public corpus* (Grezes et al., 2026), the only extra data we use: a release of 94,233 figures with their captions, paper DOIs, and the reference and citation lists of each paper. Test objects cannot be looked up there: only 2 of 5,256 test captions match a public caption, whereas all 9,325 training captions do, and 0.5% of test images match a public image by perceptual hash, against 38.5% of training images.

3 System

Figure 1 shows the system, which we describe in the order we built it. The first stage is the *cross-encoder system* (§3.1–3.3); the second averages its output with two LLM classifiers (§3.4), and we call this average the *blend*. Because the relation is symmetric, every text model reads both orders of a pair and averages the two predictions.

3.1 Representing figures as text

We read each figure with docTR (Mindee, 2021), using a DB-ResNet50 detector (Liao et al., 2020) and a PARSeq recogniser (Bautista and Atienza, 2022). The image is resized to 1024 pixels and split into tiles so that small tick labels are recovered. An image object is then represented by its OCR text and a caption object by its caption, so every pair is a pair of texts.

On an early version of the system, a cross-encoder that read only text reached 0.556 cross-validated macro-F1. Every model that received the image itself reached at most 0.533, whether as a pooled embedding or patch tokens from SigLIP 2 (Tschannen et al., 2025) or the chart encoder MatCha (Liu et al., 2023) (Appendix E). We did not test a vision–language LLM that reads the raw figure, so this finding is limited to those image inputs. Reading figures at 1024 instead of 512 pixels raised the test score by 0.011, while the score on caption–caption pairs, which contain no image and serve as a control, did not change.

3.2 Cross-encoders with citation-graph pretraining

A cross-encoder reads the two objects as one sequence of at most 512 tokens, so that each object can attend to the other, and predicts the four classes from the first token. We initialise from astroBERT (Grezes et al., 2021), whose vocabulary is built from astronomy text, and from SciNCL (Ostendorff et al., 2022), which is trained on citation neighbourhoods.

Before task fine-tuning, each cross-encoder is pretrained for one epoch on 600K pairs generated from the public corpus, 60K for each allowed combination of pair type and label. *same_figure* pairs join the OCR text of a figure with its own caption, *same_paper* pairs join two objects of one paper,

related_papers pairs join objects of two papers linked by a reference, and *unrelated_papers* pairs join objects of unlinked papers. Half of the *unrelated_papers* pairs are drawn from topically similar papers, so that similarity alone does not reveal the label. No cross-encoder is pretrained on a paper from the fold on which it is evaluated.

Pretraining, which targets generalisation to unseen papers, raised the test score by 0.022. It first imitated public image objects with *pseudo-OCR* text: captions with words dropped to resemble OCR output. Using real OCR text of the public figures instead added a further 0.038 to the cross-validated macro-F1 of a single cross-encoder, all of it in the two pair types with images.

The final *ensemble* averages 15 cross-encoders that differ in backbone, pretraining data, OCR input (whole image or tiles), and maximum length (256 or 512 tokens). The seven members pretrained on real OCR text get half of the total weight, the eight others the rest.

3.3 Stacker

The *stacker* is a small second-level classifier, one per pair type, that combines the ensemble probabilities with similarity and rank features from seven frozen embedding spaces, such as SigLIP 2 and SciNCL (Appendix H). It adds 0.004 macro-F1 over the ensemble on held-out data, a gain within noise (§4).

3.4 LLM classifiers and blending

The error analysis of the cross-encoder system found no further cue to add (§5), and adding more cross-encoders of the same kind had begun to lower the cross-validated score. We therefore added models of a different kind that read the same text input. Qwen3-8B (Yang et al., 2025) reads both objects and predicts the four classes through a new classification head; we train LoRA adapters (Hu et al., 2022) and the head for two epochs (Appendix A), without citation-graph pretraining. Gemma-4-12B (Gemma Team, 2026), from another model family, is trained the same way. These LLMs differ from the cross-encoders in size, pretraining data, and architecture at once, so we describe their contribution as making different errors, not adding capacity (Appendix G).

The final label is the argmax of the equal-weight average of the stacker and two LLM probabilities, with *same_figure* disallowed outside image–caption pairs.

Table 1: Selected development steps; each row adds to the row above (all steps in Appendix B). Below the rule: blends with stacker:Qwen:Gemma weights. CV: five-fold cross-validated macro-F1 (†: fold 1 only, where the row above scores 0.6991); Test: macro-F1 on the full test set. Shaded: gain above the test noise threshold of 0.014.

System	CV ↑	Test ↑
Frozen embeddings, rank features	0.5087	0.4972
+ OCR text features	0.5584	0.5557
+ symmetric cross-encoder	0.5908	0.5810
+ citation-graph pretraining	0.5979	0.6030
+ OCR at 1024 px, minor changes	0.6100	0.6193
+ astroBERT and SciNCL ensemble	0.6585	0.6695
+ 512-token inputs	0.6697	0.6795
+ real OCR text in pretraining	0.6976	0.7058
+ learned image projection features	0.7029	0.7072
+ Qwen3-8B (1:1)	0.7153†	0.7329
+ Gemma-4-12B (2:1:1)	–	0.7397
+ Gemma-4-12B (1:1:1, final)	–	**0.7421**

4 Experiments

Setup. We use five-fold cross-validation stratified by pair type and label, with the same folds for every component. *Held-out* predictions come from the model that did not train on the pair's fold, and CV is the macro-F1 over all of them. Unlike the test set, our folds share papers, so CV undervalues some changes (Appendix D). Qwen3-8B was validated on one fold only (fold 1, 2,000 pairs) and Gemma-4-12B not at all; each submitted LLM was trained once on all 10,000 pairs. A bootstrap gives the *noise threshold*: at the 95% level, two test scores must differ by 0.014 to be distinguishable, and two configurations compared on the same held-out pairs by about 0.008 (Appendix D). Table 1 traces the development of the system; Appendix C lists changes that did not help.

LLM blend. On fold 1, Qwen3-8B alone scores 0.6759, below the cross-encoder system (0.6991). Yet averaging the two with equal weights raises macro-F1 to 0.7153 (+0.016, paired 95% CI $[+0.002, +0.030]$) and the test score from 0.7072 to 0.7329. This is the last step supported by held-out data. Gemma-4-12B, which agrees with Qwen3-8B on only 79.4% of test labels, had no held-out predictions, so we submitted two weightings fixed in advance (Table 1) and report the higher score. That choice was made on the test set. It is worth 0.002 and the second LLM at most 0.009, both below the 0.014 threshold, so neither is established (Appendix D).

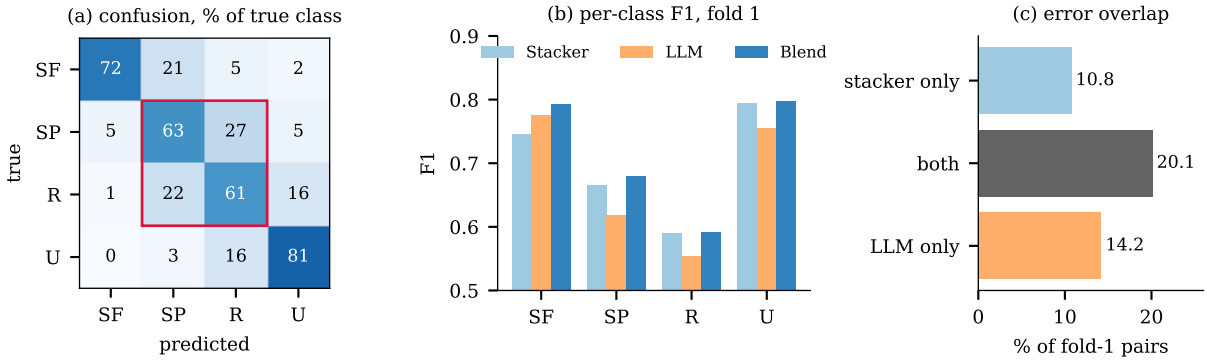


Figure 2: (a) Confusion matrix of the cross-encoder ensemble on held-out predictions for all 10,000 training pairs (% of true class; SF, SP, R, U abbreviate the labels). (b) Per-class F1 on fold 1. (c) Fold-1 pairs misclassified by one model or by both. "Stacker" is the cross-encoder system; "LLM" is Qwen3-8B.

5 Analysis

The dominant error of the first stage is *same_paper* versus *related_papers*. In Figure 2a, 27% of *same_paper* pairs are predicted as *related_papers* and 22% of *related_papers* pairs as *same_paper*; together these errors affect 14.7% of all pairs, against 9.9% for *related_papers* versus *unrelated_papers*. They are equally common in caption–caption pairs, which contain no image, so they are not a vision problem.

The errors follow similarity. We grouped the caption–caption pairs of these two classes by the word overlap of their captions, a proxy for topical similarity. From the least to the most similar group, the share of *same_paper* pairs rises from 32% to 90%, so similarity is valid evidence. The error rate falls from 42% to 10%. Errors thus occur on the pairs where this evidence misleads: *same_paper* pairs that look different and *related_papers* pairs that look alike.

No further cue we tested separates the classes. For separating *same_paper* from *related_papers*, shared object identifiers reach an area under the ROC curve (AUC) of 0.56, cross-references such as "same as Fig. N" 0.50, self-references such as "our sample" 0.50, and author–year citations 0.48. Chance is 0.50, and the cross-encoder reaches 0.85. A contrastive model trained with negatives from citing papers did remove topic from its embedding (correlation with topical similarity fell from 0.43 to 0.24), yet its *same_paper*-versus-*related_papers* AUC dropped by 0.01. Topical similarity is itself the main evidence for this distinction, so removing it discards signal (Appendix F).

The LLM makes partly different errors. On fold 1, 20.1% of pairs are misclassified by both the cross-encoder system and Qwen3-8B, 10.8% by the cross-encoder system only, and 14.2% by Qwen3-8B only (Figure 2c). The errors are correlated: independent errors would coincide on only 10.6% of pairs. Still, 35% of the cross-encoder system's errors and 41% of Qwen3-8B's are not made by the other model, and these are the pairs that averaging can correct. The blend improves *same_figure* and *same_paper* most (Figure 2b), and *related_papers* remains the hardest class.

6 Conclusion

Our AstroCLIMB system reads figures as text, pretrains cross-encoders on citation-graph pairs, and blends them with two LoRA-tuned LLMs, reaching 0.7421 test macro-F1. Its most frequent error, telling a paper from the papers it cites, was reduced by averaging with LLMs that make different errors, not by further cues. Whether a vision–language LLM would beat OCR text remains open.

Limitations

The Qwen3-8B blend was validated on a single fold of 2,000 pairs, Gemma-4-12B was not validated, and the final weights were chosen between two test submissions, so the final score is slightly optimistic. The last result supported by held-out data is the Qwen3-8B blend (0.7329). Each LLM was trained once with one random seed. We compared OCR text only with image embeddings and patch tokens from encoders up to the size of SigLIP 2. A vision–language LLM that reads the raw figure was not tested and may change that conclusion. The LLMs differ from the cross-encoders in size, pretraining

data, and architecture at once, so our results do not show which of these makes their errors different. Because figures enter the model only as OCR text, figures with little or no text, such as images of the sky, contribute little signal. Each LLM requires a 40 GB GPU, which is more than the organisers expected participants to use. DINOv3 is distributed under a licence not approved by the Open Source Initiative and was used only to compute frozen research features.

References

Darwin Bautista and Rowel Atienza. 2022. Scene text recognition with permuted autoregressive sequence models. In *Computer Vision – ECCV 2022*. Springer.

Bishwaranjan Bhattacharjee, Aashka Trivedi, Masayasu Muraoka, Muthukumaran Ramasubramanian, Takuma Udagawa, Iksha Gurung, Nishan Pantha, Rong Zhang, Bharath Dandala, Rahul Ramachandran, Manil Maskey, Kaylin Bugbee, Michael M. Little, Elizabeth Fancher, Irina Gerasimov, Armin Mehrabian, Lauren Sanders, Sylvain V. Costes, Sergi Blanco-Cuaresma, and 17 others. 2024. INDUS: Effective and efficient language models for scientific applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 98–112, Miami, Florida, US. 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.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemma Team. 2026. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

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, Anjali Pandiri, 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*, pages 5–15, Mumbai, India and virtual. Association for Computational Linguistics.

Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J. Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S. Grant, Donna M. Thompson, Roman Chyla, Stephen McDonald, Timothy W. Hostetler, Matthew R. Templeton, Kelly E. Lockhart, Nemanja Martinovic, Shinyi Chen, Chris Tanner, and Pavlos Protopapas. 2021. Building astroBERT, a language model for Astronomy & Astrophysics. *Preprint*, arXiv:2112.00590.

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. 2026. 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.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. 2020. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 12756–12770, Toronto, Canada. Association for Computational Linguistics.

Mindee. 2021. docTR: Document text recognition. GitHub repository.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, and 7 others. 2025. DINOv3. *Preprint*, arXiv:2508.10104.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *Preprint*, arXiv:2502.14786.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Hyperparameters

Compute. On one A100 GPU, a pretraining epoch takes about 12 minutes; training and test prediction take about 110 minutes for each LLM.

Citation-graph pretraining. One epoch, batch size 64, learning rate 3×10^{-5}, AdamW with weight decay 0.01, maximum length 192 tokens, 600K pairs. In pretraining and in task fine-tuning, the two objects are swapped with probability 0.5.

LLM classifier. Qwen3-8B in bfloat16, LoRA rank 16, $\alpha = 32$, dropout 0.05 on the query, key, value, and output projections; classification head trained in float32; two epochs, learning rate 10^{-4} with 5% linear warm-up, batch size 8 with two accumulation steps, gradient clipping at 1.0, maximum length 768 tokens; the two objects are presented as "Object 1 — … Object 2 — …" and swapped with probability 0.5 during training. Gemma-4-12B uses the same settings plus a linear head on its text backbone's last-token hidden state and training batches grouped by length, which cuts padding from 322 to 6 tokens per sequence.

B All development steps

Table 2 lists every submission that changed the system, in order; Table 1 merges some of these rows. Rows 18–22 have no five-fold score: row 18 changes test predictions only, Qwen3-8B was validated on fold 1 only (0.7130 for row 19 and 0.7153 for row 20), and Gemma-4-12B was not validated. Rows 6 and 18 were dropped, so rows 7 and 19 build on rows 5 and 17.

Table 2: All development steps. CV: five-fold cross-validated macro-F1; Test: macro-F1 on the full test set. Blend weights in rows 19–22 are given in Table 1.

#	Change	CV ↑	Test ↑
1	frozen embeddings, rank features	0.5087	0.4972
2	+ principal components	0.5140	0.5031
3	+ OCR text features	0.5584	0.5557
4	+ cross-encoder	0.5818	0.5671
5	+ both input orders, shared folds	0.5908	0.5810
6	+ image patches, paper clusters	0.5945	<0.5810
7	row 5 + graph pretraining	0.5979	0.6030
8	+ 6.6× citation edges	0.6047	0.6067
9	+ OCR at 1024 px	0.6088	0.6173
10	+ PNG colour-profile match	0.6100	0.6193
11	+ SciNCL/astroBERT ensemble	0.6401	0.6528
12	+ pretrained astroBERT	0.6585	0.6695
13	+ 512-token inputs	0.6697	0.6795
14	+ real OCR text in pretraining	0.6933	0.7021
15	+ 7 real-OCR members	0.6976	0.7058
16	+ image projection features	0.7015	0.7072
17	contrastive projection loss	0.7029	0.7072
18	+ transitivity repair	–	0.7067
19	row 17 ensemble + Qwen3-8B, 1:1	–	0.7292
20	row 17 stacker + Qwen3-8B, 1:1	–	0.7329
21	+ Gemma-4-12B, weights 2:1:1	–	0.7397
22	+ Gemma-4-12B, weights 1:1:1	–	0.7421

C Changes that did not help

Table 3 lists the changes we tested that did not improve the system. Each was measured against the same system without the change; details of the image inputs and of the contrastive objective are in Appendices E and F.

Table 3: Changes that did not help. Δ: change in macro-F1 relative to the same system without the change, cross-validated unless marked Test (full test set). The last row is an absolute score.

Change	Δ
Image patches in the cross-encoder (best of 4)	-0.024
Image embeddings as extra input tokens	-0.001
INDUS instead of astroBERT	-0.103
6.6× more citation edges in pretraining	$+0.003$
Self-training on test pseudo-labels	0.000
Transitivity repair on test triangles	Test -0.001
Clustering objects into paper profiles	Test < 0
Zero-shot Qwen3-8B as the classifier	0.378

D Evaluation protocol and use of the test set

Noise threshold. We have no test labels, so we estimated the sampling noise of macro-F1 from the held-out predictions of the cross-encoder ensemble for the 10,000 training pairs, a set of the same size as the test set. Resampling these pairs with replacement 2,000 times gives a standard deviation of 0.0052 for one score. The difference between two independently sampled scores then has a standard deviation of $\sqrt{2} \times 0.0052$, so two test scores must differ by $1.96 \times \sqrt{2} \times 0.0052 \approx 0.014$ to be distinguishable at the 95% level. This threshold is conservative for two systems scored on the same test pairs, because part of the sampling noise is shared, but the paired version cannot be computed without test labels. On held-out predictions it can: resampling the same pairs for both configurations gives a threshold of about 0.008 on 10,000 pairs. On fold 1 alone (2,000 pairs), intervals are wider, such as $[+0.002, +0.030]$ for the Qwen3-8B blend.

How the test set was used. The leaderboard is computed on the full test set, so every test score in Tables 1 and 2 was seen during development. These scores entered our decisions in three ways. First, changes were developed by cross-validation and then submitted, and we treated a cross-validated gain below 0.005 as unconfirmed until the test score agreed. Second, two submitted changes were dropped because the test score fell (rows 6 and 18 of Table 2). Third, the final blend weights were chosen between two weightings fixed in advance (rows 21 and 22). The weight of the Qwen3-8B blend was not tuned on the test set: equal weights were validated on fold 1. In addition, the rank features of the stacker (Appendix H) are computed among the objects of the test set itself, which uses test inputs but no test labels.

The final blend weights. Gemma-4-12B was trained only once, on all 10,000 pairs, so no held-out predictions existed for choosing its weight. We therefore fixed two weightings of the stacker, Qwen3-8B, and Gemma-4-12B before submitting either. The first, 2:1:1, treats the two LLMs as one group and keeps the 1:1 balance validated on fold 1; the second is 1:1:1. They scored 0.7397 and 0.7421, and we report the higher one. The difference of 0.002 is far below the 0.014 threshold, so the test set gives no evidence that one weighting is better. The gain of Gemma-4-12B over the Qwen3-8B blend (0.007 or 0.009) is also below the threshold. The last result supported by held-out data is therefore the Qwen3-8B blend, with a test score of 0.7329.

Why cross-validation and test scores disagree. Our cross-validation differs from the test set in two ways. First, the folds are split by pair, so a validation fold shares papers with its training folds, whereas test papers are new. A change that improves generalisation to unseen papers is therefore undervalued. Second, a test object appears in 1.79 pairs on average against 1.14 in training, so a better representation of a single object, such as better OCR, is rewarded more often on the test set. The steps we can compare are consistent with this. Citation-graph pretraining gained 3.3 times more on the test set than in cross-validation (row 7 of Table 2) and higher OCR resolution 2.6 times more (row 9), whereas adding pretraining pairs of the same kind, which changes neither property, gained only 0.55 times as much (row 8). Each ratio rests on single submissions whose differences are close to the noise threshold, so we offer this as an observation and not as a tested claim. Splitting folds by paper would remove the first difference, and we suggest it for future editions.

E Image inputs compared with OCR text

Table 4 lists every way we gave a model the image itself. All rows were measured on an early version of the system, before citation-graph pretraining, with one model per row. In a bi-encoder, each object is encoded separately into one vector by a SigLIP 2 tower, and a small classifier compares the two vectors. In the cross-encoder rows, 16 patch tokens from a frozen image encoder pass through a learned linear layer and enter the input sequence of the text model.

Table 4: Inputs for image objects on an early version of the system. CV: five-fold cross-validated macro-F1 of a single model.

Input for an image object	CV ↑
Cross-encoder	
OCR text	**0.556**
OCR text + MatCha patches	0.533
OCR text + SigLIP 2 patches	0.532
SigLIP 2 patches only	0.37
Bi-encoder	
Pooled SigLIP 2 embedding, frozen tower	0.353
Pooled SigLIP 2 embedding, fine-tuned tower	0.312

One reason for the gap in the first experiment was that the layer mapping patch tokens into the text model was trained from scratch on 8,000 pairs. We repeated the test on a final cross-encoder (0.682) with that layer pretrained on 554K public pairs that contain images. SigLIP 2 embeddings as extra input tokens changed the score by -0.001, and visual similarity written into the input as words by -0.002, both within noise.

These comparisons cover image encoders up to the size of SigLIP 2 attached to a BERT-sized text model. We did not test a vision–language LLM that reads the raw figure, such as a Qwen-VL model, which is the natural counterpart of OCR text fed to an LLM. Our finding is therefore that OCR text was the best input among those in Table 4, not that OCR is better than image models in general.

F Contrastive learning in this task

Deciding whether two objects share a paper resembles contrastive representation learning, and contrastive methods enter our system in several places. This appendix collects what each of them contributed.

Contrastively pretrained encoders as features. SigLIP 2 is trained with an image–text contrastive loss and SciNCL with a contrastive loss over citation neighbourhoods. Cosine similarities and rank features in these and five other frozen embedding spaces give 0.509 cross-validated macro-F1 (row 1 of Table 2). Comparing two separately computed embeddings is thus a weak classifier here, and reading the pair jointly with a cross-encoder is what raised the score (rows 4 and 5). As a single cross-encoder, pretrained SciNCL scores below astroBERT (0.583 against 0.620), but averaging the two gained 0.034 on the test set (row 11).

Fine-tuning a contrastive encoder. Updating the SigLIP 2 towers of a bi-encoder with the classification loss on 8,000 pairs lowered the score from 0.353 to 0.312 (Table 4), so we kept all contrastive encoders frozen.

A contrastive projection for image–image pairs. On frozen SigLIP 2 embeddings of public figures, we trained a projection that brings figures of the same paper together and used its cosine similarity as a stacker feature (§3.3). Trained with InfoNCE it gave 0.7029 cross-validated macro-F1, against 0.7015 with a binary cross-entropy loss and 0.6976 without the feature (rows 15 to 17); these differences are within noise. A projection trained directly to separate citing from unrelated papers was worse at that very distinction (AUC 0.62) than the same-paper projection (AUC 0.67).

Hard negatives in pretraining. Half of the *unrelated_papers* pretraining pairs are drawn from topically similar papers (§3.2), the cross-encoder counterpart of hard-negative sampling.

A contrastive objective aimed at the dominant error. With random negatives, a contrastive model can succeed by encoding topic alone. To make it encode which paper an object belongs to, we drew the negatives of each anchor only from papers linked to the anchor's paper by a citation. The model was a two-layer projection to 256 dimensions on frozen SigLIP 2 embeddings, trained with InfoNCE on 36,589 public figures and their captions from papers outside the training set. The positive was another object of the anchor's paper, and each anchor had 16 negatives. The control used the same anchors, architecture, and loss, with negatives from random other papers. Before training we fixed what success would look like: a lower correlation with topical similarity (the cosine of INDUS embeddings), a lower *same_paper*-versus-*unrelated_papers* AUC because topic no longer helps, and a higher *same_paper*-versus-*related_papers* AUC.

Table 5: Effect of citation-linked negatives relative to random negatives, measured on training pairs. SP, R, U abbreviate *same_paper*, *related_papers*, *unrelated_papers*. The topic correlation was measured on caption–caption pairs only.

	image–image	caption–caption
Correlation with topic	–	$0.43 \rightarrow 0.24$
AUC, SP vs. U	-0.108	-0.121
AUC, SP vs. R	-0.011	-0.010

Table 5 shows that the first two predictions held and the third did not. The objective removed topic, but topical similarity is itself the main evidence that two objects share a paper and not merely a citation link (median cosine similarity 0.54 for *same_paper* pairs against 0.42 for *related_papers* pairs). Removing it thus discarded signal without exposing another one. On caption–caption pairs the training loss with citation-linked negatives also fell to 0.13, against 0.73 with random negatives, suggesting that the model memorised the few citation neighbours of each paper.

G What the LLMs add

An LLM differs from our cross-encoders in at least three ways at once: size (8 and 12 billion against 110 million parameters), pretraining data (general-domain text against astronomy text or citation neighbourhoods), and architecture (a decoder reading up to 768 tokens against an encoder reading up to 512). We ran no experiment that separates these factors, so we do not attribute the gain of the blend to size. What we measured is the following.

Alone, the fine-tuned LLM is not stronger. Qwen3-8B scores 0.676 macro-F1 on fold 1, against 0.699 for the cross-encoder system.

Its errors differ. Of Qwen3-8B's errors, 41% are not made by the cross-encoder system, and 35% in the other direction (§5); averaging gains 0.016. The two models also differ by class and pair type: Qwen3-8B is better than the cross-encoder system on *same_figure* (F1 0.775 against 0.746) but much weaker on image–image pairs (macro-F1 0.586 against 0.679), where OCR text is often little more than axis numbers.

Pretraining knowledge does not help without fine-tuning. Zero-shot, with the four labels scored in one forward pass, Qwen3-8B reaches 0.378 macro-F1 and predicts *related_papers* for 2 of the 10,000 training pairs, so it does not recall which papers cite which. In a zero-shot binary probe on *same_paper* and *related_papers* caption–caption pairs, it reaches an AUC of 0.63 on 400 pairs from the less similar half (cross-encoder: 0.78) and 0.74 on 148 pairs from the most similar groups (cross-encoder: 0.86). Combining its ranks with those of the cross-encoder lowers the AUC on the less similar pairs (95% CI of the change $[-0.070, -0.010]$). The zero-shot LLM thus follows the same similarity as the cross-encoder and is weakest on the same pairs.

These results argue against memorised citation links as the source of the gain. They leave open whether fine-tuning draws on pretraining knowledge in ways a zero-shot test cannot reveal, and they do not separate size from architecture.

H Stacker features

For every pair we compute cosine similarities in seven frozen embedding spaces, among them SigLIP 2, DINOv3 (Siméoni et al., 2025), SciNCL, and INDUS (Bhattacharjee et al., 2024). Rank features record how close each object is to the other relative to all objects of the same split, using cross-domain similarity local scaling (CSLS; Conneau et al., 2018) to discount objects that are close to many others. Image–image pairs also get a SigLIP 2 similarity after a projection trained on the public corpus to detect figures of the same paper (Appendix F). For each pair type, a classifier chosen by cross-validation (logistic regression, a two-stage variant, or a small multilayer perceptron) maps these features and the ensemble probabilities to the four classes. On held-out data the stacker adds 0.004 macro-F1 over the ensemble, which is within noise.

I Related work

Prior work on scientific figures has mostly generated captions (Hsu et al., 2021); AstroCLIMB instead asks how figures, captions, and papers relate. Scientific language models learn from domain text (Grezes et al., 2021; Bhattacharjee et al., 2024) or from citation neighbourhoods (Cohan et al., 2020; Ostendorff et al., 2022); we use both and turn the citation graph into labelled pairs.

Local Citation Recommendation Without Graph Signals

Benjamin Basseri **Pavlos Protopapas** **Kemeng Zhang**
Harvard University
{bbasseri,kemengzhang}@fas.harvard.edu pavlos@seas.harvard.edu

Alberto Accomazzi **Stephanie Jarmak** **Anna Kelbert**
Center for Astrophysics | Harvard & Smithsonian
{aaccomazzi,stephanie.jarmak,anna.kelbert}@cfa.harvard.edu

Abstract

Resolving correct citations can be a time-consuming task for research writers. We address local citation recommendation (LCR), where the goal is to retrieve appropriate references for a given sentence. Web tools generally depend on the existing citation graph, which may bias results toward citation count or more recent publications. LCR research using full paper texts has been limited by the lack of publicly available data. We present a new dataset, the Local Citation Recommendation Corpus (LCRC), containing 14,735 scientific statements derived from astrophysics review papers, each with one or more citation targets whose full text is accessible via API. Using this dataset we systematically compare retrieval methods and develop a citation recommendation system that avoids citation graph bias while achieving 8.0% improvement in Recall@25 and 6.9% improvement in NDCG@25 over dense retrieval baselines without model fine-tuning. To support reproducibility, we make our code and dataset publicly available.

1 Introduction

Scientific integrity has long required accurately citing prior research. Comprehensive citation, besides being expected for publication in most settings, connects research efforts across disparate teams and distant times. In recent years, scientific research in general has been subject to increased public scrutiny (Economist Impact, 2023). This heightens the need for thorough citations that encourage verifiability, reproducibility, and transparency in research. However, identifying and verifying appropriate citations remains a time-consuming and error-prone process, leading many writers to rely on digital tools.

Some search engines such as Google Scholar and Semantic Scholar specialize in ranking research publications, but prior work has observed that these tools favor citation counts and author metadata (Beel and Gipp, 2009; Allen Institute for AI, 2025). These biases likely cause or exacerbate Matthew effects, where papers and authors become 'popular for being popular' (Merton, 1968). The scientific community would benefit from improved and more transparent citation recommendation tools.

This work focuses on local citation recommendation (LCR), the task of predicting appropriate citations for a given sentence from a research paper. LCR research has faced challenges making results difficult to compare, reproduce, or deploy. Open datasets for this task have lacked papers' full text, likely due to copyright constraints. Lacking this data, many systems rely on metadata and the citation graph rather than a paper's true content. LCR models trained on this metadata can quickly become obsolete as new authors and papers enter the body of scientific research.

Vector stores present a logical solution to support a constantly expanding corpus and text retrieval (Karpukhin et al., 2020). But while a query's embedding and its citation targets may likely have high similarity, so will embeddings of different sentences citing the same source. Therefore, vector similarity alone is insufficient to discriminate between true and false positives.

LCR often resembles natural language inference (NLI), where a model attempts to predict if one text implies another. Many citing sentences s cite a reference r because r supports the factuality of s, such as this one (Teufel et al., 2006). But in practice, researchers cite papers for many reasons besides 'r implies s', including to clarify, contradict, or express a unique interpretation of earlier results (Garfield, 1996). As a result, sentences can have many different relationships to the documents they cite, relationships that are not easily resolved by vector similarity, NLI scoring, or any one single method.

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 61–76
November 10, 2026 ©2026 Association for Computational Linguistics

We make the following contributions:

- LCRC, a one-to-many LCR dataset derived from expert-curated review papers.
- A systematic evaluation of retrieval-only citation recommendation without metadata or graph signals.
- A mixed query representation, ADD-PREV-02+RRF that improves Recall@25 by 8.0% over dense retrieval, without fine-tuning.

The dataset[1] and the code to reproduce our experiments[2] are both publicly available.

2 Literature Review

LCR systems have typically contained the following components:

- Document representation
- Query representation
- Document ranking against a given query
- Optional filtering or reranking of top-ranked documents

with much of the research effort focused on representation strategies.

Early LCR research used various NLP techniques to construct representations. One approach computed TF-IDF vectors from passages citing a document as that document's representation (He et al., 2010). Context-based Collaborative Filtering (Liu et al., 2015) represented queries as binary vectors computed by citation co-occurrence with other papers in the corpus. Probabilistic approaches include REFSEER (Huang et al., 2014) which adapted a translation model to produce citation likelihoods, and LOGIT-EXPANDED (Tanner and Charniak, 2015) which contextualized queries with metadata including past citations and author information.

In recent years, researchers have explored neural embeddings combined with different context augmentation strategies. Early neural approaches used LSTMs to create embeddings (Yang et al., 2018); subsequent work augmented embeddings with title and abstract (Medić and Snajder, 2020) or modeled writer intent (Ma et al., 2025). Transformer-based approaches include HATTEN(Gu et al., 2022), which uses SCIBERT (Beltagy et al., 2019) as its embedding model, and also augments inputs with their titles and abstracts. CiteBART (Çelik and Tekir, 2025) adapts BART (Lewis et al., 2020) to generate "(Author, Year)" recommendations as a sequence-to-sequence task.

Another branch of research focuses on graph-based methods. Some work models citation dynamics using temporal graph neural networks (Holm et al., 2020) or estimates citation impact (Zhang and Wu, 2024). BERT-GCN (Jeong et al., 2020) fine-tunes BERT (Devlin et al., 2019) to encode queries and a graph neural network to encode the corpus. Multi-granularity preference alignment (Li and Chen, 2025) combines signals derived from the citation graph with intent modeling. These approaches leverage structural and temporal information in citation networks, but are fundamentally oriented toward modeling existing citation graph dynamics, rather than identifying semantically relevant papers for a specific sentence.

A further limitation of prior LCR work lies in the datasets used for training and evaluation. Many studies rely on unreleased datasets or citation graphs derived from regular research papers, which may bias towards popularity, social networks, or other factors (Dougherty and Horne, 2022). Using such citation graphs as ground truth risks perpetuating these biases: when systems are evaluated against such data, it remains unclear whether they are identifying scientifically appropriate citations or simply learning to replicate existing popularity and network effects. This motivates alternative data sources, such as review papers, which reflect expert-curated citations.

3 Task Definition & Metrics

3.1 Task Definition

In the context of research writing, we define the local citation recommendation task as retrieving the most scientifically appropriate set of documents from a corpus that should be cited for an input query.

We define the query q as a sentence from a scientific manuscript, and possibly its context. Context can include additional data that helps clarify the sentence's meaning, such as preceding sentences, the paper's title, or abstract. A model performing this task aims to return the documents $D = \{d_1, \ldots, d_n\}$ which the input sentence from

[1] https://huggingface.co/datasets/adsabs/LCRC
[2] https://github.com/benjaminb/citeline

q should cite.

For the current task we assume D is non-empty and consider only nontrivial sentences, which we define as sentences that require one or more scholarly citations to support or contextualize their claims. We leave distinguishing trivial and nontrivial statements as future work.

Since the model maps queries to sets of documents, this task is a special case of document retrieval. As the task also primarily deals with natural language, it can also be thought of as a type of semantic similarity estimation. However, it is important to note that the relationship between a research statement and the set of documents it should cite is more constrained than semantic similarity alone. If paper c establishes some scientific fact which is cited by both statements s_1 and s_2 in other papers, the semantic similarity between s_1 and s_2 is likely to be high and quite possibly higher than their similarity to c.

Consider the case s_1 = "A superwind does not blow in low metallicity environments" in one paper and s_2 = "Superwinds do not blow in lower metallicity environments" in another paper, both of which should cite c = Bowen (1991) primarily for the following passage:

> "We estimate that for $Z < 0.1Z_{\odot}$ there will usually be too little dust to produce significant dynamical effects."

Although both s_1 and s_2 should cite c, their near-identical wording makes it likely that s_1 and s_2 will exhibit higher cosine similarity to one another than either does to c. As a result, a retrieval system based purely on semantic similarity is likely to rank s_2 as more relevant to s_1 than the correct citation c, despite c containing the supporting evidence. This illustrates why citation prediction cannot be reduced to sentence-level semantic similarity alone.

3.2 Metrics

A model performing the citation prediction task produces a ranked set of predicted documents $\widehat{D} = \{\widehat{d}_1, \ldots, \widehat{d}_k\}$, often referred to as the top-k documents. To evaluate the model's success, we define metrics comparing the predicted $\widehat{D}$ to the target set D, parameterized by k:

- **Hitrate@k**: 1 if *any* of the target documents appear in the top k retrieved, 0 otherwise.
- **Recall@k**: The proportion of target documents retrieved in the top k. For example, if a query has 4 target documents (i.e., four reference citations) and at $k = 20$ two of those documents appear, then Recall@20 is 0.5.
- **NDCG@k**: Normalized Discounted Cumulative Gain at rank k (Järvelin and Kekäläinen, 2002). This metric accounts for the position of relevant documents in the ranked list. Since it is common in vector-based retrieval for the same document to appear multiple times in a ranked search result due to chunking, we consider only the first occurrence of each document when computing NDCG. This is more aligned with the end goal of presenting the most likely citation candidates to the user.

Then for a dataset of N queries, the mean Hitrate@k, Recall@k, and NDCG@k are the sum of the individual per-query metric values divided by N. In Section 5 (Experiments & Results) we refer to these means simply as Hitrate@k, Recall@k, or NDCG@k when discussing results over an entire dataset.

4 Methodology

To build a citation prediction model, we constructed a dataset of queries containing sentences requiring citations, associated metadata, and their reference citations. We then formed a search corpus against which candidate citations are retrieved, enabling evaluation of each model's Hitrate@k and Recall@k. We built a labeled query dataset from review papers and a retrieval corpus from general research papers, drawn from multiple scientific literature sources. In the experiments presented in this work, these datasets are derived from the Astrophysics Data System (ADS, Kurtz et al. (2000)).

We then segmented the corpus papers into chunks and embedded them in a vector database to support dense document retrieval. We describe our experimental setup and results in Section 5.

4.1 Dataset

To build a labeled dataset, we curated 354 digitized review papers from a scientific literature corpus spanning astrophysics, earth science, and planetary science, provided by the Astrophysics Data System. Review papers synthesize prior research and are typically written with expert-curated citation practices, making them a suitable ground truth for LCR.

From these review papers we extracted individual sentences containing inline citations and, where

possible, resolved the inline citations to uniquely identifying document ID's (i.e., DOIs). We implemented an automated extraction pipeline to scale this process. First it splits the paper's body on sentence boundaries. Each sentence is then passed to an LLM agent which rejects unusable sentences such as captions, reference section entries, or sentences containing severe OCR errors. For sentences passing this filter, the agent identifies and normalizes inline citations to an array of (Author, year) tuples. The extracted (Author, year) tuples are then passed to deterministic code that maps each citation to a uniquely identifying DOI using the paper's bibliography. Sentences with any inline citations that did not fully resolve to a DOI were excluded.

For example, given the sentence:

> "Kaiser et al (1995) (see also Luppino Kaiser 1997 , Hoekstra et al 1998 for further developments) compute the second moment within a variable aperture..."

the extraction stage would yield tuples [('Kaiser et al', 1995), ('Luppino Kaiser', 1997), ('Hoekstra et al', 1998)]. These tuples are resolved to their corresponding DOIs by searching through the review's references. The resolved papers form the target set D for the query, and the inline citations are masked to form the query q.

Applying this pipeline produced 14,735 samples such that:

a) Sentences of captions, reference entries, and garbled OCR are excluded.

b) Each sample sentence contains one or more inline citations.

c) All inline citations are matched to their DOI.

The dataset is provided as a JSONL file. Each line is a JSON object containing keys for the original sentence, citation target DOIs, the sentence with inline citations masked, the sentence position within the review paper, its source DOI, title, abstract, and additional fields.[3]

4.2 Corpus

All dataset queries constructed cite a combined 9,980 distinct research papers. In our experiments we expanded the corpus to 52,618 distinct research papers to make the retrieval task more challenging. For the experiments presented in this work, we obtained these papers' full-text contents from ADS to form our corpus. From these records, we reconstructed each paper by concatenating the title, abstract, and body fields, intentionally omitting all other fields such as authors, affiliations, and references, in order to avoid unnecessary bias.

We implemented two forms of document representation: chunk-based and contribution-based representations. In the chunk-based representation, we used standard text-splitting techniques to divide the papers into a corpus of 460,801 chunks. For the contribution-based representation, we used a large language model to extract a paper's original findings, observations, and methodologies (see Section 5.2).

5 Experiments and Results

We organize our experiments into four distinct stages of semantic retrieval: choice of embedder, corpus representation, query representation, and reranking strategies. At each stage, we explored several alternatives, as detailed in the subsections below.

5.1 Embedding Model Selection

To select an embedding model, we used Milvus to implement a vector database. We chunked the research corpus and, for each embedding model evaluated, embedded the resulting chunks. Retrieval was then performed by computing cosine similarity between query sentence embeddings and chunk embeddings.

Although retrieval is performed at the chunk level using cosine similarity between query and chunk embeddings, evaluation is conducted at the paper level by mapping retrieved chunks back to their source documents via DOI, and computing Hitrate@k and Recall@k over the resulting set of unique papers.

We evaluated QWEN3-EMBEDDING-0.6B, QWEN3-EMBEDDING-8B (Zhang et al., 2025), and BGE-LARGE-EN-V1.5 (Xiao et al., 2024) models for their high-quality, general-purpose embeddings. Additionally, we evaluated several LLMs fine-tuned on astrophysics data as potential embedding models: ASTROBERT (Grezes et al., 2021), ASTROLLAMA (Nguyen et al., 2023), ASTROSAGE (de Haan et al., 2025), NASA-IBM-ST.38M (Bhattacharjee et al., 2024), and SPECTER2 (Singh et al., 2023).

We find that QWEN3-EMBEDDING-8B achieves

[3] https://huggingface.co/datasets/adsabs/LCRC

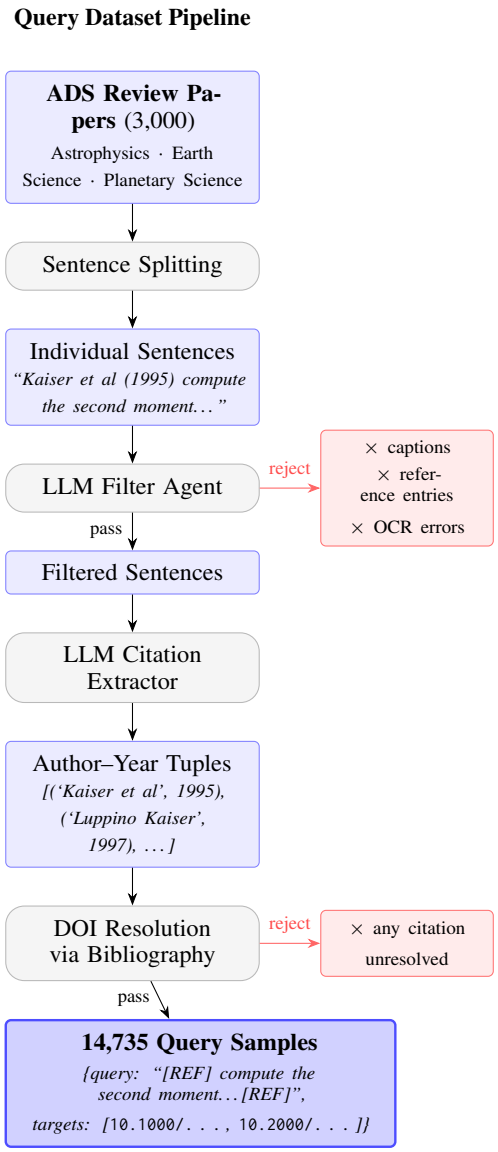


Figure 1: Query dataset construction pipeline. An automated pipeline processes 3,000 ADS review papers. Sentences are split, then an LLM agent filters out captions, reference entries, and OCR errors. Inline citations are extracted as (Author, Year) tuples and resolved to DOIs via the paper's bibliography; any sentence with an unresolved citation is dropped, yielding 14,735 labeled query samples.

the best hitrate and recall for all values of k, with QWEN3-EMBEDDING-0.6B ranking second-best despite being a significantly smaller model. These general-purpose embedding models, along with BGE-LARGE-EN-V1.5, outperform the astrophysics fine-tuned models overall (see Table 1).

5.2 Corpus Representation

We attempted to find a more effective representation for the reference corpus by tasking an LLM to write an array of sentences comprising each research paper's findings, scientific contributions, and methodologies, similar to the document expansion technique described by Nogueira et al. (2019). This sentence array aimed to provide a more concise summary of the paper's content that is most likely to be cited. We collectively refer to all such citable statements as "contributions". We used DeepSeek-v3.2 (DeepSeek-AI et al., 2025) for its advanced reasoning capabilities to produce the contributions-based document expansion, constructing a new corpus using a simple prompt (v. 1) and a more advanced prompt (v. 2). The simple prompt resulted in a corpus of 89,860 statements, while the advanced prompt resulted in a corpus of 299,286 statements. However, the baseline chunk-based representation of the corpus outperformed both contribution-based variants. Results are shown in Fig 2, and the extraction prompts are included in Appendix A.1. One possible explanation is that citation decisions are often driven by narrow, localized evidence—such as specific definitions, caveats, or technical claims—rather than high-level summaries. Chunking may better preserve these fine-grained cues, whereas contribution-level summaries could abstract away details that are important for aligning a citing sentence with its source.[4]

5.3 Query Representation

To better contextualize the input query, we attempted several types of query context management before embedding. These include prepending the title or abstract from the query's original paper, or the previous n sentences for $n \in \{1, 2, 3, 5, 7\}$. We initially found that representing the input query sentence by prepending its previous three sentences, ADD-PREV-3, was the best-performing

[4]We also experimented with summarizing individual chunks using an LLM and appending these summaries prior to embedding. This approach did not improve retrieval performance in preliminary experiments and was not pursued further.

Embedder	Hitrate			Recall			NDCG		
	@10	@25	@50	@10	@25	@50	@10	@25	@50
ASTROBERT	0.1002	0.1621	0.2256	0.0829	0.1372	0.1946	0.0472	0.0613	0.0731
ASTROLLAMA	0.3221	0.4347	0.5292	0.2844	0.3938	0.4847	0.1881	0.2163	0.2349
ASTROSAGE-8B	0.1898	0.2611	0.3237	0.1622	0.2282	0.2873	0.1068	0.1240	0.1361
BGE-LARGE-EN-V1.5	0.5623	0.6617	0.7255	0.5126	0.6160	0.6844	0.3868	0.4139	0.4281
NASA-IBM-ST.38M	0.5241	0.6269	0.7020	0.4761	0.5796	0.6585	0.3560	0.3832	0.3996
QWEN3-EMBEDDING-8B	**0.6640**	**0.7570**	**0.8126**	**0.6098**	**0.7119**	**0.7757**	**0.4801**	**0.5072**	**0.5206**
SPECTER2	0.3634	0.4706	0.5513	0.3249	0.4278	0.5070	0.2302	0.2569	0.2731

Table 1: **Embedding Model Comparison.** Hitrate, Recall and NDCG at ranks 10, 25, and 50 for evaluated embedders. General-purpose embedding models QWEN3-EMBEDDING-8B and BGE-LARGE-EN-V1.5 consistently outperform astrophysics fine-tuned LLMs.

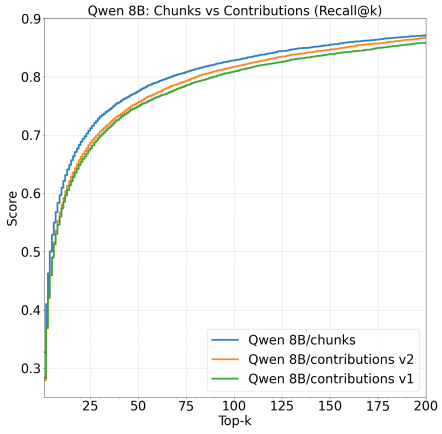


Figure 2: Comparison of document representation strategies. Baseline chunking outperformed LLM-extracted contributions (v1: simple prompt; v2: engineered prompt). Hitrate@k and NDCG@k showed similar trends.

strategy at many values of k, suggesting that this additional context helped the embedder better encode the underlying meaning of the query sentence.

After some informal error analysis of the ADD-PREV-3 strategy, we found that some queries benefited from the additional context but others did not. This led to the development of a 'mixed query representation' strategy.

In mixed query representation, we represent the input query with multiple distinct representations. Specifically, we take the original sentence and the sentence appended to its previous two sentences, then embed them in parallel. Each vector is used to perform cosine similarity search against the database, retrieving the top $k//2$ records. In other words, to retrieve a total of k results, we retrieve half from the query's original text and the other half from the augmented query. We merge the two resulting sets by interleaving by rank: for chunks $(c_1, \ldots, c_{k//2})$ returned by the unmodified query and $(e_1, \ldots, e_{k//2})$ from the augmented query, the final ranking is $(c_1, e_1, c_2, e_2, \ldots, c_{k//2}, e_{k//2})$. This interleaving outperformed sorting by cosine similarity alone.

We call this query representation strategy ADD-PREV-02 since it combines the query sentence as-is (i.e., with zero previous sentences) with the query sentence plus its previous two sentences. This strategy significantly outperformed the baseline IDENTITY and all other query representations at all values of k.

One possible explanation is that, while additional context can help disambiguate certain citation intents, it may also introduce topic drift. In review papers, preceding sentences often summarize broader background or adjacent results that are only loosely related to the specific claim being cited. As a result, uniformly expanding all queries with surrounding context may dilute the semantic signal most relevant for citation selection.

The mixed query representation may mitigate this trade-off by retaining both a high-precision signal (the isolated sentence) and a higher-recall signal (the context-augmented version). By combining these complementary views through rank interleaving or fusion, the retrieval process can preferentially surface candidates that are robust across representations, while reducing the impact of misleading context.

5.4 Reranking by Reciprocal Rank Fusion

Reciprocal rank fusion (RRF) reranks the top-k search results by forming a weighted sum of their *reciprocal* ranks (so that lower-ranked documents receive smaller contributions) with respect to multiple rerankers $R_1, \ldots, R_n$, using weights $w_1, \ldots, w_n$, plus a smoothing parameter m (Cormack and Barber, 2009).

$$\text{RRF}(q, d) = \frac{w_1}{m + R_1(q, d)} + \ldots + \frac{w_n}{m + R_n(q, d)}$$

Method	Hitrate			Recall			NDCG		
	@10	@25	@50	@10	@25	@50	@10	@25	@50
IDENTITY	0.6640	0.7570	0.8126	0.6098	0.7119	0.7757	0.4801	0.5072	0.5206
ADD-TITLE	0.6546	0.7562	0.8225	0.6012	0.7107	0.7862	0.4647	0.4936	0.5094
ADD-ABSTRACT	0.4422	0.5798	0.6807	0.3991	0.5350	0.6387	0.2725	0.3077	0.3291
ADD-TITLE+PREV-3	0.6085	0.7268	0.8012	0.5578	0.6812	0.7620	0.4091	0.4415	0.4584
ADD-PREV-3	0.5944	0.7169	0.7953	0.5442	0.6707	0.7554	0.3984	0.4317	0.4493
ADD-PREV-5	0.5389	0.6692	0.7561	0.5928	0.6238	0.7153	0.3527	0.3870	0.4059
ADD-PREV-7	0.4973	0.6289	0.7225	0.4539	0.5847	0.6823	0.3175	0.3518	0.3719
ADD-PREV-02	0.7157	0.8094	0.8681	0.6580	0.7637	0.8329	0.5069	0.5351	0.5495
ADD-PREV-02+RRF	**0.7190**	**0.8140**	**0.8735**	**0.6624**	**0.7692**	**0.8383**	**0.5135**	**0.5420**	**0.5565**

Table 2: **Query Expansion Comparison.** Hitrate, recall, and NDCG using QWEN3-EMBEDDING-8B and various query expansion methods.

We tested two rerankers, BM25 (Robertson and Zaragoza, 2009) and NUM_RETRIEVALS. BM25, a well-known text similarity function, uses term frequency and inverse document frequency to compute a similarity score between a query and a document (Robertson and Zaragoza, 2009). For NUM_RETRIEVALS, we hypothesized that a larger number of times the same source document appeared in the top-k chunks retrieved by vector similarity search may indicate stronger query–document similarity. To test this hypothesis, we defined the NUM_RETRIEVALS function as the number of times a source paper appears in the search results, as identified by its DOI.

After performing a grid search over these rerankers with various weights, we find the best performing RRF function uses BM25 with a weight of 0.4, fused with the original search results rank weighted at 0.6:

$$\text{RRF}^*(q,d) = \frac{0.6}{m + R_{\cos}(q,d)} + \frac{0.4}{m + R_{\text{BM25}}(q,d)}$$

where $R_{\cos}(q,d)$ is the document chunk's original rank with respect to the query (by cosine similarity) and $R_{\text{BM25}}(q,d)$ is the document chunk's rank when reranking by BM25 (with parameters $k_1 = 1.5, b = 0.75$), and $m = 60$. Meanwhile, rank fusion with NUM_RETRIEVALS did not improve scores.

5.5 Evaluating Citation Count or Publication Date Bias

We evaluated our dataset using rerankers based on citation count and publication age to assess whether there appeared to be a preference for higher citation counts or more recent publications in our data. To do so, we defined two rerankers:

$$\text{RECENCY} := -\log(y+1)$$
$$\text{LOG_CITATIONS} := \log(c+1)$$

where y is computed as length in years between query's publication date and the chunk's publication date, and c is the citation count of the chunk's source paper. We add 1 to each logarithm to avoid evaluating $\log(0)$.

We conducted a grid search by combining each reranker with the original ranks from ADD-PREV-02/QWEN3-EMBEDDING-0.6B results. We searched over non-negative weight combinations in increments of 0.1. At all weight combinations evaluated, RRF with either of these metrics generally caused hitrate and recall to decrease for most values of k, suggesting that the review papers did not exhibit a strong preference for higher citation counts or more recent publication among semantically similar papers.

One possible explanation is that review articles are intentionally written to cite foundational and semantically appropriate work, rather than prioritizing the most recent or most highly cited publications. As a result, rerankers based on citation count or publication recency may introduce signals that are weakly aligned—or even misaligned—with the underlying citation intent, leading to degraded retrieval performance (see Fig 3).

6 Conclusions & Future Work

In this paper, we release a new dataset for multi-label citation prediction with comprehensive metadata for retrieving citation contexts. We hope this encourages more work that focuses on the content of scientific statements rather than authorship or citation counts. We also introduce an effective

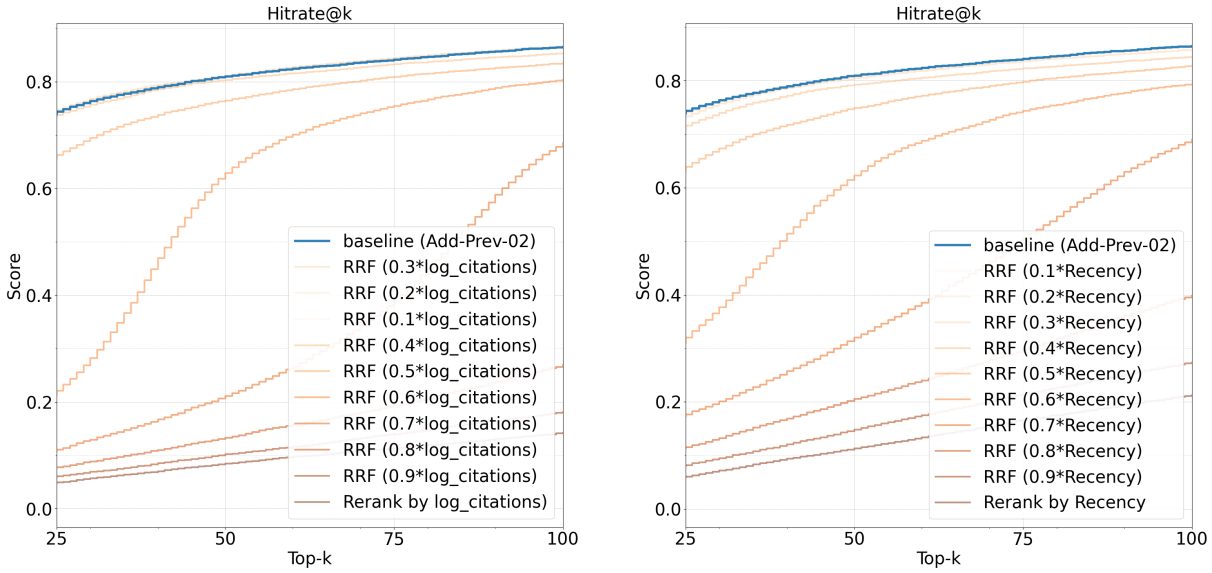


Figure 3: Grid search results for reciprocal rank fusion on original ranks from ADD-PREV-02 with LOG_CITATIONS and RECENCY. The top-scoring weights at best matched the original ranking from ADD-PREV-02 for most values of k. Applying higher weights to LOG_CITATIONS and RECENCY tended to reduce hitrate and recall.

context management technique, ADD-PREV-02, which performs citation prediction significantly better than baseline vector similarity. This is consistent with classic IR research showing multiple query representatives can improve retrieval (Belkin et al., 1993, 1995). Our strategy can be viewed as an instance of this principle in a dense retrieval setting, using natural context rather than generated query variants.

Furthermore, we identify QWEN3 as a top-performing embedding model for the citation prediction task, and we identify a query representation and rank fusion strategy that collectively improve Hitrate@25 by 7.5% and Recall@25 by 8.0% above the baseline approach. Our experiments suggest that LLMs specifically trained for embedding tended to outperform LLMs fine-tuned on astrophysics text. To represent documents in the corpus, simple chunking outperformed the summarization methods attempted. We also show that expanding the query with additional context produced limited results when used as the only input. However, when used in parallel with the original query, merged results significantly improved. Finally, our reranking experiments found that rank fusion between BM25 and the raw semantic similarity ranking achieved better results than reranking by BM25 or other methods.

We find that QWEN3 embedding models outperformed all other models fine-tuned on astrophysics data. This may be due to QWEN3's contrastive loss function, which enables it to learn high-quality document embeddings. Fine-tuning astrophysics-specific models with a contrastive loss, or further fine-tuning QWEN3 on domain-specific data, may improve performance by combining the strengths of both approaches.

The majority of our model's improvement comes from the mixed query representation strategy. Future work may explore improved combinations of query representations and evaluate whether these techniques generalize to other scientific domains.

Limitations

Our dataset excludes queries with significant amounts of math notation since the digitized text often fails to accurately represent the original expressions. Short of re-scanning the corpus with more powerful OCR, additional work is required to obtain accurate representations or address inaccurately digitized text.

In this paper, we focused on "nontrivial" inputs: sentences that require at least one citation. Not all statements require a formal citation, so additional work is needed to distinguish between such statements.

Acknowledgments

The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University. This research has made use of

NASA's Astrophysics Data System Bibliographic Services.

References

Allen Institute for AI. 2025. s2search: Semantic scholar search backend. https://github.com/allenai/s2search/blob/master/s2search/features.py. Accessed: 2025-11-20.

Joeran Beel and Bela Gipp. 2009. Google scholar's ranking algorithm: An introductory overview. In *Proceedings of the Third International Conference on Research Challenges in Information Science (RCIS)*.

Nicholas J. Belkin, C. Cool, W. Bruce Croft, and James P. Callan. 1993. The effect multiple query representations on information retrieval system performance. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '93*, pages 339–346, Pittsburgh, Pennsylvania, United States. ACM Press.

N.J. Belkin, P. Kantor, E.A. Fox, and J.A. Shaw. 1995. Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management*, 31(3):431–448.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Bishwaranjan Bhattacharjee, Aashka Trivedi, Masayasu Muraoka, Muthukumaran Ramasubramanian, Takuma Udagawa, Iksha Gurung, Nishan Pantha, Rong Zhang, Bharath Dandala, Rahul Ramachandran, Manil Maskey, Kaylin Bugbee, Mike Little, Elizabeth Fancher, Irina Gerasimov, Armin Mehrabian, Lauren Sanders, Sylvain Costes, Sergi Blanco-Cuaresma, and 17 others. 2024. INDUS: Effective and Efficient Language Models for Scientific Applications. *Preprint*, arXiv:2405.10725.

Ege Yiğit Çelik and Selma Tekir. 2025. Citebart: Learning to generate citations for local citation recommendation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Gordon V. Cormack and Maura R. Barber. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. Technical report, University of Waterloo.

Tijmen de Haan, Yuan-Sen Ting, Tirthankar Ghosal, Tuan Dung Nguyen, Alberto Accomazzi, Azton Wells, Nesar Ramachandra, Rui Pan, and Zechang Sun. 2025. AstroMLab 3: Achieving GPT-4o Level Performance in Astronomy with a Specialized 8B-Parameter Large Language Model. *Scientific Reports*, 15(1):13751.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael R. Dougherty and Zachary Horne. 2022. Citation counts and journal impact factors do not capture some indicators of research quality in the behavioural and brain sciences. *Royal Society Open Science*, 9(8):220334.

Economist Impact. 2023. Confidence in research: Researchers in the spotlight. https://impact.economist.com/projects/confidence-in-research/. Accessed: 2025-11-20.

Eugene Garfield. 1996. When to cite. *The Library quarterly (Chicago)*, 66(4):449–458.

Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J. Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S. Grant, Donna M. Thompson, Roman Chyla, Stephen McDonald, Timothy W. Hostetler, Matthew R. Templeton, Kelly E. Lockhart, Nemanja Martinovic, Shinyi Chen, Chris Tanner, and Pavlos Protopapas. 2021. Building astroBERT, a language model for Astronomy & Astrophysics. *Preprint*, arXiv:2112.00590.

Nianlong Gu, Yingqiang Gao, and Richard H. R. Hahnloser. 2022. Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking. *Preprint*, arXiv:2112.01206.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 421–430, New York, NY, USA. Association for Computing Machinery.

Nils A. Holm, Barbara Plank, Derek Wright, and Isabelle Augenstein. 2020. Longitudinal citation prediction using temporal graph neural networks. *arXiv preprint arXiv:2012.05742*.

Wenyi Huang, Zhaohui Wu, Prasenjit Mitra, and C. Lee Giles. 2014. Refseer: A citation recommendation system. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 371–374.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Chanwoo Jeong, Sion Jang, Eunjeong Park, and Sungchul Choi. 2020. A context-aware citation recommendation model with BERT and graph convolutional networks. *Scientometrics*, 124(3):1907–1922.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Michael J. Kurtz, Guenther Eichhorn, Alberto Accomazzi, Carolyn S. Grant, Stephen S. Murray, and Joyce M. Watson. 2000. The NASA Astrophysics Data System: Overview. *Astronomy and Astrophysics Supplement Series*, 143:41–59.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiangyu Li and Jingqiang Chen. 2025. Scirgc: Multi-granularity citation recommendation and citation sentence preference alignment. *Preprint*, arXiv:2505.20103.

Haifeng Liu, Xiangjie Kong, Xiaomei Bai, Wei Wang, Teshome Megersa Bekele, and Feng Xia. 2015. Context-based collaborative filtering for citation recommendation. *IEEE access*, 3:1695–1703.

Shijia Ma et al. 2025. Citation recommendation based on argumentative zoning of user queries. *arXiv preprint arXiv:2501.18292*.

Zoran Medić and Jan Snajder. 2020. Improved local citation recommendation based on context enhanced with global information. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 97–103, Online. Association for Computational Linguistics.

Robert K. Merton. 1968. The matthew effect in science. *Science*, 159(3810):56–63.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciucă, Charlie O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Li, Josh Peek, Kartheik Iyer, Tomasz Różański, Pranav Khetarpal, Sharaf Zaman, David Brodrick, Sergio J. Rodríguez Méndez, Thang Bui, Alyssa Goodman, and 5 others. 2023. AstroLLaMA: Towards Specialized Foundation Models in Astronomy. *Preprint*, arXiv:2309.06126.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. Scirepeval: A multi-format benchmark for scientific document representations. *Preprint*, arXiv:2211.13308.

Chris Tanner and Eugene Charniak. 2015. A hybrid generative/discriminative approach to citation prediction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 75–83, Denver, Colorado. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed Resources For General Chinese Embeddings. *Preprint*, arXiv:2309.07597.

Libin Yang, Yu Zheng, Xiaoyan Cai, Hang Dai, Dejun Mu, Lantian Guo, and Tao Dai. 2018. A lstm based model for personalized context-aware citation recommendation. *IEEE Access*, 6:59618–59627.

Feifei Zhang and Shuangjie Wu. 2024. Predicting citation impact of academic papers across research areas using multiple models and early citations. *Scientometrics*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

A Agent & Prompts

To construct our dataset we built an agent to implement the following pipeline:

1. Given an input sentence, identify if it is 'valid' (e.g. no serious transcription errors, not a caption)
2. If the sentence is valid, identify all substrings that constitute inline citations
3. Write out a version of the input sentence with those substrings masked
4. Write out an array of (author, year) tuples for each inline citation identified
5. Execute deterministic code that maps each (author, year) tuple to the unique DOI that represents the cited paper

The prompts used in the agent are presented below:

Sentence Validity Prompt

```
You are an expert data analyst in astrophysics and earth sciences. You are building a dataset of
snippets from research papers. Usable snippets will contain complete factual or scientific
statements and possibly an inline citation. But snippets from a research paper's reference
section or captions with little semantic content will NOT be usable.

Your task is to read the snippet below, write a reason for if the snippet appears usable, then
write True if it is usable or False if not usable.

## Rules
- Write false for snippets that are primarily captions from a figure, table, or image, with no
explanatory text.
- Write false for snippets that look like they have a lot of OCR errors or math gibberish
- Write true otherwise, REGARDLESS OF TYPOS or phrasing

## Examples
Input: "Fig. 12 shows the K -band number counts from an accumulated exposure of 1.5 h (6 h of
data) on a single tile in the ELAIS-N1 field."
Output: {{"reasoning": "Although this looks like a caption to a figure, it is a complete factual
statement", "label": true}}

Input: "In relatively uncrowded regions, the aperture photometry reaches close to the target
depths at J , H and K (see results in Warren et al. 2007)."
Output: {{"reasoning": "This refers to photometry metrics in specific regions and is a complete
factual statement", "label": true}}

Input: "Casali et al. (2007) describe the survey instrument, WFCAM."
Output: {{"reasoning": "Although this is a short sentence it states that Casali et al. (2007)
will contain a description of the WFCAM instrument. So it's a factual statement.", "label":
true}}

Input: "Figure 2 Gas temperature ( red circles ); cumulative gas mass ( blue squares )"
Output: {{"reasoning": "This appears to be a figure caption that just labels the parts of the
figure. Not a scientific or usable statement.", "label": false}}

Input: "Table 1 Observational evidence for AGN feedback"
Output: {{"reasoning": "This looks like just a title for a table, not a complete scientific
statement", "label": false}}

Input: "Allen C. W. , 1973 , Astrophysical Quantities , 3rd edn . Univ. of London , Athlone
Baraffe I."
Output: {{"reasoning": "This looks like just a citation, possibly from a reference section. Not
a complete factual or scientific statement.", "label": false}}

Input: "It also hosts a Compton-thick AGN in the Western component, observed directly in hard
X-rays (Della Ceca et al. 2002 ; Ballo et al. 2004 )."
Output: {{"reasoning": "Describes a scientific observation", "label": true}}

Based on the input sentence, provide your reasoning and a boolean 'isValid' flag.
```

```
  {format_instructions}

Input:
{sentence}
```

Inline Citation Extraction Prompt

```
System: You are a strict extractor of inline citations from snippets. You NEVER normalize,
format, or clean up the text. You only extract exact spans from the snippet, with no added or
missing characters, punctuation, or whitespace.

You are an expert data analyst in astrophysics and earth sciences. You are building a dataset of
snippets from research papers. In each snippet you must identify any substrings that are INLINE
CITATIONS.

Your task is to read the snippet below, think step by step about what substrings (if any) are
INLINE CITATIONS, then write out those substrings as an array. If the snippet has no citations,
write an empty array.

## Definition of INLINE CITATION:
An inline citation is any span of text in the snippet that refers to other authors' work,
including publication year or pre-publication status (e.g. 'submitted')

## INLINE CITATION formats:
Inline citations can appear in various formats
- Any sequence that looks like a capitalized name (optionally followed by 'et al.' or 'and
Coauthor') immediately followed by a four-digit year (with optional letter suffixes) inside
parentheses
- Parenthetical: (Smith 1994), (Smith and Jones 1994), (Smith et al. 1994)
- Non-parenethetical: Smith 1994, e.g. Smith 1994, Smith and Jones 1994
- Only year in parentheses: Smith (1994), e.g. Smith (1994 ), Smith and Jones ( 1994 )
- Multiple: (Smith 1994; Jones 1995, Brown 2000a,b)
- With "submitted" or "in prep": (Kopp et al. submitted), (Doe 2020 in preparation)
- Narrative: Smith (1994), Smith and Jones (1994), Smith et al. (1994)
- Loose year mentions in parentheses: (1994), (2000a), (2000a,b)
- Inconsistent whitespace: ( Smith 1994), ( Smith 1994 ), Smith (1994 )

## Not citations:
- Expressions not including a year: ( STS 06), ( Fig. 11), (Marine Isotopic Stage, MIS 5e at ~
122 ka)

## IMPORTANT:
In your substrings
- DO NOT alter the spacing, commas, or punctuation of any citation. You must PRESERVE THE
ORIGINAL WHITESPACE, PUNCTUATION, AND FORMATTING EXACTLY as it appears in the snippet.
- DO NOT add any parentheses not in the original snippet
- DO NOT remove any parentheses that are in the original snippet
- If the citation does not have parentheses in the original snippet, you must not add them in
the output — even if it looks like other citations that use them.
- DO NOT "clean up" spacing or insert/remove punctuation — your substring must be a **verbatim
substring** of the original snippet.
- You are **NOT** formatting citations — you are extracting them **exactly as written**.
- If the citation is ( Smith 1994 ), your output must be ["( Smith 1994 )"], **not** ["(Smith
1994)"].


## Examples:
Snippet: "These are minor species, with abundances of 1.5 10 12 , 1.7 10 9 , and 7 10 8 kg N,
respectively ( Ussiri and Lal, 2013 )."
Substring: ["( Ussiri and Lal, 2013 )"]
× WRONG: ["(Ussiri and Lal, 2013)"]

Snippet: "Focusing on 353-GHz data at high Galactic latitudes, the E and B angular power spectra
of dust polarization were constrained in the multipole range of 40l600 ( Adam et al. 2016 )."
Substring: ["( Adam et al. 2016 )"]
× WRONG: ["(Adam et al., 2016)"]
```

Snippet: "Proton blazar models (Mannheim,1993) were initially motivated by the search of sources of ultra-high-energy cosmic rays."
Substring: ["(Mannheim,1993)"]
× WRONG: ["(Mannheim, 1993)"]

Snippet: "Methods for mixing below the convection zone are well understood (Brun, Turck-Chièze Zahn 1999 , Charbonnel Talon 2005)."
Substring: ["(Brun, Turck-Chièze Zahn 1999 , Charbonnel Talon 2005)"]
× WRONG: ["Brun, Turck-Chièze Zahn 1999", "Charbonnel Talon 2005"]

Snippet: "Like Caffau et al. (2008a) , we have similar findings."
Substring: ["Caffau et al. (2008a)"]

Snippet: "Momentum balance gives an expression (Fabian 1999 ; Di Matteo, Wilman Crawford 2002 ; King 2003 , 2005)"
Substring: ["(Fabian 1999 ; Di Matteo, Wilman Crawford 2002 ; King 2003 , 2005)"]
× WRONG: ["(Fabian 1999)", "(Di Matteo, Wilman Crawford 2002)", "(King 2003 , 2005)"]

Snippet: "This first contribution describes the cluster dark matter halo (and implicitly assumes that the ICM follows a similar distribution; see Semboloni et al. (2011) for a study where this assumption is relaxed)."
Substring: ["Semboloni et al. (2011)"]
× WRONG: ["(Semboloni et al. (2011))"]

Snippet: "Sweatman (2021) ignored data from carbon spherules generated by YDIH proponents."
Substring: ["Sweatman (2021)"]
× WRONG: ["(Sweatman 2021)"]

Snippet: "In the early Universe, when the metal content was extremely low, enrichment by a single supernova could dominate preexisting metal contents (e.g., Audouse Silk 1995 ; Ryan, Norris Beers 1996)."
Substring: ["Audouse Silk 1995 ; Ryan, Norris Beers 1996"]

Snippet: "Caffau and Brun (2009) made several imaging advances primarily to identify quasars"
Substring: ["Caffau and Brun (2009)"]
× WRONG: ["(Caffau and Brun 2009)"]

Snippet: "Turck-Chieze(2005) identified when metal content was extremely low"
Substring: ["Turck-Chieze(2005)"]
× WRONG: ["(Turck-Chieze 2005)"]

Snippet: "Smith et al. (2003) noted a correspondence between the cessation of high latitude CMEs and the polar field reversal."
Substring: ["Smith et al. (2003)"]

Snippet: "In work by Jones (2000a,b) it was demonstrated that dark matter detection can be elegantly detected."
Substring: ["Jones (2000a,b)"]

Snippet: This is a process that should conclude within millions of years (e.g., Guillot et al. 1996).
Substring: ["(e.g., Guillot et al. 1996)"]

Snippet: "This paper describes the cluster (and assumes similar principles; see Boboni et al. (2011) where this assumption is relaxed)."
Substring: ["Boboni et al. (2011)"]

Snippet: "This is consistent with previous results (Pereira et al., in preparation)."
Substring: ["(Pereira et al., in preparation)"]

Snippet: "More imaging showed significant expansion (Jones, Ryan, et al., submitted)"
Substring: ["(Jones, Ryan, et al., submitted)"]

Snippet: "It also hosts a Compton-thick AGN in the Western component (Smith 2004), observed directly in hard X-rays (Della Ceca et al. 2002 ; Ballo et al. 2004).
Substring: ["(Smith 2004)", "(Della Ceca et al. 2002 ; Ballo et al. 2004)"]

Snippet: "Even prior to the Ballo et al. (2013) result on the potential rotationally supported plane in M31..."
Substring: ["Ballo et al. (2013)"]

```
Snippet: "Figure 3. Effects on N-body mass functions of different group-finding schemes."
Substring: []

##
Keeping the rules and examples above in mind, identify the citation substrings in the following
snippet:
```

Create Citation Tuples Prompt

```
Task:
Given a single input string containing one or more inline research citations, extract and output
them as a JSON array of [author(s), year] pairs. Follow these rules exactly:

1. Match patterns where an author block is followed by a 4-digit year, optionally suffixed by
letters (e.g. 2008a, 2008b).
2. Author blocks may include letters, spaces, commas, ampersands (&), "et al.", hyphens, and
accented characters.
3. Only split into separate citations at:
- Semicolons (";")
- Commas or spaces immediately preceding a 4-digit year (with optional letter suffix)
- Parentheses or brackets around the citation—but ignore parentheses after matching.
4. Do NOT split author names joined by "and" or "&". Do NOT split consecutive author names (e.g.
Smith Jones 1997, van Hoek Smith 1997)
5. For entries with multiple years after the same author block (e.g. "Smith 1997, 1998"), output
one citation per year, reusing the same author block.
6. Discard any citations without a 4-digit year, or those indicating "submitted", "in
preparation", etc.
7. Output ONLY the JSON array (no extra text or explanation).

Examples:
"""
Input: "Ussiri and Lal, 2013"
Output: [["Ussiri and Lal", "2013"]]

Input: "Ussiri Lal 2013"
Ouptut: [["Ussiri Lal", "2013"]]

Input: "Caffau et al. (2008a)"
Output: [["Caffau et al.", "2008"]]

Input: "Sweatman & Young, 2000"
Output: [["Sweatman & Young", "2000"]]

Input: "Sweatman and Young, 2000"
Output: [["Sweatman and Young", "2000"]]

Input: "van Der Monde 1997"
Output: [["van Der Monde", "1997"]]

Input: "Turck-Chieze(2005)"
Output: [["Turck-Chieze", "2005"]]

Input: "Jones (2000a,b)"
Output: [["Jones", "2000a"], ["Jones", "2000b"]]

Input: "Caffau and Brun (2009)"
Output: [["Caffau and Brun", "2009"]]

Input: "Smith and Jones (2003)"
Output: [["Smith and Jones", "2003"]]

Input: "Smith & Jones (2003)"
Output: [["Smith & Jones", "2003"]]

Input: "Smith Jones (2003)"
Output: [["Smith Jones", "2003"]]
```

```
Input: "Brun, Turck-Chièze Zahn 1999 , Charbonnel Talon 2005"
Output: [["Brun, Turck-Chièze Zahn", "1999"], ["Charbonnel Talon", "2005"]]

Input: "Brun, Turck-Chièze Zahn 1999 ; Charbonnel Talon 2005"
Output: [["Brun, Turck-Chièze Zahn", "1999"], ["Charbonnel Talon", "2005"]]

Input: "Fabian 1999 ; Di Matteo, Wilman Crawford 2002 ; King 2003 , 2005"
Output: [["Fabian", "1999"], ["Di Matteo, Wilman Crawford", "2002"], ["King", "2003"], ["King",
"2005"]]

Input: "Smith et al. (2002)", "Jones (2004)"
Output: [["Smith et al.", "2002"], ["Jones", "2004"]]

Input: "Audouse Silk 1995 ; Ryan, Norris Beers 1996"
Output: [["Audouse Silk", "1995"], ["Ryan, Norris Beers", "1996"]]

Input: "This is consistent with previous results (Pereira et al., in preparation)."
Output: []

Input: "More imaging showed significant expansion (Jones, Ryan, et al., submitted)"
Output: []
"""

Extract the inline citations from the following text:
Input:
```

A.1 Document Representation by 'Contributions'

Contribution Extraction Prompt - v1

```
You are an expert research assistant. Summarize the original scientific findings of the
following research paper. Then write them out as a list of strings.

## Task
- Write the findings as a JSON array of strings, each finding being one string element.
- Do not include acknowledgments or references: focus only on the original research
contributions made in this paper
- Only write out the JSON array, do not include any other text or formatting.

Example output:
{{
findings: ["Contribution 1", "Contribution 2", "Contribution 3"]
}}

## Paper:
{paper}
```

Contribution Extraction Prompt - v2

```
You are given the full text of a single scientific research paper. Extract all factual
statements asserted or utilized in this paper, including both novel contributions and
descriptions of methods/instruments/data used.

Hard requirements:

1. Output ONLY a single JSON object: {{"findings":[ ... ]}}. No extra text, no commentary, no
additional keys.
2. Each array item must be a single, self-contained declarative sentence (one idea per
sentence). Avoid pronouns and author-phrases (no "we", "authors", "found", "concluded",
"determined", "observed"). State the fact directly (e.g., "Stellar bulge contains significant
population of young stars").
3. Ensure each statement is fully interpretable on its own. Include definitions or clarifying
context for all domain-specific terms, abbreviations, or categories:
```

```
- Include the subject and scope (e.g., "in the galactic Bulge", "in sample of 200 patients", "in
dataset X", "in cohort aged 50-70").
- If the sentence refers to a term defined within the paper, briefly include the definition or
context so it is fully interpretable out of context
- If the sentence refers to a method/analysis, name the method and general dataset
characteristics (e.g., "High-resolution spectroscopy of F/G dwarf stars").
- Focus on general methodological approaches, broad findings, and conceptual contributions
rather than specific numeric values. Use descriptive terms like "significant", "substantial",
"moderate", "enhanced", "reduced" instead of precise measurements.
3. Ensure each statement is fully interpretable on its own. Include definitions or clarifying
context for all domain-specific terms, abbreviations, or categories:
- Include the subject and scope (e.g., "in the galactic Bulge", "in sample of 200 patients", "in
dataset X", "in cohort aged 50-70").
- If discussing a subset, subgroup, or category, ALWAYS include what it's a subset OF.
- BAD: "Outliers have masses from 9.4 to 11.1"
- GOOD: "Galaxies offset from mass-metallicity relation (MZR) have masses from 9.4 to 11.1"
- If the sentence refers to a term defined within the paper, briefly include the definition or
context so it is fully interpretable out of context.
- For well-established acronyms and abbreviations in the field, include BOTH the full term and
acronym on first use:
- Use the format: "full term (ACRONYM)"
- Examples: "mass-metallicity relation (MZR)", "cosmic microwave background (CMB)",
"supermassive black hole (SMBH)", "active galactic nucleus (AGN)", "asymptotic giant branch
(AGB)", "second generation (2G)"
- Apply this to: major concepts, standard measurements, established relations, common
astrophysical objects, and widely-recognized instruments
- If a term appears in multiple statements, include the acronym definition in each statement
(since each must be independently interpretable)
- If the sentence refers to a method/analysis, name the method and general dataset
characteristics (e.g., "High-resolution spectroscopy of F/G dwarf stars").
- Focus on general methodological approaches, broad findings, and conceptual contributions
rather than specific numeric values. Use descriptive terms like "significant", "substantial",
"moderate", "enhanced", "reduced" instead of precise measurements.
4. Prefer concise, general statements ($\le 160$ characters when possible) that capture the
essence of findings.
5. Extract these types of statements (when present):
- original findings/contributions
- methodologies & experimental design (instruments, general datasets, analytical approaches)
- qualitative and broad quantitative results
- instruments/tools used (even if established), with their purpose and capabilities
- conclusions, interpretations, and statements motivated by the paper's findigns
- method-observation pairs (e.g. [Method X] shows/detects/etc. [Observation Y])
- limitations/caveats/assumptions
- new resources (data/code/models)
- noteworthy comparisons to prior work
- open questions, unresolved problems, or acknowledged uncertainties. Prefix them with "Open
question:"
6. For each statement, write out findings at MULTIPLE levels of specificity:
- High-level interpretations: a broader, more general, citation-style statement
- Mid-level: General findings with key qualifiers ($\le 160$  chars)
- Detailed: Specific conditions, parameters, and quantitative results
- If a finding can have multiple high-level interpretations, write out all of them
- ALWAYS write statements with complete context
7. Before finalizing each statement, verify: "If I gave this sentence to someone unfamiliar with
the paper, would they understand what population/sample/reference frame is being discussed?" If
no, add the missing context.

## Research Paper

  {paper}

## Example expected output:

{{
"findings": ["finding 1...", "finding 2...", ...]
}}
```

CARDS: Citation-Aware Research Diffusion Summarization

Autumn Toney-Wails
UNU-MERIT
toney@merit.unu.edu

Katherine N. Quinn
Georgetown University
kq57@georgetown.edu

Steph Batalis
Georgetown University
sb2132@georgetown.edu

Abstract

Tracing how knowledge diffuses through scientific networks reveals how research builds on prior findings, but doing so at scale remains challenging. We present an LLM-based approach, Citation-Aware Research Diffusion Summarization (CARDS), which describes how knowledge diffuses through the scientific landscape. We trace citations from research funded by the U.S. National Institutes of Health through the landscape quantified by CSET's Map of Science cluster dataset to generate Research Diffusion Pathways, and compare LLM-generated summaries based on (1) cluster-level metadata and (2) citation-aware connections between research clusters. We find that CARDS outperforms cluster-based summaries under both human annotation and LLM-as-a-judge evaluation. Our work outlines a methodology for generating research diffusion paths, and for summarizing them with greater specificity.

1 Introduction

LLM-generated scientific summarization synthesizes information from research publications into concise descriptions of their content. Recent work has shown that incorporating citation context can improve these summaries by providing information about how research is referenced and used by subsequent work (or vice-versa) (Yasunaga et al., 2019; An et al., 2021; Syed et al., 2023; Nguyen et al., 2025a,b; Zhang et al., 2026). We combine these findings with prior work using citation network analysis to trace research diffusion (Chen and Hicks, 2004; Lewison et al., 2005; Liu and Rousseau, 2010; Gao and Guan, 2012; Yue et al., 2022; Aman and Gläser, 2025) and summarize the scientific knowledge underlying diffusion across research areas with scientific community detection and clusters (Boyack and Klavans, 2014; Traag et al., 2019; Rahkovsky et al., 2021).

We present **CARDS** (**C**itation-**A**ware **R**esearch **D**iffusion **S**ummarization), an LLM-based approach for summarizing these connections. Our pipeline has two components: (1) we construct Research Diffusion Pathways (RDPs), sequences of clusters representing distinct scientific areas, defined using citations and semantic similarity; and (2) we summarize the cited and citing publications at each inter-cluster connection (*diffusion point*) to describe the scientific knowledge linking the two research areas (*link summary*). CARDS synthesizes these link summaries into an RDP-level description of how research develops across the pathway.

To evaluate CARDS, we use research funded by the U.S. National Institutes of Health (NIH) as a case study, tracing its diffusion across scientific research areas, with data available on Zenodo at https://doi.org/10.5281/zenodo.23264540. We compare CARDS with a baseline that summarizes the same pathways using cluster-level descriptions rather than citation-level evidence. Using human and LLM-as-a-judge evaluations, we assess linkage coherence, knowledge diffusion specificity, factual consistency, and summary coverage. Both human and LLM-as-a-judge evaluations find CARDS summaries more specific than cluster-based summaries.

2 Experimental Design

Our experimental design consists of four stages: (1) linking NIH-funded publications to research clusters, (2) constructing RDPs that trace citations across clusters, (3) generating RDP summaries, and (4) evaluating a sample of RDP summaries.

2.1 Research Publication Data

We use two publicly available data sources: (1) the NIH RePORTER database[1] and (2) the Center for Security and Emerging Technology (CSET) Map of Science (Quinn et al., 2026).[2] NIH Re-

[1] https://reporter.nih.gov/
[2] https://sciencemap.eto.tech/

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 77–90
November 10, 2026 ©2026 Association for Computational Linguistics

PORTER is an online repository of NIH-funded research projects and their associated research outputs, including publications resulting from individual awards. CSET's Map of Science organizes approximately 146 million scientific articles into 91,594 research clusters using the Leiden community detection algorithm (Traag et al., 2019) over a hybrid network of citation relationships and semantic similarity links. For each research cluster, CSET provides aggregated metadata from the member publications and an LLM-generated (Gemini 3.1 Flash) title and one-sentence summary of the cluster. Using PubMed IDs, we link 2,987,570 NIH-funded publications identified through RePORTER to their corresponding research clusters (49,441 clusters in total). This linkage allows us to identify where NIH-funded research is located within the broader scientific landscape and provides the starting point for tracing its diffusion across research clusters. For our experiments, we rank research clusters by their concentration of NIH-funded publications and select the top 10% as starting points for RDP construction.

2.2 Research Diffusion Pathways

To map how research diffuses across science, we first build Research Diffusion Pathways (RDPs), sequences of research clusters connected through inter-cluster citations. Citation links can exist between clusters (via publications that cite or are cited by other work outside of their cluster) and connect publications across distinct areas of science. We refer to these inter-cluster connections as *diffusion points*, which provide observable instances in which research is cited outside the research area to which it is most closely connected.

We construct RDPs by iteratively following the publications involved in each diffusion point. The citing publications from the preceding diffusion point become the set of publications tracked in the next step. Candidate diffusion points for each subsequent step are ranked by the number of these tracked publications cited by publications in each possible destination cluster. To preserve the temporal direction of diffusion, we additionally require a minimum three-year difference in average publication year between consecutive links. Previously visited clusters are excluded, preventing cycles within a pathway. Figure 1 illustrates an RDP with two diffusion points connecting three clusters.

To construct RDPs of NIH-funded research, we use the clusters identified in Section 2.1 as starting

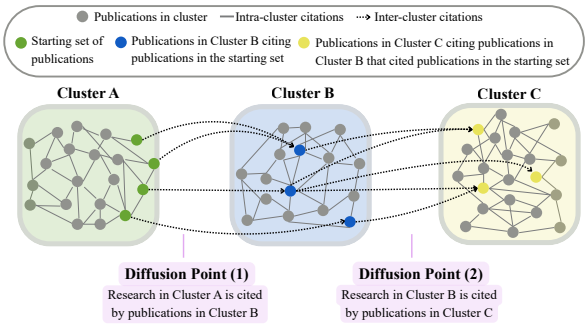


Figure 1: Example of an RDP connecting three clusters with two diffusion points.

points. For each starting NIH cluster, we select a diffusion point that maximizes the number of cited NIH-funded publications. We continue in this manner, selecting the RDP at each iteration which maximizes the number of distinct publications cited at each diffusion point (a greedy RDP), until we have an RDP for each starting cluster with four diffusion points connecting five clusters.

2.3 Research Diffusion Summarization

We test two LLM-based approaches for RDP summary generation that differ in their input: (1) cluster-based summaries and (2) citation-aware summaries (CARDS). By comparing these two approaches, we analyze whether the use of information specific to the diffusion points (i.e., the underlying citation relationships) provides a richer description of how research diffuses across science. We use Gemini 2.5 Flash for all experiments, and all summaries are restricted to 3–5 sentences with `MAX_TOKEN` $= 350$; the full prompts are listed in Appendix A. Figure 2 illustrates the full pipeline for RDP input to summary output, for the baseline cluster-based summary approach and CARDS.

The cluster-based RDP summaries are generated using the one-sentence description for each cluster in CSET's Map of Science (e.g., "Explores molecular mechanisms of cholesterol transport and the involvement of ABC transporters in human health and disease."). Thus, these summaries characterize the general diffusion based on changes in the broader research areas represented by consecutive clusters, without directly observing the publications responsible for the inter-cluster citation links.

CARDS are generated via a two-step summarization pipeline: (1) create a *link summary* for each diffusion point in a given RDP, and then (2) use the link summaries to generate an overall RDP summary. To create a link summary, we identify

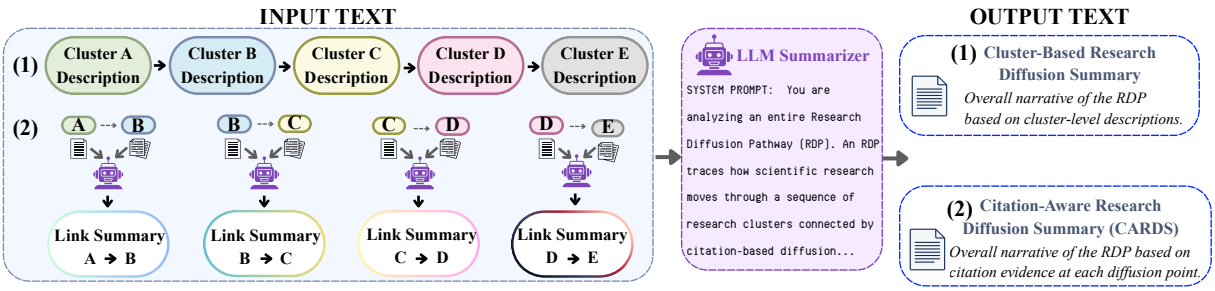


Figure 2: Experimental pipeline for cluster-based diffusion summaries and CARDS.

inter-cluster citation relationships in a diffusion point, preserving their one-to-many structure (i.e., each cited publication is paired with its set of citing publications). We provide the titles and abstracts of these publications to Gemini-2.5-Flash in a prompt to summarize the scientific content connecting the two clusters; we keep the titles but remove the abstracts of citing publications with the fewest citations until the prompt has an estimated 850K tokens to meet context window constraints. Rather than describing the broader research areas represented by the clusters, the link summaries focus specifically on the concepts, methods, findings, or applications associated with the citation relationships through which research crosses cluster boundaries.

2.4 Generative Summary Evaluation

We evaluate a random sample of 81 RDPs and their corresponding summaries with human and LLM-as-a-judge annotation. RDPs span a wide range of scientific research areas, and evaluating the factual consistency of the citation-aware summaries would require annotators to review the titles and abstracts of potentially thousands of publications underlying the diffusion points. We therefore use an automated evaluation to assess factual consistency, while focusing human annotation on two characteristics of the resulting RDP summaries: (1) linkage coherence and (2) knowledge diffusion specificity.

Human Evaluation: For linkage coherence, annotators rated whether each summary provided a coherent and plausible explanation of how research connects across the pathway on a five-point Likert scale (1 = *not coherent*; 5 = *extremely coherent*). For knowledge diffusion specificity, annotators compared the cluster-based summaries and CARDS for the same RDP and selected which more specifically described the concepts, methods, findings, or applications connecting the pathway, or indicated that they were approximately equally specific. Four annotators first evaluated 40 RDPs for linkage coherence, with either the cluster-based or citation-aware summary randomly presented, and then evaluated the remaining 41 RDPs through paired comparisons for knowledge diffusion specificity (a different subset of the sample to avoid annotator priming). For the paired comparisons, summary type was blinded and presentation order randomized. Full annotation instructions are provided in Appendix B.

LLM-as-a-Judge: We complement the human evaluation using DeepEval (Ip and Vongthongsri, 2026), an LLM-based evaluation framework, with GPT-5.4 as the LLM. For hallucination and factual consistency checks we implement the `HallucinationMetric`, which takes in the input text and the generated summary and then provides a score between 0 (hallucinations) to 1 (no hallucinations), and the `SummarizationMetric`, which provides a coverage score between 0 (no coverage) and 1 (full coverage). To run custom evaluations of linkage coherence and knowledge diffusion specificity, we use DeepEval's `GEval` (single input) and `ArenaGEval` (pairwise input) frameworks, respectively. For these custom evaluations, we provide the LLM with evaluation criteria corresponding to the instructions used in the human annotations.

3 Results

We identify 4,826 greedy RDPs originating in NIH-concentrated clusters[3] (118 of the original set of 4,944 were dropped due to incomplete paths); across this set, diffusion points contain a median of 384 citation relationships and 190,876 input tokens. Figure 3 shows three examples of representative RDPs from this set: (1) tracing how research on the immune response to viral infections provided the basis for the later development and production of next-generation CAR T-cell therapies, (2) how

[3]All link summaries and CARDS are available at: https://doi.org/10.5281/zenodo.23264540

CARDS Example #1

This RDP begins by establishing the critical role of tissue-resident memory T (TRM) cells in protection against respiratory viral infections, focusing on mechanisms of their generation, maintenance, and function, including antigen recognition and cytokine signaling. The pathway then progresses to investigate the broader mechanisms of CD8+ T cell differentiation, memory formation, and exhaustion in viral infections and cancer, emphasizing the heterogeneity of T cell subsets, particularly TCF1+ stem-like CD8+ T cells, and the influence of various transcription factors and cytokines. This understanding of T cell dysfunction and exhaustion, including specific transcription factors and epigenetic programs, is subsequently applied to develop strategies for enhancing CAR-T cell efficacy by preventing exhaustion, promoting stemness, and improving persistence through mitochondrial transfer, epigenetic reprogramming, and cytokine optimization. The pathway culminates in the development and evaluation of next-generation CAR T-cell therapies, moving from understanding limitations to designing novel approaches like combining CAR T-cells with small molecule drugs, optimizing manufacturing for less-differentiated cells, and identifying predictive biomarkers for improved clinical application.

CARDS Example #2

This RDP begins with the development and application of isotopic tracer methodologies and analytical techniques to quantify gluconeogenesis, glycogenolysis, and glucose turnover, primarily to elucidate the pathophysiology of diabetes and evaluate therapeutic strategies. This foundational understanding of glucose metabolism and insulin action then transitions into the development of mathematical and computational models of the glucose-insulin system, particularly for simulating postprandial glucose dynamics and designing control algorithms for artificial pancreas systems. Subsequently, these physiological and mathematical models are leveraged to create personalized blood glucose prediction models, integrating machine learning and deep learning techniques for advanced diabetes management. The pathway culminates in the application and optimization of these predictive models within broader diabetes management systems, including early detection, risk assessment, and personalized treatment recommendations, often integrated into mobile health applications and IoT systems.

CARDS Example #3

This Research Diffusion Pathway begins by characterizing the biochemical and functional properties of inositol polyphosphate kinases and phosphatases, establishing their roles in regulating inositol phosphate and phosphoinositide levels. This foundational understanding then progresses to elucidate the broader physiological and pathological roles of phosphoinositides and their metabolizing enzymes in diverse cellular processes like membrane trafficking, signal transduction, and metabolism, linking their dysregulation to diseases such as cancer and neurodegeneration. The pathway further narrows its focus to the phosphoinositide 3-kinase (PI3K)/Akt pathway, investigating its specific isoforms, regulatory mechanisms, and critical involvement in cancer development, progression, and drug resistance, as well as its regulation of cell growth and survival. Ultimately, the pathway culminates in the application of this knowledge to develop and evaluate PI3K/AKT/mTOR pathway inhibitors, such as capivasertib, alpelisib, and ipatasertib, for breast cancer treatment, including identifying predictive biomarkers and understanding mechanisms of therapeutic resistance.

Figure 3: Example CARDS output.

research in liver and kidney function enabled AI-driven disease prediction, and (3) how progression of basic research into single molecular functions led to breakthrough treatments for breast cancer (link summaries in Appendix C).

Figure 4 illustrates how NIH-funded research diffuses across domains. As expected given the NIH's focus on biomedical research, initial clusters are predominantly classified as biology (78.6%). However, after following the RDPs through four diffusion points, the proportion of biology clusters has decreased to 59.5%, while medicine grew from 1.9% to 13.4%, social science from 16.2% to 19.4%, materials science from 0.4% to 2.8%, and computer science from 0.3% to 1.7%. RDPs

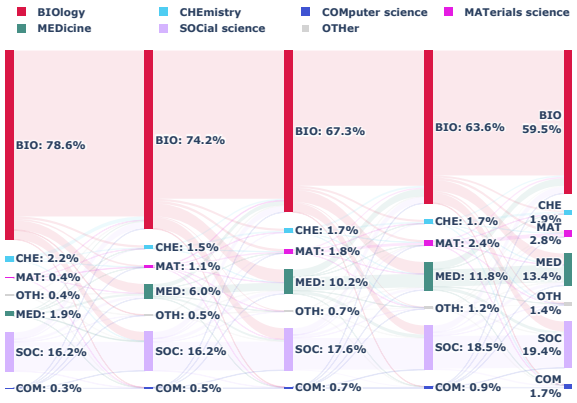


Figure 4: Field composition of the 4,826 NIH RDPs at each of the five cluster positions. Percentages give the fraction of clusters at that position assigned to each field (CSET field classification).

become more diffuse throughout the topical landscape, consistent with knowledge spreading from basic research (biology) to applied (medicine).

3.1 Evaluation

Human annotators find both cluster-based summaries and CARDS to be coherent. CARDS are rated as *Extremely coherent* or *Very coherent* 79.0% of the time, and cluster-based summaries 60.0% of the time (using a binomial test against a 2/5 null probability for random guessing with $N = 40$, obtaining $p < 0.01$ in both instances). Aggregated results are shown in Figure 5.

When comparing between two summaries for the same RDP, human annotators rated CARDS as more specific in describing the flow of information than the cluster-based summaries 70.7% of the time. A binomial test of CARDS-preferred versus all other responses gives $p = 0.017$ (with $N = 41$ majority vote against random guessing). Fleiss' κ is 0.073, indicating slight agreement between annotators. Nearly all disagreements arise when one annotator rates the citation-aware summaries as more specific and another annotator rates both summaries as equivalent.

We apply DeepEval to the full sample of LLM-generated summaries. No instances of hallucinations were detected in either case, and CARDS were uniformly rated as more specific. Mean coherence for the cluster-based summaries and CARDS is 0.880 and 0.901, respectively, indicating high coherence. Mean coverage is 0.953 and 0.719, respectively. This indicates a slightly higher overall coherence from CARDS, however less coverage

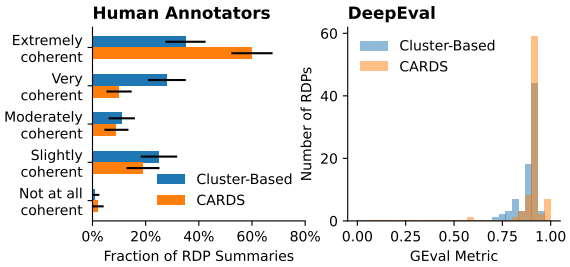


Figure 5: Coherence evaluation of cluster-based RDP summaries and CARDS.

from the input data, likely attributed to the larger input for summarization in the citation-awareness case which must condense numerous papers' titles and abstracts to a few sentences.

4 Discussion and Future Work

The CARDS examples shown in Fig. 3 indicate the type of analysis our proposed pipeline enables. In these three examples, we identify two patterns of scientific development that can be traced and summarized for end-user analysis: the more traditional progression from foundational research to practical applications and the diffusion of knowledge across scientific domains, where research in one field is taken up and extended in another. These examples demonstrate how CARDS can be used to better understand the connections between different areas of research and how scientific knowledge builds upon and extends previous findings.

Research progression from basic, foundational knowledge to application: CARDS Example #1, which starts in research cluster 22564[4] and ends in cluster 4838[5], follows CART T Cell therapy development which, a widely recognized important scientific and medical breakthrough. Along with CARDS Example #3, which starts in cluster 24449[6] and ends in cluster 17417[7] and explores glucose metabolism in the liver and kidneys through to broader health impact such as diabetes, these pathways follow foundational knowledge through to their applications.

Research diffusion *across* domains: CARDS Example #2 traces AI-enabled disease prediction. It begins in a biology-focused area, cluster 37912[8], which establishes scientific principles behind glucose metabolism, continues through social science with cluster 1156[9] that examines control systems and clinical practices for automated insulin delivery, and ends in a computer science with cluster 2002[10] to develop AI systems to predict blood glucose levels and diabetes. This RDP illustrates how knowledge can be discovered in one domain and passed through multiple other domains, where it continues to develop.

Our ongoing work uses RDPs and CARDS to analyze and evaluate the impact of funding portfolios (such as the NIH case study presented here). Specifically, we aim to better understand how funding decisions shape downstream scientific developments and the extent to which foundational research contributes to applications beyond its original scope. Take the CARDS Example #3: funding research on breast cancer alone is unlikely to have led to the development of this class of cancer drugs, because the mechanism they target was initially not known to be connected to breast cancer; it had to be discovered and characterized *before* being connected to breast cancer and leveraged as a therapeutic target.

5 Conclusion

We find that CARDS outperforms cluster-based approaches for summarizing the diffusion of knowledge through the scientific landscape. Human annotators and an LLM judge rate both summary types as coherent, but find CARDS summaries more specific. By incorporating citation-aware features in the summarization process, knowledge diffusing through the scientific landscape can be traced in a more specific manner than with cluster-level descriptions alone. These approaches can help address a range of research questions about how science develops, including how disruptive findings in one domain influence research across other fields, how breakthroughs draw upon knowledge from multiple domains, whether individual discoveries can have outsized impacts on subsequent research, and how long it takes for foundational research to develop into practical applications or contribute to multiple downstream research pathways.

[4] https://sciencemap.eto.tech/cluster/?version=4&cluster_id=22564

[5] https://sciencemap.eto.tech/cluster/?version=4&cluster_id=4838

[6] https://sciencemap.eto.tech/cluster/?version=4&cluster_id=24449

[7] https://sciencemap.eto.tech/cluster/?version=4&cluster_id=17417

[8] https://sciencemap.eto.tech/cluster/?version=4&cluster_id=37912

[9] https://sciencemap.eto.tech/cluster/?version=4&cluster_id=1156

[10] https://sciencemap.eto.tech/cluster/?version=4&cluster_id=2002

References

Valeria Aman and Jochen Gläser. 2025. Investigating knowledge flows in scientific communities: The potential of bibliometric methods. *Minerva*, 63(1):155–182.

Chenxin An, Ming Zhong, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2021. Enhancing scientific papers summarization with citation graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12498–12506.

Kevin W Boyack and Richard Klavans. 2014. Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology*, 65(4):670–685.

Chaomei Chen and Diana Hicks. 2004. Tracing knowledge diffusion. *Scientometrics*, 59(2):199–211.

Xia Gao and Jiancheng Guan. 2012. Network model of knowledge diffusion. *Scientometrics*, 90(3):749–762.

Jeffrey Ip and Kritin Vongthongsri. 2026. DeepEval: The Open-Source LLM Evaluation Framework.

Grant Lewison, Isla Rippon, and Steven Wooding. 2005. Tracking knowledge diffusion through citations. *Research Evaluation*, 14(1):5–14.

Yuxian Liu and Ronald Rousseau. 2010. Knowledge diffusion through publications and citations: A case study using esi-fields as unit of diffusion. *Journal of the American Society for Information Science and Technology*, 61(2):340–351.

Quoc-An Nguyen, Xuan-Hung Le, Hoang Quynh Le, and Hoang-Quynh Le. 2025a. Beyond the scientific document: A citation-aware multi-granular summarization approach with heterogeneous graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5034–5046.

Quoc-An Nguyen, Xuan-Hung Le, Thi-Minh-Thu Vu, and Hoang-Quynh Le. 2025b. Citearxiv: A citation-enriched dataset and heterogeneous graph-based model for scientific articles summarization. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 3084–3087.

Katherine Quinn, Rebecca Gelles, Jack Le, and Catherine Aiken. 2026. Map of science and map of patent cluster data.

Ilya Rahkovsky, Autumn Toney, Kevin W Boyack, Richard Klavans, and Dewey A Murdick. 2021. Ai research funding portfolios and extreme growth. *Frontiers in Research Metrics and Analytics*, 6:630124.

Shahbaz Syed, Ahmad Dawar Hakimi, Khalid Al Khatib, and Martin Potthast. 2023. Citance-contextualized summarization of scientific papers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8551–8568.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. 2019. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7386–7393.

Zenghui Yue, Haiyun Xu, Guoting Yuan, and Yan Qi. 2022. Modeling knowledge diffusion in the disciplinary citation network based on differential dynamics. *Scientometrics*, 127(12):7593–7613.

Hongpan Zhang, Bhummanat Sittipongpittaya, and Chongzhi Zang. 2026. Geneknow: An auditable ai framework for source-grounded biological evidence synthesis. *bioRxiv*.

A RDP Summary Prompts

Our experiments require three prompt variations: (1) cluster-based RDP summary, (2) link summary, and (3) citation-aware RDP summary.

First, the prompt shown in Figure 6 was used to generate RDP summaries from the ordered cluster-level titles and summaries provided by the CSET Map of Science. The prompt in Figure 7 was used to generate a link summary for each diffusion point in an RDP. The input preserves the citation relationships between cited publications in the source cluster and citing publications in the destination cluster. Lastly, Figure 8 displays the prompt for the CARDS output, taking the summaries generated at the RDPs diffusion points as input.

B Human Annotation Task

Four undergraduate research assistants completed the annotation task during their regular working hours using Qualtrics. Annotators were provided with a brief overview of the project and links to the underlying datasets to provide background on the study and research context, before they began their evaluations. The instructions given in the annotation interface are provided below:

Overall Instructions

In this task, you will evaluate summaries describing pathways of scientific research. Each pathway represents research that connects across multiple areas of scientific activity. The summaries are intended to explain how research connects and develops across these pathways. You will be asked to evaluate summaries based on two criteria:

- **Linkage Coherence:** How coherently the summary explains the connections across the research pathway.
- **Knowledge Diffusion Specificity:** Which of two summaries provides more specific information about the research connecting different parts of the pathway.

There are no right or wrong answers. Please base your ratings on the information presented and your judgment of how well each summary describes the research pathway.

B.1 Linkage Coherence Task

You will be shown one summary of a research diffusion pathway at a time.

Please evaluate whether the summary provides a coherent and plausible explanation of how the research connects across the pathway.

When making your judgment, consider whether:

- the relationships between different areas of research are understandable;
- the summary provides a logical account of how the research connects from one part of the pathway to another; and
- the pathway reads as a connected progression of research rather than as a collection of unrelated research topics.

Annotation Task:

How well does this summary provide a coherent and plausible explanation of how the research connects across the pathway?

Annotation Labels:

1. **Not at all coherent:** The summary does not provide an understandable explanation of how the research connects across the pathway.
2. **Slightly coherent:** Some connections are apparent, but the relationship between parts of the pathway is mostly unclear.
3. **Moderately coherent:** The general connections across the pathway are understandable, although some transitions or relationships are unclear.
4. **Very coherent:** The summary provides a clear and logical explanation of how research connects across most of the pathway.
5. **Extremely coherent:** The summary provides a clear, logical, and well-integrated explanation of how research connects throughout the pathway.

B.2 Knowledge Diffusion Specificity Task

You will now be shown two summaries describing the same research pathway. The summaries were generated using different information about the pathway. Compare the summaries based specifically on how well they describe the research or knowledge connecting different parts of the pathway.

More specific summaries may identify, for example:

- scientific concepts or ideas;

Cluster-Based RDP Summarization Prompt

Prompt

You are analyzing an entire Research Diffusion Pathway (RDP). An RDP traces how scientific research moves through a sequence of research clusters connected by citation-based diffusion.

Below are the research clusters in this RDP in pathway order, from the starting cluster to the final cluster. For each cluster, you are given its cluster title and cluster summary.

Your task is to summarize the scientific trajectory of the entire RDP. Write a concise 3--5 sentence summary explaining how the research focus develops across the ordered sequence of clusters.

Focus on:

- **The starting scientific area.** Identify the main concepts, problems, methods, mechanisms, technologies, or applications represented at the beginning of the pathway.

- **The progression across the pathway.** Explain the major scientific transitions from earlier clusters to later clusters, including changes in topic, method, scale, application, or disciplinary context.

- **Continuity and change.** Identify what scientific ideas or approaches appear to persist across multiple stages and what substantially changes as the pathway progresses.

- **The endpoint.** Describe the broader research area or application represented by the final cluster and how it differs from, or builds conceptually on, the earlier stages.

- **Scientific specificity.** Use concrete scientific terminology from the provided cluster titles and summaries rather than generic descriptions such as "related research," "advances in the field," or "different topics."

- **Evidence only.** Base the summary only on the supplied cluster titles and summaries. Do not invent paper-level citation details, specific mechanisms of transfer, or relationships that are not supported by this information.

Describe the pathway directionally, making clear how the scientific focus changes from the first cluster through the final cluster.

A successful summary should answer:

What scientific trajectory does this RDP represent, and how does the research focus evolve from its starting cluster to its endpoint?

ORDERED CLUSTERS IN THIS RDP:

<cluster_text>

Figure 6: Cluster-Based RDP Summarization Prompt

Link Summarization Prompt

You are analyzing a single link in a Research Diffusion Pathway (RDP). An RDP traces how scientific knowledge moves between clusters of research papers via citations.

The evidence below preserves the actual citation relationships for this link. Each evidence group contains:

- a **CITED SOURCE PAPER** from the source cluster, and

- the **CITING DESTINATION PAPER(S)** from the destination cluster that cite that source paper.

These citation relationships provide evidence about both what scientific knowledge is being transmitted from the source cluster and how that knowledge is being taken up, applied, extended, combined, or reframed in the destination cluster. Your task is to summarize this link only. Write a concise 3--5 sentence summary explaining the research diffusion represented by these citation relationships.

Focus on:

- **What knowledge is being transmitted.** Identify the specific scientific concepts, findings, methods, tools, mechanisms, datasets, technologies, clinical insights, or other contributions in the cited source papers that are carried across the link.

- **How the destination research draws on that knowledge.** Use the citing destination papers to explain how the transmitted knowledge is being used, applied, extended, combined with other ideas, or moved into a different research problem or application.

- **The scientific transition from source to destination.** Make the direction explicit. Prefer formulations such as The destination research draws on..." or The cited source work provides..." when supported by the evidence.

- **The nature of the connection.** When supported by the papers, identify whether the link represents transfer of a method, biological mechanism, empirical finding, conceptual framework, technology, dataset, clinical insight, or another type of scientific contribution.

- **Scientific specificity.** Use concrete scientific terminology rather than generic descriptions such as "related research," "similar methods," "advances in the field," or "a diffusion point."

- **Evidence only.** Base the summary only on the provided cited and citing papers. Do not infer a use or relationship that is not supported by their titles and abstracts.

- **Do not mention paper identifiers.** Paper IDs are provided only to preserve the citation relationships between source and destination papers. Never include paper IDs (for example, identifiers beginning with carticle_) in the summary. Refer to the scientific content of the papers rather than their identifiers.

Describe the actual scientific relationship from the source cluster to the destination cluster.

A successful summary should answer:

What scientific knowledge from the source cluster is crossing this link, and how is the destination-cluster research drawing on or carrying that knowledge forward?

Write the summary as standalone scientific prose. Do not mention paper IDs, evidence-group numbers, source-paper numbers, destination-paper numbers, or other internal identifiers from the input.

CITATION EVIDENCE FOR THIS LINK:

Figure 7: Link Summary Prompt

Citation-Aware RDP Summarization Prompt

Prompt

You are analyzing an entire Research Diffusion Pathway (RDP). An RDP traces how scientific knowledge moves through a sequence of citation-based links between research clusters.

Below are the link-level summaries for one RDP, ordered from the first link in the pathway to the final link. Each link summary describes what scientific knowledge moves from one cluster to the next and how the destination research draws on or carries that knowledge forward.

Your task is to synthesize these link summaries into a single concise summary of the entire RDP. Write a 3--5 sentence summary explaining the scientific trajectory across the pathway as a whole.

Focus on:

- **The starting scientific area.** Identify the main scientific concepts, findings, methods, mechanisms, technologies, or applications represented at the beginning of the pathway.

- **The progression across links.** Explain the major transitions in knowledge as the pathway moves from one cluster to the next. Emphasize how earlier findings, methods, or concepts are carried forward, applied, extended, combined, or reframed at later stages.

- **Continuity and change.** Identify the scientific ideas or approaches that persist across multiple links and the points where the pathway shifts into a different research problem, method, application, or disciplinary context.

- **The endpoint.** Describe the broader scientific area or application represented by the final stage of the pathway and how it relates to the earlier stages.

- **Scientific specificity.** Use concrete scientific terminology from the provided link summaries rather than generic descriptions such as "related research," "advances in the field," or "different topics."

- **Evidence only.** Base the RDP summary only on the supplied link summaries. Do not invent paper-level details, cluster characteristics, citation relationships, or mechanisms of transfer that are not supported by those summaries.

Synthesize the pathway rather than simply listing or repeating the four link summaries. Describe the pathway directionally, making clear how scientific knowledge evolves from the first link through the final link.

A successful summary should answer:

What scientific trajectory does this RDP represent, and how does knowledge evolve as it moves across the sequence of links?

ORDERED LINK SUMMARIES:

<link_summaries>

Figure 8: Citation-Aware Research Diffusion Summarization Prompt

- methods or techniques;
- findings or results;
- data or tools;
- applications; or
- other forms of research knowledge that connect one area of research to another.

Do not select a summary simply because it is longer or contains more detail overall. Focus on whether the additional detail helps explain what research or knowledge connects across the pathway.

Annotation Task:

Which summary more specifically describes the concepts, methods, findings, applications, or other research knowledge connecting different parts of the pathway?

- Summary A is more specific
- The summaries are about equally specific
- Summary B is more specific

C Example RDP Link Summaries

C.1 CARDS Example #1

Link Summary 1: The cited research highlights the critical role of tissue-resident memory T (TRM) cells, particularly in the lung airways, in providing rapid and effective protection against respiratory viral infections. The destination papers draw on this knowledge to investigate the mechanisms governing TRM cell generation, maintenance, and function, emphasizing factors like antigen recognition, cytokine signaling (e.g., IL-2, IL-7, IL-15, TGF-β), and the influence of the tissue microenvironment. This research aims to understand how these factors contribute to the heterogeneity and protective capacity of memory T cell subsets, ultimately informing the development of more effective T cell-based vaccines and immunotherapies for both infectious diseases and cancer.

Link Summary 2: The cited research investigates the complex mechanisms governing CD8+ T cell differentiation, memory formation, and exhaustion in the context of viral infections and cancer. These papers collectively highlight the heterogeneity of CD8+ T cell subsets, the molecular and epigenetic factors that drive their fate decisions, and the impact of the microenvironment on their function and persistence. The destination research draws on this foundational knowledge to explore specific regulatory pathways and therapeutic strategies. For instance, several destination papers utilize the understanding of TCF1+ stem-like CD8+ T cells, a population critical for sustained immune responses and checkpoint blockade efficacy, to investigate factors like TGF-β signaling, IL-15, and various transcription factors (e.g., KLF2, TOX, NR4A3, ID3, BATF, Gfi1, Tle1, PTMA) that influence their maintenance, differentiation, and anti-tumor activity. Other destination papers leverage insights into epigenetic programming, metabolic regulation, and the role of cytokines (e.g., IL-2, IL-10, IL-15, IL-21, IL-33) to develop novel immunotherapeutic approaches, such as targeting specific signaling pathways or using natural products to enhance T cell function and overcome exhaustion in cancer.

Link Summary 3: This RDP link illustrates the transmission of knowledge regarding T-cell dysfunction and exhaustion, particularly in the context of chronic infections and cancer, and its implications for CAR-T cell therapy. The source papers characterize T-cell exhaustion as a state of reduced function, altered transcriptional and epigenetic programs, and highlight the role of specific transcription factors (e.g., TCF1, TOX, BATF, NR4A) and cytokines (e.g., IL-2, TGF-β, IFN-I) in its development and maintenance, as well as the importance of stem-like T-cell populations. The destination papers draw on this knowledge to develop strategies for enhancing CAR-T cell efficacy by preventing or reversing exhaustion, promoting stemness and memory phenotypes, and improving persistence. This includes approaches such as mitochondrial transfer, modulating transcription factors (e.g., FOXO1, BACH2, TET2, NR4A, GSNOR, FOS), epigenetic reprogramming (e.g., targeting SUV39H1, HDACs, TET2, DNMT3A), optimizing cytokine signaling (e.g., IL-10, IL-15, IL-21, CXCL13), and engineering CAR-T cells to resist immunosuppressive microenvironments or improve infiltration.

Link Summary 4: This RDP link illustrates the ongoing efforts to enhance the efficacy and safety of CAR T-cell therapy, particularly in addressing challenges like T-cell exhaustion, tumor microenvironment suppression, and treatment-related toxicities. The source papers contribute foundational knowledge on mechanisms of CAR T-cell dysfunction, strategies to improve CAR T-cell persistence and anti-tumor activity (e.g., through transient "rest," cytokine release, or epigenetic repro-

gramming), and the impact of CAR design and manufacturing protocols on product quality. Destination papers draw on this knowledge to develop and evaluate novel approaches, such as combining CAR T-cell therapy with small molecule targeted drugs, optimizing manufacturing to enrich for less-differentiated T cells, and exploring predictive biomarkers for treatment outcomes and adverse events. This diffusion highlights a scientific transition from understanding the limitations of current CAR T-cell therapies to developing and testing next-generation strategies for improved clinical application.

C.2 CARDS Example #2

Link Summary 1: The cited source papers primarily focus on methods for quantifying gluconeogenesis, glycogenolysis, and glucose turnover in humans using various isotopic tracers (e.g., 2H2O, 13C-labeled glucose, 3H-labeled glucose, 14C-labeled lactate/acetate/alanine, 2H5-glycerol) and techniques like nuclear magnetic resonance spectroscopy and gas chromatography-mass spectrometry. They also investigate the relative contributions of these metabolic pathways to overall glucose production in different physiological states (fasting, postprandial) and in conditions like type 2 diabetes.

The destination research draws on this knowledge by applying and extending these isotopic tracer methodologies and analytical techniques to further investigate glucose metabolism in the context of type 1 and type 2 diabetes, obesity, and pre-diabetes. Specifically, the destination papers use these methods to:

- Assess the mechanisms of fasting and postprandial hyperglycemia in type 2 diabetes, including the roles of increased gluconeogenesis, impaired glycogen synthesis, and altered hepatic autoregulation.
- Quantify insulin sensitivity, beta-cell function, and glucose effectiveness in various diabetic and pre-diabetic populations.
- Evaluate the impact of interventions like caloric restriction and insulin therapy on glucose fluxes and metabolic pathways.
- Develop and validate new mathematical models for estimating glucose metabolism parameters from oral glucose tolerance tests and intravenous glucose tolerance tests.
- Investigate the role of specific organs (liver, kidney, muscle) in glucose homeostasis and their dysfunction in diabetes.

This represents a scientific transition from developing and validating fundamental metabolic measurement techniques to applying these advanced tools to elucidate the complex pathophysiology of diabetes and evaluate therapeutic strategies.

Link Summary 2: The destination research draws on methods and findings related to glucose metabolism and insulin action in both healthy individuals and those with type 1 and type 2 diabetes. Specifically, the cited source papers provide experimental protocols and insights into glucose-induced suppression of glucose production, impaired glucose disposal in type 2 diabetes, the relationship between hepatic glucose production and fasting plasma glucose, and the impact of insulin delivery patterns on carbohydrate tolerance. This knowledge is used by the destination papers to develop and refine mathematical and computational models of the glucose-insulin system, particularly for simulating postprandial glucose dynamics, assessing insulin sensitivity, and designing control algorithms for artificial pancreas systems in diabetes management. The diffusion highlights a transition from physiological studies using tracer techniques and clamp methods to the application and extension of these findings in in silico modeling and control system development for diabetes therapy.

Link Summary 3: The source papers provide a comprehensive foundation in mathematical modeling of glucose-insulin dynamics, including whole-body glucose kinetics, meal absorption patterns, insulin action, and the development of in silico simulators for type 1 and type 2 diabetes. This knowledge is transmitted to the destination research, which focuses on developing and evaluating advanced algorithms for blood glucose prediction and diabetes management. Specifically, the destination papers leverage these physiological and mathematical models to create personalized prediction models, integrate machine learning and deep learning techniques for improved accuracy, and design decision support systems for artificial pancreas applications, often validating their approaches using data generated from the established simulators. This represents a transition from foundational physiological modeling to the application and refinement of these models within predictive and control systems for diabetes.

Link Summary 4: The source cluster introduces various machine learning and deep learning models, including neural networks, support vector machines, and ensemble methods, for predicting blood glucose levels and hypoglycemic/hyperglycemic events in diabetic patients, often incorporating continuous glucose monitoring (CGM) data, meal information, insulin dosages, and physical activity. The destination cluster draws on this knowledge to develop and evaluate new predictive models for diabetes diagnosis, progression, and glucose level fluctuations, frequently employing and comparing different machine learning algorithms, feature selection techniques, and ensemble approaches. This diffusion highlights a transition from foundational glucose prediction models to their application and optimization within broader diabetes management systems, including early detection, risk assessment, and personalized treatment recommendations. The destination research also extends the application of these predictive techniques to different types of diabetes and integrates them into mobile health applications and IoT systems for enhanced patient self-management and clinical decision support.

C.3 CARDS Example #3

Link Summary 1: The cited source papers establish the biochemical and functional characteristics of various inositol polyphosphate kinases and phosphatases, including their substrate specificities, cellular localization, and roles in regulating inositol phosphate and phosphoinositide levels. This knowledge is transmitted to the destination research, which draws on these foundational findings to investigate the broader physiological and pathological roles of phosphoinositides and inositol polyphosphates. Specifically, the destination papers utilize this understanding to explore the involvement of these molecules in diverse cellular processes such as membrane trafficking, cytoskeletal organization, signal transduction, gene expression, and their implications in human diseases like cancer, neurodegeneration, and metabolic disorders. The scientific transition is from identifying and characterizing individual enzymes and their immediate products to elucidating their complex interplay within cellular networks and their impact on overall cell function and disease.

Link Summary 2: The source papers establish the critical roles of phosphoinositides (PIs) and their metabolizing enzymes (kinases and phosphatases) in regulating diverse cellular processes, including membrane trafficking, signal transduction, cell growth, proliferation, and metabolism. They highlight how different PI species are spatially and temporally regulated and how their dysregulation is linked to various human diseases, including cancer and neurodegeneration. The destination research draws on this foundational knowledge to investigate the specific functions of different PI3K isoforms and PI phosphatases in various physiological and pathological contexts. This includes exploring their roles in cancer development and therapy, immune responses, metabolic regulation, and endocytic pathways, often focusing on the therapeutic potential of targeting these enzymes. **Link Summary 3:** The destination research draws on the source cluster's foundational knowledge of phosphoinositide 3-kinases (PI3Ks) and their role in activating Akt (also known as protein kinase B or PKB), a key signaling molecule in various cellular processes. Specifically, the destination papers utilize the understanding of PI3K isoforms, their regulatory subunits, and the generation of phosphoinositide lipids like PI(3,4,5)P3 and PI(3,4)P2, which are crucial for Akt activation and localization. This knowledge is applied to investigate the PI3K/Akt pathway's involvement in cancer development, progression, and drug resistance, as well as its regulation of cell growth, survival, metabolism, and migration. The destination research also explores the isoform-specific functions of Akt and the spatial and temporal control of Akt activation within different cellular compartments, including the nucleus and various membrane structures, to better understand its diverse roles in both physiological and pathological contexts.

Link Summary 4: The cited research establishes the critical role of the PI3K/AKT/mTOR signaling pathway and its components (PI3K, AKT isoforms, PTEN, PDK1) in breast cancer development, progression, and therapeutic resistance. Specifically, it highlights the oncogenic potential of PIK3CA and AKT1 mutations, the distinct functions of Akt isoforms, and the importance of PDK1 in activating downstream kinases. The destination research draws on this knowledge to investigate the efficacy and safety of PI3K/AKT/mTOR pathway inhibitors, such as capivasertib, alpelisib, and ipatasertib, in breast cancer treatment. This includes exploring their use in various breast cancer subtypes, identifying predictive biomarkers for treatment response (e.g., PIK3CA mutations,

PTEN loss, AKT1 E17K mutation), understanding mechanisms of resistance (e.g., mTORC1 activation, feedback loops), and evaluating combination therapies to enhance anti-tumor activity and overcome resistance while managing adverse events like hyperglycemia.

Bridging the Data Bottleneck: A Multi-Modal Generative Extraction Pipeline for Legacy Scientific Literature

Gaurav Kumar
New York University
gk2839@nyu.edu

Anasse Bari
New York University
abari@nyu.edu

David J. Nagel
George Washington University
nagel@gwu.edu

Abstract

The majority of historical experimental data across the empirical sciences remains locked in unstructured legacy formats, which slows data-driven discovery. We introduce an automated multi-modal pipeline that processes large literature corpora, filters non-empirical noise, and outputs structured datasets. Layout-aware document parsing is paired with a lightweight semantic triage gatekeeper to substantially reduce the computational overhead of applying frontier multi-modal models at scale. A generative model reconstructs experimental trajectories from text and figures, while a schema-constrained prompt and a strictly typed programmatic validation layer check physical bounds before serialization. We evaluate the pipeline on a 60-document corpus of extreme-condition thermodynamic literature drawn from Low Energy Nuclear Reactions (LENR) research, reporting concrete throughput, triage, and constraint-violation statistics. We distinguish safeguards enforced in code (physical-bound validators) from those requested through the prompt, such as source grounding, and outline the evaluation needed to verify the latter. The architecture is domain-agnostic by design, and we outline how it could transfer to catalysis, energy storage, and preclinical biomedical research.

1 Introduction

Data-driven science, from battery discovery to drug design, increasingly depends on machine learning. Yet decades of experimental findings sit inside unstructured legacy PDFs. Modern methods such as Graph Neural Networks and Bayesian optimizers need dense numeric tensors, while historical literature buries experimental conditions and kinetic rates inside prose, multi-column tables, and rasterized plots that are effectively invisible to large-scale computational analysis.

Naive OCR and text-extraction libraries destroy spatial structure, collapsing tables and charts into incoherent strings. LLMs parse this text more fluently, but applying them naively introduces new failure modes: hallucinated decimal points (Huang et al., 2025), inconsistent units, and a positivity bias that can silently drop failed experimental runs, precisely the negative examples a reinforcement-learning agent needs to learn the boundaries of a physical system.

We present a multi-modal, automated extraction pipeline designed for large, disordered collections of scientific papers. Its front end has two stages. First, layout-aware Intelligent Document Parsing (IDP) converts legacy publications into spatial Markdown that preserves the position of tables and figures through bounding-box data. Next, a two-stage semantic triage and visual screening step removes papers without scientific substance, along with decorative images, before the core generative model sees them.

To constrain hallucination in the generative step, we combine a prompt-level framework, requesting verbatim source-quote grounding and converting visual inflection points into temporal arrays, with validators that check simple physical bounds on whatever does get extracted. As the Limitations section discusses, we distinguish the two layers explicitly.

Our contributions:

- A multi-modal, dual-LLM architecture that reconstructs legacy scientific PDFs into spatial Markdown and visual tensors, cutting extraction cost through automated semantic triage.
- An anti-hallucination framework whose schema requests exact source-quote grounding for numeric claims and captures negative (null) results so that failed runs stay in the dataset.
- A bifurcated extraction design in which a

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 91–98
November 10, 2026 ©2026 Association for Computational Linguistics

schema-constrained generative model normalizes categorical entities and SI units, while a strictly typed validation layer independently checks physical bounds on the extracted numeric fields, rejecting grossly impossible values.

2 Related Work

Intelligent Document Processing (IDP) OCR-based extraction degrades sharply on documents with complex multi-column layouts or embedded figures. Early layout-aware models such as LayoutLM (Xu et al., 2020), together with datasets such as PubLayNet (Zhong et al., 2019) and DocBank (Li et al., 2020), showed that combining text and spatial information substantially improves document understanding. More recent multi-modal models, including Nougat (Blecher et al., 2024), LayoutLMv3 (Huang et al., 2022), and the OCR-free Donut transformer (Kim et al., 2022), preserve structure such as mathematical formulas and tables, and benchmarks such as DocGenome (Xia et al., 2024) extend this to large-scale scientific documents. At scale, however, they trade speed against accuracy. We combine layout-aware parsing (Docling, Auer et al., 2024) with a lightweight triage stage that keeps a large share of documents away from the expensive model.

LLMs in Scientific Extraction and Hallucination Large language models are increasingly used to extract structured data from unstructured text in materials science, chemistry, and pharmaceutical research, but they can produce factual errors (Huang et al., 2025). HalluMat (Vangala et al., 2025), for example, showed that LLMs often fabricate plausible but incorrect material formulations, and that multiple verification stages reduce this tendency without eliminating it. We call the analogous error "physical hallucination": extracting a value that contradicts established thermodynamics because the model has no firm grounding in physical law. Fact-checking systems such as FActScore (Min et al., 2023) and claim-verification benchmarks such as SciFact (Wadden et al., 2020) address a related problem by retrieving supporting evidence, and retrieval-augmented fact-checking (Lewis et al., 2020) is the prevailing approach, though it can struggle with very sparse datasets. Our approach instead relies on a prompt-level instruction to ground values in source quotes; the current state of independent verification of this step is described in the Limitations section.

Structured Schemas and Constraint Validation Schema-driven generation, such as asking an LLM for JSON or constraining the decoding process itself (Willard and Louf, 2023), yields a consistent format but does not guarantee physically correct values. We therefore add Python validators built on Pydantic (Colvin et al., 2024) that reject any output violating a small set of predefined physical limits, regardless of what the generative model produced.

Domain-Specific Scientific Extraction Closest to our setting: ChemDataExtractor mines chemical properties with classical NLP (Swain and Cole, 2016), MatSciBERT adapts a domain language model for materials-science text mining (Gupta et al., 2022), and Dagdelen et al. (2024) extract structured records from materials papers with instruction-tuned LLMs and a lightweight validation step. We target a harder regime on two axes at once: historical, multi-modal source documents (prose, tables, rasterized plots), and a target domain, Low Energy Nuclear Reactions, with unusually severe reporting variance and positivity bias.

3 The Generative Extraction Architecture

The system follows the six-step process shown in Figure 1, which separates spatial document handling from the LLM's reasoning. As a result, the expensive generative model sees only filtered input that has passed triage.

3.1 Document Parsing and Spatial Preservation

Simple extraction tools such as PyPDF2 and PDFMiner read documents as flat character streams, which merges table cells, destroys multi-column layouts, and causes errors in later inference stages. We instead use a layout-aware Document Layout Analysis system, IBM's Docling (Auer et al., 2024), to identify titles, paragraphs, and tables, and rebuild them as Markdown, keeping tables as Markdown tables rather than loose text strings.

The same pass performs visual extraction: rather than asking the LLM to infer data from a text description of a graph, the DLA engine locates every figure's bounding box and crops it to a standalone image. This bifurcated output, structured

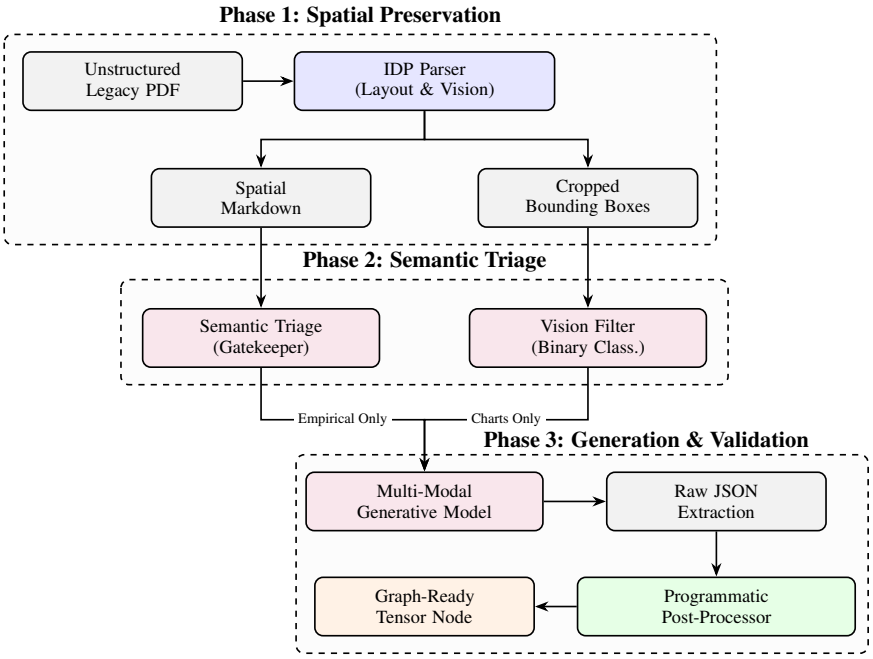


Figure 1: The Generative Extraction Architecture. Legacy PDFs are spatially decomposed into structural Markdown and visual elements. A dual-model triage system filters non-empirical noise before schema-guided multimodal extraction, and a typed post-processing layer rejects values that violate simple physical bounds before serialization.

Markdown plus isolated images, is what feeds the generative stage.

3.2 Semantic Triage (Cost and Token Optimization)

Running a frontier multi-modal model over an entire scraped corpus is expensive, and most scraped papers are theoretical or lack quantitative data, so they are useless to a downstream ML pipeline. We use a two-layer, low-cost triage gatekeeper to avoid paying for that noise.

The first layer performs document-level triage: a lightweight, high-speed LLM (`gemini-3.5-flash`) reads the first 30,000 characters (roughly 10 pages) and answers one boolean question: does this text contain extractable experimental data? Irrelevant or purely theoretical documents are routed out before the expensive model ever sees them.

The second layer filters images. Docling extracts every embedded rectangular region, including logos, headers, and portraits. A lightweight vision model (also `gemini-3.5-flash`), run at temperature 0.0 with output limited to 5 tokens, answers one yes/no question per image: is it a graph or a table? On our 60-document pilot (Section 6), the filter removed 79 of 332 extracted images (23.8%) as unsuitable for scientific analysis, which reduced the visual input volume (Table 1).

3.3 Multi-Modal Generative Extraction

The core generative model, Gemini (Gemini Team, Google, 2023), receives the structured Markdown together with the filtered high-resolution images and links the two modalities, for example when a value appears only on a plot's axis label or in a table's caption, which text-only extraction cannot handle. Its output is a structured but not yet validated JSON payload, which moves the information from natural language into a form that code can check. In the full pipeline, triage uses a lightweight Flash-class model and extraction uses a larger model (`gemini-3-pro-preview`); the 60-paper pilot reported here used `gemini-3.5-flash` for all stages.

4 The Anti-Hallucination Framework

LLMs can struggle with zero-shot numerical extraction: they may misplace decimal points, reverse logical relationships, and lose track of dependencies between variables in complex multi-dimensional settings. Figure 2 shows our approach, which combines three prompt-level instructions (source grounding, negative-result capture, and entity anchoring) with the programmatic validation described in Section 5. The Limitations

section separates what the model is only asked to do from what is checked independently in code.

4.1 Source Grounding and Confidence Scoring

The schema requires a `source_quote` for each key parameter. When the LLM proposes a value, such as a peak temperature of 450 K, it must also supply the exact wording that supports it. This requirement is implemented through the prompt and schema. The programmatic validation layer (Section 5) checks physical limits and data types, but it does not check that each `source_quote` is an exact excerpt of the source document. We describe the status of our audit in the Limitations section; a dedicated substring validator is the natural next step.

The model also emits a single `extraction_confidence_score` (0.0–1.0) per processed paper, reflecting its overall confidence in that document's legibility and completeness. This document-level score lets downstream models down-weight poorly digitized papers, but it cannot distinguish a confidently extracted value from an uncertain one within the same document; we discuss the case for per-parameter confidence in the Limitations section.

4.2 Negative Result Capture and Entity Anchoring

Scientific publications favor successful outcomes, and generative models inherit that bias by underreporting failed configurations, a serious problem when a reinforcement-learning agent needs negative boundaries to understand a system's limits. We counter this with an explicit Negative Bias Trap: the model is prompted to hunt for null runs and baseline controls, setting an `is_failed_replication` flag so downstream models learn from failures as well as successes.

Historical literature also re-publishes the same experiment across venues over many years; treating each paper as an isolated event inflates dataset scale and causes entity-resolution bleed. Entity Anchoring addresses this by capturing citations linking a run to prior publications into a `related_prior_experiments` array, letting the downstream graph merge duplicate runs.

5 Programmatic Validation and Strict Type Enforcement

A syntactically valid JSON payload is not necessarily a physically valid one: the LLM remains stochastic and can produce impossible numeric values even when formatting is perfect. We separate the two concerns: the generative model performs initial normalization under prompt constraints, and a strictly typed validation library, Pydantic (Colvin et al., 2024), independently rechecks the result before it is allowed into the database.

5.1 Entity Normalization and Categorical Enums

Different labs describe the same material differently (e.g., "heavy water bath" vs. "D_2O medium"); naive extraction preserves this fragmentation and floods a knowledge graph with duplicate nodes. We instead bind the model to predefined schema Enums, and Pydantic rejects any unmapped raw string in a categorical field. Since science continuously introduces materials no enum anticipates, unknown values fall back to a generic `OTHER` category paired with a free-text `_notes` field (e.g., `medium: OTHER`, `medium_notes: "5% D2SO4 in D2O"`), so novel materials are preserved without polluting the categorical structure.

5.2 Trajectory Serialization

Experimental parameters like temperature or pressure are rarely static, but naive extraction collapses them to a single scalar, discarding the temporal context time-series models need. We model these instead as `Trajectory` objects: the generative model reads inflection points off a plot (or computes them from a stated rate, e.g., 1 K/min) and serializes them as `[time_offset_seconds, value]` pairs. The validation layer then rejects any trajectory containing a negative `time_offset_seconds`, ruling out causally impossible timelines.

5.3 Enforcing Physical Constraints

Unconstrained LLMs can produce thermodynamically infeasible values or mishandle units, for example by writing a Celsius value into a Kelvin field. We use two layers. First, the prompt instructs the model to convert every measurement to SI units before emitting JSON. Second, Pydantic `@field_validator` functions independently

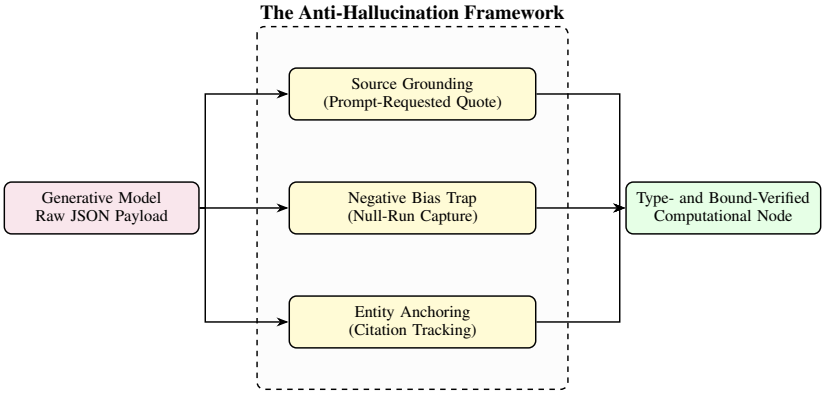


Figure 2: The Anti-Hallucination Framework. Source grounding, the negative-result trap, and entity anchoring are all implemented through the prompt and schema; only physical bounds and types are checked in code (Section 5).

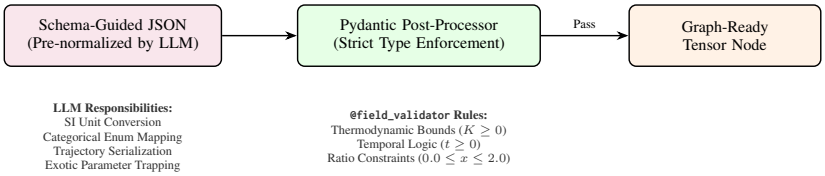


Figure 3: The Programmatic Validation Layer. The generative model handles unit conversion, time-series serialization, and entity normalization via prompt constraints; the Pydantic post-processor independently rejects values that violate physical bounds before serialization.

reject negative temperatures, pressures, and time offsets (along with negative molarity and coefficient of performance) and keep the deuterium-to-palladium loading ratio (`loading_ratio_D_Pd`) within 0.0–2.0. A violation raises a `ValueError`, the payload is rejected, and it is not inserted into the database. These are plausibility checks that catch grossly impossible values, such as a negative temperature in kelvin.

5.4 The Exotic Parameter Escape Hatch

A rigid schema risks silently discarding genuinely novel variables: an unusual stimulation method or dopant with no predefined field. Our `exotic_parameters` escape hatch routes such values into key-value pairs (e.g., `{"key": "electron_beam_energy_MeV", "value": "1.0"}`) instead of dropping them. To keep this from becoming a dumping ground, the prompt explicitly forbids placing standard variables (temperature, voltage, current, pressure) here; a stated "temperature range 9 to 270 K" must be formatted as a `Trajectory`, not an exotic key-value pair.

6 Case Study: Extreme-Condition Thermodynamic Literature

We applied the pipeline to Low Energy Nuclear Reactions (LENR) literature (Fleischmann and Pons, 1989; Storms, 2010), which spans three decades. The field has high reporting variance, non-standard units, and strong positivity bias, which makes it a hard test of the "data bottleneck" problem. We benchmarked throughput and constraint enforcement on a randomly selected, unannotated subset of 60 historical papers.

Table 1: Empirical Evaluation Metrics (60-Document Corpus)

Category	Metric	Value
Semantic Triage	Documents Processed	60
	Theoretical Papers Filtered	35 (58.3%)
	Bounding Boxes Processed	332
	Non-Sci. Images Pruned	79 (23.8%)
	Tokens Saved (est.)	**314,416**
Validation	Payloads Rejected (Pydantic)	1
Throughput	Total Latency	48.5 min
	Latency / Document	48.5 sec

6.1 Throughput and Token Optimization

Triage flagged 35 of the 60 papers (58.3%) as theoretical and removed them before generative extraction. The vision filter removed 79 of 332 images (23.8%) as unsuitable for scientific analysis (Sec-

tion 3.2). We estimate that the two filters together avoided about 314,416 tokens. This is a heuristic estimate, not billed usage: it counts 258 tokens per rejected image plus, for each skipped paper, the smaller of one quarter of its character count and 32,000, and about 94% of the total comes from skipped papers. Processing took 48.5 seconds per document on average (2,910.7 seconds in total). Triage, image filtering, and pilot extraction all used `gemini-3.5-flash`.

6.2 Anti-Hallucination and Constraint Enforcement

In the pilot, the validation layer rejected one extraction payload. Source grounding is requested through the prompt. Our evaluator checks it by substring match for two `Measurement` fields (peak coefficient of performance and total excess energy), but neither was populated in this sample, so the check is inconclusive (0 of 0). Extending it to all `Measurement` fields and making grounding a validation rule is the next step (see the Limitations section).

The enum-constrained schema also mapped descriptions like "heavy water bath" and "deuterium oxide medium" onto a single `D2O` value.

6.3 Temporal Reconstruction and Exotic Variables

Trajectory serialization matters for LENR because the field relies on temporal thermodynamics: the model reconstructed temporal arrays by interpolating calorimetry charts and by deducing offsets from stated rates in text. The Negative Bias Trap is designed to flag `is_failed_replication` on null runs so the resulting graph captures negative boundaries as well as successes. The `exotic_parameters` field was used to store non-standard conditions such as external magnetic fields, laser stimulation, and unusual lattice dopants as key-value pairs, keeping them out of the strictly typed continuous variables.

7 Generalization to Broader Scientific Domains

The pipeline separates spatial parsing (IDP) and programmatic validation (Pydantic) from the generative model, so adapting it to another field would in principle require new enumerations and new dimensional limits rather than model retraining. This is a design argument: we have not yet run the pipeline outside LENR, and the examples below are possible applications rather than results.

- **Energy Storage and Battery Materials:** Trajectory serialization could extract capacity and Coulombic-efficiency timelines from cycling plots, with the validation layer enforcing voltage bounds.
- **Catalysis and Chemical Synthesis:** Kinetic parameters such as turnover frequency, reaction temperature, and yield often appear in tables, which spatial preservation is meant to keep intact (Swain and Cole, 2016); novel dopants could go into the exotic-parameter field.
- **Preclinical Biomedical Research:** Dosage schedules and pharmacokinetic profiles could in principle be extracted the same way, with validation appropriate to clinical use.

Limitations

We note the following limitations.

Source grounding is prompt-level, not validator-enforced. The schema requests a verbatim `source_quote` alongside every numeric claim (Section 4.1), but the validation layer does not yet run an automatic substring check confirming the quote is a literal excerpt; only physical bounds and types are independently enforced (Section 5). Our evaluator currently audits grounding for two `Measurement` fields (peak coefficient of performance, total excess energy); on our 60-paper pilot both were unpopulated, so the check returned 0 of 0 rather than a meaningful pass rate. Extending programmatic grounding verification to all `Measurement` fields, and re-running the audit where those fields are populated, is the most important follow-up implied by this paper.

Confidence scoring is document-level, not parameter-level. One `extraction_confidence_score` per paper is enough to down-weight an entire document, but cannot distinguish a confidently extracted value from an uncertain one within it. Per-parameter confidence would need a schema change and additional prompt engineering, left to future work.

Evaluation is a small, single-domain pilot. The quantitative results come from a pilot of 60 randomly selected LENR papers (Section 6), and

the pipeline has been run only on LENR (Section 7). We did not compare extracted values against a human-verified reference set, so we report no precision or recall; the metrics measure throughput and constraint behavior rather than extraction correctness. The validator bounds are coarse, and the token savings are a heuristic estimate.

Single generative backbone. All pilot stages used one model family (Gemini, `gemini-3.5-flash`); we have not tested whether the prompt-level safeguards hold under a different backbone LLM.

8 Conclusion

We presented an architecture for the "data bottleneck" in the empirical sciences: legacy PDFs digitized via decoupled spatial parsing, schema-guided extraction, and independent programmatic validation. The two-layer triage and vision-filtering system keeps the expensive generative model off irrelevant material, and the typed validation layer independently rejects values that violate simple physical bounds (negative temperature, pressure, or time, and an out-of-range loading ratio). Source-quote grounding is currently requested through the prompt, and verifying it programmatically is the natural next step.

Applied to extreme-condition LENR thermodynamics, this pipeline turns non-standardized legacy literature into structured data at about 49 seconds per document on average, including papers removed by triage. We believe that separating what is programmatically checked from what is merely requested of the model is useful more broadly for reliable LLM-based extraction from the scientific literature.

Acknowledgments

The research presented in this paper is part of a broader research effort led by Professor David J. Nagel of George Washington University and Professor Anasse Bari of New York University, under whose joint supervision this work was carried out. The overarching project aims to advance Artificial Intelligence (AI) and Predictive Analytics tools to support the understanding of Low Energy Nuclear Reactions (LENR) and, more broadly, scientific discovery. The program builds an end-to-end pipeline that turns unstructured scientific literature into structured, machine-readable data, organizes that data into knowledge graphs, and trains predictive models on it, with the longer-term goal of supporting autonomous scientific discovery. This paper presents the data-extraction stage of that work. Gaurav Kumar, lead researcher at the Predictive Analytics and AI Research Lab within the Courant Institute of Mathematical Sciences at New York University, played a key role in the design and development of the pipeline described here. This study was also supported in part by the Suzanne McIntosh Master's Research Fellowship awarded to him. We also acknowledge the entire research team at the Predictive Analytics and AI Research Lab at NYU for their continued collaboration and assistance, and we thank the anonymous reviewers for their constructive feedback.

References

Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Nikolaos Livathinos, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Fabian Lindlbauer, Kasper Dinkla, Lokesh Mishra, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. 2024. Docling technical report. *Preprint*, arXiv:2408.09869.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2024. Nougat: Neural optical understanding for academic documents. In *The Twelfth International Conference on Learning Representations*.

Samuel Colvin et al. 2024. Pydantic: Data validation using python type hints. `https://docs.pydantic.dev`.

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.

Martin Fleischmann and Stanley Pons. 1989. Electrochemically induced nuclear fusion of deuterium. *Journal of Electroanalytical Chemistry and Interfacial Electrochemistry*, 261(2):301–308.

Gemini Team, Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tanishq Gupta, Mohd Zaki, N M Anoop Krishnan, and Mausam. 2022. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8:102.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen,

Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for document AI with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020. DocBank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 949–960.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei 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.

Edmund Storms. 2010. Status of cold fusion (2010). *Naturwissenschaften*, 97(10):861–881.

Matthew C Swain and Jacqueline M Cole. 2016. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904.

Bhanu Prakash Vangala, Sajid Mahmud, Pawan Neupane, Joel Selvaraj, and Jianlin Cheng. 2025. HalluMat: Detecting hallucinations in LLM-generated materials science content through multi-stage verification. *arXiv preprint arXiv:2512.22396*.

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*, pages 7534–7550.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*.

Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. 2024. DocGenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. PubLayNet: Largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE.

GRAIL: Compact Multimodal Similarities for Astronomy Citation Linking

Egor Fedosov
Ural Federal University

Abstract

We present GRAIL, a system for classifying relations between astronomy figures and captions in the AstroCLIMB shared task. GRAIL combines frozen CLIP and SciNCL representations with optical character recognition (OCR), lexical overlap, and figure layout. A factorized gradient-boosted classifier predicts four relation classes from 190 symmetric pair features. The fixed system achieves 0.48 macro-F1 under cross-validation that separates exact objects across folds and 0.48 on Kaggle. With the same boosted classifier, the full features improve local macro-F1 by 0.068 over raw CLIP/SciNCL cosines. Nested classifier selection gives no observed aggregate gain. Removing OCR or corpus-weighted lexical features lowers local macro-F1 by 0.0081 and 0.013. Direct citation relations remain the hardest class.

1 Introduction

Scientific figures and captions contain evidence about what a paper studies: object names, measured quantities, instruments, and methods. Recovering relations between papers from this evidence requires more than recognizing a shared topic. Two papers may describe similar observations without citing one another. AstroCLIMB (Grezes et al., 2026) tests this distinction by asking systems to classify pairs of figures and captions according to their source relationship.

We ask how far a compact classifier can go when it combines several views of the same pair. CLIP (Radford et al., 2021) provides a shared space for images and text, while SciNCL (Ostendorff et al., 2022) provides representations learned from scientific documents and citation neighborhoods. We supplement these representations with explicit agreement between words, numbers, and identifiers. OCR makes some of this textual evidence available inside figures. The resulting system, GRAIL, learns from 190 pair features without fine-tuning either encoder.

Our experiments test whether explicit pair features add to raw cosine similarities, whether both CLIP backbones contribute, and whether additional model selection helps. Controlled feature ablations support OCR, corpus-weighted lexical evidence, and both CLIP blocks in the fixed classifier. Nested cross-validation evaluates selecting a classifier per modality and adjusting its class decisions. The submitted variants receive similar Kaggle scores, with the fixed model slightly ahead, so the added selection procedure gives no observed aggregate gain.

2 Task and Data

AstroCLIMB provides 10,000 training pairs and 10,000 scoring pairs. Each object is a figure or a caption. The four mutually exclusive labels identify a matching figure and caption (`same_figure`), different objects from the same paper (`same_paper`), objects from papers linked by a citation in either direction (`related_papers`), or unrelated papers (`unrelated_papers`). The evaluation metric is macro-F1, the unweighted mean of the four class F1 scores.

The training set contains 4,000 figure-caption, 3,000 figure-figure, and 3,000 caption-caption pairs. Each feasible class contributes 1,000 examples within its modality. The same-figure label is possible only for figure-caption pairs. Relations are symmetric, so predictions should not depend on object order.

We process the CSV files incrementally and identify objects by hashes of their original serialized strings. Across both splits, the cache contains 14,598 images and 14,581 captions. Images are resized to fit within 384 by 384 pixels without changing their aspect ratio and saved as JPEG at quality 88; transparent regions are composited

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 99–103
November 10, 2026 ©2026 Association for Computational Linguistics

onto white. Pair IDs and row order are excluded from the predictors. We use no scoring labels or external lookups of paper titles, authors, DOIs, or citation links. Frozen encoders process objects independently. Learned corpus statistics use only objects in each fitting partition; the scoring split is never used to fit them.

3 GRAIL

GRAIL stands for Group-safe Relational Alignment for Illustrations and Language. Here, grouping refers to exact objects: it does not guarantee separation by source paper. The pipeline extracts object representations, computes symmetric pair features, and classifies each pair with a model for its modality.

3.1 Object representations

We use frozen CLIP ViT-B/16 and ViT-B/32 encoders. Each produces normalized 512-dimensional image and caption vectors. Their different patch sizes provide two visual representations; controlled ablations quantify each block's contribution in the final classifier (Section 5). A white square canvas preserves the full figure before CLIP preprocessing. CLIP captions use at most 77 tokens. SciNCL supplies 768-dimensional caption vectors from sequences of up to 512 tokens.

We also compute 33 image descriptors that summarize dimensions, aspect ratio, color and grayscale statistics, histograms, pixel fractions, and edges. Apple's Vision recognizer extracts English text in fast mode from the cached JPEGs, whose longest side is at most 384 pixels. Reusing this compact cache limited local storage and preprocessing cost; it was an engineering choice, not an OCR accuracy optimum. The recognizer returns nonempty text for 8,859 of the 14,598 images. This is coverage, not a transcription-accuracy measure. Full-resolution or tiled OCR was not evaluated. For lexical comparisons, a figure is represented by its OCR text and a caption by its text after HTML unescaping, Unicode NFKC normalization, and whitespace normalization. OCR uses no task labels.

3.2 Symmetric pair features

The 190-dimensional representation consists of three modality indicators, 34 lexical features, 14 TF-IDF features, five serialized-object length features, 68 representation statistics, and 66 layout coordinates. Every feature is invariant to swapping the two objects. The five length features use original CSV strings, including base64 strings for images.

Lexical features measure exact or normalized matches, substring containment, and overlap between words, informative terms, numbers, acronyms, and alphanumeric identifiers. Jaccard similarity measures agreement across both texts; containment measures how much of the shorter text appears in the longer one. This distinction is useful when a short OCR string matches part of a caption. Inverse document frequency (IDF) gives more weight to uncommon shared terms. We retain at most 4,000 characters from each textual view.

Two TF-IDF spaces capture word unigrams and bigrams, and character trigrams and four-grams within word boundaries. Each vocabulary contains at most 40,000 entries; term frequencies are scaled sublinearly. Seven descriptors per space summarize cosine similarity, weighted Jaccard similarity, containment, sparsity, and text availability. Vocabularies and IDF weights are fitted on unique objects from the fitting partition.

Each of the two CLIP spaces, the SciNCL space, and the layout descriptor space contributes 17 summary statistics. These include cosine similarity, distances, quantiles of absolute differences, product statistics, sign agreement, and centered or variance-scaled cosine similarity. Means and standard deviations are estimated from fitting objects only, with standard deviations floored at 0.01. SciNCL comparisons apply to caption pairs. Missing representations are flagged explicitly. The layout block also contains the coordinate-wise absolute difference and mean of the two 33-dimensional descriptors.

3.3 Relation classifiers

We train a separate classifier for each pair modality. The primary system uses a factorized classifier with three binary decisions: whether the objects are the same figure, whether they come from the same paper otherwise, and whether different papers are citation-related. Let the corresponding conditional probabilities be q_f, q_p, and q_r. The

four class probabilities are

$$p_f = q_f, \qquad p_p = (1 - q_f)q_p, \tag{1}$$

$$p_r = (1 - q_f)(1 - q_p)q_r, \tag{2}$$

$$p_u = (1 - q_f)(1 - q_p)(1 - q_r). \tag{3}$$

For figure-figure and caption-caption pairs, we set $q_f = 0$. Each binary model uses histogram gradient boosting with at most seven leaves per tree.

We compare this factorization with flat multiclass models using seven or fifteen leaves. All models use 180 boosting iterations, a learning rate of 0.055, and a minimum of 35 examples per leaf. L2 regularization is 10 for the seven-leaf models and 15 for the fifteen-leaf model. We use neither early stopping nor class reweighting. The fixed system chooses the class with the largest predicted probability and applies no learned decision offsets.

4 Experimental Protocol

4.1 Grouping and model selection

Repeated objects can connect several pairs, so random row splits could expose a validation object during training. We instead construct a graph whose vertices are object hashes and whose edges are training pairs. Stratified group splitting assigns whole connected components to five outer folds. Each validation fold contains 2,000 pairs. Automated checks confirm zero exact-object overlap with the corresponding fitting partition.

All fixed classifiers use these same folds. To evaluate the additional selection procedure, we divide each outer training partition into three inner folds. Within every inner fit, we refit the text vocabularies, IDF weights, and representation statistics before training the candidate classifiers. Inner out-of-fold (OOF) macro-F1 over feasible classes selects one classifier per modality. We then fit four additive log-probability offsets to maximize overall macro-F1 on these inner predictions. Coordinate search considers offsets from -0.8 to 0.8 in steps of 0.05, for at most four rounds.

The chosen models and feature transformations are refitted on the outer training partition. Only then do they predict the held-out outer fold. Thus, outer labels do not determine either model selection or offsets. Pooled outer predictions give the nested development score. A 95% interval is estimated from 1,000 bootstrap resamples of exact-object components, keeping the predictions fixed. The outer-fold and estimator seed is 20260919; the inner-fold seed is 20260920.

4.2 Submissions and ablations

Both Kaggle submissions average class probabilities from five outer models. The fixed submission uses the factorized seven-leaf model throughout, without selection or offsets. The nested submission uses the models and offsets chosen within each outer training partition. It also averages probabilities across exact duplicate unordered scoring pairs before taking the highest-probability class; the fixed submission does not apply this step.

Feature ablations retain the fixed factorized model and the same outer folds. We remove either OCR-derived text or the TF-IDF and IDF-weighted lexical features, then refit the remaining pipeline. These comparisons assess feature sets without model selection or decision offsets. Removing OCR also changes the corpus used to estimate IDF, so this ablation measures the effect of the OCR pipeline as a whole.

Camera-ready comparisons use raw cosine similarities from L2-normalized CLIP and SciNCL vectors. SciNCL cosine applies only to caption-caption pairs and is zero otherwise; classifiers remain separate by modality. Logistic references use the same factorized decisions, fitting a standard scaler and L2-regularized logistic regression ($C = 1$, `liblinear`, at most 2,000 iterations) within each training fold. A matched-head reference replaces all 190 features with at most three cosines while retaining the seven-leaf boosted classifier and its settings.

These comparisons regenerate frozen representations for the 10,000 validated training pairs and reuse the saved outer folds and label-free OCR. CLIP ablations remove only the corresponding 17 summary features, keeping all other blocks and fitting procedures unchanged. The replayed full model is reported alongside these comparisons; its OOF score differs from the saved fixed run by 0.00022. Table 1 retains the original submission results. No comparison uses scoring data, model selection, or decision offsets.

5 Results and Discussion

We report scores and absolute differences to two significant digits. Differences are computed before rounding, so displayed scores may coincide despite a nonzero difference.

Additional selection gives no observed gain. The fixed model has a small numerical advantage: 0.0033 macro-F1 locally and 0.00071 on Kaggle,

System	Local OOF	Kaggle
Fixed factorized	0.48	0.48
Nested selection	0.48	0.48
Fixed minus nested	0.0033	0.00071

Table 1: Macro-F1 and absolute differences for the submitted systems. Kaggle values are the scores displayed on September 19, 2026, not final private-leaderboard results.

Classifier / features	Macro-F1	Drop
Flat, 7 leaves	0.47	0.013
Flat, 15 leaves	0.47	0.013
Factorized, 7 leaves	0.48	0
without OCR	0.47	0.0081
without TF-IDF / rarity	0.47	0.013

Table 2: Local macro-F1 on the same outer folds, without routing or decision offsets. Drop is the absolute decrease from the full factorized model, computed before rounding.

although the rounded scores in Table 1 coincide. These differences provide no evidence that the additional selection procedure improves aggregate performance. We do not claim a statistically established difference between the Kaggle scores. The nested local estimate has a conditional bootstrap interval of [0.47, 0.49]; that interval applies only to the nested predictions and must not be transferred to the fixed model.

Explicit textual evidence helps the fixed model. The factorized model outperforms both flat alternatives in local validation (Table 2). Removing OCR lowers its macro-F1 by 0.0081. Removing TF-IDF and rarity features lowers it by 0.013. These controlled comparisons support retaining both feature groups in this system. They do not establish that every individual feature is useful or that the same gains would hold for other classifiers. The OCR ablation compares cached low-resolution OCR with no OCR; it does not compare image resolutions or show that downsampling preserves all useful text.

Citation relations remain difficult. Related papers have the lowest class F1 in both systems (Table 3). The nested variant improves this class and same-figure detection, but lowers F1 for the other two classes. For the fixed model, feasible-class

Relation	Fixed	Nested
Same figure	0.51	0.52
Same paper	0.51	0.48
Related papers	0.37	0.39
Unrelated papers	0.53	0.52

Table 3: Per-class F1 from local OOF predictions. The small aggregate difference hides changes in which relations each system predicts well.

CLIP blocks	Features	F1	Drop
B/16 + B/32	190	0.48	0
B/32 only	173	0.47	0.0096
B/16 only	173	0.47	0.0076

Table 4: Controlled CLIP ablations of the replayed full model. The classifier, outer folds, and non-CLIP inputs are fixed. Drop is computed before rounding.

macro-F1 is 0.63 for caption-caption pairs, 0.47 for figure-figure pairs, and 0.35 for figure-caption pairs. These modality scores average over different feasible label sets, so they are neither directly comparable nor components of the overall four-class metric.

Both CLIP blocks contribute in the final classifier. Table 4 removes each CLIP block from the final boosted model. Excluding B/16 lowers macro-F1 by 0.0096; excluding B/32 lowers it by 0.0076. Each ablation retains 173 features, including SciNCL, OCR, lexical comparisons, and layout. The observed decreases support complementary predictive information from the two backbones in this development comparison, rather than complete redundancy. The gains are modest and are not a measured runtime or energy advantage.

Explicit pair features improve over raw cosine similarity. Table 5 reports simple cosine-only references. Combining B/16 and SciNCL with logistic heads gives 0.42 overall macro-F1; adding B/32 gives the same rounded score. SciNCL alone reaches 0.50 on caption-caption pairs, compared with 0.63 for the full features on that same three-class scope. Holding the boosted classifier fixed, replacing at most three cosines with the 190-feature representation improves overall macro-F1 from 0.41 to 0.48, an absolute gain of 0.068. This matched-head comparison isolates the combined feature design from a change of classifier within the stated development protocol.

Inputs / head	All	CC
B/16 cosine / LR	0.40	0.44
B/32 cosine / LR	0.39	0.43
SciNCL cosine / LR	-	0.50
B/16 + SciNCL / LR	0.42	0.52
Dual CLIP + SciNCL / LR	0.42	0.52
Dual CLIP + SciNCL / GB	0.41	0.50
190 features / GB	0.48	0.63

Table 5: Cosine-only references and the replayed full model. LR: logistic heads; GB: fixed seven-leaf boosted heads. All averages four labels over 10,000 pairs; CC averages three feasible labels over 3,000 caption-caption pairs. All inputs except the last row are raw cosines.

6 Conclusion

GRAIL combines frozen semantic representations with explicit textual and visual comparisons in a compact classifier for AstroCLIMB. The fixed factorized system reaches 0.48 macro-F1 on Kaggle. A matched-head cosine reference and controlled ablations support the richer pair features, both CLIP blocks, OCR, and corpus-weighted lexical evidence. Additional model selection yields no observed aggregate gain. The remaining difficulty is linking content similarity to a direct citation relation. Future evaluation should test whether these findings persist when source papers, rather than exact objects, are separated across folds.

References

Egor Fedosov. 2026. GRAIL for AstroCLIMB: Source code and experiment reports. GitHub repository. https://github.com/denzasikora-lab/astroclimb-grail.

Felix Grezes, Kelly Lockhart, Alberto Accomazzi, Atilla Alkan, Ethan Seefried, and Tirthankar Ghosal. 2026. 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.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 11670-11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, page 8748-8763. PMLR.

Parameter-Efficient Adaptation of Qwen3-VL for Astronomy Citation Linking at AstroCLIMB 2026

Syed Mohaiminul Hoque
Center for Computational and Data Sciences Lab
Independent University, Bangladesh

Abstract

We describe a submission to the AstroCLIMB shared task at WASP 2026. The task requires classifying a pair of astronomy figures and/or captions as the same figure, from the same paper, from citation-related papers, or from unrelated papers. Our system casts classification as constrained next-token prediction with Qwen3-VL and adapts the model using 4-bit quantization and low-rank adapters. On the 10,000-pair Kaggle test set, three Qwen3-VL-8B models with language-attention adapters obtain 0.73230–0.73307 macro-F1 individually. Averaging their four-class probabilities raises the score to 0.73604, and masking the impossible same-figure class for caption–caption and image–image pairs produces 0.73725. This paper documents empirical system engineering rather than proposing a new learning algorithm; broader visual adaptation, restricted-loss blending, and citation-focused auxiliary training do not improve the result, and most remaining errors involve the `related_papers` class.

1 Introduction

Scientific figures condense results that are difficult to recover with text-only indexing. The AstroCLIMB shared task asks systems to partially reconstruct an astronomy citation graph using only figures and captions (Grezes et al., 2026). Unlike conventional document classification, the input is a pair of heterogeneous objects: either object may be an image or a caption, and the target is a symmetric relation between their source papers.

Modern vision–language models (VLMs) offer a natural interface for such mixed inputs. However, full fine-tuning remains expensive, especially on the two NVIDIA T4 GPUs available in a standard Kaggle environment. We therefore formulate the task as instruction-following with four possible answer tokens and use QLoRA (Dettmers et al., 2023) to adapt Qwen3-VL (Bai et al., 2025a).

We develop a uniform multimodal pipeline with random object swapping; compare scale, loss normalization, adapter coverage, and seed variation; and analyze both a balanced validation split and a negative auxiliary-training result. Our three-seed probability ensemble with a modality constraint reaches 0.73725 test macro-F1, with `related_papers`, which is confused with both neighbouring classes, remaining the main difficulty. Code, notebooks, and artifacts are available at `https://github.com/SyedT1/Astroclimb-sharedtask`.

2 Related Work

Scientific-figure research commonly studies caption generation or cross-modal retrieval. SciCap introduced a large figure–caption corpus and captioning baselines (Hsu et al., 2021); SciMMIR benchmarks retrieval over 530K scientific image–text pairs (Wu et al., 2024). Chart retrieval also shows value in combining visual, textual, and derendered table evidence (Nowak et al., 2024). AstroCLIMB instead predicts provenance relations rather than retrieving a matching caption.

SPECTER2 learns citation-informed scientific-document representations (Cohan et al., 2020; Singh et al., 2023); SigLIP2 and DINOv3 provide cross-modal and self-supervised visual features (Tschannen et al., 2025; Siméoni et al., 2025). We compare their frozen features with our VLM approach. Casting labels as answer tokens is related to pattern-based classification (Schick and Schütze, 2021); our contribution is a carefully evaluated shared-task system combining this interface with QLoRA, ensembling, and a task-valid constraint.

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 104–111
November 10, 2026 ©2026 Association for Computational Linguistics

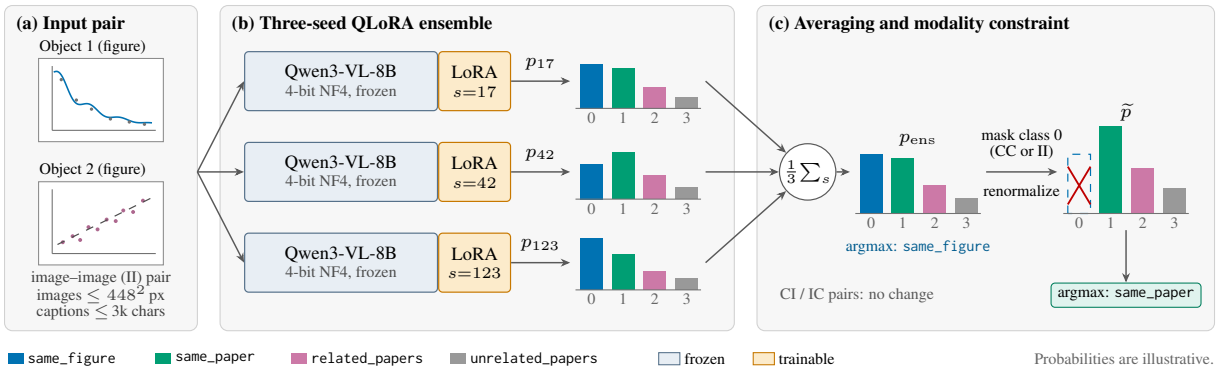


Figure 1: Inference pipeline of the best system, traced on an illustrative image–image pair. (b) Three full-vocabulary Qwen3-VL-8B models, each a frozen 4-bit backbone with a seed-specific language-attention LoRA adapter, produce four-class probabilities. (c) Equal averaging reduces seed variance; the deterministic modality mask then removes the impossible `same_figure` option from caption–caption and image–image pairs and renormalizes, here changing the prediction to `same_paper`.

3 Task and Data

Each example is a pair $x_i = (o_{i1}, o_{i2})$. An object is either an English figure caption or a base64-encoded scientific figure. The four mutually exclusive classes are: (0) `same_figure`, when the objects are the image and caption of one figure; (1) `same_paper`, when they are different objects from one paper; (2) `related_papers`, when one source paper cites the other; and (3) `unrelated_papers`. The relation is symmetric under exchanging the two objects.

The Kaggle release contains 10,000 labeled training pairs and 10,000 test pairs. Training labels are distributed $(1000, 3000, 3000, 3000)$ in class order. The official metric is macro-averaged F1,

$$F_{1,\text{macro}} = \frac{1}{4}\sum_{c=0}^{3} F_{1,c}, \tag{1}$$

so performance on the less frequent `same_figure` class contributes equally. The broader collection contains 94,233 figure–caption records with citation metadata (NASA Astrophysics Data System, 2026); we use it only for the auxiliary experiment in Section 5.4.

For controlled experiments, we reserve 200 examples per class using seed 42, producing a balanced 800-pair validation set and 9,200 training pairs. Full-data systems use all 10,000 labeled pairs. We do not use test metadata, external DOI lookup, or hidden labels.

4 System Description

4.1 Input representation

We use instruction-tuned Qwen3-VL-4B and Qwen3-VL-8B checkpoints. A system message defines the four relations and assigns them digits 0–3. The user message then presents Object 1 and Object 2 in their native modalities and asks for one digit. Captions are limited to 3,000 characters by retaining both their beginning and end. Images are decoded to RGB and processed within an area range of 256^2 to 448^2 pixels. This cap is important for fitting the 8B model on T4 hardware.

Because the relation is symmetric, each training pair is randomly swapped with probability 0.5 without changing its label. We do not use swap test-time augmentation in the submitted systems.

4.2 QLoRA adaptation

We load the backbone in 4-bit NormalFloat (NF4) format with double quantization and FP16 computation. Low-rank adaptation (Hu et al., 2022) modifies a frozen weight W_0 as

$$W_{\text{eff}} = Q_{\text{NF4}}(W_0) + \frac{\alpha}{r} BA, \tag{2}$$

where $r = 16$, $\alpha = 32$, and LoRA dropout is 0.05. Our main systems adapt language-attention q/k/v/o projections; we also test adding visual-merger projectors or adapting all eligible linear layers.

Training uses two-process distributed data parallelism, one process per T4. The per-device batch size is one with eight gradient-accumulation steps,

Backbone	Training setup	Epochs	Adapter coverage	Test macro-F1
Frozen encoders	SPECTER2/SigLIP2/DINO + features	—	modality CatBoost	0.48279
Qwen3-VL-4B	balanced 5K, $\mathcal{L}_{\text{vocab}}$	1.0	language attention	0.67856
Qwen3-VL-4B	full 10K, $\mathcal{L}_{\text{vocab}}$	1.0	language attention	0.70751
Qwen3-VL-4B	9.2K, $\mathcal{L}_4$	2.0	language attention	0.71453
Qwen3-VL-4B	full 10K, $\mathcal{L}_{\text{vocab}}$	1.0	language + visual projectors	0.71016
Qwen3-VL-4B	full 10K, $\mathcal{L}_{\text{vocab}}$	1.0	all linear layers	0.71202
Qwen3-VL-8B	full 10K, $\mathcal{L}_{\text{vocab}}$, seed 42	1.0	language attention	0.73230
Qwen3-VL-8B	full 10K, $\mathcal{L}_{\text{vocab}}$, seed 123	1.0	language attention	0.73250
Qwen3-VL-8B	full 10K, $\mathcal{L}_{\text{vocab}}$, seed 17	1.0	language attention	0.73307
Qwen3-VL-8B	full 10K, $\mathcal{L}_4$	1.5	language attention	0.73032
Qwen3-VL-8B	three-seed probability ensemble	1.0	language attention	0.73604
Qwen3-VL-8B	75% ensemble + 25% restricted	—	probability blend	0.73590
Qwen3-VL-8B	ensemble + modality mask	—	deterministic constraint	**0.73725**

Table 1: Final Kaggle leaderboard results. "Full 10K" denotes all labeled pairs; the 9.2K run retains 800 validation pairs. The final three rows are derived systems.

giving a global batch size of 16. The primary full-data systems train for one epoch with learning rate 10^{-4}. The 8B restricted-loss refit uses 5×10^{-5} for 1.5 epochs, selected on the fixed validation split.

4.3 Classification objectives

Let t_c be the tokenizer ID for digit c, and let $z_{i,v}$ be the logit immediately before the answer. The full-vocabulary objective supervises only that position:

$$\mathcal{L}_{\text{vocab}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp z_{i,t_{y_i}}}{\sum_{v\in\mathcal{V}}\exp z_{i,v}}. \quad (3)$$

We also test $\mathcal{L}_4$, replacing the denominator by the four valid digit tokens. Inference always selects the largest of those four logits, avoiding unconstrained generation. Appendix A gives the full restricted-loss, inference, and swap definitions.

4.4 Probability ensemble and modality constraint

Figure 1 summarizes the best system. We independently train the full-vocabulary 8B configuration with seeds 17, 42, and 123. If $p_s(c \mid x)$ is the normalized four-class distribution from seed s, their equal-weight ensemble is

$$p_{\text{ens}}(c \mid x) = \frac{1}{3}\sum_{s\in\{17,42,123\}} p_s(c \mid x). \quad (4)$$

This is probability averaging rather than majority voting over hard predictions.

The `same_figure` label is valid only for a cross-modal figure–caption pair. For caption–caption (CC) and image–image (II) pairs, we impose the task constraint after ensembling:

$$\widetilde{p}_0 = 0, \qquad \widetilde{p}_c = \frac{p_{\text{ens},c}}{\sum_{j=1}^{3} p_{\text{ens},j}} \quad (c = 1, 2, 3). \quad (5)$$

Cross-modal probabilities remain unchanged. The mask modifies vectors for 6,000 of 10,000 test rows but changes only seven hard predictions. We do not use swap test-time augmentation, class-bias calibration, or the restricted-loss model in the final system.

5 Experiments

5.1 Zero-shot baselines

With the same prompt and 4-bit loading, Qwen3-VL-4B is the strongest zero-shot model at 0.47956 macro-F1; Qwen2.5-VL (Bai et al., 2025b) scores 0.46745. The frozen-encoder baseline in Table 1 reaches 0.48279; Appendix B details both baselines.

5.2 Fine-tuned systems

Table 1 summarizes the final-leaderboard results. Full-data 4B training raises macro-F1 from 0.47956 to 0.70751; the analogous 8B seed-42 model reaches 0.73230. Replications score 0.73250 and 0.73307, motivating probability averaging.

The best 8B model exceeds the comparable 4B model by 0.02556. Across three seeds, the mean is 0.73262 ± 0.00040 sample standard deviation (range 0.00077). Broader 4B adapter coverage adds only 0.00265–0.00451; a visual-projector-only run scores 0.48182 because caption–caption pairs bypass its trainable path.

The restricted-loss comparison also changes the split and epochs. Its selected 8B checkpoint ob-

tains 0.729164 validation macro-F1; a full-data refit scores 0.73032. Blending 25% of its probabilities into the ensemble scores 0.73590, 0.00014 below the raw ensemble.

Probability averaging improves the best seed by 0.00297; the mask adds 0.00121, reaching 0.73725. Seven predictions move from the impossible `same_figure` label to `same_paper`, with no extra inference.

Test labels are unavailable, so these scores carry no exact confidence intervals. Rescaling the validation bootstrap of Appendix B to 10,000 rows suggests a 95% half-width of roughly 0.008. The 4B-to-8B gain is therefore resolvable, whereas seed differences and the ensemble gain should be read as indicative. The mask gain differs in kind: all seven changed predictions were necessarily wrong before masking.

5.3 Controlled validation analysis

On the balanced 800-pair split, the selected restricted-loss checkpoint obtains 0.72916 macro-F1. `same_figure` is easiest and `related_papers` weakest (Appendix B). The split is small: a stratified bootstrap over the recorded validation confusion matrix gives a 95% interval half-width of 0.029. Adding language MLP targets reaches 0.729019, a difference of 0.00015 that this split cannot resolve; we keep attention-only adaptation because it is cheaper, not because it is more accurate. The split is also pair-level, so it may overestimate performance on unseen source papers.

5.4 Citation-focused auxiliary training

We construct 20,000 auxiliary examples from the public AstroCLIMB collection (NASA Astrophysics Data System, 2026), distributed as $(2000, 6000, 8000, 4000)$ in class order. Class 0 replays training examples with replacement; classes 1–3 pair captions from one DOI, directly citation-linked DOIs, or non-adjacent papers sharing a meaningful title token, with metadata used only for selection. UUID-pair keys enforce pair-level uniqueness for classes 1–3, and exact normalized-caption matches exclude recoverable validation and test DOIs, although image-only papers cannot be mapped. This auxiliary stage lowers macro-F1 from 0.72916 to 0.71405, losing 0.035 and 0.042 F1 on `same_figure` and `same_paper` while gaining at most 0.013 on the different-paper classes.

Four properties of the auxiliary data can explain this pattern. First, *modality shift*: all 18,000 auxiliary pairs in classes 1–3 are caption–caption, whereas competition pairs mix captions and images. Second, *prior shift*: the auxiliary labels follow a 10/30/40/20% split rather than 10/30/30/30%, over-weighting `related_papers`. Third, *invisible labels*: a citation edge does not imply that two randomly drawn captions from the linked papers share visible evidence, and title-token negatives are topically similar by construction, so classes 2 and 3 are often separable only through metadata. Fourth, *replay*: class 0 reuses each of the 800 training `same_figure` rows about 2.5 times. We did not ablate these factors individually. A better sampler would retain only pairs with shared visible evidence, such as object names or numerical values; include image objects; follow the competition label prior; and interleave auxiliary and competition batches rather than training them in sequence. Appendix B gives detailed results.

6 Error Analysis

Where errors concentrate. Of the 234 validation errors in the recorded confusion matrix, 113 (48%) confuse `same_paper` with `related_papers` and 74 (32%) confuse `related_papers` with `unrelated_papers`; 34 involve `same_figure` and `same_paper`, and only 13 involve other pairs. `related_papers` is thus confused with both neighbouring classes. Both confusions are nearly symmetric (62 vs. 51 and 36 vs. 38), which points to genuine ambiguity rather than a class-prior bias that recalibration could remove. On the unlabeled test set, the three seeds of the final system give a label-free view (Table 6). They disagree on 13.3% of rows; 47.8% of these disagreements are between `same_paper` and `related_papers`, and 42.1% are between `related_papers` and `unrelated_papers`. `related_papers` predictions are also the least confident, with a median maximum probability of 0.566 against 0.739–0.956 for the other classes.

Why `same_paper` and `related_papers` collide. Both relations are defined by metadata the model never sees: a shared DOI or a citation edge. Two figures from one paper can cover different sub-analyses, while figures from citing papers often share instruments, targets, and notation. Topical similarity therefore indicates that a pair is related, but says little about which relation holds. The decisive cues are sparse: an identical dataset, a

repeated object identifier, or a matching plotting style. Example-level validation predictions were not retained, so we cannot show annotated examples.

Use of visual evidence. Exact figure–caption matching remains tractable because shared entities, values, and panel labels align directly. The raw ensemble nevertheless predicts `same_figure` for seven same-modality pairs, which the mask corrects. Whether the model reads content inside figures is less clear. Validation macro-F1 falls from 0.728 on caption–caption to 0.671 on caption–image and 0.651 on image–image pairs. On the test set, seeds agree less often on cross-modal pairs (85.6%) than on same-modality pairs (87.5%). Our frozen-feature baseline already uses OCR-token overlap and perceptual hashing, yet reaches only 0.48279, so shallow figure-text matching is far from sufficient. That language-side adaptation beats broader visual adaptation does not show that figures are unimportant: 10,000 labels may be too few to tune many visual parameters, while the pretrained encoder already supplies useful features. Image-blanking and OCR-substitution ablations are needed to measure how much predictions depend on figure content.

7 Efficiency and Reproducibility

The restricted 8B run trains 15,335,424 parameters for 938 steps in 320.45 minutes, peaking at 11.53 GiB on rank 0. Two-T4 inference over 10,000 pairs takes 78.15 minutes. The final system requires three independently trained and inferred adapters, but ensembling and masking are CPU-only. We use Transformers 4.57.1, PEFT 0.17.1, Accelerate 1.10.1, and bitsandbytes 0.47.0. The split seed is 42; full-data seeds are 17, 42, and 123. Appendix C reports the optimizer, scheduler, checkpoint policy, and output validation.

8 Conclusion

Our language-attention QLoRA system constrains Qwen3-VL prediction to four tokens. One 8B model reaches 0.73307 test macro-F1, three-seed averaging reaches 0.73604, and the modality mask yields 0.73725. Remaining errors centre on `related_papers`, which is confused with both `same_paper` and `unrelated_papers`. Broader visual adaptation, restricted-loss blending, and citation-graph supervision do not improve the result. Future work should evaluate on source-paper-grouped splits, ablate figure content directly, and use evidence-visible, modality-matched auxiliary examples.

Limitations

Most full-data configurations lack local validation, and differences are measured by single final-leaderboard scores. Test labels and example-level validation predictions are unavailable, so paired significance tests cannot be computed. Our confidence intervals come from the single recorded validation confusion matrix, and test intervals are approximate; the seed standard deviation measures run variability, not a confidence interval. The controlled split is pair-level and uses one seed. Kaggle pairs carry no paper identifiers, and DOIs can be recovered only for caption objects by exact caption matching, so we could not build a source-paper-grouped split. We did not run image-blanking or OCR-substitution ablations, so reliance on figure content is not quantified, and the proposed causes of the auxiliary-training drop are not individually ablated. Some comparisons vary multiple factors, including data size and epoch count. Caption truncation and image resizing can remove evidence. Finally, metadata-defined citation relations may be invisible in the presented objects, imposing a task-level ceiling that we do not estimate.

Ethics Statement

The system processes public scientific figures and captions supplied for the shared task. It is a benchmark model, not an authoritative bibliographic database: predicted links can be wrong and should not assign credit, evaluate researchers, or alter citation records without verification. We use only released data and public task metadata and infer no sensitive personal attributes. GPU computation is the main environmental cost; 4-bit quantization, LoRA, and constrained one-step inference reduce it relative to full-model training and free-form generation.

Acknowledgments

The author thanks the WASP 2026 organizers, NASA ADS, and AstroExplorer for the ASTRO-CLIMB benchmark.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025a. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025b. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36.

Felix Grezes, Kelly Lockhart, Alberto Accomazzi, Atilla Alkan, Ethan Seefried, and Tirthankar Ghosal. 2026. 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.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

NASA Astrophysics Data System. 2026. AstroCLIMB dataset. Hugging Face Datasets. Accessed 2026-09-18.

Averi Nowak, Francesco Piccinno, and Yasemin Altun. 2024. Multimodal chart retrieval: A comparison of text, table and image based approaches. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5488–5505.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–269.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. 2025. DINOv3. *arXiv preprint arXiv:2508.10104*.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. 2025. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, Noura Al Moubayed, Jie Fu, and Chenghua Lin. 2024. SciMMIR: Benchmarking scientific multi-modal information retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*.

A Objective and Prompt Details

Because the ground-truth relation is symmetric, each training pair is swapped with probability 0.5:

$$(\tilde{o}_{i1}, \tilde{o}_{i2}) = \begin{cases} (o_{i2}, o_{i1}) & b_i = 1, \\ (o_{i1}, o_{i2}) & b_i = 0, \end{cases} \qquad (6)$$
$$b_i \sim \text{Bernoulli}(0.5).$$

The label is unchanged. For the restricted objective, the denominator contains only the four valid digit tokens:

$$\mathcal{L}_4 = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp z_{i,t_{y_i}}}{\sum_{c=0}^{3} \exp z_{i,t_c}}. \qquad (7)$$

Regardless of the training objective, inference reads only those digit logits:

$$\hat{y}_i = \arg \max_{c \in \{0,1,2,3\}} z_{i,t_c}. \qquad (8)$$

The system instruction used in the main runs is:

> You classify the relationship between two objects from astronomy papers.
> 0: The objects are the figure and caption of the same scientific figure.
> 1: The objects are from different figures in the same paper.
> 2: The objects are from different papers and one paper cites the other.
> 3: The objects are from unrelated papers.
> The relationship is symmetric. Output only one digit: 0, 1, 2, or 3.

Each object is introduced as either a "scientific figure" or a "figure caption." The final user instruction is: "Classify their relationship. Reply with one digit only."

B Additional Experimental Results

Our non-generative baseline combines frozen SPECTER2 caption embeddings (Singh et al., 2023), SigLIP2 image–text embeddings (Tschannen et al., 2025), and DINO visual embeddings (Siméoni et al., 2025) with word/character TF–IDF, OCR overlap, and perceptual-hash similarities. Three modality-specific CatBoost classifiers handle caption–caption, mixed, and image–image pairs. It is therefore a frozen-feature classifier pipeline, not a SPECTER-only model. Its submission scores 0.48279 final-leaderboard macro-F1.

The auxiliary stage contains 20,000 pairs distributed as $(2000, 6000, 8000, 4000)$, followed by restricted-loss tuning on 9,200 competition pairs. Its confusion matrix (gold rows, predicted columns) is

Model	Test macro-F1
Qwen3-VL-4B-Instruct	**0.47956**
Qwen2.5-VL-7B-Instruct	0.46745
Qwen3-VL-8B-Instruct	0.40112

Table 2: Zero-shot results on the 10,000-pair test set.

Metric	F1
`same_figure`	0.92913
`same_paper`	0.66988
`related_papers`	0.53790
`unrelated_papers`	0.77975
Macro average	**0.72916**

Table 3: Fixed-split results for restricted-loss 8B QLoRA.

$$\begin{bmatrix} 165 & 30 & 5 & 0 \\ 4 & 131 & 62 & 3 \\ 0 & 51 & 113 & 36 \\ 0 & 5 & 38 & 157 \end{bmatrix}. \qquad (9)$$

The largest confusion is bidirectional between `same_paper` and `related_papers` (62 and 51 cases), followed by `related_papers` and `unrelated_papers` (36 and 38); together these account for 187 of the 234 errors. Auxiliary training improves the two different-paper classes but degrades `same_figure` and `same_paper`. Modality-level macro-F1 is 0.72845 on caption–caption, 0.67108 on caption–image, and 0.65102 on image–image pairs. These values diagnose the auxiliary model, not the full-data submission.

Validation uncertainty. Macro-F1 depends only on the confusion matrix. Resampling validation examples within each gold class (200 per class) is therefore equivalent to resampling each row of Eq. (9) multinomially. Over 20,000 replicates, the auxiliary model's macro-F1 has a 95% percentile interval of $[0.6845, 0.7431]$ (standard deviation 0.0149). The per-class intervals are $[0.860, 0.926]$, $[0.578, 0.678]$, $[0.487, 0.593]$, and $[0.751, 0.833]$. Confusion matrices of the other validation runs were not recorded, but models of similar accuracy on this split should have intervals of similar width. Because the interval is unpaired, it is conservative for comparisons between two models evaluated on the same examples. Scaling the standard deviation by $\sqrt{800/10{,}000}$ gives the approximate test half-width of 0.008 used in

Metric	Baseline	+Auxiliary
Macro-F1	**0.72916**	0.71405
`same_figure`	**0.92913**	0.89431
`same_paper`	**0.66988**	0.62830
`related_papers`	0.53790	**0.54067**
`unrelated_papers`	0.77975	**0.79293**

Table 4: Per-class F1 before and after auxiliary training.

Setting	Value
Backbone	Qwen3-VL-8B-Instruct
Component seeds	17, 42, 123
Training pairs	10,000 per seed
Epochs	1 per seed
Learning rate	10^{-4}
Optimizer	paged AdamW, 8-bit
Scheduler	cosine; 5% warm-up
Weight decay	0.01
Gradient clipping	maximum norm 1.0
Quantization	4-bit NF4, double quantization
Compute dtype	FP16
LoRA (r, α, p)	$(16, 32, 0.05)$
LoRA targets	language $q/k/v/o$ projections
Global batch size	16
Maximum image area	448×448
Maximum caption length	3,000 characters
Training augmentation	object swap, $p = 0.5$
Ensemble	equal mean of three probability vectors
Postprocessing	mask class 0 for CC and II pairs
Inference	four digit logits, ensemble, mask, argmax
Hardware	2× NVIDIA T4; CPU postprocessing
Checkpoint policy	final epoch; no validation or early stopping

Table 5: Configuration of the best leaderboard system.

Section 5.2. This ignores the different test class prior and is only a rough guide.

Auxiliary data properties. Classes 1–3 of the auxiliary set contain only caption–caption pairs, and the auxiliary epoch precedes, rather than mixes with, competition tuning. Class 0 draws 2,000 rows with replacement from the 800 `same_figure` rows of the 9,200-pair training split. Relative to the baseline, the auxiliary model gains 0.003 and 0.013 F1 on `related_papers` and `unrelated_papers`, but loses 0.035 and 0.042 on `same_figure` and `same_paper`. This is consistent with a shift toward caption-level topical cues.

Prediction	Share	Agree	Med. $p_{\max}$	<0.6
`same_figure`	9.9%	92.7%	0.956	6.8%
`same_paper`	24.9%	85.4%	0.739	26.4%
`related_papers`	35.0%	83.4%	0.566	65.5%
`unrelated_papers`	30.3%	89.7%	0.854	17.8%
All rows	100%	86.7%		

Classes involved	Seed disagreements	Low-margin rows
1 and 2	634 (47.8%)	1,335 (58.2%)
2 and 3	558 (42.1%)	874 (38.1%)
0 and 1	133 (10.0%)	86 (3.7%)
0, 1, and 2	1 (0.1%)	—
Total	1,326	2,295

Table 6: Label-free diagnostics on the 10,000-pair test set. Top: final-system predictions by class, with the share of rows where all three seeds agree, median maximum probability, and the share below 0.6. Bottom: classes among the three seeds' hard predictions when they disagree, and the top-two classes of rows whose final top-two margin is below 0.2.

C Primary Configuration and Reproducibility

The full-data seed models use the final one-epoch adapter without validation-based selection; intermediate recovery checkpoints are saved every 100 steps. The restricted validation experiment evaluates at quarter-run milestones and selects checkpoint 863 (1.5 epochs) by maximum validation macro-F1 before the locked full-data refit. The restricted 8B run trains 15,335,424 parameters for 938 optimizer steps in 320.45 minutes, with 11.53 GiB peak allocated memory on rank 0. Two-GPU inference over 10,000 pairs takes 78.15 minutes. Combining stored probabilities and applying the modality mask require no model inference. The implementation uses Transformers 4.57.1, PEFT 0.17.1, Accelerate 1.10.1, and bitsandbytes 0.47.0, and validates unique IDs, exact column order, finite normalized probabilities, binary targets, one prediction per row, and 10,000 output rows.

Modality-Conditioned Ensembles and Zero-Shot Cross-Encoder Reranking for Relation Classification Between Astronomy-Paper Objects

M. Tahsan Zahid and Most. Atkia Farzana and Tonmoy Sarker Sourav
Rajshahi University of Engineering and Technology (RUET), Rajshahi, Bangladesh

Abstract

The WASP 2026 (AstroCLIMB) shared task requires classifying the relation between a pair of objects drawn from astronomy papers — *same_figure*, *same_paper*, *related_papers*, or *unrelated_papers* — where each pair combines text (captions) and/or images (figures) in one of three disjoint modality configurations: image+text, text+text, and image+image. Since *same_figure* cannot occur outside the image+text subset, a single unified classifier is structurally sub-optimal; we instead adopt three independent per-modality classifiers whose predictions are pooled for the competition metric (pooled macro-F1). Starting from a hand-engineered baseline (CLIP, SciNCL and DINOv2 embeddings with lexical/OCR features and a LightGBM ensemble), we find that replacing the encoder with a joint vision-language embedding model (Qwen3-VL-Embedding-2B) produces the largest single improvement among the experiments we tested (public leaderboard 0.4808 → 0.62319), exceeding the combined effect of six feature-engineering ideas that were implemented and measured but produced less than 0.01 macro-F1 each. We further show that *zero-shot* cross-encoder rerankers (Qwen3-VL-Reranker-2B, Qwen3-Reranker-4B), applied with no task-specific fine-tuning, add a second large improvement on top of the engineered ensemble (leaderboard 0.64210 → 0.66299; a later configuration with high-resolution inputs and an out-of-fold-trained VL cross-encoder reaches 0.68759), and that an untrained text+text reranker alone already outperforms the fully engineered pipeline on that subset (macro-F1 0.7111 vs. 0.6704). We report a parallel distant-supervision branch that trains on 120,000 externally-labeled pairs mined from the competition's source corpus and reaches a comparable but lower leaderboard score (0.57713) through a different mechanism; under our tested data and modeling regimes, a better encoder helped more than either additional hand-engineered features or our particular distant-supervision recipe (a controlled data-scaling study is left to future work). Reported numbers are traced to raw out-of-fold prediction arrays or literal notebook output under a shared row-level `StratifiedKFold(5)` protocol (blend weights, class multipliers and the cascade are tuned on the same OOF predictions), and we document two negative results (a broken nested-honest validation split, and several feature-engineering ideas that measurably hurt) as concrete engineering findings.

1 Introduction

Astronomy papers make heavy use of figures — plots, spectra, images — that are frequently discussed, reused, and cross-referenced across papers. The WASP 2026 (AstroCLIMB) shared task (Grezes et al., 2026) poses this as a four-way relation classification problem over pairs of objects (a caption, a figure, or both): *same_figure* (the pair refers to the same physical figure), *same_paper* (different figures/captions from the same paper), *related_papers* (the source papers cite one another), and *unrelated_papers*. Object pairs are drawn from three disjoint modality configurations — image+text, text+text, image+image — and the competition metric pools predictions across all three subsets into a single macro-F1 score.

This paper documents an end-to-end study of that task: a staged, gate-driven data pipeline (§2), a progression of modeling choices from a hand-engineered baseline to a joint vision-language embedding model with zero-shot cross-encoder reranking (§3, §4), a parallel distant-supervision branch pursued independently on the same infrastructure (§5), and an error analysis that examines why, in our tested configurations, representation quality appeared to matter more than both feature engineering and additional supervision for this task (§6, §7). We deliberately report several negative results (§7.2) and one validation bug we discovered

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 112–121
November 10, 2026 ©2026 Association for Computational Linguistics

and fixed (§9.1), since we believe the negative evidence is as informative as the positive result for anyone attempting a similar task.

Our main claims are:

1. Swapping the visual/text encoder for a joint vision-language embedding model (Qwen3-VL-Embedding-2B) produced a larger single improvement (leaderboard $+0.1424$) than the sum of every hand-engineered feature we tried (six ideas, each < 0.01 macro-F1, one of them negative).

2. Zero-shot cross-encoder reranking, applied with *no* task-specific training, is a second independent lever of comparable importance: it moves the leaderboard score from 0.64210 to 0.66299 and, on the text+text subset alone, an untrained reranker outperforms the entire engineered pipeline (0.7111 vs. 0.6704 macro-F1).

3. A parallel branch that traded a better encoder for $12\times$ more (distantly supervised) training data reached a lower score (0.57713) through a different mechanism. Because that branch changes several factors at once (label source, sampling, pseudo-OCR augmentation, data size), we treat it as suggestive comparative evidence only, not an isolated test of supervision quantity versus representation quality.

4. Two of our reports are negative/engineering findings that we keep because they generalize: a nested-honest validation split silently broken by an unshuffled, label-sorted source CSV, and several feature-engineering ideas that measurably hurt.

All accuracy figures in this paper are traced to one of four sources: a number typed directly from a Kaggle leaderboard screen, a number independently recomputed from a raw out-of-fold prediction array shipped with a result archive, a number printed by an executed notebook cell, or a number recorded in a shared results log. Where a number could not be independently re-derived from a raw array in this pass, we mark it explicitly in the corresponding table caption.

2 Task and Data

2.1 Task definition

The task is 4-class relation classification between pairs of objects drawn from astronomy papers.

Subset	Train	Test	Classes
image+text	4000	4000	all 4
text+text	3000	3000	3
image+image	3000	3000	3

Table 1: Task subsets. The competition metric is *pooled* macro-F1 over all 10,000 test rows, *not* the mean of the three subset scores (§7).

Each pair falls into exactly one of three disjoint modality subsets, summarized in Table 1. *same_figure* is only possible in the image+text subset, since it requires exactly one image and one caption from the same physical figure; the other two subsets are 3-way.

2.2 A gate-driven, not architecture-first, pipeline

Rather than committing to one architecture up front, the preprocessing pipeline was organized into sequential stages (load $\rightarrow$ modality audit $\rightarrow$ imbalance profiling $\rightarrow$ decode/clean $\rightarrow$ encode $\rightarrow$ pair-features $\rightarrow$ split), each ending in a decision gate whose answer determined the next stage's design. Several architecture-defining facts were not knowable until the data was actually inspected; exploratory data analysis (EDA) overturned four a-priori plans: **(1)** SMOTE (Chawla et al., 2002) oversampling for rare classes was dropped, since the imbalance ratio is only $\sim$3 and class weighting was used instead (we report no SMOTE ablation); **(2)** object-grouped CV was not used, since train/test object overlap is only 1 object out of $\sim$29,000 (paper-level dependence between folds is not excluded, §8); **(3)** harvesting rare-class pairs from the HF source corpus was repurposed as an *external validation* set for generalization claims instead of training data (§5 uses it differently, for training); and **(4)** a single unified 4-way classifier was replaced by three independent per-modality classifiers, since *same_figure* is structurally impossible outside image+text.

2.3 Standing constraints

Three constraints, established during the modality/structural audit stage, hold for every experiment reported in this paper:

- `id` is never used as a feature: both `train.csv` and `test.csv` are sorted by label block, so `id` alone leaks the label.

- `solution.csv` (full test-set ground truth)

is never used for training, tuning, or thresholding — only for the offline analysis reported in this paper, and disclosed as such throughout.

- The floor baseline is **0.2896** macro-F1, the score of a modality-constraint-aware random predictor (one that respects *same_figure* being impossible outside image+text) — not 0.25, which ignores that constraint. Every model below is compared against 0.2896.

2.4 Caching architecture

Kaggle's GPU quota and session-crash constraints motivated a two-stage cache: a CPU stage decodes and resizes all images once to a shared 384px JPEG cache plus parquet tables of hashes, text, and metadata, persisted as a Kaggle Dataset; each subsequent embedding experiment reads from this cache rather than re-decoding the original base64-encoded CSVs (>10 GB), so swapping encoders costs only GPU embedding time. All embeddings are deduplicated by content hash (unique objects, not unique rows, since a caption or figure can recur across pairs) and L2-normalized on save so downstream cosine similarity reduces to a dot product.

3 Method

3.1 Modeling architecture

For each of the three modality subsets we train an independent ensemble of two LightGBM (Ke et al., 2017) models (one on scalar pairwise features, one on the full feature block) and a logistic-regression model, blended with tuned weights, followed by a per-class threshold correction (coordinate ascent) and, where beneficial, a binary related-vs-unrelated cascade model blended into the multiclass output. All three subsets share an identical `StratifiedKFold(5, shuffle=True, random_state=0)` protocol (Pedregosa et al., 2011), which keeps every stage's out-of-fold (OOF) predictions blendable (row-level folds; see §8 for the tuning caveats).

Implementation details. Per subset we build 45/31/42 scalar pair features (image+text/text+text/image+image) and, separately, embedding-pair blocks ($|e_1 - e_2|$ and $e_1 \odot e_2$ per embedding space; 5,632/7,168/7,168 dimensions), for 5,677/7,199/7,210 inputs in the full block. Scalar features comprise cosine and CSLS similarity ($2\cos - r_a - r_b$, where r is the mean top-10 cross-pool cosine), log-rank of the partner within the pool (hubness), word- and character-n-gram (3–5) TF-IDF cosine, rare-token (document frequency ≤ 10), entity, number and IDF-weighted overlap, 2-hop proxy similarities through each object's nearest cross-modal neighbors, text/OCR length, and reranker features (raw score, percentile rank, deviation from the object's mean score, and object degree). LightGBM on scalar features: 600 trees, learning rate 0.03, 15 leaves, column subsample 0.7, 3 bagging seeds; on the full block: 700 trees, learning rate 0.05, 31 leaves, column subsample 0.3, 2 bagging seeds; class-balanced weights. Logistic regression uses standardized inputs with $C = 0.05$. Blend weights are searched on a 0.1 grid, class multipliers by coordinate ascent, and the cascade (a binary LightGBM on related/unrelated rows) by a grid over blend weight $\times$ threshold.

3.2 Encoder progression

The pipeline's encoder stack evolved through the following stages, each evaluated with the same downstream LightGBM ensemble so that the encoder change is the only varying factor:

1. **Baseline**: CLIP ViT-L/14 (Radford et al., 2021) for images/cross-modal similarity, SciNCL (Ostendorff et al., 2022) for scientific-text embeddings, and DINOv2 (Oquab et al., 2023) for image-only similarity.

2. **+ lexical/OCR features**: TF-IDF, hubness-corrected similarity, and 2-hop graph features, plus OCR text extracted from figures.

3. **Encoder swap**: CLIP/SciNCL/DINOv2 replaced by a single joint image+text embedding model, Qwen3-VL-Embedding-2B (Qwen Team, 2026), used for all three modality subsets.

4. **+ zero-shot cross-encoder reranking**: scalar relevance scores from Qwen3-VL-Reranker-2B (image+text) and, in a later run, Qwen3-Reranker-4B (text+text) and Qwen3-VL-Reranker-2B (image+image) (Qwen Team, 2026; Zhang et al., 2025), none of which are fine-tuned on this task's training data (each pair is scored in both orders and averaged under a task instruction prompt; the image+image reranker receives each figure together with its OCR text), added as additional input features to the same LightGBM ensemble.

We additionally tested SigLIP-so400m (Zhai et al., 2023) as an added image/text encoder at several points in this progression; its effect was subset-dependent and is reported as a controlled ablation in §A, not as part of the main progression, since it was ultimately not carried into the best-performing pipeline.

4 Experimental Results

4.1 Headline progression

Table 2 traces the pooled macro-F1 and public leaderboard score across the encoder and reranking progression described above. Leaderboard (LB) scores are typed from the Kaggle screen. Pooled values marked [b]/[c] are the pipeline's *self-reported* scores (including the cascade re-decision); the footnotes give the separately *recomputed* value from the raw OOF arrays (`sklearn.metrics`, without the cascade re-decision). Other values come from notebook output or our results log without an independent re-check.

The largest jump among the stages we tested is the encoder swap (CLIP → Qwen3-VL-Embedding-2B): leaderboard 0.4808 → 0.62319. This exceeds the combined effect of every feature-engineering idea attempted before the swap (§7.2), suggesting that representation, not hand-built features, was the stronger bottleneck at that stage. The second largest jump is zero-shot cross-encoder reranking (§7), added with no task-specific training on top of the fully engineered pipeline.

Final submission configuration. Our best leaderboard score (0.68759) comes from a later run of the same ensemble with three additions: high-resolution (1280px) OCR text and a Qwen3-VL embedding of the OCR text, a second Qwen3-VL-Embedding image embedding computed from high-resolution images, and class-probability features from a Qwen3-VL cross-encoder trained with 2-fold out-of-fold predictions (in addition to the zero-shot rerankers). The run prints per-subset tuned macro-F1 of 0.6407/0.7310/0.5847 (image+text/text+text/image+image) and pooled OOF 0.6861, consistent with the leaderboard score; boundary AUCs (sp/rel, rel/unrel) are 0.749/0.783, 0.825/0.911 and 0.735/0.752. Its half-split check (class multipliers re-tuned on one half) gives 0.625/0.720/0.559. These numbers are taken from the run's printed output, not re-derived from raw arrays, and the individual contribution of each addition is not ablated.

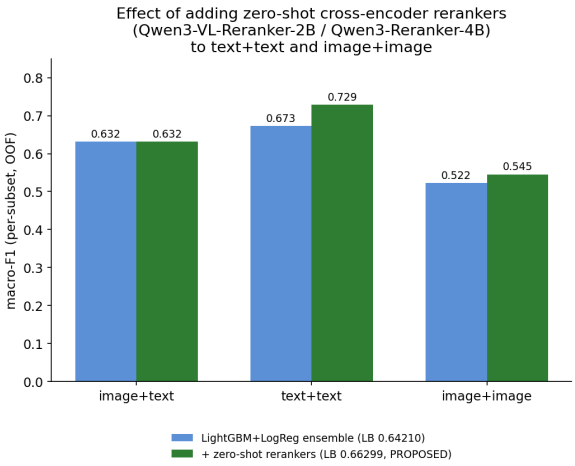


Figure 1: Per-subset macro-F1, with vs. without zero-shot cross-encoder reranking.

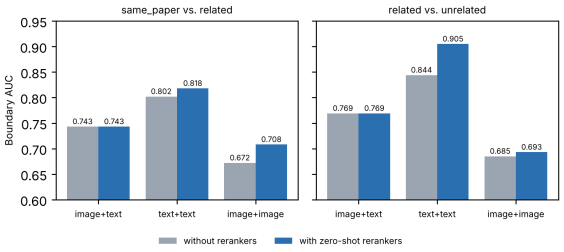


Figure 2: Boundary AUC per subset, with vs. without zero-shot cross-encoder reranking (values in Table 4).

4.2 Verified model trace

We trace exactly three result archives back to the notebooks that produced them (Table 3). The pooled and per-class numbers in Table 2 (footnotes b, c) and Table 5 are recomputed from the raw arrays inside those archives; all other numbers come from notebook output or our results log (§8).

Boundary AUC. Table 4 reports the two decision-boundary AUCs (*same_paper* vs. *related*, *related* vs. *unrelated*) from the proposed model's own report file, with and without the zero-shot reranker features.

Per-class results. Table 5 gives pooled per-class precision/recall/F1 across all 10,000 test rows for the proposed model, recomputed directly from `F_oof_{combo}.npy` arrays with `sklearn.metrics.classification_report`; Table 6 gives the per-class scores of the final configuration and Figure 3 the confusion matrix of the proposed model. Figure 1 shows per-subset F1 with and without reranking.

Stage	img+txt	txt+txt	img+img	Pooled	LB
Constraint-aware random floor	–	–	–	0.2896	–
Baseline: CLIP+SciNCL+DINOv2+LightGBM	0.3964	0.5529	0.4548	0.4806	0.4808
+ OCR/TF-IDF/hubness features	0.4031	0.6196	0.4705	–	–
+ per-class threshold	0.4186	0.6327	0.4792	0.5010[a]	–
Alt. branch: distant supervision (§5)	–	–	–	–	0.57713
Qwen3-VL-Embedding-2B + cascade	0.6031	0.6736	0.5242	0.6006[b]	0.62319
+ engineered feats. + reranker (img+txt)	0.6317	0.6728	0.5223	0.6511[c]	0.64210
+ zero-shot rerankers (txt+txt, img+img) — **proposed**	0.6317	**0.7286**	**0.5455**	0.6707[c]	**0.66299**
+ hi-res OCR/embeddings + OOF-trained VL cross-encoder	0.6407	0.7310	0.5847	0.6861[d]	**0.68759**

Table 2: Headline progression. "–" marks a value that does not apply (the random floor and the distant-supervision branch have no per-subset breakdown in our records) or was not recorded (the OCR/TF-IDF/hubness stage was not separately pooled or submitted to the leaderboard). [a]Weighted-average artifact from an early bookkeeping bug, kept for the historical record, not the pooled metric used elsewhere. [b]Self-reported pooled macro-F1; recomputed from raw OOF arrays (threshold applied, no cascade re-decision): 0.5965. [c]Self-reported pooled macro-F1; recomputed from raw OOF arrays: 0.6496 and 0.6683 respectively (see §4.2); these are ∼0.001–0.002 lower than the pipeline's self-reported number because the self-reported number additionally applies a cascade re-decision step whose own OOF was not saved to disk and so cannot be replayed here — the two numbers agreeing this closely is itself a sanity check. [d]Pooled OOF printed by the run (not re-derived from raw arrays).

Model	Result archive	What changed
Qwen3-VL-Embedding-2B + cascade	`report_A.json` (LB 0.62319)	Joint image/text embedding space, replacing CLIP+SciNCL+DINOv2; LightGBM×2 + logistic-regression blend; per-class threshold; related-vs-unrelated cascade.
+ engineered features + image+text reranker	`report_final.json` (run 2; LB 0.64210)	Adds the full engineered-feature block (CSLS-style local scaling, hubness, lexical/2-hop) on top of the above embeddings, plus the Qwen3-VL-Reranker-2B image+text score.
+ zero-shot rerankers on all subsets (**proposed**)	`report_final.json` (run 3; LB 0.66299)	Identical to the row above, plus zero-shot reranker scores added for text+text (Qwen3-Reranker-4B) and image+image (Qwen3-VL-Reranker-2B).

Table 3: Verified model trace. All three runs share the identical `StratifiedKFold(5, shuffle=True, random_state=0)` protocol per combination, so OOF arrays are directly comparable and blendable.

Subset	sp vs. rel	rel vs. unrel
image+text	0.743 → 0.743[*]	0.769 → 0.769
text+text	0.802 → **0.818**	0.844 → **0.905**
image+image	0.672 → **0.708**	0.685 → **0.693**

Table 4: Boundary AUC, without → with zero-shot reranker features. [*]image+text is unchanged since the zero-shot reranker addition in this stage targets text+text and image+image only.

Class	P	R	F1
same_figure	0.894	0.808	0.849
same_paper	0.678	0.597	0.635
related_papers	0.501	0.485	0.493
unrelated_papers	0.642	0.760	0.696
pooled macro-F1			**0.6683**

Table 5: Pooled per-class precision/recall/F1 computed from the proposed model's saved out-of-fold predictions (Qwen3-VL-Embedding-2B ensemble + zero-shot rerankers). The corresponding submission scored 0.66299 on the public leaderboard; these class-level values are OOF, not leaderboard test-set, metrics.

No result archive was generated for the Qwen3-VL-Embedding-2B-only stage that contains a raw OOF array (only summary JSON), so no confusion matrix or per-class chart is produced for that intermediate stage; Table 2 reports its summary numbers verbatim.

5 Alternative Branch: Distant Supervision from an External Corpus

In parallel with the encoder-progression branch above, a second branch pursued a different lever entirely: *more labeled supervision*, on the same feature-engineering base (CLIP/SciNCL/DINOv2 + OCR + lexical/2-hop features), rather than a better encoder. We report it as a separate branch rather than folding it into Table 2 because it does not build on the Qwen3-VL embeddings and reaches a comparable-but-lower leaderboard score through

Class	P	R	F1
same_figure	0.863	0.859	0.861
same_paper	0.708	0.597	0.648
related_papers	0.508	0.540	0.524
unrelated_papers	0.681	0.746	0.712
pooled macro-F1			**0.686**

Table 6: Pooled OOF per-class precision/recall/F1 of the 0.68759 configuration, as printed by the run (class multipliers applied, no cascade re-decision; not re-derived from raw arrays).

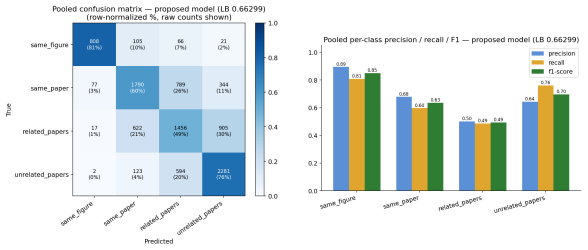


Figure 3: Pooled 4×4 confusion matrix (left) and pooled per-class precision/recall/F1 (right), both computed from out-of-fold predictions for the proposed model (Qwen3-VL-Embedding-2B ensemble + zero-shot rerankers; public leaderboard score 0.66299).

entirely different means — the comparison itself is informative.

Core idea. The competition's source corpus, `adsabs/AstroCLIMB` on HuggingFace, allows the same four relation labels to be reconstructed from metadata: *same_paper* from shared DOI, *related_papers* from an existing citation edge, and *unrelated_papers* from the absence of one. As downloaded and filtered, the corpus contains 94,059 figure rows over 9,998 papers with 26,995 citation edges; two DOIs touching a test caption were dropped before sampling.

Leakage audit. An audit printed by the run itself found 81.8% of Kaggle training captions present in the HF corpus, and 0.0% of Kaggle test captions present in it — the latter, safety-relevant figure is confirmed independently by the explicit DOI-drop step that runs regardless of the audit result.

Training data. 40,000 pairs were sampled per class (*same_paper* / *related_papers* / *unrelated_papers*), 120,000 pairs total ($12\times$ the size of the Kaggle training set). 16,000 of the *unrelated_papers* pairs are title-vocabulary hard negatives; 59,911 pairs (roughly half) have at least one side replaced with a pseudo-OCR noisy view, to better match the OCR-derived text used on the image side of the Kaggle task.

Metric	Value
Held-out macro-F1 (on HF pairs)	0.5929
Zero-shot transfer to Kaggle text+text	0.6242
Zero-shot transfer to Kaggle image+image	0.4158

Table 7: HF-corpus-trained model, transfer to the Kaggle task.

The 0.21 macro-F1 gap between clean-caption transfer (text+text) and OCR-represented transfer (image+image, Table 7) is exactly the motivation for the pseudo-OCR augmentation described above: a model trained only on clean captions transfers markedly worse to noisy OCR text.

Final result. The full per-subset blend (LightGBM on scalar and full features, logistic regression, and the HF-corpus-trained model's own prediction as a fourth blend arm for text+text/image+image) reaches an overall OOF macro-F1 of 0.5692 and a leaderboard score of **0.57713** — in close agreement (within 0.008), the same kind of local/leaderboard consistency seen in the main branch (§4.2).

Transductive post-processing (not scored). A test-time paper-clustering step (union-find over confident *same_paper*/ *same_figure* edges, pooling probabilities across test pairs connecting the same two clusters) found 7,922 paper clusters in the test set, of which 131 cross-cluster groups had more than one test pair between them (285 pairs total); 2,679 test pairs fell inside a single cluster. This was shipped as separate A/B submission variants rather than trusted into the base submission, and is not part of the reported 0.57713 score.

Comparative interpretation. $12\times$ more (distantly supervised) training data and several additional signal sources did not reach the level of a single better encoder: the encoder swap alone (§4.1) reached leaderboard 0.62319 using only the original 10,000 Kaggle pairs and one embedding model. This is consistent with our central claim (§7) that representation quality was the stronger bottleneck in our setting, but it is not a controlled comparison: the branch varies label source, sampling, augmentation and data size simultaneously, and its distant labels are noisy (a missing citation edge is treated as "unrelated", which can be wrong). We did not estimate that label noise; a learning-curve study at fixed representation (e.g., 10K/30K/60K/120K pairs) would be needed to isolate the effect of su-

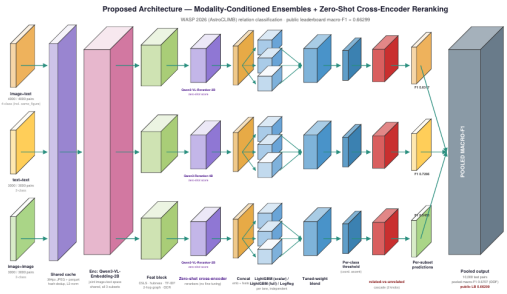


Figure 4: Proposed architecture: per-modality Qwen3-VL-Embedding-2B encoding, engineered features, zero-shot cross-encoder reranking, and a LightGBM/logistic-regression ensemble with per-class threshold and cascade post-processing, pooled into the leaderboard metric (LB 0.66299).

Boundary	**20%**	**40%**	**60%**
same_paper ↔ related	0.654	0.679	0.704
related ↔ unrelated	0.657	0.684	0.710

Table 8: Oracle analysis: pooled macro-F1 after fixing the given fraction of errors at one boundary.

sp/rel AUC	**rel/unrel AUC**	**Pooled F1**
0.730 (baseline)	0.766 (baseline)	0.6293
0.810	0.816	0.6531
0.869	0.862	0.6891
0.916	0.907	0.7369

Table 9: AUC-to-pooled-F1 mapping used to set feature/model priorities.

pervision quantity.

6 Error Analysis

At the Qwen3-VL-embedding stage (pooled OOF 0.6293), 78% of the 4041 total OOF errors fall in exactly two class boundaries: *same_paper*↔*related_papers* (1471 errors) and *related_papers*↔*unrelated_papers* (1690 errors); *same_figure* is substantially easier than the paper-level relation classes at this stage (F1 0.789), though not perfectly solved. Table 8 shows an oracle analysis: fixing a random fraction of errors at one boundary and re-measuring pooled macro-F1. These are diagnostic upper-bound sensitivity estimates, not scores attainable by any model we built.

Mapping this onto a target AUC (Table 9) turned the oracle analysis into an actionable target: any candidate feature or model was judged by whether it moved these two boundary AUCs, independent of how principled it looked on paper.

Per-subset AUC breakdown shows image+image as the consistently weakest link: text+text reaches 0.797/0.846 (sp/rel, rel/unrel), image+text 0.722/0.768, and image+image only 0.672/0.683 — the same subset that zero-shot reranking (§7) improves the least in absolute AUC terms (Table 4), suggesting that a stronger image-only encoder, rather than reranking, is the more promising future lever for that subset specifically.

7 Discussion

7.1 Untrained cross-encoders rival a hand-engineered pipeline

Three independent measurements motivated a fine-tuning phase for cross-encoder rerankers; apart from the out-of-fold-trained VL cross-encoder in the final configuration (§4.1, not independently verified), it was not completed to a verified result (§8): on image+text (*same_paper* vs. *related*), a single zero-shot reranker scalar reaches AUC 0.709, close to the full ∼5.7K-input engineered pipeline's 0.722, and combining both reaches 0.741; on text+text, Qwen3-Reranker-4B alone, with zero training, reaches macro-F1 0.7111, *exceeding* the full engineered pipeline's 0.6704; on image+image, the reranker alone (F1 0.509) roughly ties the pipeline (0.520). If an untrained cross-attention model matches or beats a heavily engineered bi-encoder-plus-tree-ensemble pipeline, a plausible reading is that much of the remaining signal is representational (requiring token-level cross-attention) rather than feature-space — this is our central interpretive hypothesis, not a proven conclusion, and the reason the encoder swap and zero-shot reranking additions were prioritized over further feature engineering.

Table 10 gives a partial ablation: the zero-shot reranker score alone (AUC on all training pairs; no fitting involved), the engineered embedding pipeline without rerankers (OOF AUC), and the

Subset	Boundary	Rerank.	Pipeline	Both
txt+txt	sp/rel	0.824	0.802	0.818
txt+txt	rel/unrel	0.892	0.844	0.905
img+img	sp/rel	0.710	0.672	0.708
img+img	rel/unrel	0.653	0.685	0.693

Table 10: Boundary AUC: zero-shot reranker alone (Qwen3-Reranker-4B for text+text; Qwen3-VL-Reranker-2B on figure+OCR for image+image), engineered pipeline without rerankers, and the final ensemble. Reranker AUCs are over all 3,000 training pairs per subset; pipeline and ensemble AUCs are OOF.

combined ensemble. The reranker alone beats the engineered pipeline on three of four text+text and image+image boundaries, and the ensemble improves on both single sources only at related-vs.-unrelated; at same_paper-vs.-related the reranker alone is as good as or slightly better than the ensemble (0.824 vs. 0.818 text+text; 0.710 vs. 0.708 image+image). This suggests the reranker largely supplants rather than complements the similarity features at that boundary. An embedding-only arm and a full factorial are not available.

7.2 Rejected feature-engineering ideas

Every idea in Table 11 was implemented, measured on real OOF predictions, and reverted; we report them as direct evidence against "just add more features" as a strategy for this task.

7.3 Model-side (structural) findings

Two structural changes, orthogonal to both the encoder and the feature set, measurably helped: a per-class multiplier/threshold via coordinate ascent ($+0.011$ honest, weighted; plain argmax is not macro-F1-optimal), and a binary related-vs-unrelated cascade blended into the multiclass output ($+0.006$ to $+0.011$ nested-honest) — the only successful structural change we found, and notably the simpler two-knob cascade beat a five-knob one-vs-rest variant, consistent with the overfitting pattern in Table 11's last free-parameter row.

8 Limitations

Validation and tuning. Blend weights, class multipliers and cascade parameters were tuned on the same row-level OOF predictions that are then scored, with one seed (`random_state=0`) and no confidence intervals, so the pooled OOF figures (0.6496/0.6683) are mildly optimistic. A half-split check that re-tunes only the class multipliers gives 0.626/0.663/0.509 (image+text/text+text/image+image) against 0.631/0.670/0.520 tuned, i.e., 0.004–0.011 lower (on a different half-sample); the final cascade OOF was not saved. Folds are not paper-grouped, so paper-level dependence between folds is not excluded. Our team made 17 leaderboard submissions across all branches, so leaderboard feedback may have influenced stage selection, and differences of $\sim$0.01 or less cannot be distinguished from noise.

Ablations. We lack a full factorial comparison (embedding only, reranker only, features only, and their combinations); Table 10 gives a partial one. The 0.68759 configuration (high-resolution OCR/embeddings and an out-of-fold-trained VL cross-encoder) is not ablated, and the cross-encoder's training is documented only in our released notebooks.

Verification and scope. Only the OOF-recomputed numbers marked in Table 2 are reproducible from saved arrays; the error analysis, rejected ideas and distant-supervision ablations come from logs or notebook output. Two fine-tuning notebooks (DeBERTa-v3-large; one early attempt) were not verified to completion and carry no claim. Distant-supervision label noise was not estimated. Feature definitions are in Section 3, and code and notebooks are in our public repository. All models are open-weight and run locally; no external LLM API was used.

Planned experiments. Our results would be strengthened by (i) a fully nested evaluation of all tuned parameters, (ii) paper- or citation-component-grouped cross-validation, (iii) the full reranker ablation, (iv) multiple seeds with confidence intervals, and (v) a controlled learning curve (e.g., 10K/30K/60K/120K pairs) at fixed representation for the distant-supervision branch. We did not run these before the camera-ready deadline.

9 Reproducibility and Data Statement

The following rules were applied to every result in this paper: `solution.csv` was never used for training, tuning, or threshold selection; `id` was never used as a feature (both CSVs are label-sorted, so `id` alone would leak the label); an identical `StratifiedKFold(5, shuffle=True, random_state=0)` (or its matching 2-fold variant) was applied across every

Idea	Measured effect	Why it failed
IPF/Sinkhorn class-count matching	-0.0087 pooled	Macro-F1's optimum is not class-balanced.
Graph/label-transitivity propagation	Would have hurt	Zero closed triangles in train graph; label correlation significant ($\chi^2 = 59.5$, $p < 10^{-4}$) but *anti-transitive* ($P(\textit{same_paper} \mid \textit{neighbor same_paper}) = 0.72$).
OCR text as a *same_paper* signal	AUC 0.51–0.53 (chance)	Mean OCR length only 54 chars.
SigLIP added on CLIP+SciNCL, text+text	$0.5529 \rightarrow 0.5420$ (-0.0109)	64-token limit truncates captions.
MLP blend instead of LightGBM	Always below LightGBM	Tree inductive bias fits this feature space.
One-vs-rest (5 knobs) vs. cascade (2 knobs)	$+0.0025$ vs. cascade's $+0.0063$ (honest)	More free parameters overfit the OOF used for tuning.
Zero-shot Qwen3-Reranker-4B on proxy captions (nearest retrieved caption) for image+image	AUC 0.575/0.701 (sp/rel, rel/unrel) vs. pipeline 0.672/0.685	Retrieved captions are noisy stand-ins; helps rel/unrel only and was not used in the proposed model.
Train/test object-overlap exploitation	1 shared object of $\sim$29,000	No leakage path available.

Table 11: Feature-engineering ideas that were implemented, measured, and reverted.

notebook contributing an OOF feature, so that OOF predictions can be blended on a shared protocol (the tuned scores are not unbiased estimates, §8); and, where nested-honest tuning was implemented and retained, threshold/blend weights were tuned on one random half and measured on the other. One exception: the cascade re-decision step's own OOF was not saved to disk, so its exact honest score cannot be replayed (Table 2).

9.1 Engineering findings

Two findings generalize beyond this competition. First, naively calling a vision-language model's full generation head (`Qwen3VLForConditionalGeneration.forward()`) for hidden states materializes a logits tensor sized for a 151,936-token vocabulary at every sequence position (212 MB for a 731-token sequence here) that is never used — the cause of every GPU OOM we observed; the fix is to call the inner transformer directly and read `last_hidden_state`. Second: an early nested-honest split naively dividing a CSV into "first half"/"second half" produced implausible macro-F1 of 0.1087/0.1478/0.1205 (image+text/text+text/image+image) against "final" scores of 0.60/0.67/0.52 on the identical run — a $\sim 5\times$ collapse revealing the source CSV was label-sorted, making the halves near-disjoint by class. The fix was to permute indices with a seeded RNG before splitting — a concrete illustration of why a nested-honest check needs its own sanity check.

10 Related Work

Our encoder progression draws on CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), DINOv2 (Oquab et al., 2023) and SciNCL (Ostendorff et al., 2022), combined via gradient-boosted trees (Ke et al., 2017). Our strongest results build on the Qwen3 embedding and reranker models (Qwen Team, 2026; Zhang et al., 2025; Yang et al., 2025), used zero-shot; the distant-supervision branch uses a vision-language model (Bai et al., 2025) for an (ultimately disabled) captioning path and similarity local scaling (Conneau et al., 2018). Cross-validation uses scikit-learn (Pedregosa et al., 2011).

11 Conclusion

We presented a multi-branch effort on the WASP 2026 (AstroCLIMB) relation-classification task, tracing every reported number to notebook output, a raw OOF array, or our results log. Under our tested regimes, representation quality was the stronger bottleneck: a single encoder swap outperformed six feature-engineering ideas combined, and zero-shot cross-encoder reranking rivaled or beat the fully engineered pipeline on two of three subsets. Our proposed submission reaches a public leaderboard score of 0.66299 (0.68759 with the later high-resolution and fine-tuned cross-encoder configuration). Task-specific cross-encoder fine-tuning (§8) is the most promising next step.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dai, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-VL technical report. ArXiv preprint.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Cross-domain similarity local scaling for bilingual lexicon induction. In *International Conference on Learning Representations (ICLR)*.

Felix Grezes, Kelly Lockhart, Alberto Accomazzi, Atilla Alkan, Ethan Seefried, and Tirthankar Ghosal. 2026. 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.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Qwen Team. 2026. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *Preprint*, arXiv:2601.04720.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Subset	Space(s)	Macro-F1
image+text	clip	0.3964
image+text	sig	0.4155
image+text	**clip+sig**	**0.4351** (+0.0387)
text+text	clip+sci	**0.5529**
text+text	sig	0.4327
text+text	clip+sci+sig	0.5420 (−0.0109)
image+image	clip+dino	0.4548

Table 12: SigLIP ablation. SigLIP helps the cross-modal subset but hurts the text-only subset when stacked on an already-good text space, consistent with a 64-token truncation explanation (text+text captions are the longest of the three subsets).

An Yang et al. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, et al. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

A SigLIP: a subset-dependent additional encoder

We separately measured SigLIP-so400m (Zhai et al., 2023) as an additional encoder, holding the LightGBM pipeline fixed and swapping one encoder combination at a time (Table 12).

Fusing Sentence-Level and Document-Level Representations for Keyphrase Extraction

Marko Pranjić[1,2] **Kristijan Armeni**[3] **Senja Pollak**[1]
[1]Jožef Stefan Institute, Ljubljana, Slovenia
[2]Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[3]Independent Researcher
marko.pranjic@ijs.si senja.pollak@ijs.si
armeni@validateresearch.com

Abstract

Supervised sequence-labeling approaches to keyphrase extraction operate at token level, but whether a phrase is relevant may depend on context beyond its sentence. Processing an entire document at token resolution can increase sequence length and introduce irrelevant information. We investigate a mixed-resolution context-fusion architecture that retains token representations for the target sentence while compressing available surrounding sentences into pooled representations. Both representation types are placed in document order and processed jointly by a self-attention module before IOB prediction. We evaluate the model on kp20k, inspec, and kptimes against SEKE, a recent state-of-the-art supervised system with approximately 60% more total parameters. The model achieves comparable F1 scores on the two scientific-domain benchmarks, while a larger performance gap remains on the journalistic benchmark. Error analysis shows that the model generates substantially more candidates than there are gold keyphrases. However, many predictions counted as false positives under exact-match evaluation partially overlap with gold keyphrases, while exact matches tend to receive higher scores than partial and non-matches.

1 Introduction

Keyphrases provide compact descriptions of a document's main topics and support indexing, categorization, and information retrieval (Turney, 2002). In scientific publishing, automatically extracting such phrases can help organize and retrieve information from rapidly growing literature collections. When these collections are large, manual annotation is impractical, motivating automated keyphrase extraction.

The "keywordiness" of a word or a phrase is largely determined by the global document context. In natural language processing, representations that depend on the word's contextual role in the text, known as contextualized word embeddings, are a hallmark of transformer language models pretrained on large text corpora (Devlin et al., 2019). This is underpinned by the model's ability to weigh the importance of all left and right context words when learning the representations for the current word, a process known as self-attention (Vaswani et al., 2017; Bahdanau et al., 2014). Contextual embeddings and self-attention have proven powerful architectural components in a wide array of downstream NLP tasks, including sequence labeling for keyphrase extraction (Sahrawat et al., 2020; Martinc et al., 2022, 2025).

Although in principle transformer large language models (LLMs) process the entire context when performing a task (e.g. a complete document when extracting keyphrases), *in practice* this risks burying the relevant signal among distracting context, especially as document length increases (Liu et al., 2023; Martínez-Cruz et al., 2024; Hong et al., 2025). Alternatively, extracting keyphrases sentence by sentence means transformers would discard document-level context altogether. Keyphrase-extraction approaches have addressed representation and context limitations through recurrent classification heads (Sahrawat et al., 2020; Martinc et al., 2022), task-specific training objectives (Kulkarni et al., 2022), document-attention and graph-based augmentation (Santosh et al., 2020; Martínez-Cruz et al., 2025), and mixture-of-experts (MoE) components (Martinc et al., 2025).

We investigate a particular context-augmentation construction for supervised keyphrase extraction. Instead of processing the entire document at token resolution, we retain token embeddings for the target sentence, compress each surrounding sentence into one pooled vector, and combine both representation types in a single ordered input to a three-layer self-attention module (Figure 1). This provides an explicit representation of the document

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 122–132
November 10, 2026 ©2026 Association for Computational Linguistics

context without sacrificing the token resolution of the target sentence.

Our contributions are threefold:

1. We investigate a mixed-resolution context-fusion architecture for keyphrase extraction that preserves token-level representations for the target sentence while compressing surrounding document context into sentence-level representations. A three-layer self-attention module jointly contextualizes target-token representations with sentence-level representations of the surrounding context before keyphrase prediction.

2. We evaluate the complete 170.9M-parameter architecture on three established keyphrase-extraction benchmarks under fine-tuning settings aligned with SEKE, a recent state-of-the-art supervised keyphrase-extraction system with 274M parameters. The architecture achieves comparable performance on the scientific-domain benchmarks while exhibiting substantially weaker performance on the journalistic benchmark, characterizing both the feasibility and current limits of the approach.

3. We characterize the model's principal extraction errors and show that low precision is associated with systematic over-generation. The candidate sets nevertheless contain substantial partial overlap with gold keyphrases, while exact matches tend to receive higher scores than partial and non-matches, suggesting that candidate selection and phrase delimitation are important targets for further improvement.

Code availability. Code and experiment configurations are available at `https://github.com/knowledge-technologies/context-fusion-kpe`.

2 Related work

Our work takes as a starting point the approaches leveraging transformer-based language models (Vaswani et al., 2017) for keyphrase extraction. In one of the earlier works, Sahrawat et al. (2020) combined pretrained transformer language models with recurrent neural networks and showed that using contextualized tokens improves keyphrase extraction performance across several benchmarks and models over static embeddings. Martinc et al. (2022) proposed a transformer-only architecture consisting of a next-word prediction pretraining followed by a keyphrase classification objective as a competitive alternative to unsupervised graph-based and recurrent neural network based keyphrase extractors at the time.

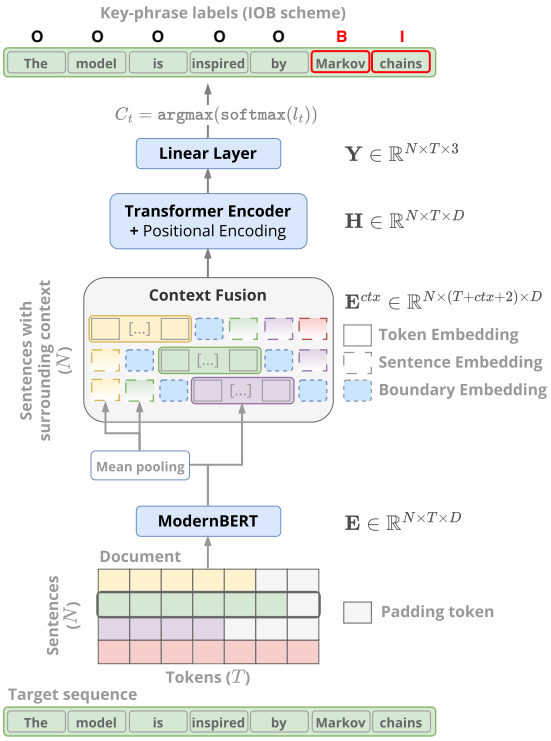


Figure 1: Model architecture overview. Token embeddings of the target sentence and pooled embeddings of the surrounding sentences form a single ordered input to the context-fusion module.

The use of broader document context for keyphrase extraction is well established. In unsupervised extraction, Liang et al. (2021) jointly modeled phrase–document relevance and local graph structure. More directly, DAKE (Santosh et al., 2020) enriched sentence-level BiLSTM word representations with supporting information from other sentences through document-level attention and gating before CRF decoding. Martínez-Cruz et al. (2025) instead augmented contextualized language-model token representations with embeddings of a document-wide co-occurrence graph. These studies establish the value of augmenting local representations with document information. Our architecture differs in the fusion operation: each surrounding sentence becomes one coarse-grained position alongside the fine-grained target-token positions in a single ordered sequence, enabling joint self-attention between both representation types rather

than explicit document attention and gating or concatenated graph features.

Most recently, Martinc et al. (2025) introduced SEKE, a state-of-the-art supervised keyphrase-extraction system that combines a DeBERTa backbone with a mixture-of-experts (MoE) (Jacobs et al., 1991) and BiLSTM token classification head. They froze the original backbone weights while training low-rank adaptations to its query and value projections (Hu et al., 2022), together with all parameters of the custom classification head. During inference, the full DeBERTa backbone remains active, while a learned router selects the top-k experts that process each token.

3 Methods

3.1 Benchmarks

We fine-tuned the model on the validation splits of three keyphrase benchmarks: kp20k dataset (Meng et al., 2017), kptimes (Gallina et al., 2019), and inspec (Hulth, 2003) (as in (Martinc et al., 2025); see Section 3.5). Summary statistics for validation datasets are presented in Table 1.

3.2 Preprocessing

We needed to prepare the benchmark data to conform to our annotation and extraction scheme. The document title, abstract, and full text (if present) were first segmented into sentences using the SpaCy[1] library with the en_core_web_sm model. Each sentence was tokenized using the Huggingface tokenizer for the pretrained model we used (answerdotai/ModernBERT).

Annotating target tokens. To annotate class labels to the tokenized dataset, we performed matching of the tokens in the document with the ground truth keyphrases provided in the 'keyword' (or 'keywords') field. We first stemmed both document and ground truth keyphrases using the nltk Porter stemmer. For each sentence, we then iterated over the stemmed ground truth keyphrases and checked for matches with stemmed document words. This stem-based matching allowed us to boost the relevant training signal by ignoring the variance due to word surface forms.

3.3 Task definition

Keyphrase extraction is a 3-way classification problem. The task involves classifying each token at position t in the target sentence j either as a start of a new keyphrase (B - beginning), continuation of a multi-token keyphrase (I - inside) or a non-keyphrase token (O - outside). This is known as the IOB labeling scheme. An example is shown in Table 3.

3.4 Model architecture

The overall architecture consists of two stages: a document encoder and a classification head. The document encoder provides the surrounding context for the target sentence for which the candidate keyphrases are extracted. The final output is a 3-way prediction for each token in the target sequence.

Token embedding layer (ModernBERT) The document encoder is the pretrained bidirectional encoder ModernBERT (Warner et al., 2024), applied independently to each sentence. Every sentence is tokenized with the matching tokenizer and truncated to at most 512 subtokens, and the last hidden state of the encoder provides the token-level embeddings. For a target sentence j, these are the embeddings $\mathbf{w}_{j,t} \in \mathbb{R}^d$ used in the fused sequence below; for context sentences, the same embeddings are mean-pooled into the sentence embeddings $\mathbf{s}_k$. The embedding dimension is $d = 768$, matching the hidden size of `ModernBERT-base`.

Context fusion with sentence embeddings For each target sentence j in a document, we construct a fused input sequence containing, in order: the *sentence* embeddings of the preceding document context, a learnable boundary embedding, the *token* embeddings of the target sentence, a second learnable boundary embedding, and the *sentence* embeddings of the following document context:

$$\mathbf{x}_j = \big[\mathbf{s}_{j-c_{\text{pre}}}, \ldots, \mathbf{s}_{j-1},\ \mathbf{m}_{\text{pre}},\ \mathbf{w}_{j,1}, \ldots, \mathbf{w}_{j,T_j},\ \mathbf{m}_{\text{post}},\ \mathbf{s}_{j+1}, \ldots, \mathbf{s}_{j+c_{\text{post}}}\big]$$

where: $\mathbf{s}_k \in \mathbb{R}^d$ is the mean-pooled sentence embedding of sentence k (before or after the current target sentence j); $c_{\text{pre}} = \min(j, n_{\text{left}})$ and $c_{\text{post}} = \min(N - j - 1, n_{\text{right}})$ are the context window sizes, bounded by document length N; $\mathbf{w}_{j,t} \in \mathbb{R}^d$ is the token embedding of token t in target sentence j; $\mathbf{m}_{\text{pre}}, \mathbf{m}_{\text{post}} \in \mathbb{R}^d$ are learnable boundary embeddings. During training, sentences are processed in document-ordered batches of 50;

[1] https://spacy.io/

Statistic	kp20k	kptimes	inspec
Domain	CS	news	CS
Nr. Documents	19,992	10,000	1,500
Mean abstract length (words)	146.5	776.3	117.1
Max abstract length (words)	1,547	5,617	490
Mean keyphrases/doc (all/in doc)	5.3 / 3.8	5.0 / 2.7	9.6 / 7.7
Vocabulary size (first 1k docs)	11,659	34,875	10,277
Words per keyphrase (all/in doc)	1.9 / 1.8	2.2 / 1.6	2.3 / 2.4

Table 1: Validation dataset statistics for the three benchmarks

for documents longer than one batch (primarily kptimes), the context is drawn from the sentences within the current batch. Fused sequence lengths are therefore bounded well below the 768 positional embeddings of the context-fusion head.

Context fusion and output projection. Learnable positional embeddings are added to each position: $\tilde{\mathbf{x}}_{j,i} = \mathbf{x}_{j,i} + \mathbf{p}_i, \quad 0 \le i < |\mathbf{x}_j|$. The resulting sequence is passed through a three-layer self-attention module with a 768-dimensional representation split across four attention heads:

$$\mathbf{h}_j = \text{ContextFusion}(\tilde{\mathbf{x}}_j)$$

Token-level logits for sentence j are extracted from each position t in the sequence through a linear layer: $\ell_{j,t} = \mathbf{W}_{\text{out}}\,\mathbf{h}_{j,t}$. All parameters are fine-tuned: the ModernBERT encoder has 149.0M parameters and the context-fusion head 21.9M, for 170.9M parameters in total.

Inference computation. A batch of sentences from the same document is encoded once by ModernBERT to obtain token-level representations. Average pooling yields one sentence-level representation per sentence, without another encoder pass. For each target sentence, we construct an input combining its fine-resolution token representations with the coarse-resolution sentence representations of the surrounding sentences. Each resulting input is processed only by the context-fusion head, which contains approximately 13% of the model's parameters (21.9M of 170.9M).

3.5 Fine-tuning

Following Martinc et al. (2025), we fine-tuned our model in two regimes: *validation set fine-tuning* and *training set fine-tuning* (kp20k only). The main difference between the two is the size of the dataset used in fine-tuning. In validation set fine-tuning, we fine-tuned the model on the validation set split for the respective benchmark. We randomly sampled 80% of the documents from the validation split[2]. We fine-tuned the model for up to 20 epochs and used early stopping (metric: F1@10, delta: 0.001, patience: 5 epochs). In training-set fine-tuning, we used the full training set split for fine-tuning. We fine-tuned for 20 epochs without early stopping. This corresponds to the regime dubbed "pretraining" in (Martinc et al., 2025, Appendix E, Table 5).

In both cases, we used AdamW with learning rates of 2×10^{-5} for the base encoder and 5×10^{-5} for the context-fusion head. The effective batch size was 100 sentences, implemented as two 50-sentence forward passes. We used cross-entropy loss with class weights inversely proportional to class frequency: $w_c = \frac{n_samples}{(n_classes * n_samples_c)}$. Special tokens (CLS, SEP, PAD) and continuation subtokens were masked with `IGNORE_IDX` and excluded from loss computation.

3.6 Keyphrase extraction

Token-level tag assignment. At inference time, we converted the `IOB` class logits for token t in sentence j into a normalized score vector

$$\boldsymbol{\pi}_{j,t} = \text{softmax}(\boldsymbol{\ell}_{j,t}).$$

The predicted tag and its score were then defined separately as

$$\hat{y}_{j,t} = \arg\max_c \pi_{j,t,c}, \qquad q_{j,t} = \pi_{j,t,\hat{y}_{j,t}}.$$

For words tokenized into multiple subtokens, only the first subtoken's predicted tag $\hat{y}_{j,t}$ and score $q_{j,t}$ are retained during keyphrase extraction; subsequent subtokens of the same word are ignored (Ta-

[2]This was to match the dataset size used in (Martinc et al., 2025), where 80% of each validation split was used for fine-tuning and the remaining 20% for hyperparameter optimisation.

Model/Benchmark	kp20k		kptimes		inspec	
F1	@5	@10	@5	@10	@5	@10
Validation Set Fine-tuning						
SEKE (Martinc et al., 2025)	32.7	33.0	44.9	45.1	47.2	56.5
Ours (ModernBERT)	31.3	30.3	40.8	33.0	48.7	54.3
Training Set Fine-tuning						
SEKE (Martinc et al., 2025)	36.3	36.5	54.8	55.0	49.4	58.8
Ours (ModernBERT)	37.5	35.9	—	—	—	—

(a)

Metric	kp20k		kptimes		inspec	
	@5	@10	@5	@10	@5	@10
Validation Set Fine-tuning						
Precision	26.4	22.6	31.7	21.8	57.1	47.5
Recall	38.3	45.9	57.6	68.2	42.5	63.4
Training Set Fine-tuning						
Precision	31.9	27.0	—	—	—	—
Recall	45.5	53.9	—	—	—	—

(b)

Table 2: Performance by benchmark and fine-tuning regime: (a) F1 for our model and SEKE; (b) precision and recall for our model. SEKE scores are from Martinc et al. (2025, Tables 2 and 5); validation-set values are means over five runs and training-set values are from a single run.

ble 4). This step resulted in one `I/O/B` tag and associated score for each word in the sentence.

Keyphrase assembly and scoring. To assemble the multi-word keyphrases, we iterated over all words in the sentence and started a new keyphrase if the predicted tag was `B` with $q \geq 0.5$, and appended a word to the current keyphrase if its predicted tag was `I` with $q \geq 0.5$. Current keyphrase construction was terminated if a `B` or `I` tag had $q < 0.5$, or if an `O` tag was encountered. Each keyphrase was then assigned a score by averaging the scores of its constituent words.

3.7 Evaluation

For evaluation we largely followed Martinc et al. (2025). To ensure robustness against surface-form variations, we replaced hyphens (−), slashes (/), and asterisks ($\star$) with spaces, and lowercased and Porter-stemmed both gold and candidate keyphrases prior to evaluation. Candidate keyphrases were additionally deduplicated by retaining only the highest-scoring occurrence among duplicates. For each document d, we computed precision $P_d@k$ and recall $R_d@k$ from the top-k ranked candidates. If fewer than k candidates were available, precision was computed over the available candidates. We then averaged precision and recall separately across the D evaluated documents:

$$\overline{P}@k = \frac{1}{D}\sum_{d=1}^{D} P_d@k, \qquad \overline{R}@k = \frac{1}{D}\sum_{d=1}^{D} R_d@k.$$

The reported F1 score was their harmonic mean:

$$F_1@k = \frac{2\,\overline{P}@k\,\overline{R}@k}{\overline{P}@k + \overline{R}@k}.$$

This aggregation follows the SEKE evaluation implementation, ensuring direct comparability with its reported results. Gold keyphrases were taken as the union of those stem-matched within the document text, and documents with no remaining gold keyphrases were excluded from metric computation.

4 Results

4.1 Benchmark performance

General benchmark performance. Overall, our model performs comparably with the state-of-the-art SEKE model (MoE BiLSTM architecture) proposed by Martinc et al. (2025). When fine-tuned on the smaller validation sets of the kp20k and inspec benchmarks, our model performed within the margin of +/- 1–3 percentage points (Table 2(a)) relative to SEKE. When fine-tuned on the larger

training set of the kp20k benchmark, our model showed an F1@5 score of 37.5 (Table 2(a)) which is competitive with the reported SEKE score of 36.3. By far the largest performance divergence between the two models is seen on the kptimes benchmark where SEKE outperformed our model by 4.1 (F1@5) and 12.1 (F1@10) percentage points.

Precision. Precision ranged from 21.8% to 57.1% across the evaluated settings (Table 2(b)), indicating that many top-ranked candidates did not match the annotated gold keyphrases. Recall generally exceeded precision across the evaluated settings, consistent with a tendency to over-generate keyphrase candidates. Performance was weakest on the journalistic kptimes benchmark, whose documents are longer, have a larger vocabulary, and cover a broader range of topics than the scientific abstracts in kp20k and inspec (Table 1).

4.2 Characteristics of extracted keyphrases

Eager keyphrase extraction. We examined overgeneration by comparing the numbers of predicted and gold keyphrases in 100 randomly sampled test documents per benchmark. Figure 4 in the Appendix reports the median counts. Across all three benchmarks, the median predicted count substantially exceeded the median gold count, with the largest relative difference on kptimes.

Error analysis. With the model generating a large number of keyphrases, are the generated candidate keyphrases relevant or do they mostly represent irrelevant noise? In Table 5 in the Appendix, we show a sample of gold and extracted keyphrases from six documents in the test set stratified by document-specific F1 scores into high, mid, and low accuracy tiers. Each predicted keyphrase is marked using the surface-match categories described below.

Score distribution. To examine whether keyphrase scores distinguish correct from incorrect predictions, we analysed 100 randomly sampled kp20k test documents and all 784 keyphrase candidates produced for them by the model fine-tuned on the full kp20k training set. For this descriptive analysis, we classified a candidate as an exact surface-form match, a partial match if it shared at least one word with a gold keyphrase, and a non-match otherwise. These categories were used only for the error analysis; benchmark evaluation used normalized and stemmed exact matching. As Figure 2 shows, exact matches generally received higher scores: 18.5% scored 0.9 or lower, compared with 47.2% of partial matches and 55.5% of non-matches.

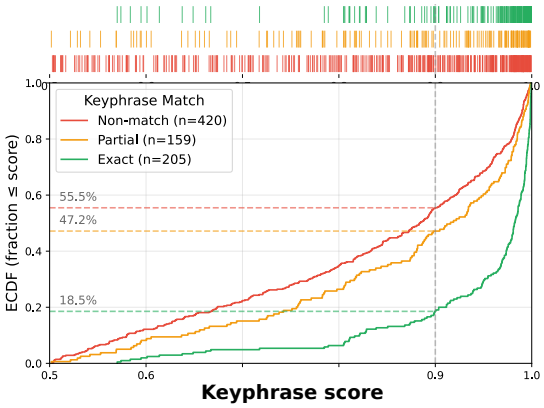


Figure 2: Keyphrase-score distributions by surface-match type. Curves show empirical cumulative distributions; dashed guides indicate the proportions with scores of 0.9 or lower, and rug marks represent individual predictions.

Partial matches Examples show that, apart from exact matches and non-matching candidates, the model frequently extracted one or more partial matches. For example, in a document with the gold keyphrase *conventional surfactant*, the model extracted partial matching candidates *surfactant* and *surfactant systems*. The general pattern of partial matches across all performance tiers is shown in Figure 3. Partial matches account for approximately 20% of candidates in the mid-performance tier and 25% in the low-performance tier. These candidates receive no credit under exact-match evaluation.

Acronym variants. Qualitative inspection also revealed candidates that were acronyms of longer gold keyphrases. Because the evaluation requires an exact match after normalization and stemming, these pairs receive no credit: the predicted acronym is counted as a false positive and the unretrieved gold keyphrase as a false negative. Our partial-match criterion also classifies such cases as non-matches because the acronym and its expanded form share no words. These cases illustrate that exact-match evaluation may understate the usefulness of some predictions, although we did not quantify their prevalence.

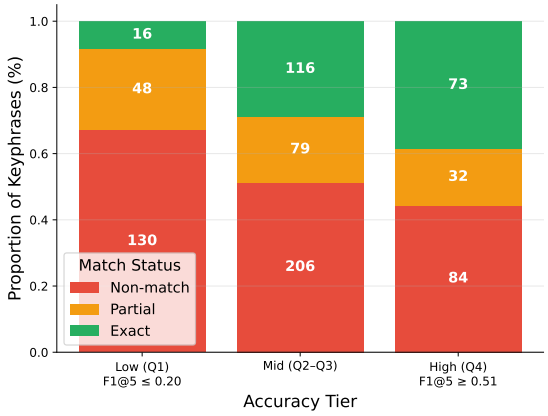


Figure 3: Keyphrase-candidate composition across performance tiers on the kp20k test set, based on per-document F1@5. Partial matches account for approximately 20% and 25% of candidates in the mid- and low-performance tiers, respectively. These candidates share at least one word with a gold keyphrase but receive no credit under exact-match evaluation. Numbers inside the bars indicate counts.

5 Discussion

In this paper, we presented an initial evaluation of a transformer-based keyphrase extraction model leveraging explicit fusion of local (sentence) and global (document) context. The complete architecture is smaller overall than the state-of-the-art comparison system developed by Martinc et al. (2025) (170.9M versus 274M parameters) and achieves comparable performance on the two scientific-domain benchmarks without its mixture-of-experts and BiLSTM components, although it performs substantially worse on the journalistic benchmark. This parameter-count difference distinguishes the architectures but does not by itself establish lower computational cost.

Despite the overall promising extraction performance, the model showed low precision and (over)eager extraction. It tended to generate large keyphrase candidate pools, especially on longer documents, leading to an increase in false positives. While we confirmed that the candidate pools often contained relevant tokens (e.g. substrings of target keyphrases), this is an area where further improvement is needed. One possible way to address this would be to tune the score threshold, which was fixed at 0.5 in our experiments, using held-out data.

More broadly, why train dedicated small language models in an era of large language models with increasing scale? Small language models (sLMs) fine-tuned for dedicated tasks can still outperform their larger generic pre-trained counterparts under certain settings (Martínez-Cruz et al., 2024). This is noteworthy especially considering the smaller computational footprint and greater potential for interpreting the models (Chen and Varoquaux, 2026). Because keyphrase extraction is a starting point for several downstream tasks like indexing, categorization, and summarization, ongoing development of language models that balance flexibility and computational footprint can outweigh the benefits of larger models.

6 Conclusion

We investigated a mixed-resolution context-fusion architecture that combines token-level information from the target sentence with sentence-level information from the surrounding context in a single ordered sequence. Across three benchmarks, the full model achieved performance comparable to SEKE, a recent state-of-the-art supervised system, on the two scientific-domain datasets, although a larger gap remained on the journalistic dataset. Notably, SEKE's overall parameter count is approximately 60% larger than that of our model. This competitive performance is consistent with the hypothesis that compressed document context can support keyphrase extraction, but it does not establish that context fusion itself improves performance because the architecture was evaluated only as a whole. Error analysis showed systematic over-generation, but many predictions counted as false positives under exact-match evaluation partially overlapped with gold keyphrases, and exact matches tended to receive higher scores than partial and non-matches. These findings identify candidate selection and phrase delimitation as promising targets for improvement. Controlled ablations and direct measurements of computational cost remain necessary to isolate the contribution of context fusion and assess its efficiency.

7 Limitations

Our empirical comparison is limited to SEKE (Martinc et al., 2025). Although SEKE reports comparisons with a broader set of supervised, unsupervised, and generative systems, reproducing those systems under our preprocessing and training protocol was beyond the scope of this study. Broader controlled comparisons are needed to establish the model's position relative to other approaches. Our evaluation focuses on three English-

language benchmarks spanning scientific and journalistic text. While this selection enables comparison across domains and document lengths, evaluation on additional domains and languages would provide a broader account of generalization. The reported scores are based on a single training run; repeated runs are needed to estimate variance and determine the reliability of small performance differences between models. Our model is smaller overall than SEKE (170.9M versus 274M parameters), although all 170.9M of its parameters are fine-tuned, compared with 45M trainable parameters reported for SEKE. Because total and trainable parameter counts reflect different computational costs, direct measurements of training cost, inference latency, and memory use would provide a more complete efficiency comparison. Finally, the present experiments evaluate the architecture as a whole. They establish the feasibility of combining token-level target representations with compressed document context, but do not isolate the effect of that context. Comparisons with a target-sentence-only model, together with controls that vary the context window or replace document context with unrelated sentences, are needed to determine whether and how document context affects extraction performance.

8 Ethical Considerations

This study uses publicly available benchmark datasets and involves no human participants or private data. Incorrectly extracted keyphrases could lead to misleading semantic indexing or retrieval, particularly if model outputs are used without human verification.

Acknowledgments

This work was supported by the Slovenian Research and Innovation Agency (ARIS) through the core research program Knowledge Technologies (P2-0103), the project Embeddings-based techniques for Media Monitoring Applications (EMMA; L2-50070), and the project Linguistic Accessibility of Social Assistance Rights in Slovenia (LASAR; J5-50169).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Lihu Chen and Gaël Varoquaux. 2026. What is the Role of Small Models in the LLM Era: A Survey. *arXiv preprint*. ArXiv:2409.06857 [cs.CL], v8.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ygor Gallina, Florian Boudin, and Beatrice Daille. 2019. KPTimes: A Large-Scale Dataset for Keyphrase Generation on News Documents. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 130–135, Tokyo, Japan. Association for Computational Linguistics.

Kelly Hong, Anton Troynikov, and Jeff Huber. 2025. Context Rot: How Increasing Input Tokens Impacts LLM Performance. Technical report, Chroma.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Anette Hulth. 2003. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 216–223.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

Mayank Kulkarni, Debanjan Mahata, Ravneet Arora, and Rajarshi Bhowmik. 2022. Learning Rich Representation of Keyphrases from Text. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 891–906, Seattle, United States. Association for Computational Linguistics.

Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li. 2021. Unsupervised Keyphrase Extraction by Jointly Modeling Local and Global Context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 155–164, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint*. ArXiv:2307.03172 [cs].

Matej Martinc, Thi Hong Hanh Tran, Senja Pollak, and Boshko Koloski. 2025. SEKE: Specialised Experts for Keyword Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2025*,

pages 14191–14205, Suzhou, China. Association for Computational Linguistics.

Matej Martinc, Blaž Škrlj, and Senja Pollak. 2022. TNT-KID: Transformer-based neural tagger for keyword identification. *Natural Language Engineering*, 28(4):409–448.

Roberto Martínez-Cruz, Debanjan Mahata, Alvaro J. López-López, and José Portela. 2025. Enhancing keyphrase extraction from long scientific documents using graph embeddings. *Applied Intelligence*, 55(10):711.

Roberto Martínez-Cruz, Alvaro J. López-López, and José Portela. 2024. ChatGPT vs state-of-the-art models: a benchmarking study in keyphrase generation task. *Applied Intelligence*, 55(1):50.

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592, Vancouver, Canada. Association for Computational Linguistics.

Dhruva Sahrawat, Debanjan Mahata, Haimin Zhang, Mayank Kulkarni, Agniv Sharma, Rakesh Gosangi, Amanda Stent, Yaman Kumar, Rajiv Ratn Shah, and Roger Zimmermann. 2020. Keyphrase extraction as sequence labeling using contextualized embeddings. In *Advances in Information Retrieval*, pages 328–335, Cham. Springer International Publishing.

Tokala Yaswanth Sri Sai Santosh, Debarshi Kumar Sanyal, Plaban Kumar Bhowmick, and Partha Pratim Das. 2020. DAKE: Document-Level Attention for Keyphrase Extraction. In *Advances in Information Retrieval*, pages 392–401, Cham. Springer International Publishing.

Peter D. Turney. 2002. Learning to Extract Keyphrases from Text. *arXiv preprint*. ArXiv:cs/0212013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. _eprint: 2412.13663.

A Additional implementation details and analyses

Token	Label	Reason
`[CLS]`	`IGNORE`	Special token
`neural`	B	Starts keyphrase
`net`	I	Continuation (first subtoken)
`##works`	`IGNORE`	Continuation subtoken
`are`	O	Outside keyword
`powerful`	O	Outside keyword
`[SEP]`	`IGNORE`	Special token

Table 3: Example IOB labeling with subtoken splitting. Continuation subtokens (e.g. ##works) are ignored during training, but inherit the predicted label of the first subtoken during inference to allow word-level predictions.

Word	Subtokens	Subtoken tags (scores)	Decoded word tag
deep	deep	B (0.92)	B
learning	learn, ##ing	I (0.88), O (0.40)	I
models	models	I (0.85)	I
are	are	O (0.95)	O
weak	weak	B (0.40)	O ($q < 0.5$)

Table 4: Example of subtoken-to-word I/O/B tag assignment. For words split into multiple subtokens (learning $\rightarrow$ learn, ##ing), only the first subtoken's score and predicted tag are retained; continuation subtokens (##ing) are ignored. Predictions with $q < 0.5$ are treated as outside (O) during keyphrase generation.

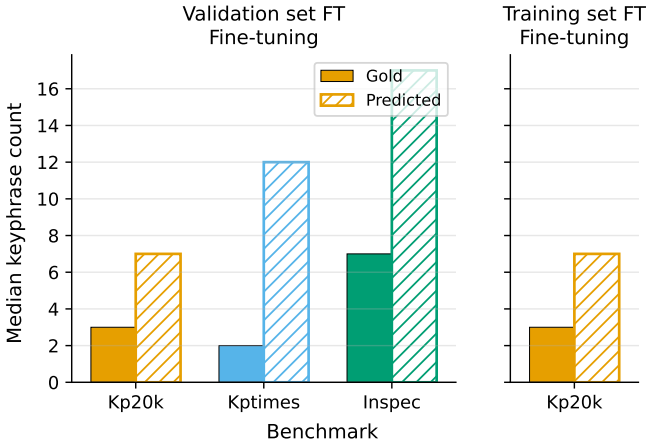


Figure 4: Median numbers of gold and predicted keyphrases across 100 randomly sampled test documents per benchmark. The model consistently predicts more keyphrases than are present in the gold annotations, ranging from approximately 2.3 times as many for kp20k and inspec to 6 times as many for kptimes.

ID	F1@10	Gold Keywords	Extracted Keywords
High Accuracy (in top quartile)			
1	66.7%	boundary element method, object oriented programming, subregions	boundary element method (1.00), object oriented programming (0.99), boundary integral equation methods (0.97)
2	40.0%	cellular structure, fill material, fill materials, retaining wall, shear resistance	distinct element method (1.00), shear resistance (0.99), retaining wall (0.99), displacement incremental analysis (0.98), cellular structure (0.97), distinct element method analysis (0.85), shear behavior (0.83), steel frame (0.81), friction (0.71), wall (0.59)
Mid Accuracy (in mid quartiles)			
3	30.8%	conventional surfactant, lattice monte carlo, mixed micellization, synergism	lattice monte carlo (0.99), surfactant (0.97), aqueous solution (0.96), mixed gemini (0.95), mixed micellization (0.94), surfactant systems (0.90), monte carlo simulation (0.87), gemini (0.83), simulation (0.62)
4	30.8%	deteriorating jobs, multitasking, tabu search	scheduling (1.00), tabu search (0.99), algorithms (0.99), deteriorating jobs (0.98), complexity (0.98), framework (0.96), multitasking environments (0.95), mathematical programming (0.95), lagrangian relaxation (0.94), optimal (0.93), search algorithm (0.81), planning (0.79)
Low Accuracy (in bottom quartiles)			
5	22.2%	wireless sensor networks	adapt (0.98), interpreted (0.97), adaptlets (0.97), wireless (0.93), customization (0.92), wireless sensor networks (0.83), performance (0.82), sensor networks (0.78)
6	0.0%	bareiss algorithm, bulk synchronous parallel, toeplitz system, toeplitz systems	bsp (0.97), load balance (0.97), systolic array (0.95), scheduling (0.92), toeplitz (0.90), data distribution (0.78), speedup (0.78)

Table 5: Examples of ground-truth and extracted keyphrases for six randomly sampled documents from the kp20k test set stratified by document-level accuracy. Color legend: Exact match, Partial match (dashed border). Numbers in parentheses indicate model scores.

Beyond Fluency: A Framework for Analyzing Epistemic and Argumentative Shifts in AI-Assisted Scientific Text Reformulation

Laura Centonze
Department of Human and Social Sciences
Universitas Mercatorum, Italy
laura.centonze@unimercatorum.it

Abstract

Although generative AI increasingly reformulates scientific texts, evaluations based on fluency, readability, and semantic similarity may overlook changes in the epistemic and argumentative properties of the reformulated texts. This paper introduces a *Claim-Preservation Framework* (CPF) to assess whether AI-generated reformulations preserve not only propositional content, but also epistemic force, evidential relation, argumentative function, and qualification and attribution. The framework distinguishes stylistic variation from consequential shifts in modality, hedging, evidentiality, causal interpretation, and argumentative function; its analytical potential is illustrated through an exploratory proof-of-concept application to 40 SciFact claims reformulated under a standardized instruction. This application demonstrates how the proposed dimensions can be used to identify and classify potential shifts in scientific meaning, but it is not intended as a full validation of the framework or as a general estimate of LLM behavior. By conceptualizing reformulation as a transformation of scientific meaning, the paper proposes claim preservation as a distinct dimension of LLM evaluation and outlines future research on computational methods for detecting epistemic and argumentative shifts in scientific publications.

1 Introduction

The rapid adoption of Large Language Models (LLMs) has introduced new forms of computational assistance into scientific writing. Researchers increasingly use generative systems to draft, edit, summarize, translate, and reformulate scholarly texts, making AI-assisted writing a growing component of the scientific publication process (Liang et al., 2024). These applications offer clear benefits in terms of efficiency, linguistic accessibility, textual readability, and stylistic refinement; however, they also raise a less obvious problem: a reformulated scientific statement may become linguistically improved while becoming scientifically different.

The issue is particularly important in scholarly communication because scientific texts do not merely convey propositions: they also encode degrees of epistemic commitment, evidential relationships, qualifications, attribution, causal interpretations, and argumentative functions through which authors position claims in relation to evidence and disciplinary knowledge. Consider the difference between the statements "The findings suggest that the intervention may improve academic writing" and "The findings demonstrate that the intervention improves academic writing." Although the two sentences refer to the same general intervention and outcome, the second formulation increases the apparent certainty of the claim and removes the original qualification. Similarly, changing "the results indicate an association between X and Y" to "the results indicate that X affects Y" introduces a causal interpretation that is not necessarily justified by the source. Replacing "the effect was observed primarily among advanced learners" with "the effect was observed among learners" may likewise broaden the population to which the finding appears to apply. These transformations alter the scientific interpretation of a statement without necessarily reducing lexical or semantic similarity: in each case, the output remains fluent and semantically recognizable while altering an aspect of the information conveyed by the source.

These issues intersect with several established areas of scientific NLP. Research on scientific claim verification, for instance, has examined whether claims are supported or contradicted by available evidence (Wadden et al., 2020), while subsequent studies have extended this line of inquiry to claim–evidence reasoning and citation-based verification with large language models (Alvarez et al., 2024; Javaji et al., 2025). Argument min-

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 133–141
November 10, 2026 ©2026 Association for Computational Linguistics

ing has similarly examined the identification of claims, evidence, and argumentative relations in scientific discourse (Fergadis et al., 2021). Such approaches provide valuable computational representations of scientific knowledge, but generally conceptualize claims as objects to be extracted, classified, retrieved, or verified. AI-assisted reformulation raises a complementary question: how are the scientific properties of a claim affected when an LLM transforms its linguistic realization?

The present paper conceptualizes AI-assisted scientific reformulation as a claim-preservation problem. An epistemic or argumentative shift occurs when an AI-generated reformulation changes the epistemic force, evidential relation, qualification, attribution, causal interpretation, or argumentative function of the original statement, even when its broad propositional content remains recognizable. Semantic similarity is therefore necessary but not sufficient for scientific equivalence: two statements may express broadly related content while differing in the certainty, evidential strength, scope, or argumentative role they convey. To operationalize this distinction, the paper introduces the *Claim-Preservation Framework* (CPF), which represents scientific claims along five analytical dimensions: propositional content (P), epistemic force (E), evidential relation (V), argumentative function (A), and qualification and attribution (Q). These dimensions make it possible to distinguish reformulations that preserve the relevant scientific properties of the source from those that alter one or more of them; although analytically distinct, they may interact within a single reformulation, affecting several aspects of scientific meaning simultaneously. For aggregate evaluation, the framework also assigns an overall outcome label to each pair (Section 4).

The paper makes three main contributions. First, it conceptualizes scientific reformulation as a structured transformation problem rather than exclusively as a language-generation or stylistic task. Second, it proposes the CPF as a multidimensional framework for identifying changes that may remain invisible to conventional semantic similarity measures. Third, it illustrates the applicability of the proposed framework through an exploratory proof of concept involving a purposively sampled set of scientific claims from the SciFact benchmark and a standardized AI reformulation procedure. The empirical analysis assesses the operational feasibility of the framework and illustrates how claim-level changes can be identified and annotated (see Section 7 for its limitations).

The study positions claim preservation as a distinct dimension of LLM evaluation for scientific writing. Rather than assuming that successful reformulation is equivalent to fluency, readability, or semantic correspondence, the proposed approach asks whether the output continues to express the scientific claim with the same epistemic force, evidential relation, argumentative function, and qualifications as the source. This perspective provides a basis for future larger-scale annotation studies, inter-annotator agreement analysis, automated detection, and comparison with conventional semantic similarity, factuality, and claim-verification measures (see Section 7).

2 Scientific Text Reformulation Beyond Fluency

AI-assisted reformulation is generally understood as a constrained form of language generation in which an existing text serves as the basis for a new linguistic realization. In scientific writing, reformulation may be requested to improve clarity, concision, formality, readability, grammatical accuracy, or disciplinary appropriateness. Because the user provides the source text, such operations may appear relatively low risk compared with open-ended generation: the system is ostensibly transforming existing content rather than creating scientific content from scratch. This assumption, however, depends on a strong equivalence premise: that changes in linguistic realization leave the scientific properties of the source statement intact. The premise becomes problematic when scientific meaning is considered at the discourse level. Scientific statements encode not only what is being claimed, but also how strongly it is being claimed, what kind of evidential relationship is associated with it, under which conditions the claim applies, who or what is presented as its source, and what argumentative role the statement performs.

A distinction is needed between linguistic reformulation and scientific claim preservation. Linguistic reformulation concerns changes in form that improve or modify the expression of a statement. Claim preservation concerns whether the relevant scientific properties of that statement remain stable after transformation. The two notions are related but not equivalent. This distinction also clarifies the relationship between fluency, semantic similarity, factuality, and claim preservation. Fluency-

oriented evaluation primarily concerns the linguistic quality of the generated text, whereas semantic similarity measures estimate the degree of correspondence between source and output and are useful for detecting whether two texts express broadly related content. Factuality and claim-verification approaches, in turn, ask whether a claim is supported by available evidence or whether the generated statement is consistent with a source or knowledge base. Claim preservation addresses a related but distinct question: has the transformation maintained the scientific status of the original statement? The distinction is consequential because epistemic and argumentative changes can occur without a substantial loss of semantic correspondence. The two examples in Section 1 illustrate this: the first preserves the broad proposition while increasing epistemic commitment and removing a qualification, and the second modifies the evidential and causal interpretation of the relationship. In both cases, a similarity-based system may identify substantial correspondence between the sentences while failing to represent the nature of the transformation.

Scientific writing provides numerous linguistic resources through which such distinctions are encoded. Hedging and boosting, for example, allow authors to calibrate the strength of claims and negotiate degrees of certainty within disciplinary communities (Hyland, 1998). Evidential predicates and reporting expressions distinguish between observations, interpretations, previous findings, and stronger forms of evidential commitment. Causal and associative constructions likewise encode different relations between scientific entities. Finally, qualifications relating to population, methodology, temporal scope, or experimental conditions constrain the interpretation of a proposition. These properties may also interact (Section 3), so claim preservation cannot be reduced to a single linguistic feature or lexical substitution. Accordingly, the present study treats AI-assisted reformulation as a structured transformation problem: the relevant comparison is not Source → Output → Similarity Score, but *Source Claim → AI reformulation → Multidimensional Comparison → Shift Classification*.

This perspective complements existing research on scientific claim verification, claim–evidence reasoning, and argument mining. Within it, the central distinction is between surface-preserving reformulation, which may modify lexical choice, syntax, or discourse organization while maintaining the relevant scientific properties of the source, and consequential reformulation, which changes one or more properties that contribute to the interpretation of the claim, even when the resulting sentence remains fluent, plausible, and semantically related to the source. The CPF introduced in Section 4 is designed to make this distinction explicit and operational.

3 Epistemic and Argumentative Shifts in AI-Generated Reformulations

The distinction between surface-level and consequential reformulation can be developed through two related dimensions: epistemic shifts and argumentative shifts. Epistemic shifts concern changes in the degree of commitment associated with a proposition. They include strengthening, weakening, modification of hedging and modality, and changes in the way evidence is presented as supporting a claim. The first example in Section 1 illustrates this: the same relationship is presented as tentative and inferential in one formulation and as established in the other. Epistemic shifts can also occur through the modification of evidential predicates. Expressions such as "suggest," "indicate," "support," "demonstrate," and "confirm" position a proposition differently with respect to the available evidence. A reformulation that systematically replaces weaker expressions with stronger ones may produce a form of epistemic inflation even when no new factual information has been introduced. Conversely, introducing additional modal or hedging expressions may weaken an originally well-supported conclusion.

Argumentative shifts concern changes in what a sentence does within the structure of a scientific argument. Scientific discourse contains propositions that function as claims, evidence, interpretations, qualifications, contrasts, limitations, or conclusions. A reformulation may preserve the general topic while changing this function. For example, "Previous studies have identified several limitations of this approach" reports limitations, whereas "Previous studies have demonstrated the ineffectiveness of this approach" converts those limitations into a stronger evaluative conclusion. The second formulation does more than intensify vocabulary: it changes the argumentative work performed by the sentence.

The two dimensions are often intertwined. Strengthening a claim can change its argumenta-

tive role; removing a qualification can increase the apparent generality of a conclusion; replacing an associative relation with a causal one can transform evidence into a stronger explanatory claim. This interaction makes epistemic and argumentative preservation especially important for scientific NLP, where claims are increasingly represented in relation to evidence, citations, discourse structures, and knowledge graphs. Against this background, the task addressed in this paper is not to determine whether AI-generated text is "good" in a general linguistic sense, but whether the transformation preserves the scientific configuration of the source statement.

4 The Claim-Preservation Framework

The Claim-Preservation Framework (CPF) represents a scientific claim as a five-dimensional structure:

$$C = <P, E, V, A, Q>$$

where P represents propositional content, E epistemic force, V evidential relation, A argumentative function, and Q qualification and attribution. Table 1 summarizes the dimensions and the values used for their annotation. Adequate preservation requires that the proposition remain stable together with the epistemic, evidential, argumentative, and qualifying information that determines its interpretation, and a single modal verb, evidential predicate, causal expression, or omitted qualifier can be sufficient to alter the scientific status of a claim. The dimensions are described in turn below.

The first dimension, *propositional content*, concerns whether the entities and relations described by the original statement remain stable. This dimension corresponds most closely to conventional semantic preservation and provides the foundation for subsequent analysis, but it does not by itself determine scientific equivalence. The second dimension, *epistemic force*, concerns whether the degree of commitment associated with the proposition remains comparable. Relevant linguistic resources include modal verbs, evidential predicates, hedges, boosters, and expressions of probability or certainty. The criterion is preservation rather than maximization: a reformulation is not better for being more confident, nor worse for being more cautious. A problem arises only when the degree of commitment of the reformulation differs from that of the source without clear justification.

The third dimension, *evidential relation*, concerns the relation between a claim and the information presented as supporting it, including distinctions between empirical observation, previous research, author interpretation, theoretical inference, and speculation. It also captures changes in the relation asserted between scientific variables, such as the distinction between association and causation (see the second example in Section 1). This dimension connects the framework with scientific claim-verification research (Wadden et al., 2020). The fourth dimension, *argumentative function*, reflects the fact that a sentence may operate as evidence, interpretation, qualification, contrast, limitation, or conclusion depending on its relation to surrounding claims. Thus, preservation requires consideration of local context rather than isolated sentence similarity. This dimension is closely related to scientific argument mining, which seeks to identify argumentative components and relations within scholarly texts (Fergadis et al., 2021). Because the argumentative function of an isolated claim may not always be recoverable from the available text, cases in which the function cannot be established reliably are treated as indeterminate rather than being assigned a forced classification. The final dimension, *qualification and attribution*, is relevant because scientific findings are frequently restricted by population, methodology, temporal scope, disciplinary context, or evidential conditions. Explicit attribution also identifies the source of a proposition and can affect its epistemic interpretation. Removing such information can broaden or strengthen a claim even when its central proposition remains intact. To limit overlap, E is restricted to the degree of commitment expressed through modals, hedges, and boosters, whereas V concerns the type of relation claimed and its support, including association versus causation and observation, prior research, or inference. Because a single reformulation can affect more than one dimension (see Section 3), the dimensions may co-occur and are therefore annotated independently. For descriptive analysis, each source–reformulation pair is assigned a single overall label, derived from the dimension-level annotation as specified in Table 2. The labels are mutually exclusive for aggregate analysis, whereas the underlying dimension-level changes may co-occur.

Dimension	Core question	Annotation values
Proposition	Is the underlying relation preserved?	preserved/ shifted
Epistemic force	Is the degree of commitment preserved?	preserved/ strengthened/ weakened
Evidential relation	Is the evidence-claim relation preserved?	preserved/ strengthened/ weakened
Argumentative function	Does the statement perform the same function?	preserved/ shifted / indeterminate
Qualification & attribution	Are limits, conditions and sources preserved?	preserved/ removed/ added

Table 1: Dimensions of claim preservation.

5 Proof-of-Concept Analysis

To show how the framework can be operationalized, we conducted an exploratory proof-of-concept analysis on claims from the SciFact benchmark (Wadden et al., 2020). SciFact was developed for scientific claim verification and contains claims paired with evidence-bearing scientific abstracts, available through the official GitHub repository (https://github.com/allenai/scifact). Its use here serves a different purpose: rather than asking whether a claim is supported or contradicted by the available evidence, we treat the claim as the source text for an AI-mediated reformulation task, which creates a direct methodological connection between established scientific claim-verification research and the proposed problem of claim preservation under linguistic transformation. The analysis draws on a purposive sample of 40 claims, selected to cover five areas relevant to the CPF: epistemic modality and hedging, evidential relation and strength, association versus causation, qualification and scope, and attribution or argumentative function. A final group comprised claims that potentially involve several dimensions. The sample was designed to provide sufficient variation for testing whether the dimensions can be applied to naturally occurring scientific claims, not to estimate how frequently epistemic or argumentative shifts occur. Each claim was submitted, in a separate conversation, to the same model (ChatGPT, GPT-6, web interface with default settings, accessed in October 2026) under a standardized instruction. The prompt, reported in full in Appendix B, asked the system to improve clarity, concision, and academic style while preserving the scientific properties of the source claim and prohibiting any addition of information or strengthening or weakening of the claim. Keeping the prompt constant isolates reformulation effects from differences in prompting conditions.

The resulting source–reformulation pairs were annotated manually by the author alone, which precluded the calculation of inter-annotator agreement (see Section 7). Each dimension was coded independently using the values in Table 1. An overall label was then assigned according to the criteria in Table 2. The framework supports two descriptive measures: the shift rate for each dimension, defined as the proportion of pairs in which that dimension changes, and the any-shift rate, defined as the proportion of pairs in which at least one dimension changes. Given the small, purposively selected sample, aggregate frequencies are not reported; the analysis focuses on the qualitative examination of representative pairs.

Table 3 presents six illustrative pairs from the sample, reporting the reformulation, the dimensions examined, and the overall label. The corresponding source claims are given in Appendix A and can be retrieved from SciFact through the listed identifiers. The remaining pairs are withheld because they are used in ongoing research. Five of the six pairs are classified as PRESERVED. The exception is #563: the source states that, in British men, haplogroup I "decreases risk" of cardiovascular disease, a causal formulation, whereas the reformulation states that it "is associated with a decreased risk." Population and direction of the relationship are preserved, but the causal commitment is weakened, despite the instruction prohibiting any weakening. The two sentences differ by only a few words, a change that a similarity measure may fail to register; this shows why semantic correspondence alone cannot establish claim preservation.

These observations should be read as a test of applicability rather than as validation. The analysis assesses whether the annotation scheme can be applied to naturally occurring claims, whether it identifies transformations that are consequential for scientific interpretation, and which categories may need refinement before larger-scale annotation. A natural next step is to compare the dimension-level annotations with conventional similarity and factuality measures, in order to test whether high similarity reliably predicts preservation of epistemic and argumentative structure. Pairs that score

Label	Derivation rule	Typical signal
PRESERVED	All relevant dimensions preserved	Stylistic/ lexical reformulation
STRENGTHENED	E and/or V strengthened, with no qualifying shift in another dimension	*suggest* → *demonstrate; associated with* → *affects*
WEAKENED	E and/or V weakened, with no qualifying shift in another dimension	*demonstrate* → *indicate*; *causal* → *associative*
SHIFTED	P, A, or Q changed, or multiple dimensions change in ways not reducible to strengthening/weakening	Qualifier or source removed; function changed
AMBIGUOUS	Reliable classification cannot be established from the available context	Context-dependent interpretation

Table 2: Summary labels derived from per-dimension annotation.

SciFact ID	AI reformulation	CPF dimension and label
#285	Continued HHV-8 transmission among MSM in San Francisco may be attributable to urogenital contact.	Epistemic force; evidential relation; propositional content; PRESERVED
#494	HSCT-T dosage may not be diagnostic when symptom onset occurs less than 3 hours before AMI.	Epistemic force; qualification; PRESERVED
#553	IgA plasma cells specific for transglutaminase 2 are scarce in the duodenal mucosa of patients with active celiac disease.	Propositional content; qualification/scope; PRESERVED
#563	In British men, haplogroup I is associated with a decreased risk of cardiovascular disease.	Evidential relation; WEAKENED
#760	Mathematical models predict that using artemisinin-based combination therapy instead of nongametocytocidal drugs has only a modest impact on reducing malaria transmission.	Evidential relation; propositional content; PRESERVED
#771	In elderly patients with metastatic colorectal cancer, treatment with single-agent fluoropyrimidines resulted in similar efficacy and better quality of life compared with oxaliplatin-based chemotherapy.	Propositional content; qualification/scope; argumentative function; PRESERVED

Table 3: Illustrative examples of claim-preservation outcomes. Source claims are provided in Appendix A.

high on similarity yet are classified as strengthened, weakened, or shifted would support treating claim preservation as a distinct evaluation problem. The SciFact-based analysis provides a structured basis for quantifying the distribution of preservation and shift types in larger samples. Its principal purpose here is to assess whether the proposed annotation scheme captures transformations that conventional semantic evaluation might overlook. The analysis also motivates a direct research question for future benchmark development: does high semantic similarity between an original scientific claim and its AI reformulation reliably predict preservation of epistemic and argumentative structure? If the answer is negative, claim preservation becomes a distinct evaluation problem rather than a secondary component of semantic similarity.

6 Implications for Scientific NLP and LLM Evaluation

The proposed framework positions AI-assisted scientific reformulation as a distinct transformation task within scientific NLP. Existing research has examined argumentative structures in scientific discourse (Fergadis et al., 2021) and the verification of scientific claims against supporting evidence (Wadden et al., 2020; Alvarez et al., 2024; Javaji et al., 2025). The present problem is complementary because it asks what happens when an already structured scientific statement is transformed by an AI system. The relevant question is no longer simply whether a claim exists or whether evidence supports it, but whether the transformation preserves the properties that make the claim scientifically interpretable. This has direct implications for LLM evaluation. Fluency, readability, coherence, and semantic similarity remain useful indicators of output quality, but they should not be treated as sufficient evidence of scientific preservation (Wysocka et al., 2024). Likewise, scientific factuality and claim preservation should be treated as related but distinct dimensions of trustworthy scientific text generation.

The framework also suggests a practical direction for AI-assisted scientific writing systems. Instead of silently replacing a phrase such as "may indicate" with "demonstrates," an AI assistant could

detect the transformation and flag it for user review. Similarly, a system could identify when an associative relationship has become causal or when a methodological or population restriction has been removed. Such functionality would move writing assistance from optimizing text to monitoring how the edits affect the argument. The proposal is also compatible with emerging research on the rhetorical behavior of LLMs, which has reported that LLM-generated representations of scientific research can exhibit increased causal language and rhetorical confidence relative to source texts (Peters and Chin-Yee, 2025). Such findings reinforce the importance of evaluating not only factual correctness but also changes in how scientific knowledge is framed. The present framework addresses this issue at the level of controlled reformulation and may provide a basis for integrating epistemic and argumentative preservation into broader scientific-text evaluation pipelines.

A future benchmark could pair original scientific claims with multiple AI-generated reformulations and require systems to identify whether each transformation preserves, strengthens, weakens, or shifts the scientific configuration of the source. Such a benchmark could be evaluated against expert annotation and compared with semantic similarity, factuality, and claim-verification metrics. The resulting evaluation would provide a more comprehensive account of the reliability of AI-assisted scientific writing.

7 Limitations and Future Research

The framework presented here is conceptual and exploratory and should not be read as a validated detection method. The proof-of-concept analysis is deliberately small and is intended to establish the feasibility of the annotation approach rather than provide generalizable estimates of AI behavior. The analysis relies on a single model and a single prompt, which limits the extent to which the observed behavior generalizes to other systems or instructions. Annotation was carried out by a single annotator, and the reliability of the proposed categories remains to be established. The complete annotated sample is not publicly released at this stage, which limits independent verification of the reported observations. Moreover, epistemic force and argumentative function can be context-sensitive, meaning that reliable annotation may require information extending beyond the individual sentence. Future research should therefore construct a larger corpus of original scientific claims and AI-generated reformulations across disciplines, publication sections, models, and prompting conditions. Expert annotation should establish the reliability of the proposed categories, with particular attention to inter-annotator agreement for borderline cases and to whether the five dimensions are jointly sufficient. Subsequent computational experiments could then test whether existing semantic similarity metrics detect the same transformations identified by human experts and whether LLM-based classifiers can reliably identify specific forms of epistemic and argumentative shifts.

An important direction would be to integrate claim-preservation analysis with existing scientific claim-verification and citation-based evidence systems. This would enable a two-stage evaluation in which a system first establishes whether a scientific claim is adequately supported and then determines whether an AI transformation has preserved the relationship between the claim and its evidential basis. Such integration could contribute to more trustworthy scientific text generation and automated quality assessment.

8 Conclusion

Efforts in AI-assisted scientific writing increasingly focus on improving the linguistic realization of scholarly texts. However, scientific communication depends on more than fluency, readability, or semantic similarity. Scientific claims are also characterized by degrees of epistemic commitment, evidential grounding, qualification, attribution, and argumentative function. This paper has proposed a Claim-Preservation Framework for examining whether these properties remain stable when scientific statements are reformulated by AI. By distinguishing propositional content from epistemic force, evidential relation, argumentative function, and qualification or attribution, the framework conceptualizes scientific reformulation as a structured transformation problem rather than as a purely stylistic operation.

The proof-of-concept analysis illustrates how this perspective can be operationalized by comparing original scientific claims with AI-generated reformulations. In one example, the reformulation weakens the evidential relation of a claim by turning a causal formulation into an associative one, a change that evaluation based primarily on

semantic similarity may leave undetected. The framework is not intended to replace existing factuality, similarity, or claim-verification methods, but to complement them by addressing the preservation of scientific discourse properties. Future research should validate the proposed dimensions at scale, as outlined in Section 7. More broadly, detecting epistemic and argumentative shifts could become an important component of trustworthy AI-assisted scientific writing, helping ensure that improvements in linguistic quality do not come at the cost of scientific meaning and evidential integrity.

References

Carlos Alvarez, Maxwell Bennett, and Lucy Wang. 2024. Zero-shot scientific claim verification using LLMs and citation text. In *Proceedings of the 2024 Workshop on Scientific Document Processing*. Association for Computational Linguistics.

Alexandros Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Harris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111. Association for Computational Linguistics.

Ken Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins.

S. R. Javaji, Y. Cao, H. Li, Y. Yu, N. Muralidhar, and Z. Zhu. 2025. Can AI validate science? benchmarking LLMs on claim–evidence reasoning in AI papers. In *Proceedings of IJCNLP-AACL 2025*, pages 2355–2379. Association for Computational Linguistics.

Weixin Liang, Y. Zhang, Z. Wu, H. Lepp, W. Ji, X. Zhao, H. Cao, S. Liu, S. He, Z. Huang, D. Yang, Christopher Potts, Christopher D. Manning, and James Y. Zou. 2024. Mapping the increasing use of LLMs in scientific papers. *arXiv preprint arXiv:2404.01268*.

U. Peters and B. Chin-Yee. 2025. Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12:241776.

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*, pages 7534–7550. Association for Computational Linguistics.

Magdalena Wysocka, Oskar Wysocki, Maxime Delmas, Vincent Mutel, and André Freitas. 2024. Large language models, scientific knowledge and factuality: A framework to streamline human expert evaluation. *Journal of Biomedical Informatics*, 158:104724.

A. Source claims for the examples in Table 3

ID	Source claim
#285	Continued HHV-8 transmission among MSM in San Francisco may be explained by urogenital contact.
#494	HSCT-T dosage may not be diagnostic if the onset of symptoms occurs less than 3 hours before AMI.
#553	IgA plasma cells that are specific for transglutaminase 2 are scarce in the duodenal mucosa of celiac disease patients when the disease is active.
#563	In British Men, haplogroup I decreases risk of cardiovascular disease.
#760	Mathematical models predict that using Artemisinin-based combination therapy over nongametocytocidal drugs have only a modest impact in reducing malaria transmission.
#771	Metastatic colorectal cancer treated with a single agent fluoropyrimidines resulted in similar efficacy and better quality of life when compared with oxaliplatin-based chemotherapy in elderly patients.

From the SciFact dataset available at https://github.com/allenai/scifact (accessed on October 9, 2026)

B. Reformulation prompt

The following instruction was used for all 40 claims:

> Rewrite the following scientific claim to improve clarity, concision, and academic style. Preserve exactly the original propositional content, degree of certainty, evidential status, causal or associative relation, argumentative function, qualifications, population restrictions, temporal restrictions, quantitative information, and attribution. Do not add information or strengthen/weaken the claim. Do not change associations into causal relationships. Output only the reformulated claim.

TRACS: Telescope Reference & Astronomy Categorization Dataset and Shared Task

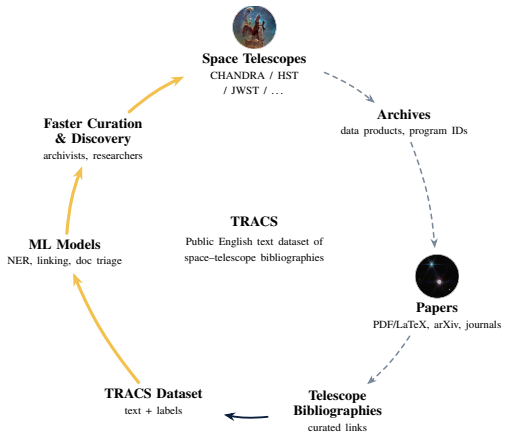
Felix Grezes[1], **Kelly Lockhart**[1], **Alberto Accomazzi**[1], **Jennifer Lynn Bartlett**[1], **Ethan Seefried**[2], **Anjali Pandiri**[3], **Tirthankar Ghosal**[2]
[1]Center for Astrophysics | Harvard & Smithsonian, USA
[2]Oak Ridge National Laboratory, USA
[3]Florida State University, USA
[1]{felix.grezes,kelly.lockhart,aaccomazzi,jennifer.bartlett}@cfa.harvard.edu
[2]{seefriedej,ghosalt}@ornl.gov [3]ap23bl@fsu.edu

Abstract

To assess the scientific impact of observational facilities, Astronomers analyze publications which have made use of the data by the facility. This requires bibliographies annotating and linking data products with the literature, allowing bibliometrics to measure impact of the data. Creating bibliographies is laborious, involves specialists searching the literature for names, acronyms and identifiers, then determining how/if observations were used in those publications. These bibliographies are useful beyond metrics: from the perspective of a research scientist, the data-literature connections provide a critical path to data discovery and access. Thus, by leveraging the efforts of librarians and archivists, we can use telescope bibliographies to support the scientific inquiry process. To this end, we introduce **TRACS**: **T**elescope **R**eference & **A**stronomy **C**ategorization dataset and **S**hared task, consisting of over 89K public English texts from space telescope bibliographies, categorized using a new, simple taxonomy based on discussion with established bibliographers. We hope TRACS will be used to train modern Machine Learning and Artificial Intelligence models that in turn will help data curators build bibliographies. We evaluate how off-the-shelf Large Language Models perform on automatic bibliography curation as a baseline analysis. Results show the difficulty of the task and the need for custom tools.

1 Introduction

Astronomical instruments generate a wealth of data, not just directly with measurements, but indirectly as well, in the form publications that make use of these measurements or describe software created to handle them. To properly credit the teams behind the telescopes, bibliographies linking the software and research back to the telescope are needed.



Since its launch as the NASA Astrophysics Data System (Good, 1992; Kurtz et al., 2000), the Science eXplorer[1] (SciX) (Bartlett et al., 2025) has aimed to help astronomers with bibliographic tools for both discovery and impact measurement. For example, users are not only able to filter by papers in the Hubble Space Telescope (HST) bibliography, a list of papers manually curated by the Space Telescope Science Institute (2025), but also able to see cited/citing paper for the bibliography, which authors or institutions contribute the most, activity over time, and many more advanced second-order operators (Henneken and Kurtz, 2019). While SciX already offers best practices for building and maintaining bibliographies (Observatory Bibliographers Collaboration et al., 2024), and some have automated part of of the process (Grothkopf and Treumann, 2003), it typically remains labor intensive.

Typical Curation Process & Instructions While different groups use different approaches and criteria to the problem of bibliography creation and maintenance, the steps involved typically consist of the following:

1. Use a set of full-text queries to the ADS bibli-

[1]scixplorer.org

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 142–153
November 10, 2026 ©2026 Association for Computational Linguistics

ographic database in order to find all possible relevant papers. This first step aims to identify articles that contain mention of the telescope/instrument of interest so that they can be further analyzed. For instance, the set of query terms used to find papers related to the Chandra X-Ray telescope may be "Chandra," "CXC," "CXO," "AXAF," etc.

2. Analyze the text containing mentions of the telescope/instrument and its variations in order to disambiguate the use of the terms of interest. For the Chandra example, this includes teasing apart the different entities associated with "Chandra," which may correspond to a person, a ground-based telescope, or a space-based telescope.

3. Identify whether the paper in question shows evidence of the use of datasets generated by the telescope or hosted by the archive of interest. The mention of data use may be explicit (e.g. the listing of dataset identifiers), or implied in the text (e.g. mention of analysis and results without identification of the actual dataset). Whenever dataset ids are used, they should be extracted and identified.

4. In some cases, additional classification of the dataset may be collected, such as the instrument used in the observations. This information is also correlated with the kind of data that was used (e.g. image vs. spectra vs. catalog) and its characteristics. In the case of Chandra, there are 7 different instruments that can be used for the data collection (ACIS, HRC, HETG, LETG, HRMA, PCAD, EPHIN), and their use, if explicitly mentioned in the paper, should be reported.

5. For some bibliographies, additional information is collected, such as the relevance of the paper to the scientific use of the data archive. For example, for the Chandra bibliography, the following categories are defined:

   (a) Direct use of Chandra data
   (b) Refers to published results
   (c) Predicts Chandra results
   (d) Paper on Chandra software, operations, and/or instrumentation
   (e) General reference to Chandra

Goals With modern Large Language Models (LLM) capable of ingesting ever larger quantities of text, for ever more sophisticated tasks (Minaee et al., 2024), we at SciX decided to create a dataset to help the community build tools to facilitate the creation and curation of bibliographies. This dataset is TRACS: Telescope Reference & Astronomy Categorization dataset and Shared task, a collection of texts from open access astronomy papers, categorized into three space telescope bibliographies (Chandra X-ray Observatory, Hubble Space Telescope, James Webb Space Telescope), as well as how the papers use the data from the telescope.

Contributions

- a bibliographic taxonomy based on discussion with established bibliographers
- an open dataset of space telescope bibliographies, adapted to our taxonomy from human-curated bibliographies
- a baseline analysis, evaluating off-the-shelf LLMs on the task of automating bibliographic curation

TRACS is available publicly on HuggingFace[2] and was used for the shared task challenge at the 3rd WASP @ IJCNLP-AACL 2025[3]. The scoring evaluation was run on the Kaggle platform[4] .

2 Related Work

Astronomy bibliographies and telescope curation. Telescope bibliographies have been maintained for decades as a means of measuring the scientific return of major observational facilities and of connecting published results back to the data that enabled them. Early efforts at the Space Telescope Science Institute and the Chandra X-ray Center established conventions for what counts as "use" of a facility, how to disambiguate instrument names, and how to track data products through the literature (Space Telescope Science Institute, 2025; Chandra X-ray Center, 2025). The Observatory Bibliographers Collaboration has more recently worked toward harmonizing these practices across observatories, producing shared guidelines for bibliography construction and classification (Observatory Bibliographers Collaboration et al., 2024).

[2] huggingface.co/datasets/adsabs/TRACS
[3] ui.adsabs.harvard.edu/WIESP/2025
[4] kaggle.com/competitions/tracs-wasp-2025

telescope	title/author/year	Sc/In/Me/NT Labels	abstract / body (truncated)
CHANDRA	Chandra X-Ray Observatory Observation of the High-Redshift Cluster MS 1054-0321, (Jeltema et al., 2001), 2014	true, false, false, false	Using **Chandra**, we make a more accurate temperature determination; we examine substructure in the X-ray distribution, and estimate mass/velocity dispersion of MS 1054 to assess cosmological constraints...
HST	Supernova 1996cr: SN 1987A's Wild Cousin?, (Bauer et al., 2008), 2008	true, false, false, false	Note that increasing the aperture radius in the **HST** F656N band to 1.2 yields a magnitude of 17.5 (i.e., an increase of 25% over the pointlike magnitude from SN 1996cr alone), which we attribute to the flux of the underlying H ii region...
JWST	Warm Jupiters in TESS Full-frame Images: A Catalog and Observed Eccentricity Distribution for Year 1, (Dong et al., 2021), 2021	false, false, true, false	The confirmation of these targets will help to select ideal candidates for Warm Jupiter atmospheric characterization for future missions (e.g., **JWST**). Follow-up observations on candidates with missing information listed in Table 5 are also important....
NONE	Tidal adaptive softening and artificial fragmentation in cosmological simulations, (Mostoghiu Paun et al., 2025), 2025	false, false, false, false	Traditional N-body methods introduce localized perturbations in the gravitational forces governing their evolution. These perturbations lead to an artificial fragmentation in the filamentary network of the large-scale structure...

Table 1: **Core fields view of sample records.** The full dataset contains additional fields (see §3.2); here we show the core subset: *telescope*, *author/title/year*, a short *excerpt*, and *annotation flags*. Labels are booleans in the order **Sc**ience, **In**strument, **Me**ntion of telescope, and **N**ot-**T**elescope. Each excerpt is chosen to illustrate how a telescope is referenced in context and is lightly normalized and truncated for fit.

Complementary infrastructure on the discovery side, most notably the SciX (Bartlett et al., 2025) and earlier generations of the same system (Kurtz et al., 2000), has made it possible to query the literature at the scale required for these bibliographies, but the curation itself remains predominantly manual. Prior work on (partially) automating curation has focused on full-text keyword retrieval and rule-based filtering (Grothkopf and Treumann, 2003), which scales poorly to ambiguous mentions and to the fine-grained distinctions (e.g., *science* vs. *mention*) that bibliographers actually make. TRACS differs from these efforts in that it exposes the curatorial decision itself, not just the retrieval step, as a supervised learning problem over public text.

Dataset and software mention detection in scientific text. TRACS is closely related to a growing body of work on identifying mentions of datasets, instruments, and software in scientific publications. The Rich Context / Coleridge Initiative competitions framed dataset mention detection as a named-entity and linking problem over social-science articles (Pan et al., 2023), and SciREX extended this idea to document-level extraction of datasets, methods, metrics, and tasks in machine-learning papers (Jain et al., 2020). Software mention detection has been pursued in similar terms, with corpora such as SoMeSci and the Softcite dataset providing span-level annotations of software references and their usage type (Schindler et al., 2021; Du et al., 2021). A common thread across these resources is the distinction between *mentioning* an artifact and *using* it, essentially the same distinction that motivates our `science`, `instrumentation`, `mention`, and `not_telescope` labels. TRACS adapts this framing to astronomical facilities, where the "artifact" is a telescope (and, implicitly, its data archive) rather than a dataset file or software package, and where disambiguation must contend with shared acronyms, personal names, and grant references.

Scientific NLP in astronomy. Within astronomy specifically, NLP resources have historically lagged behind those in biomedicine and computer science, though this has begun to change (Grezes et al., 2022, 2023). Domain-adapted language models such as astroBERT (Grezes et al., 2021), have established that generic encoders underperform on astronomical text and that domain-specific annotated data is a bottleneck. Most existing astronomy NLP corpora target entity recognition over celes-

tial objects, instruments, or measurements, rather than the document-level, curation-oriented labels we consider here. To our knowledge, TRACS is the first public dataset that pairs full-text astronomy articles with bibliographer-grade labels indicating both *which* facility a paper concerns and *how* the facility's data is used, making it complementary to existing span-level resources and directly usable for evaluating curation assistants.

3 Dataset Description

3.1 Data Collection and Creation

The TRACS dataset consists of papers associated with a telescope and four categories likely to be of interest to bibliographers. We have drawn the categories from a simplification of those discussed by the Observatory Bibliographers Collaboration (2024). These are `science`, `instrumentation`, `mention`, `not_telescope`. Broadly, `science` papers use data from the designated telescope to obtain new results; `instrumentation` papers describe the technical aspects of the telescope; `mention` papers do reference the designated telescope but do not produce new scientific results; and `not_telescope` are papers that include a reference that might otherwise be confused with the designated telescope, i.e. false positives. Full details are available in 3.2 below.

Bibliographic data for the Chandra X-ray Observatory (`CHANDRA`) was provided by the Chandra X-ray Center (Chandra X-ray Center, 2025), while data for the Hubble and James Webb Space Telescopes (`HST`, `JWST`) was provided by the Space Telescope Science Institute (Space Telescope Science Institute, 2025). These curated, human-verified bibliographies include more information than the scope of this dataset (ex: sub-instrument data use, links to grants) and had to be pre-processed into the categories of interest. Furthermore, the papers in this data set do not represent the full corpus of any of these human-curated bibliographies and are not an adequate substitute for them for scientific or administrative purposes. In addition, a small set of papers unrelated to any of these three nor any other space telescope was provided, labeled as `None` telescope in the dataset. This set allows users of the TRACS dataset to easily verify that their models correctly predict that a paper does not relate to a space telescope.

To validate the quality of the dataset, we performed our own manual annotations. The authors, ranging from expert astronomers to non-experts, labeled 10 papers each, giving us a Cohen's kappa score of 0.72 when comparing to the references provided by the `CHANDRA`, `HST`, and `JWST` experts.

3.2 Technical Details

TRACS entries are astronomy papers with the following features:

- `bibcode`: unique string that identifies the entry in the SciX database
- `telescope`: the telescope referenced
- `science`, `instrumentation`, `mention`, `not_telescope`: boolean labels
- `author`, `year`: metadata for the entry
- `title`, `abstract`, `body`, `acknowledgments`, `grants`: the relevant textual information for the entry.

On Kaggle, an additional `Id` column is present for automatic scoring purposes.

`science` New science papers use data from the designated telescope to obtain new results. The authors may be using new observations, using archival observations, or reanalyzing previous results. However, papers that merely refer to previous results for comparison or suggest what might be possible with future observations are Mentions, rather than Science papers. Science papers may use observations directly or indirectly, such as through a published source catalog. Indirect use must be substantive. Papers that overlay new data over images from the designated telescope without discussing the underlying image are Mentions, rather than Science papers. Papers that use catalog data, such as positions or measurements, without further discussion are Mentions, rather than Science papers. Papers that reference a grant associated with the designated telescope but provide no evidence of using data from it are Not Telescope papers, rather than Science papers or Mentions.

`instrumentation` Instrumentation papers describe the technical aspects of the telescope, its calibration activities, its data processing pipeline, or its archival procedures. These papers can discuss hardware, software, or methodologies. A paper that includes new science facilitated by use of the hardware, software, or methodology described in the paper may be both a Science and an Instrumentation paper. A paper that describes a novel

technique or software to achieve its scientific conclusions may be a Science and an Instrumentation paper. A paper that uses calibration, alignment, or engineering data to produce new results may be a Science and an Instrumentation paper.

`mention` Mentions are papers that do reference the designated telescope but do not produce new scientific results (Science) or contribute to understanding it (Instrumentation). If a paper meets the criteria for a Science paper or an Instrumentation paper anywhere, then the paper is a Science paper, even if it also contains statements that would otherwise be considered a Mention. Papers that discuss the designated telescope as part of their introductory overview of the issue, of the history of a problem, or their survey of current relevant research are Mentions. Papers that discuss the designated telescope and its scientific contributions as part of an in-depth review of a research topic are Mentions. Papers that merely refer to previous results for comparison or suggest what might be possible with future observations are Mentions, rather than Science papers. Papers that overlay new data over images from the designated telescope without discussing the underlying image are Mentions, rather than Science papers. Papers that use catalog data, such as positions or measurements without further discussion are Mentions, rather than Science papers. Papers that use a secondary catalog that incorporates data from a catalog produced directly by the designated telescope are Mentions, even if that paper acknowledges the telescope. Papers that reference a grant associated with the designated telescope but provide no evidence of using data from it are Not Telescope papers, rather than Science papers or Mentions.

`not_telescope` Not Telescope papers are papers that include a reference that might otherwise be confused with one or more designations used for the telescope of interest. An telescope may share part of their name with a historical figure for which several things are named. An telescope may share an acronym with an unrelated program. Papers that reference a grant associated with the designated telescope but provide no evidence of using data from it are Not Telescope papers, rather than Science papers or Mentions. If a paper meets the criteria for a Science paper, an Instrumentation paper, or Mention anywhere, then the paper belongs to that category, even if it also contains references to other items that share names in common with the designated telescope or instrument.

	CHANDRA	HST	JWST	None
`train`	31275	37118	11698	294
`test`	3475	4125	1300	294

Table 2: Distribution of dataset entries.

3.3 Data Segmentation for Baseline Task

The TRACS dataset comprises of 89579 entries in total. Table 2 gives the distribution of entries across the three space telescopes and the training and testing dataset splits, as well as across papers that do not feature any space telescopes.

4 Baseline Evaluation Tasks

An automated assistant able to emulate the supervised curation activities listed in 5 above would provide a valuable contribution to the human effort involved. LLMs have shown flexibility in interpreting and classifying scientific articles (Beltagy et al., 2019), which is the basis for this curation activity. They have also been successfully used for information extraction tasks (Grezes and Bartlett), which would help identify the specific datasets mentioned in the papers. This baseline task aims at improving the state of the art technologies to support these curation efforts.

4.1 Task Definition

The TRACS baseline task is composed of two subtasks: *Telescope Classification* and *Usage Classification*, each evaluating a distinct dimension of model understanding over scientific publications.

4.1.1 Telescope Classification

Given the textual fields `title`, `abstract`, `body`, `acknowledgments`, and `grants`, participants were required to predict which telescope was referenced or used in each paper. Valid predictions are limited to `CHANDRA`, `HST`, `JWST`, or `None`. This subtask focuses on assessing the model's ability to correctly identify telescope mentions and usage contexts within natural language.

4.1.2 Usage Classification

The second subtask evaluates how each paper utilizes telescope data. As defined in Section 3.2, each entry includes four boolean labels: `science`, `instrumentation`, `mention`, and `not_telescope`. Each system must output a struc-

tured CSV prediction containing one telescope label and four usage flags for every paper.
In the official Kaggle competition, participants submitted predictions as a single CSV file named `sample_submission.csv` with the following column headers:

```
Id,telescope,science, instrumentation,
mention, not_telescope
```

The `Id` uniquely identifies each paper and is used to align predictions with gold labels during scoring.

4.2 Evaluation Metrics

Each submission is evaluated by matching predictions to reference labels via the `Id` field. Because the test set is partitioned by gold telescope label into four reference splits (`CHANDRA`, `HST`, `JWST`, `NONE`), we report metrics *per split*: within each split, only one of the four telescope classes is populated, so per-split scores measure how well a model recognizes that specific telescope (or correctly abstains, in the `NONE` split). We use F1 scores adapted from the Scikit-Learn library (2011): for telescope classification, the in-class F1 against the gold label; for usage classification, the macro-average over the four boolean usage flags. We also report classification accuracy on the telescope subtask. A model's overall ranking is the average of its per-split scores. A valid example submission is shown below. The first block corresponds to the *Telescope Classification* task, and the second block lists the binary labels for the *Usage Classification* task:

```
Id,telescope
2012A&A...537A..18M,CHANDRA

sci,inst,men,not_tel
True,False,False,False
```

4.2.1 Reproducibility

All models were run with deterministic (greedy) decoding, with an identical instruction prompt and a fixed field-concatenation order (`title`, `abstract`, `body`, `acknowledgments`, `grants`). Per-field input was truncated to fixed token or character budgets calibrated to each model's context window. The full system prompt, per-model context budgets, exact model identifiers, and decoding parameters are reported in Appendix C. All code is publicly available [5] ; All inference scripts, prompt templates, and evaluation notebooks are publicly available [6] .

[5] github.com/eseefried/WASP-dataset-paper26
[6] huggingface.co/datasets/adsabs/TRACS

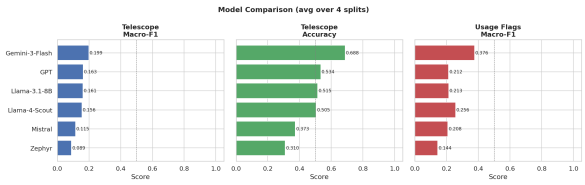


Figure 1: Model comparison averaged over the four reference splits. Gemini-3-Flash leads on every metric; the open 7–20B baselines cluster well below it.

4.3 Baseline Experiments

To establish initial performance benchmarks for the TRACS baseline task, we evaluated six large language models (LLMs) spanning a range of sizes, vendors, and training regimes: *Gemini-3-Flash* (Google, 2025), *GPT-OSS-20B* (OpenAI et al., 2025), *LLaMA-3.1-8B-Instruct* (Grattafiori et al., 2024), *LLaMA-4-Scout* (et.al. , Redacted by arXiv), *Mistral-7B-Instruct* (Jiang et al., 2023), and *Zephyr-7B-Beta* (Tunstall et al., 2023). Each model was run *out of the box*, that is, without any task-specific fine-tuning, using the same instruction set, prompt template, and token limit across all test splits to ensure comparability. Table 3 summarizes the key characteristics of each baseline.

4.3.1 Results and Analysis

Table 4 reports per-split telescope F1 and usage macro-F1 for each model, ordered by average telescope F1. Figure 1 summarizes the same data as average scores across the three primary metrics.

A strong frontier model substantially exceeds open baselines. Gemini-3-Flash dominates on every split and every metric. It reaches 82.4% F1 on `CHANDRA`, 79.0% on `HST`, and 57.1% on `JWST` for telescope recognition, and 64.4% / 46.4% / 39.6% macro-F1 on the corresponding usage subtasks. The next-best models: GPT-OSS-20B and LLaMA-3.1-8B, trail by roughly 15–20 absolute points on telescope F1 and by an even larger margin on usage classification. This indicates that the task is not intrinsically unsolvable in a zero-shot regime, but that it requires the broader world knowledge, instruction-following, and reading comprehension that currently distinguish frontier models from mid-scale open ones.

Telescope difficulty follows corpus prevalence. Across all six models, `CHANDRA` is the easiest telescope to recognize and `JWST` the hardest, with `HST` in between. This ordering matches the relative volume of `CHANDRA`, `HST`, and `JWST` content in the open

Model	Architecture Summary	Motivation	Params
Gemini-3-Flash	Proprietary mixture-of-experts decoder, accessed via API, optimized for fast multimodal reasoning.	Represents a current frontier closed model, providing an upper-bound reference for zero-shot performance.	N/A
GPT-OSS-20B	Decoder-only transformer trained on diverse web and technical text, representing a general-purpose open-source GPT design.	Closest open analog to proprietary GPT-series models; strong general baseline.	20B
LLaMA-3.1-8B-Instruct	Transformer decoder with rotary embeddings and optimized tokenization.	Balances compactness with state-of-the-art reasoning and factuality for 8B-scale models.	8B
LLaMA-4-Scout	Mixture-of-experts decoder from the LLaMA-4 family with long-context support.	Tests whether a newer-generation MoE architecture improves over its dense 8B predecessor.	17B active
Mistral-7B-Instruct	Decoder-only dense transformer using grouped-query attention (GQA) and sliding-window context.	Efficient context handling and strong instruction tuning despite small parameter size.	7B
Zephyr-7B-Beta	Transformer decoder fine-tuned via reinforcement learning from human feedback (RLHF).	Represents the Hugging Face community's open instruction-tuned family emphasizing dialogue coherence.	7B

Table 3: Baseline models evaluated on the TRACS dataset. Each model was run "out-of-the-box" with identical prompts and token limits.

	Telescope F1 (%)				Usage Macro-F1 (%)			
Model	**CH.**	**HST**	**JWST**	**NONE**	**CH.**	**HST**	**JWST**	**NONE**
Gemini-3-Flash	**82.4**	**79.0**	**57.1**	99.8	**64.4**	**46.4**	**39.6**	0.0
GPT-OSS-20B	66.9	59.1	35.4	**100.0**	34.1	27.6	23.1	0.0
LLaMA-3.1-8B	68.6	56.2	37.6	95.5	33.4	27.2	24.4	0.0
LLaMA-4-Scout	65.8	48.1	35.0	**100.0**	44.3	25.6	32.6	0.0
Mistral-7B	36.5	24.2	24.3	99.7	31.2	30.6	21.1	0.0
Zephyr-7B	27.2	7.1	8.8	**100.0**	28.9	20.1	8.6	0.0

Table 4: **Per-split baseline performance on TRACS.** Telescope F1 is computed within each gold telescope split. Usage Macro-F1 averages the four usage labels within each split. Models are ordered by average telescope F1.

scientific literature predating each model's training cutoff: Chandra has been operating since 1999 and HST since 1990, while JWST began producing science only in mid-2022. Notably, JWST recognition is still the weakest class even for Gemini-3-Flash, suggesting that recency rather than ambiguity is the limiting factor.

The `NONE` split is nearly saturated. All six models achieve $\geq 95\%$ accuracy on the `NONE` reference split (Table 4, last column), and four of them score 100%. This is consistent with the intuition that abstaining—predicting that a paper involves none of the three target telescopes—is much easier than identifying which specific telescope is involved. However, this strong `NONE` performance comes at a cost: weaker models default to `NONE` indiscriminately. The confusion matrices in Appendix D show that Mistral and Zephyr misclassify 71–94% of true telescope papers as `NONE`, while Gemini-3-Flash misclassifies only 22–29%.

Usage classification is qualitatively harder. For every model, usage macro-F1 is well below the corresponding telescope F1 (Figure 2, Appendix). Within the usage labels, `science` is consistently the strongest signal (reaching 83.1% F1 for Gemini-3-Flash on HST), while `not_telescope` is the hardest—two of the six models score 0.0% on `not_telescope` across every split. This mirrors the distinction observed in earlier dataset-mention work between identifying that an artifact is referenced versus characterizing how it is used, and reinforces the value of *negative* training examples for future work.

Model agreement decreases on hard examples. On the JWST split, pairwise agreement between models ranges from 0.34 (Gemini-3-Flash vs. Zephyr) to 0.86 (Mistral vs. Zephyr), with high agreement clustering among models that perform similarly poorly. A per-row difficulty analysis (Appendix Fig. 4) shows that roughly 60% of JWST papers are mislabeled by every model in the cohort,

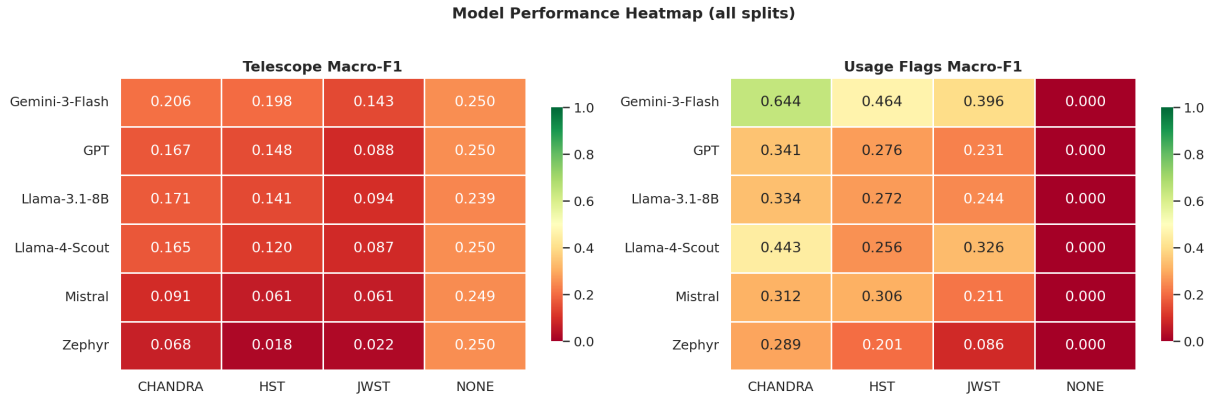


Figure 2: Per-split per-metric heatmap. Left: telescope macro-F1 (computed on the four-class label space, hence diluted by structural zeros from out-of-split classes). Right: usage macro-F1. Gemini-3-Flash is the clear top row across both panels; the NONE column of the usage panel is zero by construction since NONE papers carry no positive usage flags.

indicating that a meaningful fraction of TRACS examples require capabilities that the entire current generation of open models lacks.

Implications for future work. These results carry three practical implications. First, fine-tuning on TRACS is likely to yield substantial gains: even the best zero-shot model leaves $\geq$ 17 points of headroom on CHANDRA telescope F1 and roughly 35 points on the harder usage subtasks. Second, addressing the NONE-bias of smaller models requires either balanced sampling or explicit negative supervision, since several models effectively collapse to a single-class predictor. Third, the hardest residual examples—those that no current model handles—are the most valuable to surface to human curators, suggesting an active-learning workflow in which models triage easy cases and route ambiguous ones to bibliographers.

5 Conclusion and Future Directions

In this paper, we present three contributions to automated bibliographic curation for astronomy. First, we introduce a bibliographic taxonomy developed in collaboration with established bibliographers, grounded in real-world curatorial practices. Second, we provide an open dataset of space telescope bibliographies, manually adapted to align with our taxonomy. Third, we conduct a thorough baseline analysis evaluating the performance of off-the-shelf large language models on a bibliographic curation task. The baseline experiments on TRACS, with the best off-the-shelf LLMs achieving 38% accuracy and 11.5% F1-score on the bibliographic classification task, show that creating bibliographies for space telescopes is not a trivial task to solve, and requires dedicated tools. By releasing the TRACS dataset and taxonomy, we aim to enable further research in this specialized but critical area of scholarly infrastructure. As astronomy archives continue to grow, tools that augment curator expertise will become increasingly essential for maintaining comprehensive and accurate bibliographic records. In the future, we plan to continue to expand TRACS with as many human-curated bibliographies as possible, including ground telescopes. We have already started coordinating with curators at the European Southern Observatory[7] to add the Very Large Telescope[8] to the dataset. In addition to more data, we also would like to refine the evaluation tools. In particular, we would like to use unsupervised evaluation metrics to measure how good models are at recognizing telescope bibliographies from unseen telescopes, evaluating models on:

- Can models generalize and be used to create bibliographies from a new telescope given just a list of names and synonyms for that telescope?
- Can models cluster and detect telescopes in unlabeled astronomy data?
- Can these models be deployed and used by current curators alongside or replacing existing tools?

[7] https://www.eso.org/
[8] https://www.eso.org/public/teles-instr/paranal-observatory/vlt/

References

Jennifer Bartlett, Mugdha Polimera, Kelly Lockhart, Alberto Accomazzi, Michael Kurtz, and Science Explorer Team. 2025. ADS and SciX: Pioneering the Next Generation of Interdisciplinary Research Discovery. In *American Astronomical Society Meeting Abstracts #245*, volume 245 of *American Astronomical Society Meeting Abstracts*, page 442.04.

Jennifer Bartlett, Mugdha Polimera, Kelly Lockhart, Alberto Accomazzi, Michael Kurtz, and Science Explorer Team. 2025. Ads and scix: Pioneering the next generation of interdisciplinary research discovery.

F. E. Bauer, V. V. Dwarkadas, W. N. Brandt, S. Immler, S. Smartt, N. Bartel, and M. F. Bietenholz. 2008. Supernova 1996cr: SN 1987A's Wild Cousin? , 688(2):1210–1234.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv e-prints*, arXiv:1903.10676.

Chandra X-ray Center. 2025. Chandra bibliographic statistics. cxc.harvard.edu/cda/bibstats/bibstats.html. [Online; accessed 14-October-2025].

Jiayin Dong, Chelsea X. Huang, Rebekah I. Dawson, Daniel Foreman-Mackey, Karen A. Collins, Samuel N. Quinn, Jack J. Lissauer, Thomas Beatty, Billy Quarles, Lizhou Sha, Avi Shporer, Zhao Guo, Stephen R. Kane, Lyu Abe, Khalid Barkaoui, Zouhair Benkhaldoun, Rafael Brahm, François Bouchy, Theron W. Carmichael, and 41 others. 2021. Warm Jupiters in TESS Full-frame Images: A Catalog and Observed Eccentricity Distribution for Year 1. , 255(1):6.

Caifan Du, Johanna Cohoon, Patrice Lopez, and James Howison. 2021. Softcite dataset: A dataset of software mentions in biomedical and economic research publications. *Journal of the Association for Information Science and Technology*, 72(7):870–884.

Aaron Adcock et.al. (Redacted by arXiv). 2026. The Llama 4 Herd: Architecture, Training, Evaluation, and Deployment Notes. *arXiv e-prints*, arXiv:2601.11659.

J. C. Good. 1992. Overview of the Astrophysics Data System (ADS). In *Astronomical Data Analysis Software and Systems I*, volume 25 of *Astronomical Society of the Pacific Conference Series*, page 35.

Google. 2025. Gemini 3 flash. `https://deepmind.google/technologies/gemini/flash`. Computer software / Large language model.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 181 others. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, arXiv:2407.21783.

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*, page 143.

Felix Grezes and Jennifer Bartlett. KAILAS: Keyword Labeler At SciX. huggingface.co/adsabs/KAILAS.

Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S Grant, Donna M Thompson, Roman Chyla, Stephen McDonald, and 1 others. 2021. Building astrobert, a language model for astronomy & astrophysics. *arXiv preprint arXiv:2112.00590*.

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 *First Workshop on Information Extraction from Scientific Publications*, pages 1–7.

Uta Grothkopf and Angelika Treumann. 2003. Towards an Automated Retrieval of Publications based on Telescope Observations. In *Library and Information Services in Astronomy IV (LISA IV)*, page 193.

Edwin A. Henneken and Michael J. Kurtz. 2019. Usage Bibliometrics as a Tool to Measure Research Activity. In *Usage Bibliometrics as a Tool to Measure Research Activity. In: Glänzel W*, pages 819–834. Springer International Publishing, Cham.

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.

Tesla E. Jeltema, Claude R. Canizares, Mark W. Bautz, Michael R. Malm, Megan Donahue, and Gordon P. Garmire. 2001. Chandra X-Ray Observatory Observation of the High-Redshift Cluster MS 1054-0321. , 562(1):124–132.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv e-prints*, arXiv:2310.06825.

Michael J Kurtz, Guenther Eichhorn, Alberto Accomazzi, Carolyn S Grant, Stephen S Murray, and Joyce M Watson. 2000. The nasa astrophysics data system: Overview. *Astronomy and astrophysics supplement series*, 143(1):41–59.

Michael J. Kurtz, Guenther Eichhorn, Alberto Accomazzi, Carolyn S. Grant, Stephen S. Murray, and Joyce M. Watson. 2000. The NASA Astrophysics Data System: Overview. , 143:41–59.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large Language Models: A Survey. *arXiv e-prints*, arXiv:2402.06196.

Robert A. Mostoghiu Paun, Darren Croton, Chris Power, Alexander Knebe, Adam J. Ussing, and Alan R. Duffy. 2025. Tidal adaptive softening and artificial fragmentation in cosmological simulations. , 542(2):735–746.

Observatory Bibliographers Collaboration, Raffaele D'Abrusco, Monique Gomez, Uta Grothkopf, Sharon Hunt, Ruth Kneale, Mika Konuma, Jenny Novacescu, Luisa Rebull, Elena Scire, and et al. 2024. Assessing your Observatory's Impact: Best Practices in Establishing and Maintaining Observatory Bibliographies. *The Open Journal of Astrophysics*, 7:85.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv e-prints*, arXiv:2508.10925.

Huitong Pan, Qi Zhang, Eduard Dragut, Cornelia Caragea, and Longin Jan Latecki. 2023. Dmdd: A large-scale dataset for dataset mentions detection. *Transactions of the Association for Computational Linguistics*, 11:1132–1146.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

David Schindler, Felix Bensmann, Stefan Dietze, and Frank Krüger. 2021. Somesci-a 5 star open data gold standard knowledge graph of software mentions in scientific articles. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4574–4583.

Space Telescope Science Institute. 2025. Stsci library and institutional archive. stsci.edu/scientific-community/stsci-library-and-institutional-archive. [Online; accessed 14-October-2025].

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct Distillation of LM Alignment. *arXiv e-prints*, arXiv:2310.16944.

Appendix

A Ethics Statement

The authors of this paper follow principles of transparency and reproducibility. The dataset and code described are publicly available and open source, released under the permissive MIT license, ensuring accessibility for verification and future research. Large language models were employed solely as baseline comparisons in our experiments, in a non-generative mode only, as classifiers. We acknowledge that LLMs may carry inherent biases present in their training data, and we have taken care to document these limitations in our analysis. The use of LLMs as baselines does not constitute endorsement of their outputs, but rather provides a standardized benchmark for evaluating our proposed methods. We are committed to responsible research practices and have considered the potential societal impacts of this work throughout the research process.

B Instructions for Human Annotators

`For each row, select which telescope it relates to (CHANDRA, HST, JWST or None) and how it's used (science, instrumentation, mention, or not_telescope) (if telescope=None, leave blank. not_telescope means it has the string in it but it's about something else: ex a paper citing a person named Chandra).`

C Full Reproducibility Details

This appendix provides the complete prompt template, input handling, and decoding configuration used for the baseline experiments in §4.2.1.

C.1 System Prompt

All models received an identical instruction set. For the API model (Gemini-3-Flash) the instructions were passed as a system message; for the local open-weights models (GPT-OSS, Mistral, Zephyr, LLaMA-3.1-8B, LLaMA-4-Scout) they were prepended to the user message. The verbatim instruction text follows:

```
You are an expert research assistant tasked
with analyzing scientific papers from the
TRACS dataset. Your job is to read the
provided fields and produce structured
predictions.

Tasks:
1. Identify the Telescope:
```

```
   - Output exactly one telescope from:
     CHANDRA, JWST, HST, or None.
   - If no telescope is identifiable,
     output None.
2. Classify the Paper's Usage:
   - science -> True if the paper uses
     telescope data to produce new scientific
     results, else False.
   - instrumentation -> True if the paper
     describes technical or engineering
     aspects of the telescope, else False.
   - mention -> True if the paper references
     the telescope but does not generate new
     results, else False.
   - not_telescope -> True if the reference
     is unrelated or a false positive, else
     False.
   Multiple labels can be True.
3. Output Format:
   Return exactly one CSV row and nothing
   else. Do not include the header. The row
   must have exactly these six fields:
   Id,telescope,science,instrumentation,
   mention,not_telescope
   Example row:
   2012A&A...537A..18M_CHANDRA,JWST,True,
   False,False,False
```

C.2 User Message Template

Each paper was rendered into a user message using a fixed field- concatenation order. Fields are inserted verbatim from the dataset entry, subject to the per-model truncation budgets in Appendix C.3.

```
Id={bibcode}
Title: {title}
Abstract: {abstract}
Body: {body}
Acknowledgments: {acknowledgments}
Grants: {grants}
```

C.3 Input Truncation Budgets

Paper bodies routinely exceed model context windows, so each field is truncated to a fixed budget before concatenation. Local open-weights models are budgeted in tokens (using each model's own tokenizer); Gemini-3-Flash uses a character-based budget. Truncation is applied field by field from the end, preserving leading content.

C.4 Decoding Configuration

All baseline models were run with deterministic (greedy) decoding to remove sampling variance from the comparison. For local models we set `do_sample=False`; for Gemini-3-Flash we set `temperature=0` and `thinking_budget=0` to suppress chain-of-thought tokens from the output budget. Nucleus sampling (`top_p`) was not used by any model.

Field	Budget	Formula
GPT-OSS, Mistral, Zephyr, LLaMA-3.1-8B (ctx=8,192 tokens)		
Title	682	$\lfloor ctx/12 \rfloor$
Abstract	2,048	$\lfloor ctx/4 \rfloor$
Body	2,048	$\lfloor ctx/4 \rfloor$
Acknowledgments	512	$\lfloor ctx/16 \rfloor$
Grants	512	$\lfloor ctx/16 \rfloor$
LLaMA-4-Scout (ctx=4,096 tokens)		
Title	256	$\min(256,\ \max(64, ctx/16))$
Abstract	1,024	$\min(1024,\ \max(128, ctx/4))$
Acknowledgments	341	$\min(512,\ \max(64, ctx/12))$
Grants	341	$\min(512,\ \max(64, ctx/12))$
Body	1,622	$ctx - above - 512$ (remainder)
Gemini-3-Flash (ctx=24,000 characters)		
Title	2,000	fixed
Abstract	6,000	fixed
Acknowledgments	3,000	fixed
Grants	3,000	fixed
Body	remainder	see caption

Table 5: Per-field input budgets for every baseline model. Local models share a uniform fractional allocation of an 8,192-token context; LLaMA-4-Scout's tighter context allocates a residual to Body; Gemini-3-Flash uses fixed character budgets, with the Body field receiving $\max(1{,}000,\ 24{,}000 - used - 2{,}000)$ characters.

Model	Temp.	Max out	Sampling
Gemini-3-Flash	0.0	1,024	thinking disabled
GPT-OSS-20B	—	10,000	Greedy
LLaMA-3.1-8B	—	256	Greedy
LLaMA-4-Scout	—	64	Greedy
Mistral-7B	0.0	256	Greedy
Zephyr-7B	0.0	256	Greedy

Table 6: Decoding parameters used for each baseline model. All local models used greedy decoding (`do_sample=False`). GPT-OSS-20B was given a larger output budget to accommodate its reasoning trace preceding the final CSV row.

D Inter-Model Agreement and Per-Paper Difficulty

The aggregate metrics in §4.3.1 report *how often* each model is correct, but not *which papers* each model gets right. To probe this, we examine two complementary views on the JWST split, the hardest of the three telescope subsets.

Pairwise model agreement. Figure 3 shows the fraction of JWST papers on which each pair of models produces the same telescope label, regardless of whether that label is correct. Two patterns stand out. First, agreement is high *among the weaker models*: Mistral and Zephyr agree on 86% of papers, and the four open mid-scale baselines (Llama-3.1-8B, Llama-4-Scout, Mistral, Zephyr) form a visibly high-agreement cluster. Inspection of the

Model	Model ID
Gemini-3-Flash	`gemini-3-flash-preview`
GPT-OSS-20B	`openai/gpt-oss-20b`
LLaMA-3.1-8B	`Llama-3.1-8B-Instruct`
LLaMA-4-Scout	`Llama-4-Scout-17B-16E-Instruct`
Mistral-7B	`OpenHermes-2.5-Mistral-7B`
Zephyr-7B	`HuggingFaceH4/zephyr-7b-beta`

Table 7: Exact model identifiers used for the baseline experiments, matching the configurations in Table 6.

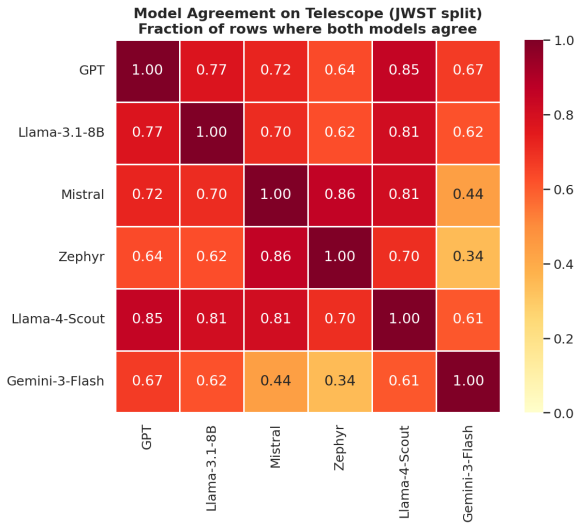


Figure 3: Pairwise model agreement on the JWST split, measured as the fraction of rows on which both models produce the same telescope prediction. Agreement clusters among the weaker models, which tend to default to NONE together; Gemini-3-Flash, the strongest model, shows the lowest agreement with every other model.

underlying predictions shows that this shared behavior is dominated by *co-abstention*: when in doubt, these models converge on NONE. Second, Gemini-3-Flash—the strongest model—shows the *lowest* agreement with every other model in the cohort (0.34 with Zephyr, 0.44 with Mistral). This is consistent with Gemini-3-Flash recovering telescope labels that the other models miss, rather than making independent errors.

Per-paper difficulty. Figure 4 reframes the same predictions from the paper's perspective. For each JWST paper, we count the fraction of the six baseline models that produce the correct telescope label, and plot the distribution. The distribution is sharply bimodal: a large left peak of roughly 750 papers (about 60% of the split) for which *no model* predicts correctly, a small right tail of papers that most or all models handle, and a sparse middle. The left peak is the practically important observation: a substantial fraction of TRACS examples lie outside the current capability frontier of the evaluated open models, and any future system that closes this gap would be providing genuinely new signal rather than re-discovering what easier methods already find. These are the examples that most warrant inclusion in a human-in-the-loop curation workflow.

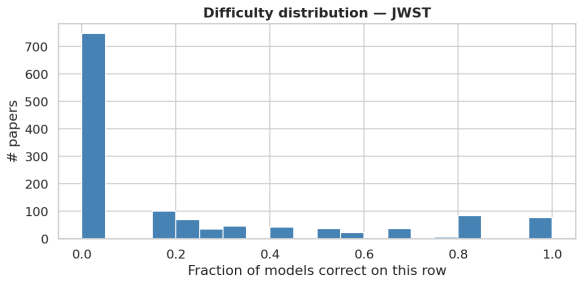


Figure 4: Per-paper difficulty distribution on the JWST split. Each bar counts the number of papers for which a given fraction of the six baseline models predicted the correct telescope. Roughly 60% of JWST papers sit in the leftmost bin, indicating that no model in the current cohort correctly identifies them.

Public Scientific-Figure Adaptation for AstroCLIMB with a 4B Vision–Language Model

Irving Ernesto Quezada Ramírez
Independent Researcher
research@irvingernesto.com

Abstract

We adapt Qwen3.5-4B to the AstroCLIMB task, which asks whether two scientific figures or captions come from the same figure, the same paper, citation-linked papers, or unrelated papers. For caption–image pairs, we first adapt the model on 80,000 public pairs labeled automatically from figure, paper, and citation structure in the public corpus, then train on gold data. Across three gold-training seeds that share one public checkpoint, this initialization improves caption–image development macro-F1 by 0.067 on average over gold-only training under matched gold schedules, although with more total compute. Answer generation and the reinforcement-learning variants we tested do not beat the classifier. The combined system of three modality-specific classifiers reaches a pooled development macro-F1 of 0.756 and a Kaggle leaderboard score of 0.751. Repeated development selection and earlier project exposure to competition materials confine these findings to the evaluated setting, and we document both.

1 Introduction

Much of the evidence in an astronomy paper lives in its figures, yet figure content is hard to search: a light curve or a density cross-section is indexed, if at all, only through its caption. The AstroCLIMB shared task (Grezes et al., 2026; WASP Organizers, 2026) benchmarks whether multimodal models can recover relations between papers from figures and captions alone. Given two objects, a system must decide whether they come from the same figure, the same paper, two papers connected by a citation in either direction, or unrelated papers. Inputs come in three forms: caption–image (CXI), image–image (IXI), and caption–caption (CXC). The organizers frame this as a multimodal benchmark rather than a deployed citation-reconstruction tool, and we make no deployment claim.

The task is harder than its four labels suggest. A same-paper pair and a cited-paper pair can share a topic, the same astronomical objects, and the same plotting conventions, so surface similarity separates neither from the other. At the same time, the task offers an unusual resource. The public AstroCLIMB corpus records figure, paper, and citation structure for tens of thousands of figures, so relation labels can be derived from the graph without manual annotation. Our question is whether that free supervision helps a compact vision–language model before it sees the gold training data.

We answer this with a system of three modality-specific classifiers built on Qwen3.5-4B (Qwen Team, 2026) with low-rank adaptation (LoRA; Hu et al., 2021). Our contributions are:

- A public adaptation stage for the caption–image condition whose labels come from the public corpus graph, with content-matched exclusion of gold and test papers, followed by gold training. In a three-seed matched comparison it improves development macro-F1 by 0.067 on average over gold-only training (Section 4).
- Per-class development results for every input type, showing that related-paper pairs are the weakest class throughout and that same-paper versus related-paper confusion accounts for half of all errors.
- Comparisons that bound the recipe: longer public adaptation trades accuracy between classes, and answer generation with supervised or reinforcement-learning objectives does not surpass the classifier.
- An explicit account of prior project exposure to competition materials and of the dataset's label noise (Section 5), together with public source, prompts, and adapter weights.

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 154–165
November 10, 2026 ©2026 Association for Computational Linguistics

Input type	Training	Development
Caption–image (CXI)	2,230	794
Image–image (IXI)	1,670	590
Caption–caption (CXC)	1,649	602
Total	5,549	1,986

Table 1: Gold partitions after leakage filtering.

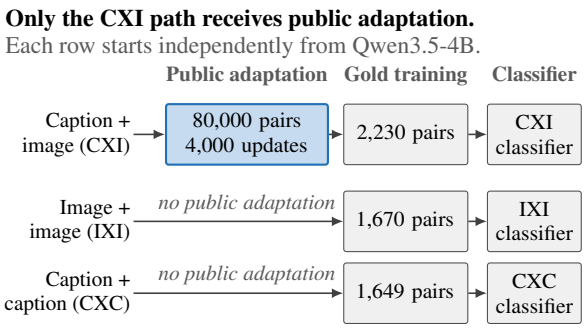


Figure 1: Training overview. One classifier per input type, each trained independently from the same frozen Qwen3.5-4B base (rank-16 LoRA adapters, four-class head). Only the caption–image (CXI) classifier receives public adaptation (80,000 public pairs; gold and test papers excluded) before gold training; IXI and CXC train on gold data alone.

Related work. Prior work has learned from scientific figures and their captions for correspondence (Gómez-Pérez and Ortega, 2019), aligned medical images with text (Subramanian et al., 2020), generated captions for scientific figures (Hsu et al., 2021), and pretrained multimodal scientific models at scale (Tarsi et al., 2024). SPECTER learns document representations from citation signals (Cohan et al., 2020); we instead derive figure-level pair labels from citation structure and use them to adapt a generative vision–language model.

2 Task and Data

Gold partitions. We inherit a gold split made earlier in the project and do not rebuild it. The split groups pairs by paper component, so pairs connected through shared papers form one gold group, and each group lies entirely in one of five folds. We train on folds 2–4 and select models on fold 1, the development set. Fold 0, with 2,001 pairs, was exposed during earlier project work; we exclude it from all fitting, selection, and evaluation, but it is not an untouched test set. After the duplicate and leakage checks in Appendix A, 5,549 training and 1,986 development pairs remain (Table 1).

Public supervision. We index all 114 shards of the public `adsabs/AstroCLIMB` release (94,233 figures) and label caption–image pairs from its recorded figure identity, paper identity, and citation edges. Because a paper can appear more than once in the release, we connect public papers with matching figure content or identifiers into groups, which we use together for exclusion and for the public train–validation split; pair labels use the original paper identities (Appendix A). Before drawing any pair we match public papers, by normalized caption and image hash, against every gold split and against the test figures and captions, and we drop every public component that matches. From the 21,258 figures eligible under the conditions in Appendix A, we draw 80,000 class-balanced training pairs and 2,000 validation pairs from disjoint paper components. Unrelated pairs mix random-paper pairs, same-journal pairs, and two-hop pairs: papers joined through a common third paper in the citation graph but with no citation edge between them. The same matches also prune gold groups that public papers connect to reserved gold, meaning gold held out of both training and development, including fold 0, or to test content. Test content serves only for exclusion and never carries labels. These labels are not manually verified (Section 5). Appendix A gives the procedure and counts.

Selection and reporting. The task metric is four-class macro-F1 computed from the pooled confusion matrix over all development pairs. Because IXI and CXC contain no same-figure pairs, we also report an active-class macro over the three classes present; the four-class macro scores the absent class as zero. Exploratory feasibility trials (Appendix G) were screened on fixed subsets of the gold training set declared before each trial, which we call TRAIN monitors. Public validation pairs never decided when adaptation stopped. All development numbers here are selection results rather than independent tests (Section 5).

3 System

Figure 1 shows the pipeline; the submission concatenates the three classifiers' predictions.

Classifiers. Each classifier adds rank-16 LoRA adapters and a four-class head to a frozen BF16 Qwen3.5-4B base, implemented with Transformers (Wolf et al., 2020) and PEFT (Mangrulkar et al., 2022). The adapters cover language-model MLP

and attention projections, about 21.2M trainable parameters; the vision tower stays frozen. The head reads the final non-padding representation and is trained with cross-entropy. No prompt contains paper identifiers or retrieved labels, and for IXI and CXC the same-figure logit is masked in training and prediction. Appendix B lists adapter placement, prompt handling, and hyperparameters.

Public adaptation for CXI. We adapt only the CXI classifier on public data because CXI is the cross-modal condition and the only one containing all four labels, so it permits one controlled comparison of public initialization against gold-only training without a masked class. The classifier is first trained on the 80,000 public pairs with a 10,000-update warmup–cosine schedule, and the checkpoint saved at update 4,000 is the parent from which gold training starts. Public adaptation for IXI and CXC was not evaluated in this compute-limited study; its absence reflects scope, not a demonstrated optimum.

Gold training and prediction. CXI gold training starts from the parent's adapter and head with a fresh optimizer; IXI and CXC start from the pristine base. All three use the same three-epoch gold schedule at seed 7, and we keep whichever end-of-epoch checkpoint has the highest development macro-F1. Each classifier predicts by argmax without imposing test class counts. For IXI we average class probabilities over both image orders, a choice fixed before evaluation; caption-order averaging did not help CXC, which uses the original order. The submission uses the development-selected models directly, without refitting on all labels. Inference runs at batch size one, because a larger batch changed a few predictions in an earlier model (Appendix C).

4 Results

Submitted system. Table 2 reports per-class development F1 for the submitted classifiers; the pooled four-class macro-F1 is 0.756. Same-figure pairs, which occur only in CXI, are nearly solved. Related-paper pairs are the weakest class in every input type, and the pooled confusion matrix (Appendix D) shows that same-paper versus related-paper confusions make up 281 of the 549 errors, or 51%. Frozen-feature baselines built before training the IXI and CXC classifiers score lower: a blend of a frozen SigLIP 2 (Tschannen et al., 2025) classifier and an OCR-topic model reaches an active-class macro of 0.541 on IXI, against 0.678 for the classifier, and logistic regression on frozen caption embeddings (Zhang et al., 2025; Beijing Academy of Artificial Intelligence, 2023; Xiao et al., 2023) reaches 0.673 on CXC, against 0.728 (Appendix F).

On the Kaggle leaderboard the submission scores 0.751, rank 2 among 17 ranked teams as of 12 October 2026. The leaderboard page states that it is calculated with all of the test data (Kaggle, 2026), and no separate private score is displayed, so no withheld private Kaggle subset awaits evaluation. Because entries were still being accepted, the rank is a dated snapshot rather than an organizer-confirmed final placement.

Matched gold-only controls. The central question is whether public adaptation helps. Under one CXI gold schedule, on one RTX 3090 runtime, with gold seeds 7, 19, and 37, the public arm starts gold training from the parent (the public checkpoint at update 4,000) and the gold-only arm from the pristine base. A fresh export of each run reproduces its selected and final predictions exactly. Public initialization improves every seed pair (Table 3; Figure 2b), by 0.067 on average at the development-selected checkpoint and 0.068 at the final epoch, and within each seed both arms select the same gold update. Several boundaries apply. The three public-arm runs share one parent, so the seeds vary only the gold stage. Total compute is not matched, since the public arm adds its adaptation to the same gold schedule. That parent was itself chosen on this development set among three durations, a choice the gold-only arm lacked, and the shared gold schedule was inherited from the public-initialized recipe rather than retuned for gold-only training, so the gain holds under this fixed schedule, not over an optimized gold-only baseline.

Historical versus fresh runs. The submitted CXI checkpoint, trained earlier on an L40S, scores 0.748, whereas the fresh seed 7 public arm on the RTX 3090 scores 0.756 from identical parent files, data hashes, and hyperparameters. Each run reproduces its own predictions exactly, so the gap lies between separate optimizer runs. Without a controlled cross-hardware test we cannot attribute it to hardware rather than to software or runtime (Appendix C), so we compare fresh arms only within one runtime, and the 0.008 gap shows that this single-seed setup cannot establish differences of a few thousandths as reliable improvements.

Input	SF	SP	REL	UNR	Macro (active)	Macro (4)	N
Caption–image	0.927	0.654	0.610	0.803	0.748	0.748	794
Image–image	—	0.712	0.554	0.768	0.678	0.508	590
Caption–caption	—	0.751	0.593	0.840	0.728	0.546	602
Pooled	0.927	0.708	0.586	0.804	0.756	0.756	1,986

Table 2: Per-class F1 of the submitted system on the development set (not training data). SF, SP, REL, and UNR denote same figure, same paper, related papers, and unrelated papers. A dash marks a class absent from that input type. Macro (active) averages over the classes present; Macro (4) averages over all four classes and scores an absent class as zero. The pooled row is computed from the pooled confusion matrix over all 1,986 pairs and is not the mean of the per-type macros.

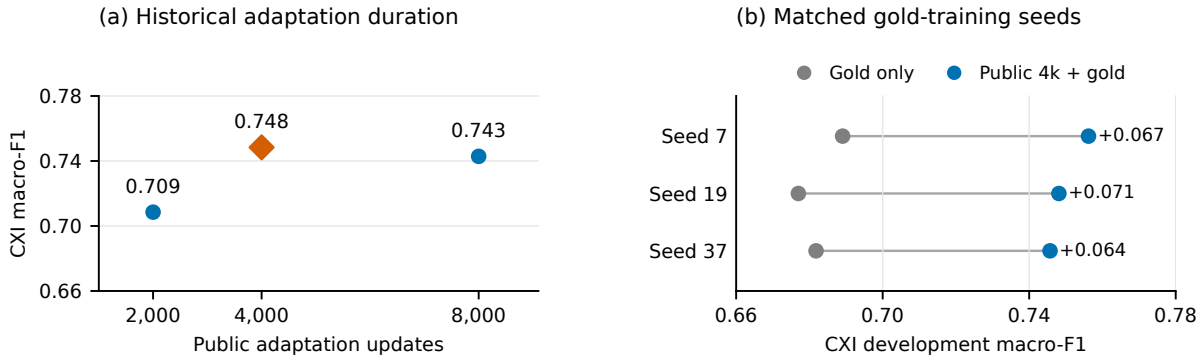


Figure 2: CXI development macro-F1 (794 pairs; zoomed axes). (a) Historical comparison of public adaptation duration with the gold schedule fixed; the diamond marks the submitted model. (b) Fresh runs paired by gold seed; public-initialized runs share one parent. Gold schedules match, total compute does not. Both panels use development-selected checkpoints.

Gold seed	Gold only	Public 4k	Gain
7	0.689	0.756	+0.067
19	0.677	0.748	+0.071
37	0.682	0.746	+0.064
Mean (selected)			+0.067
Mean (final epoch)			+0.068

Table 3: Matched CXI controls on 794 development pairs with epoch-boundary selection and one fixed public parent. These are fresh runs on one runtime, distinct from the submitted model.

Public adaptation duration. A historical single-seed comparison fixes the gold schedule and varies the number of public updates (Figure 2a). Moving from 2,000 to 4,000 updates raises CXI development macro-F1 from 0.709 to 0.748, mainly through the paper-relation classes, with 81 corrections against 49 regressions on the same 794 pairs. Doubling to 8,000 updates gives 0.743: same-paper F1 falls by 0.049 while related and unrelated F1 rise, with 45 corrections and 47 regressions. This is a redistribution across classes, not evidence of a capacity ceiling. At the pooled level the 8,000-update branch is less than 0.001 below the selected system, far less than the cross-runtime gap above, so we retained the 4,000-update branch on its CXI score alone. Appendix E gives the per-class changes and group-resampling intervals for these fixed historical predictions.

Generative and reinforcement-learning alternatives. We also asked whether answer generation could beat the classification head. From the classifier-adapted model, we taught canonical answer tags in a short supervised stage and then applied Group Sequence Policy Optimization (GSPO; Zheng et al., 2025a) with verifiable rewards. Neither the 32- nor the 512-update GSPO model beats the classifier on the pooled development set: GSPO at 512 updates scores 0.751, above answer-supervised fine-tuning at 0.746 and below the classifier at 0.756. Its corrections concentrate on related-paper cases, but without a supervised control matched at 512 updates the gain over answer-supervised training cannot be credited to reinforcement learning. In our group-relative reward setup, a prompt contributes a learning signal

only when its sampled answers disagree in reward, so we also tried selecting such mixed-outcome prompts beforehand, as in dynamic sampling (Yu et al., 2025); this raised the share of informative rollout groups but left a fixed TRAIN monitor essentially unchanged. OCR features added little. Appendices F and G give these experiments in full.

5 Dataset and Exposure Limitations

What the development numbers mean. Every development score here comes from a set consulted repeatedly for selection across the project, and the TRAIN monitors used for screening are training pairs already used to fit the classifier from which the screened models descend. Neither is an independent test. The Kaggle score is computed on all of the test data, but the earlier exposure described below limits how independent even that score is. All comparisons other than the three-seed matched controls rest on a single seed.

Prior project exposure. Earlier work on this project, predating the reported system, was exposed to competition materials in two ways that cannot be undone. First, a scoring-label file named `solution.csv`, distributed in the competition bundle, was inspected during earlier project work, and an identifier-only exploration inferred labels from the class-block ordering of pair identifiers. Second, an earlier synthetic caption corpus of 966,646 pairs overlapped 968 gold caption pairs and used journal-sampling proportions estimated from all gold unrelated examples, including held-out gold. The reported recipe starts from a pristine pretrained base, a new public corpus with gold and test content excluded by component, and gold training data only; every earlier fine-tuned model, feature set, and prediction affected by that exposure is excluded. Pair identifiers serve only to join records, never for prediction, and no test-label supervision or forced test class counts enter the system. We therefore do not claim a clean, independent benchmark result.

Public label noise. The public labels derive from an observed citation graph without manual verification, so a missing edge can make a truly related pair look unrelated, and the duplicate-identity mapping used for exclusion is incomplete. Zero detected overlap after filtering is not proof of independence. All three matched public arms descend from one parent, since we ran no independent public-stage seeds, and exposure of the base model to the underlying astronomy papers during pretraining is unknown.

Reproducibility. We hash source, data, model, and prediction artifacts, and we accept an export only after a fresh inference run reproduces it exactly. Source, prompts, configurations, and adapter and head weights are public at `https://github.com/Sekinal/AstroCLIMB`, which links the model downloads. The packaged inference script verifies the released files against a hash manifest before running but has not been validated on a GPU, and full retraining from the release has not been independently verified.

6 Conclusion

Public initialization from graph-derived caption–image labels improves a 4B vision–language classifier over gold-only training across three gold-stage seeds, with one public parent and extra compute. In our single-seed sweep, extending public adaptation from 4,000 to 8,000 updates traded performance between classes; generative and reinforcement-learning alternatives do not surpass classification, and related-paper pairs remain the hardest class in every input type. Future work includes public-stage seed replication, a compute-matched gold-only arm, public adaptation for the image–image and caption–caption conditions, a long supervised control for the 512-update GSPO run, and independent evaluation and retraining from the release.

Acknowledgments

GLM 5.3 Flash, GLM 5.3, DeepSeek Flash V4, GPT 6 Astra, and Muse Spark 1.3 Contributor assisted with literature discovery, experiment planning, code generation and debugging, experiment orchestration, evidence review, and drafting and revising all sections of the manuscript. Claude Fable 5.1 assisted with manuscript review and prose revision. Responsibility for the submitted claims, citations, code, and interpretation of results rests solely with the author.

The results figure was produced with Matplotlib (Hunter, 2007).

I thank Lium for providing GPU credits used in this work.

References

Beijing Academy of Artificial Intelligence. 2023. bge-base-en-v1.5 Model Card. Official software or model resource; accessed 2026-09-14.

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.

José Manuel Gómez-Pérez and Raúl Ortega. 2019. Look, read and enrich - learning from scientific figures and their captions. In *Proceedings of the 10th International Conference on Knowledge Capture*, pages 101–108. ACM.

Felix Grezes, Kelly Lockhart, Alberto Accomazzi, Atilla Alkan, Ethan Seefried, and Tirthankar Ghosal. 2026. 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.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Kaggle. 2026. AstroCLIMB leaderboard. Accessed 12 October 2026. The page states that the leaderboard is calculated with all of the test data.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. *Preprint*, arXiv:2309.06180.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025. Visual-RFT: Visual Reinforcement Fine-Tuning. *Preprint*, arXiv:2503.01785.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Qwen Team. 2026. Qwen3.5-4B model card. Model revision 851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. 2025. VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model. *Preprint*, arXiv:2504.07615.

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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4548–4559.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *Preprint*, arXiv:2502.14786.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning.

WASP Organizers. 2026. AstroCLIMB: Shared task at WASP 2026.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. 2025. Reinforcement Learning with Verifiable Rewards Implicitly Incentivizes Correct Reasoning in Base LLMs. *Preprint*, arXiv:2506.14245.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2023. C-Pack: Packed Resources For General Chinese Embeddings. *Preprint*, arXiv:2309.07597.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. DAPO: An open-source LLM reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? *Preprint*, arXiv:2504.13837.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *Preprint*, arXiv:2506.05176.

Yubo Zhang, Xueqing Wang, Manhui Lin, Yue Zhang, Penglongyi Deng, Ting Sun, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Changda Zhou, Hongen Liu, Suyin Liang, Cheng Cui, Yi Liu, Dianhai Yu, and Yanjun Ma. 2026. PP-OCRv6: From 1.5M to 34.5M parameters, surpassing billion-scale VLMs on OCR tasks. *Preprint*, arXiv:2606.13108.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025a. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025b. DeepEyes: Incentivizing "Thinking with Images" via Reinforcement Learning. *Preprint*, arXiv:2505.14362.

A Public Supervision Construction

Matching and exclusion. The goal of matching is to find every public figure that is also a gold or test figure, so that public supervision never reuses that content. We compare every public paper against every gold split and against the test figures and captions, using their content but not their labels. Two figures match if their normalized captions are equal, if their image bytes or decoded pixels hash to the same value, or if their 256-bit perceptual difference hashes differ in at most eight bits. Because one paper can appear more than once in the public release, we first connect public papers into groups, which we call components, whenever their figures share an exact normalized caption, image hash, or figure identifier, or have nearly matching perceptual hashes. These components serve only for exclusion and for splitting public papers between training and validation; pair labels use the original paper identities. We then forbid every component that contains a match. This forbids 7,500 public papers. A figure is eligible for pair sampling only if it meets four conditions: it lies outside the forbidden components, its paper has a nonempty DOI, its image-byte and difference hashes were computed successfully, and its normalized caption has at least 40 characters. Exclusion and these eligibility conditions together leave 21,258 eligible figures out of the 94,233 indexed in the 114 shards of `adsabs/AstroCLIMB`. Public training and validation pairs are drawn from disjoint paper components.

Removing overlap between splits. The same matches can link gold groups to content they must not share. A gold object that matches a public paper joins that paper's component, so gold groups can be linked indirectly through chains of public papers and other groups. We therefore take connected components of the combined graph of gold groups, gold objects, and public components, and remove whole gold groups by split (Figure 3):

- a training group is removed if its component contains any nontraining gold object (development, fold 0, or other reserved gold) or any test object;
- a development group is removed if its component contains any reserved gold object (gold outside both training and development, including the exposed fold 0) or any test object;
- if a training and a development group share a component with no other forbidden link, only the training group is removed.

This removes 450 training and seven development pairs, leaving 5,549 and 1,986, with no remaining intersections under these checks. The identity map behind the matches is incomplete, so zero detected overlap is not proof of independence.

Pair sampling. Same-figure, same-paper, and related pairs are labeled directly from figure identity, paper identity, and observed citation edges. Unrelated pairs are drawn from three sources: random pairs of papers, pairs of papers from the same journal, and two-hop pairs, meaning papers joined through a common third paper in the citation graph but with no citation edge between them. The same-journal and two-hop sources are intended to supply negatives that are closer in subject than random pairs; we did not test whether they are harder or whether they share a topic. The 80,000 training pairs are class-balanced. The 2,000 validation pairs only monitored training and never decided when adaptation stopped.

B Training Details

Adapters. Each classifier adds rank-16 LoRA adapters (alpha 32, effective scaling 2) and a four-class head on the final non-padding representation of a frozen BF16 Qwen3.5-4B base. The adapters cover the language-model MLP projections in all 32 layers and the query, key, value, and output projections in the eight full-attention layers, about 21.2M parameters; the vision tower and the linear-attention projections stay frozen. Training minimizes cross-entropy without adapter dropout.

Prompts. CXI prompts hold one image and one caption, IXI prompts hold Figure A and Figure B, and CXC prompts hold two separated captions. Public captions are lowercased and whitespace-normalized, gold captions retain their original case, and both are truncated at 2,400 characters. Images are decoded to RGB from verified original bytes and capped at 262,144 pixels. No prompt contains paper identifiers or retrieved labels. For IXI and CXC the same-figure logit is masked in training and prediction.

Public adaptation. The CXI public stage uses batch size four with two gradient-accumulation steps, so each update sees eight pairs. The adapter

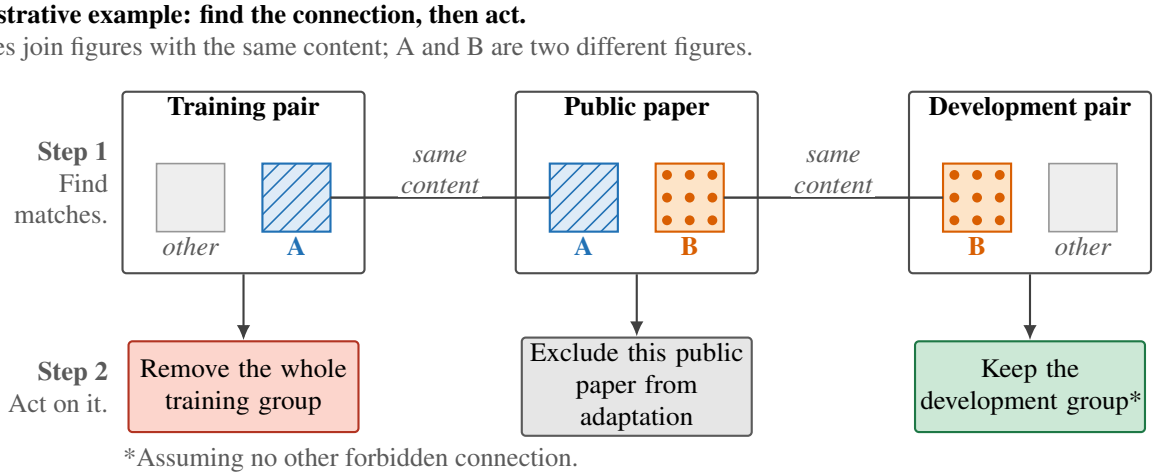


Figure 3: Illustrative example. A training pair and a development pair can overlap indirectly through different figures of one public paper: schematic tile A matches training to the paper, tile B the paper to development. The whole training group is removed, the public paper is excluded from adaptation, and the development group is kept only if no other forbidden link exists. Transitive checks and other removal rules are in the text.

and head learning rates are 10^{-4} and 3×10^{-4}, following a warmup–cosine schedule laid out over 10,000 updates. The submitted parent is the checkpoint saved at update 4,000 of that schedule, after 32,000 pair presentations. We did not rerun a schedule shortened to 4,000 updates, so the parent was taken partway through the cosine decay.

Gold training. CXI gold training starts from the parent's adapter and head with a fresh optimizer; IXI and CXC start from the pristine base. All three use batch size one with eight accumulation steps, three epochs, seed 7, and adapter and head learning rates of 3×10^{-5} and 10^{-4}. With these settings, three epochs amount to 837 updates for CXI, 627 for IXI, and 621 for CXC. We evaluate on the development set at the end of each epoch and keep the checkpoint with the highest development macro-F1, choosing the earlier checkpoint on ties. The submitted checkpoints are CXI at update 837, IXI at update 627, and CXC at update 414. The matched controls in Section 4 reuse this CXI schedule with gold seeds 7, 19, and 37 (Figure 4). All six runs complete 837 updates and are evaluated at the epoch boundaries, updates 279, 558, and 837. For seed 19 both arms select update 558; for seeds 7 and 37 both select update 837.

C Prediction and Reproducibility

Batch size. We run inference at batch size one, the same batch size used in training, because batching changed predictions in an earlier model. That model, trained from the 2,000-update public parent to gold update 837, changed 11 of its 794 CXI development predictions when scored at batch size two, with a class probability moving by as much as 0.380 for one pair. We did not isolate the numerical cause. Exporting at batch size one does not explain the effect; it removes it from our pipeline so that our exports replay exactly, which is not a general guarantee of reproducibility.

Cross-runtime difference. Two runs with the same recipe on different machines gave different scores. The submitted CXI checkpoint was trained on an NVIDIA L40S and scores 0.748 on the 794 development pairs. The fresh seed 7 public arm of the matched controls was trained on an RTX 3090 from the same public parent files, the same data hashes, and the same gold hyperparameters, and scores 0.756; the two configurations differ only in file paths and a wall-clock safety deadline. Both runs complete 837 updates. For each run, a fresh inference pass reproduces that run's own predictions exactly, which shows that each run is internally consistent but does not show that training is reproducible bit for bit across hardware. Different floating-point kernels and reduction orders in BF16, together with nondeterministic GPU execution, are plausible mechanisms. We ran no controlled cross-hardware test, so we cannot attribute the 0.008 gap to hardware rather than to software or runtime, and we do not claim that GPU type alone causes it.

Verification and release. We hash source, data, model, and prediction artifacts, meaning the code, inputs, weights, and prediction files on which an

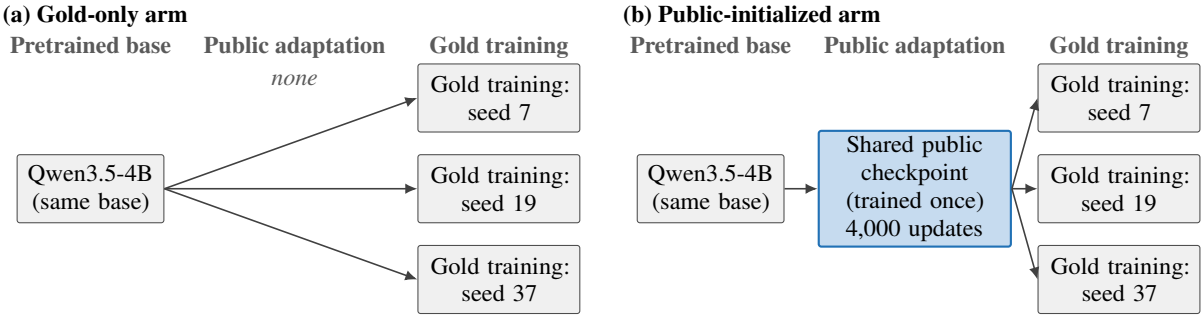


Figure 4: Design of the three-seed matched control (Section 4, Table 3). Both arms start from the same pretrained base and run the same CXI gold schedule on the same runtime with gold seeds 7, 19, and 37. (a) The gold-only arm trains each seed directly from the base. (b) The public-initialized arm first trains one public adaptation checkpoint, saved at update 4,000, and all three gold runs branch from that single checkpoint, so the seeds replicate only the gold stage, not public adaptation. Gold schedule and runtime match across paired seeds; the public arm uses extra adaptation compute. Arrows denote training progression.

export depends, and we accept an export of predictions only after a fresh inference run reproduces it exactly. Source, prompts, configurations, and adapter and head weights are released. The packaged inference script checks every released weight and configuration file against a hash manifest before it runs, but we have not validated it on a GPU from a clean install, and full retraining from the release has not been independently verified. Kaggle displays the leaderboard score as 0.75149.

D Confusion Matrices and Examples

Figure 5 gives the development confusion matrices of the submitted classifiers, with true classes in rows and predicted classes in columns. These describe repeatedly selected development results and are not an independent evaluation.

Two caption–caption errors. To illustrate the dominant confusion we selected, deterministically, the development pair with the smallest identifier in each direction of the same-paper versus related-paper confusion for the caption–caption classifier. We paraphrase the captions rather than quote them. In the first pair, the gold label is same paper and the classifier predicted related papers. One caption describes a horizontal cross-section of the initial density in a simulation; the other describes the time-integrated Poynting flux through a boundary in tuned and detuned simulations. The two captions concern different physical quantities from the same study, which illustrates how judging a same-paper relationship can be difficult when the captions describe different measurements. In the second pair, the gold label is related papers and the classifier predicted same paper. One caption discusses the radial structure, pressures, and luminosity of a simulated accretion disk; the other plots black-hole spin-up against spin and accretion rate from general-relativistic magnetohydrodynamic simulations. The captions share an accretion theme and a simulation vocabulary, but a shared theme does not establish a shared source paper. These are descriptive selected errors; they do not explain the model's internal reasoning or confirm that the gold labels are correct, and we draw no claim about the underlying papers from them.

E Public Adaptation Duration

The duration comparison in Section 4 is a historical single-seed sweep on an L40S with the gold schedule fixed. Development selection chose gold update 837 for the 2,000- and 4,000-update parents and update 279 for the 8,000-update parent. Per-class CXI F1 changes from 2,000 to 4,000 public updates are -0.005 for same figure, $+0.052$ for same paper, $+0.072$ for related papers, and $+0.041$ for unrelated papers, with 81 corrections against 49 regressions on the same 794 pairs. From 4,000 to 8,000, same-paper F1 falls by 0.049 while related and unrelated F1 rise by 0.018 and 0.011, with 45 corrections and 47 regressions. Both pooled scores round to 0.756; the 8,000-update branch is lower by less than 0.001.

To gauge how much these differences depend on which development pairs happen to be present, we resample the development set at the level of its 278 inherited gold groups. We draw the groups with re-

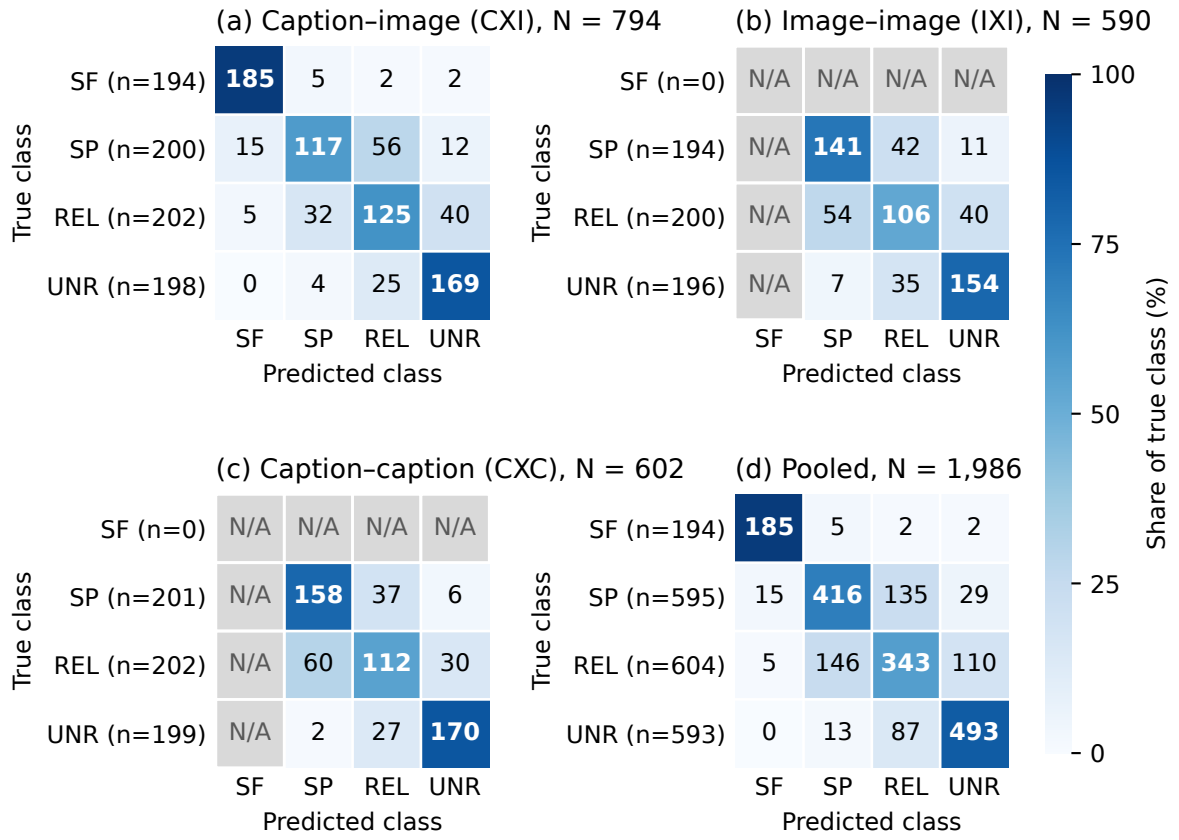


Figure 5: Development confusion matrices for (a) caption–image, (b) image–image, (c) caption–caption, and (d) all pairs pooled. Rows are true classes with support n; columns are predicted classes. Color is the row-normalized share of each true class; numbers are raw pair counts. Gray N/A marks the same-figure class, absent from IXI and CXC and masked at prediction. SF, SP, REL, and UNR denote same figure, same paper, related papers, and unrelated papers. The pooled panel sums counts rather than averaging them; the task metric is its four-class macro-F1.

placement 10,000 times, keep every model's predictions fixed, recompute CXI macro-F1 from each resampled confusion matrix, and report the 2.5th and 97.5th percentiles of the difference between models. The resulting 95% intervals are $[0.007, 0.074]$ for the 2,000-to-4,000 change, $[-0.036, 0.023]$ for the 4,000-to-8,000 change, and $[-0.029, 0.002]$ for GSPO-512 minus the classifier. These intervals capture only the variation from resampling fixed historical predictions. They do not account for model selection, for training variation across runs, for the pooled scores, or for the three-seed gain. The inherited groups may miss dependencies between pairs, and an interval that spans zero does not establish that two models are equivalent.

F Frozen-Feature Baselines, OCR, and Auxiliary Transfer

Frozen-feature baselines. Before training the IXI and CXC classifiers we built frozen-feature baselines. For IXI, a 75/25 blend of a multilayer perceptron on frozen SigLIP 2 (Tschannen et al., 2025) features and a logistic model on OCR-topic features reaches an active-class macro-F1 of 0.541. The IXI classifier reaches 0.655 with a single image order and 0.678 with image-order averaging. For CXC, logistic regression on frozen Qwen3-Embedding-8B (Zhang et al., 2025) and BGE-base-en-v1.5 (Beijing Academy of Artificial Intelligence, 2023; Xiao et al., 2023) caption features with TF-IDF and lexical statistics reaches 0.673; the CXC classifier reaches 0.728 in the original caption order and 0.727 with caption-order averaging, so we keep the original order.

OCR. OCR adds little. PP-OCRv6 (Zhang et al., 2026) features alone score 0.441 on CXI, and OCR-conditioned adaptation of the vision–language model falls slightly below the pristine gold-only model, although a fixed 50/50 combination lifts a superseded whole-system baseline from 0.690 to

0.695.

Frozen-feature IXI transfer. In a separate experiment with a different architecture and training procedure, public-to-gold transfer did not help a frozen-feature IXI classifier, even in a matched-duration follow-up. This negative result concerns frozen features, not the adapted vision–language model, and it did not determine the CXI-only scope of public adaptation in Section 4; it says nothing about whether adapting the IXI vision–language classifier on public data would help.

G Generative Training and Reinforcement Learning

Motivation. Visual-RFT and VLM-R1 (Liu et al., 2025; Shen et al., 2025) motivated trying verifiable visual rewards, studies of coverage and reasoning under verifiable rewards (Yue et al., 2025; Wen et al., 2025) motivated our sampled versus single-answer evaluation, and DeepEyes (Zheng et al., 2025b) motivated fixed-view inspection trials, which were negative. None of these pipelines was reproduced in full.

Answer supervision. The generative variants predict by writing a short answer tag instead of reading the classifier head. The classifier-adapted model initially produced only 13 strictly valid answer tags in 64 completions, so we first taught it the canonical tags with supervised fine-tuning on 1,024 gold training pairs for 128 updates; we call this stage answer-SFT. All generative models reported below predict from the generated answer, not from the classifier head.

GSPO. Group Sequence Policy Optimization (Zheng et al., 2025a) is a reinforcement-learning method that samples several answers for each prompt, which we call a context, scores each answer with a reward, and updates the model toward answers that score better than the others sampled for the same context. We apply it with TRL (von Werra et al., 2020) using batch size 16, four sampled completions per context, two uses of each rollout, learning rate 10^{-6}, rewards normalized within each context's group of samples, no penalty for drifting from a reference model, and asymmetric clipping bounds of 0.0003 and 0.0004 on the sequence-level importance ratio. The reward is one only for a correct canonical answer terminated by the end-of-sequence token, and zero otherwise. On the pooled development set with IXI and CXC held fixed, the submitted classifier scores 0.756, answer-SFT 0.746, GSPO after 32 updates 0.745, and GSPO after 512 updates 0.751. Of the 512-update model's 26 corrections relative to answer-SFT, 25 are related-paper cases, and all 19 regressions lie elsewhere. A short supervised control matches the 32-update GSPO run in parent, 64 TRAIN contexts, 32 updates, batch size, learning rate, and seed 7, but not in tokens, FLOPs, or wall time, and scores 0.747. Without a supervised control matched at 512 updates, the gain of the 512-update model over answer-SFT cannot be credited to reinforcement learning. The 32- and 512-update runs also differ in context pool, not only in update count.

Mixed-outcome selection. In our setup, when every sampled answer for a context receives the same reward, the group-relative advantage is zero and that context contributes no learning signal. Dynamic sampling (Yu et al., 2025) exploits this by discarding such contexts during training. We tested an offline version of the idea on training data only. We sampled eight answers per candidate context and selected 16 contexts whose answers had mixed outcomes, some correct and some not. As a control we selected 16 random contexts with the same class distribution. Each set received 32 GSPO updates with vLLM-assisted rollouts. Mixed selection raised the number of informative groups, those with nonzero advantage, from 14 to 24 of 32. The fixed TRAIN monitor of 256 contexts barely moved: 1,874 correct answers of 2,048, against 1,875 for the control. Neither trial met the prespecified screen of a two-percentage-point improvement. This was an offline selection motivated by dynamic sampling, not a full online implementation of that method.

Backends. vLLM (Kwon et al., 2023) generates the training rollouts, but all reported evaluation uses the native Transformers inference path, because the two backends do not produce identical predictions. For scoring and for checks that a reloaded checkpoint reproduces its predictions, we match the batch shape and the BF16 autocast setting.

Accelerating Literature Data Extraction using AI at the IPAC Archives

Nicholas Susemiehl Nima Chartab Meca Lynn Ricky Nilsson
IPAC, California Institute of Technology
1200 E. California Blvd., Pasadena, CA 91125, USA
{nsuse, nchartab, mlynn, ricky}@ipac.caltech.edu

Abstract

The repositories of the NASA Extragalactic Database (NED) and NASA Exoplanet Archive (NEA) provide the scientific community with regularly-updated records of published data on observed astronomical objects. Over their years of operation, both archives have seen increases in the rate of publications that are relevant to their holdings. We are developing an AI-powered data extraction tool to increase the rate at which NED and NEA can process published articles into their databases. This is accomplished through a multi-stage pipeline that combines large language models (LLMs) with programmatic techniques and human supervision to create formatted database load files. Pipeline evaluations on sampled human-made load files estimate a macro-level recall of 0.89 for NED and 0.83 for NEA with typical runtimes of 20 minutes per-paper. While this tool is still being developed (aided by human feedback), present results suggest that contemporary LLM-based approaches are useful for complex structured data extraction. Future work remains to improve the pipeline's precision and further adapt it to the wide variety of articles present in the literature.

1 Introduction

The NASA Extragalactic Database (NED)[1] and NASA Exoplanet Archive (NEA)[2] are two astronomical data repositories operated by IPAC at the California Institute of Technology. NED has collected over 1.1 billion distinct objects, including galaxies, quasars, and groups/clusters. NEA seeks to provide a complete list of confirmed exoplanets, which now number over 6,300, and their stellar hosts. Both archives ingest their holdings from the text and tables of published astronomical literature which present the discovery of new objects or new measurements of existing ones. Examples of data the archives extracts include the object's type, coordinates, and redshift for NED and a planet's mass, transit midpoint, and detection method for NEA. Data extracted from articles must be transformed to an archive-specific database load file. Both archives face challenges ingesting relevant articles at the rate they are published. Creating load files is a significant manual effort at NED, with hundreds of papers processed each month. Anticipated exoplanet yields from future missions have the potential to increase NEA's holdings by hundreds of thousands of objects (Perryman et al., 2014; Penny et al., 2019; Wilson et al., 2023; Matuszewski et al., 2023; Ge et al., 2024). Creating load files by hand can take minutes or hours depending on article complexity. Given these realities, it has become important for the data archives at IPAC to enhance the throughput of their data ingestion pipelines.

The field of natural language processing (NLP) has yielded tools that are increasingly capable of mining data from the text of scientific journal articles. Transformer-architecture large language models (LLMs), which use attention mechanisms to create dynamic contextual understanding of text (Vaswani et al., 2017), have been shown to be applicable to a wide range of NLP tasks even without expensive fine-tuning (Brown et al., 2020). Recent advances in LLMs have greatly expanded the possibilities of an AI-driven data extraction pipeline: enlarged context limits (now ranging from 250 thousand to over 1 million tokens) allow for the inclusion of multiple full articles and their tables in a single prompt (Gemini Team et al., 2024), increased instruction-following and thinking capabilities help the models reason over the numerous extraction rules and complex article structures (Wei et al., 2022; Guo et al., 2025), and tool usage enables dynamic context-driven unit conversions (Yao et al., 2022; Schick et al., 2023).

The goal of this work is to produce a tool which accelerates the processes of data extraction and

[1] https://ned.ipac.caltech.edu/
[2] https://exoplanetarchive.ipac.caltech.edu/

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 166–173
November 10, 2026 ©2026 Association for Computational Linguistics

load file creation for NED and NEA. There is no expectation that the load files created using these tools will be perfect, so human supervision and feedback is integral to this process.

2 Methods

The data ingestion pipelines for NED and NEA contain a variety of AI, programmatic, and human-driven components. At a high level, this pipeline is identical for both NED and NEA: newly published articles are discovered through listing services, relevant articles are identified, necessary data is extracted from relevant articles, and load files are created to ingest these data into the archives' holdings.

2.1 Data Retrieval

Each module of this work uses the cleaned body text of a journal article and/or tables published either in the article itself or externally. Articles are retrieved via the Science Explorer[3] (SciX) API according to a format priority order as accessible: first the publisher version of the article's HTML, then the publisher PDF, then the pre-print HTML, and finally the pre-print PDF. HTML files are converted to strings using Beautiful Soup (Richardson, 2007) and PDF files with Docling (Livathinos et al., 2025). Tables are converted to Pandas dataframes (Wes McKinney, 2010), images (Artifex Software, Inc., 2026), and markdown key-value format[4]. Preprocessing normalizes characters and strips extraneous formatting. External data tables hosted on publisher websites are accessed when possible. The impacts of text and table parsing errors are discussed in the Appendix (7.4).

2.2 Relevance Classification

Both NED and NEA use trained machine learning classifiers to predict the probability that an article is relevant to their holdings. The relevance classifier used by NED (Chen et al., 2022) is based on the Stanford Classifier (Finkel et al., 2005), while the NEA tool uses a finetuned INDUS (Bhattacharjee et al., 2024) model. Both of these tools successfully accelerate the task of scientists to identify relevant papers.

[3]https://scixplorer.org/

[4]https://www.improvingagents.com/blog/best-input-data-format-for-llms/

2.3 Data Extraction

Once a relevant article is identified, the data it presents in texts and tables must be extracted to structured outputs. Data ingestion follows a sequential pipeline involving four main LLM-backed stages. This begins with the pre-processing of article text/tables and ending with a draft database load file. An article's text and tables follow parallel paths: LLM prompts only include the body text and smaller tables while all tables go through a programmatic extraction code. Pre-processing and extraction currently use the LLM Qwen 3.8 27B (BF16 quantization; Qwen Team, 2026) while validation and normalization are performed with Gemma 4 31B (Q4_K_M quantization; Gemma Team et al., 2026). These be swapped with any tool-capable decoder LLM. These models are distributed over 6 x NVIDIA L40S (46 GB VRAM) for production and development, with one article running on a single card in the cluster. The prompting of these models employs in-context learning by providing examples of extracted data structures to accomplish this task without expensive finetuning. A diagram of the pipeline is available in the Appendix (Figure 1).

2.3.1 Pre-processing

The article text and tables go through initial preprocessing steps to prepare for extraction. The two main components identify the locations of object names and parameter labels (table column/row headings) before full extraction. Pre-processing includes a first pass at object name detection by searching the article text for matches to the NED/NEA accepted regular expression name patterns. This is inexpensive but not robust to variations from expected object name spellings (e.g. abbreviations), so name detection is performed more robustly in Stage 1 (Section 2.3.2). Parameter labels are identified two ways: similarity matching to a dictionary with keys as the preferred labels and values with lists of previously-seen labels and an LLM call prompted to find labels. Found names and labels are tagged within the text.

2.3.2 Stage 1: Name Extraction

The first data extraction step identifies the names of astronomical objects mentioned in the article and classifies them according to their relevance. The LLM is prompted to return verified object names as well as their basic type (extragalactic/galactic/solar system/exclude for NED and exoplanet/star/solar

system/exclude for NEA). This prompt includes names tagged via regular expressions in the previous step as likely object names, but not necessarily the only ones. Objects classified as extragalactic for NED and exoplanet/star for NEA (along with both primary and secondary objects) are then passed to the next stage for data extraction.

2.3.3 Stage 2: Parameter Extraction

The second primary extraction stage is an LLM call with the processed article text, small tables, and relevant names identified in Stage 1 included in the prompt. Tables with fewer than 200 rows (selected to preserve context usage and written in markdown key-value format) are included in the prompt in the markdown key-value format. The model is instructed to extract all NED and NEA parameters for the provided objects simultaneously. Descriptions are provided for each parameter along with examples. The text and previously-extracted data from the most similar article selected via retrieval augmented generation (RAG; Lewis et al., 2020) are included in the Stage 2 prompt as a task example. The RAG process uses the Qwen 3 embedding model (Zhang et al., 2025) to vectorize the first 32,000 tokens of an article and selects the most similar article to the present case by computing cosine similarities against a repository of previously-embedded articles. Stage 2 extracts the parameter data exactly how they appear in the text and returns them in a structured JSON. The Stage 2 outputs also include source texts for each extracted value, snippets of text from the article the LLM used as evidence to support the extraction of a particular data point. The source texts improve the extraction quality and are useful later for validation, normalization, and provenance tracing.

2.3.4 Tabular Data Extraction

A separate path exists for the extraction of data from tables treated as text for two reasons: LLM's required input of strings of texts flattens the two dimensional structure of tables to one dimension, losing associative information, and large tables (some having thousands of rows) consume significant fractions of a model's total token limit. A programmatic approach leverages the structure of tables to deterministically and efficiently extract data. This relies on the identification of object names (from regular expressions and Stage 1) and parameter labels (from dictionary matching and LLM label pre-pass). Once object names and parameter labels are found, the value in the cell that lies at the row/column interception of the labels is extracted. This method is capable of handling common structures such as objects in rows and parameters across columns, the transpose of this, single-object tables without an object name, and object names as row headings with parameters underneath. More generally, any object name that aligns in a row/column with a parameter label can be extracted, making this technique robust to a greater variety of interpreted table formats. Data extracted from this programmatic table code are later merged with the LLM-extracted data, with values from the tabular path favored in conflicting cases.

2.3.5 Stage 3: Validation

Once the main data extraction phases are complete, the results are passed to a series of three LLM calls to validate the previous work using an independent model. Stage 3A validates the recall by double checking that all data from the article that should have been included in the pipeline outputs are present, and adds any missing data (along with source texts). Stage 3B validates the precision by removing or reassigning any data that should not be associated with an extracted object. Stage 3C provides a subjective estimate of the performance in terms of precision/recall/F1 as a 1-10 confidence score. This is estimated according to a set of fifteen counted metrics from the pipeline generation, including what data were added/removed in stages 3A and 3B, names from Stage 1 missing in the generated files, and values without sufficient article evidence. If the confidence score is 6 or lower, one additional loop of stage 3 is performed (this does not repeat further). The confidence scores are used to give operators an estimate of pipeline performance when a human-made truth file is not available for direct score calculation. Future work will develop a more empirical means of deriving a confidence estimate and increase reliability.

2.3.6 Stage 4: Normalization & Load File Creation

All preceding steps dealt with the extraction of data exactly as they appear in the article, but NED and NEA have certain formats and units they require data to be in for ingestion. Stage 4 of the pipeline rewrites the extracted data to the correct format and performs any necessary unit conversions. This includes narrative forms of labels to archive-specific

	N.	Precision		Recall		F1		Runtime (min.)	
		macro	micro	macro	micro	macro	micro	median	max
NED	100	0.6363	0.7793	**0.8930**	**0.8867**	0.6462	0.8295	7.4	108.1
NEA	100	0.6072	0.5100	**0.8265**	**0.8127**	0.6131	0.6267	33.4	90.0

Table 1: Extraction pipeline performance on random samples of NED and NEA articles. Batches of size 100 were chosen to limit maximum runtime. Total NED runtime - 29.1 h. Total NEA runtime - 56.2 h. Development has thus far focused on maximizing recall.

encoding (e.g. 'galaxy' to 'G') and the conversion of certain values in their published unit to their archive-expected unit. First, names extracted by the AI pipeline are matched to archive-canonical forms by querying each archive's web-accessible name resolution service (this is useful for de-duplication and later scoring). Next is a deterministic pre-pass which attempts to establish the unit a value was published in from the extracted source text and convert it to the correct unit programmatically. All extracted values are then added to the fourth stage's LLM prompt along with the unit they are inferred to be presented in and the final required unit. The LLM then uses tool calling to perform deterministic conversions as it deems necessary, based on the article's context. After this final stage, all extracted data are written to a database load file.

2.4 Load File Scoring

Both archives hold repositories of historical database load files which contribute to an evaluation set of 9,758 unique bibcodes for NED and 907 for NEA. These records are used as the ground truths for comparisons against load files generated by the AI-driven extraction pipeline. Precision, recall, and F1 scores are computed to estimate the performance for each individual parameter (including names) and load files as a whole. Scores are calculated by matching object names in the test and truth load files via their queried canonical forms and comparing each AI-extracted value for strict equality with the respective human-extracted value. These are aggregated as the macro-average to assess file-level performance and micro-average to provide a score balanced by the amount of data in each file. Initial development has focused on maximizing the recall over the precision, owing to the relative ease of deleting unneeded data from a generated load file over searching for missed data.

2.5 Web User Interface

We developed a browser-accessible user interface available to NED and NEA curators to work with the data extraction tool. The functionalities of this site include bibcode/article upload for pipeline extraction, truth file upload for scoring, feedback submission, an interactive editor which shows the trace between extracted data and their locations within the original article text, and a chat interface to a Claude Sonnet model to allow for targeted actions (e.g. extract a specific table). The data extraction pipeline runs automatically on the accepted results from the paper relevance classifiers and these outputs are listed on the pipeline webpage daily for operator review.

3 Results

Development on this data extraction pipeline remains active and involves regular benchmarking on larger sets of sampled truth data. We present the results on sets of 100 articles drawn from the NED and NEA truth file repositories following distributions in number of objects, number of parameter types, and number of parameters per object (parameter density) proportional to the truth distributions. Table 1 shows the latest precision, recall, and F1 metrics as well as runtimes for both archives.

3.1 Discussion of Results

Overall performance varies noticeably between NED and NEA. NED articles achieve higher precision and recall, at both macro- and micro-aggregate levels. This is largely due to differences in the types and quantities of data included in load files for each archive. NEA load files often contain more parameters per object (both additional values and their accompanying uncertainties) than the typical NED load file. The latest NED benchmark set contains an average of 6 parameters per object while the NEA articles have an average of 27 parameters per object. Extracting the correct subset of data

while not including extraneous data and assigning the right data to the right objects therefore become more challenging in the NEA case. Differences in the macro- and micro-level aggregates can be attributed to the types of articles included in the samples: articles with more data extracted from them weigh more heavily in the macro-level score than articles with fewer amounts of object data, while the micro-level score aggregates extractions across articles.

3.1.1 File Creation Efficiency

The amount of time it takes a human curator to extract data and write a load file by hand varies according to factors such as the number of objects and parameters in the article (like the AI pipeline), and is typically within fifteen to forty-five minutes. The total runtime of the AI file generation process for a typical article (Table 1) is on a comparable scale to that of a human curator. The AI tool's runtime is also affected by stage 3C-triggered retries, which occurred in 6% of NED articles and 39% in NEA articles, primarily due to low precision confidence. The runtime of the NEA pipeline is further increased by tool calls for unit conversion, which the NED side does not use. While the AI tool writes load files at a similar speed as a human, it is more capable of efficiently processing large batches of articles. The reported runtimes are for processing on a single GPU, while multiple are available for NED and NEA production. In addition, these can run outside of working hours and be made available for a human curators to review later.

3.1.2 Operator Feedback

NED and NEA curators have been actively using the tool in their regular work and providing feedback that has been invaluable in revealing gaps in pipeline capabilities. This includes assistance towards aligning instructions given to the AI with the archives' conventions. While pipeline errors can be caused by parsing errors or convention misalignment, some data points scored in disagreement with the recorded truth may actually represent correct extractions. We observed that the AI tool may extract real data from an article that a human chose not to retrieve (lowering precision) and that the contents of the truth load file may not match the contents of the article because the author made some unreplicable transformation (lowering recall). Instances of this occurring have been noticed anecdotally through feedback but not quantified.

4 Conclusions

Pipeline performance depends heavily on the formatting and verbiage of an article. The articles that perform best are those with small numbers of objects or machine-readable tables. Atypical object name aliases, parameter labels, and table structures can present challenges. The heterogeneity of article format and necessity of careful scientific decisions for data inclusion suggest that AI-driven information retrieval will require a human to remain in the loop for these tasks. However, historical human-made load files and quantitative metrics make improving the pipeline tractable. While there are improvements yet to be made to the extraction performance and the tool's runtime can be significant, it is approaching a level of reliability that will allow for the ready adoption of load files generated autonomously.

4.1 Next Steps

Future development will focus on maximizing pipeline precision and recall while introducing more comprehensive evaluations. We will seek to improve the reliability of article text/table parsing and the stage 3C confidence scores. Larger article sets, performance comparisons between different LLMs, and a characterization the error rates of different components of the tool by running an ablation study will be included in future evaluations. Feedback will continue to be collected and incorporated.

5 AI Usage Disclosure

Almost the entirety of the code used in the implementation of the data extraction pipeline was written by an AI (Claude Opus/Fable) and reviewed by a human developer. None of the text in this article was written by an AI.

6 Acknowledgments

We thank Douglas McElroy, Denis Zaytsev, and Joseph Mazzarella for their important early contributions to this work.

7 Appendix

7.1 Pipeline Diagram

A visualization of the data extraction pipeline is presented in Figure 1 (Appendix).

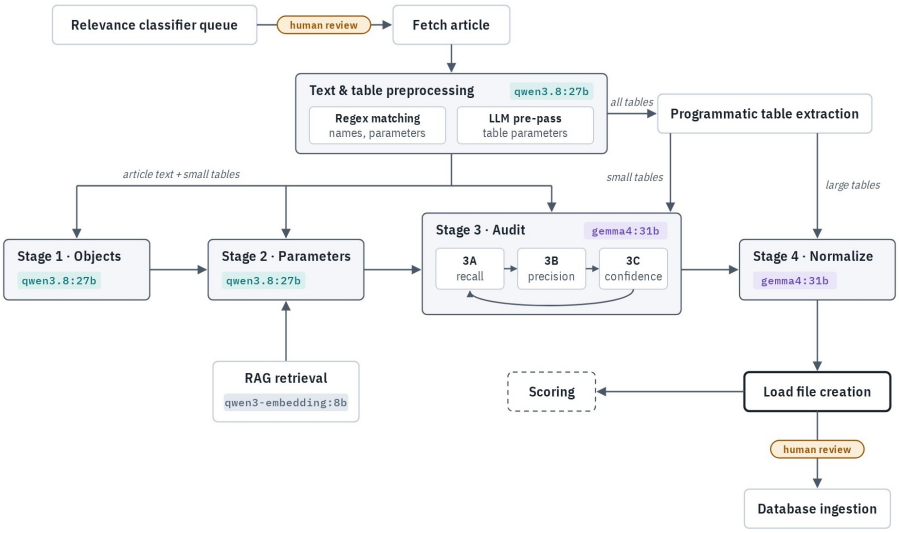


Figure 1: Visualization of pipeline stages. The model stages read the article text and only the small tables. Every table is also parsed programmatically: rows from small tables join the frame before Stage 3 audits it, and rows from large tables join after Stage 3, at Stage 4. Stages 1–3 keep values as printed. Only Stage 4 converts them to archive form. People review the queue before articles are fetched, and each load file before it is ingested.

7.2 User Interface

Figure 2 (Appendix) shows a screenshot of the AI pipeline's web user interface.

7.3 Error Source Analysis

We characterize the sources of errors for each archive according to several categories by counting the false positives and false negatives and examining the differences from truth for each article in the NED and NEA benchmarks (Table 1). These are presented in Table 2 (Appendix). For both archives, precision errors are caused by extra objects and parameters included in the generated file compared to the truth. NED false negatives came more from missing parameters while NEA false negatives were more from missing objects, suggesting object detection is stronger for NED.

	Category	NEA	NED
FP	Objects not in truth	53.5%	62.1%
	Extra parameters or values not in truth	36.1%	21.6%
	Wrong value	10.4%	16.3%
FN	Missed objects	37.2%	19.3%
	Parameter not emitted for a found object	31.0%	41.5%
	Wrong value	31.9%	39.3%

Table 2: Categorized errors for the 100-article NEA and NED benchmarks. Shares are of each archive's false positives (NEA 15,050; NED 3,962) or false negatives (NEA 4,589; NED 1,984).

7.4 Impact of Text/Table Processing on Scores

There is a possibility for errors to be introduced in the pipeline through the imperfect parsing of articles' texts and tables (either from the HTML or PDF source), even though these are performed using reliable packages (See Section 2.1). Conversion text artifacts that have been observed to cause errors include inline math rendered as images and dropped footnotes. Zero NED articles in the latest benchmark had any extraction mistakes due to text parsing errors, but 20% of NEA articles had at least one parsing error. These errors were mostly from periods and Latex formatting dropped from PDF conversions. Examples of table formats that often lead to parsing errors are multi-level headers and object/parameter labels split across multiple rows. In the latest benchmark set, 1% of NED failed to obtain an external table and zero had parsing errors. For NEA, 14% of articles had inaccessible external data and 1% of articles had a PDF parsing error. Parsing errors may be specific to the particular formats used by some publishers, which appear at different rates in NED and NEA holdings. Most NED articles were processed using their HTML while most NEA articles used the PDF due varying publisher access.

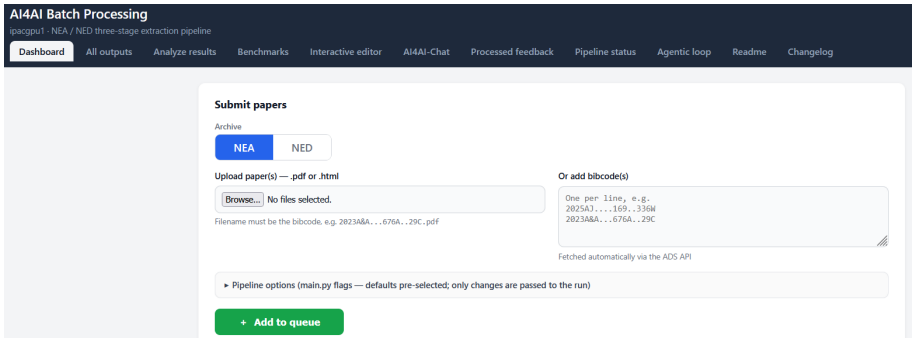


Figure 2: Screenshot the web UI's main page for uploading articles to process. Additional functionalities are in other tabs include viewing results from automated queues, uploading new truth files for scoring, interacting with an AI chat and visual editor interfaces, viewing latest benchmarks, and logging feedback.

References

Artifex Software, Inc. 2026. Pymupdf. Python package for high-performance document processing.

Bishwaranjan Bhattacharjee, Aashka Trivedi, Masayasu Muraoka, Muthukumaran Ramasubramanian, Takuma Udagawa, Iksha Gurung, Nishan Pantha, Rong Zhang, Bharath Dandala, Rahul Ramachandran, Manil Maskey, Kaylin Bugbee, Mike Little, Elizabeth Fancher, Irina Gerasimov, Armin Mehrabian, Lauren Sanders, Sylvain Costes, Sergi Blanco-Cuaresma, and 17 others. 2024. INDUS: Effective and Efficient Language Models for Scientific Applications. *arXiv e-prints*, arXiv:2405.10725.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. *arXiv e-prints*, arXiv:2005.14165.

Tracy X. Chen, Rick Ebert, Joseph M. Mazzarella, Cren Frayer, Scott Terek, Ben H. P. Chan, David Cook, Tak Lo, Marion Schmitz, and Xiuqin Wu. 2022. Classification of Astrophysics Journal Articles with Machine Learning to Identify Data for NED. , 134(1031):014501.

Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370. Association for Computational Linguistics.

Jian Ge, Hui Zhang, Yongshuai Zhang, Yan Li, Dan Zhou, Haijiao Jiang, Pengjun Zhang, Xinyu Yao, Jiapeng Zhu, Yong Yu, Congcong Zhang, Zhenghong Tang, Jianqing Cai, Chaoyan Wang, Hongping Deng, Wen Chen, Kun Chen, Yingquan Yang, Xuliang Duan, and 50 others. 2024. Progress in the Earth 2.0 (ET) space mission. In *Space Telescopes and Instrumentation 2024: Optical, Infrared, and Millimeter Wave*, volume 13092, page 1309218. International Society for Optics and Photonics, SPIE.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 181 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv e-prints*, arXiv:2403.05530.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 181 others. 2026. Gemma 4 Technical Report. *arXiv e-prints*, arXiv:2607.02770.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. , 645(8081):633–638.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv e-prints*, arXiv:2005.11401.

Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. 2025.

Docling: An Efficient Open-Source Toolkit for AI-driven Document Conversion. *arXiv e-prints*, arXiv:2501.17887.

F. Matuszewski, N. Nettelmann, J. Cabrera, A. Börner, and H. Rauer. 2023. Estimating the number of planets that PLATO can detect. , 677:A133.

Matthew T. Penny, B. Scott Gaudi, Eamonn Kerins, Nicholas J. Rattenbury, Shude Mao, Annie C. Robin, and Sebastiano Calchi Novati. 2019. Predictions of the WFIRST Microlensing Survey. I. Bound Planet Detection Rates. , 241(1):3.

Michael Perryman, Joel Hartman, Gáspár Á. Bakos, and Lennart Lindegren. 2014. Astrometric Exoplanet Detection with Gaia. , 797(1):14.

Qwen Team. 2026. Qwen3.8-Max: A new bar for coding and cowork.

Leonard Richardson. 2007. Beautiful Soup Documentation.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *arXiv e-prints*, arXiv:2302.04761.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv e-prints*, arXiv:1706.03762.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv e-prints*, arXiv:2201.11903.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Robert F. Wilson, Thomas Barclay, Brian P. Powell, Joshua Schlieder, Christina Hedges, Benjamin T. Montet, Elisa Quintana, Iain Mcdonald, Matthew T. Penny, Néstor Espinoza, and Eamonn Kerins. 2023. Transiting Exoplanet Yields for the Roman Galactic Bulge Time Domain Survey Predicted from Pixel-level Simulations. , 269(1):5.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv e-prints*, arXiv:2210.03629.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Extracting Published Discussions of 13,616 Galaxies with Language Models

Nolan Koblischke and Jo Bovy
David A. Dunlap Department of Astronomy and Astrophysics
Dunlap Institute for Astronomy & Astrophysics
University of Toronto, 50 St. George Street, Toronto, ON, M5S 3H4, Canada
nolan.koblischke@mail.utoronto.ca

Abstract

Astronomers often have difficulty determining whether an object, such as a star or galaxy of interest, has been discussed in the literature. Existing tools enable object identification and linking in papers but do not record whether an object was genuinely discussed in a paper or just appears in a table or catalog. We address this by using language model agents to identify galaxies discussed in more than two sentences across a corpus of 156,006 arXiv astro-ph papers and extract those discussions. We extract 44,332 discussions of individual galaxies containing 85,314 quotes from 18,641 papers. With sky coordinates for 98.2% of discussions, we group them into 13,616 galaxies. We bring these results together into GalaxyWiki, where users can find literature discussing galaxies, read what was said about them, and follow links to the source papers. The code, dataset, and webpage are available at https://github.com/NolanKoblischke/galaxywiki-paper.

1 Introduction

The NASA/IPAC Extragalactic Database (NED) combines identifiers, physical measurements, and literature information for galaxies and other extragalactic objects (Mazzarella, 2020). Similarly, SIMBAD covers galaxies and stars, providing identifiers, basic data, selected measurements, and bibliographic references (Wenger et al., 2000). NED and SIMBAD also provide catalog-level object classifications like whether a galaxy is a spiral or an elliptical (Kuhn et al., 2022). Here we seek a complementary resource: a collected account of how individual galaxies have been interpreted scientifically, grounded in extended quotations from the literature.

Astronomers searching for previous discussions of a galaxy can find the object bibliographies provided by these databases (Mazzarella, 2020; Wenger et al., 2000). However, an object-to-paper link alone does not distinguish an extended discussion from an appearance in a catalog. Readers may therefore need to inspect the linked papers to determine what has actually been said about the galaxy.

Wikipedia provides this kind of scientific interpretation, although its coverage is limited. We count 3,379 English Wikipedia article pages that use the *galaxy* infobox, a proxy for its coverage of galaxies.[1] However, these pages might not provide a complete account of the literature about a given galaxy. Galaxy Zoo offers another setting for discussion, allowing volunteers to classify galaxies (Lintott et al., 2008) and discuss them on forums that bring together citizen scientists and experts (Mankowski et al., 2011).

To simplify this search, we aim to extract every galaxy discussed in more than two sentences in our corpus of arXiv papers, along with quotes. We bring these discussions together in a database that captures scientific interpretations. Astronomers can then immediately check whether a galaxy has been discussed in the extracted literature, read what was said, and identify the papers containing those discussions, as illustrated in Figure 1.

1.1 Related Work

Recent work has used language models to extract information from scientific papers and organize it into resources that are easier to search and read (Lála et al., 2023; Skarlinski et al., 2024). WikiCrow generates cited articles about human genes from the research literature using PaperQA2 agents that search papers and synthesize their findings (Skarlinski et al., 2024). In astronomy, *Lodestone* uses language models to bring together findings from multiple papers in cited encyclopedia

[1] https://linkcount.toolforge.org/?project=en.wikipedia.org&page=Template%3AInfobox+galaxy

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 174–179
November 10, 2026 ©2026 Association for Computational Linguistics

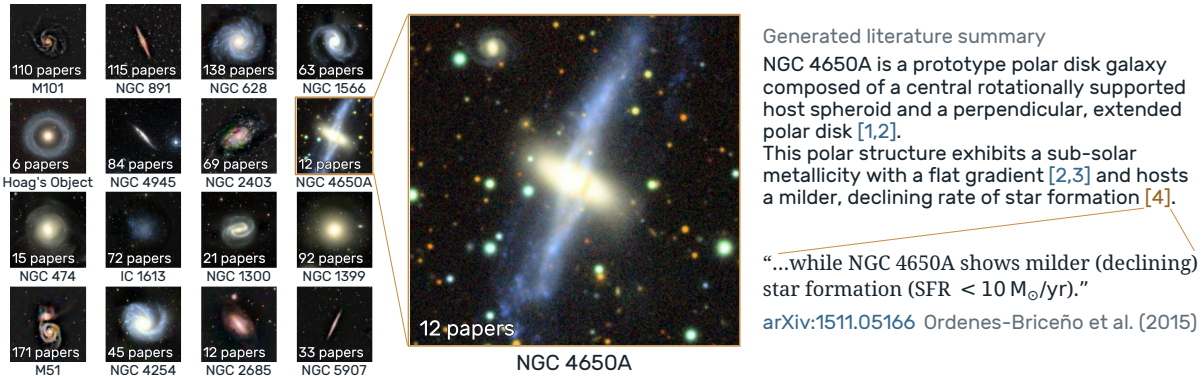


Figure 1: GalaxyWiki brings together published discussions of individual galaxies. *Left:* Selected galaxies and the number of papers discussing each galaxy in our dataset. *Right:* The two-sentence summary of the literature discussions of NGC 4650A (Iodice, 2010, 2014; Spavone et al., 2011; Ordenes-Briceño et al., 2016), along with a citation, extracted quote, and source paper.

articles (Iyer, 2026). Shao et al. (2024) extracts celestial object identifiers and telescope names from journal articles. Wang et al. (2026) constructs a galaxy knowledge graph from literature abstracts, while Li et al. (2026) resolves object mentions to SIMBAD identifiers and link them to scientific concepts. With GalaxyWiki, we focus on the literature discussions of individual galaxies, rather than just their properties or identifiers.

2 Methods

Source corpus We use preprocessed arXiv paper text from the MINT-1T dataset (Awadalla et al., 2024) that is created by parsing the source LaTeX into text (Awadalla et al., 2024). Specifically, we use `minty-astro-ph`,[2] a subset of MINT-1T focused on astro-ph papers. This contains 156,006 astro-ph papers from June 1993 to May 2024. For comparison, we count 299,824 astro-ph postings in the same time period according to the Kaggle arXiv metadata dataset.[3] This incomplete coverage appears largely attributable to the MINT-1T figure-extraction requirements, since the original corpus was designed to pair text with images. In future work, we will attempt to construct a complete astro-ph corpus. As a first processing step, we apply a filter that retains only papers containing the standalone word "galaxy" or "galaxies," leaving 87,899 papers.

Galaxy mention extraction We process each paper with the GPT-5.5[4] language model in the open-source Codex[5] agent harness, and task it with identifying mentioned galaxies and extracting quotes. In the Codex harness, the language model has access to terminal-based tools that allow it to copy quotes verbatim and selectively read the paper. For each galaxy identified, we request one to three quotes totaling three to eight sentences, and ask the language model to generate a 20-to-40-word summary supported by those quotes. We also ask it to choose up to five relevant topics drawn from a fixed subset of the Unified Astronomy Thesaurus (UAT; Frey and Accomazzi, 2018), enabling thesaurus-driven search in our GalaxyWiki. We further task the language model with extracting the galaxy's sky coordinates when explicitly provided in the paper, which is required for associating galaxies across multiple papers. We explicitly instruct the language model to ignore mentions of the Milky Way Galaxy and objects within it and instruct it to extract at most 30 galaxies per paper.

To reduce the setup and take-down overhead of the agents, we assign each agent a batch of 10 papers. We run 20 concurrent agents for approximately 71 hours with token usage corresponding to approximately $12,600 USD at standard API prices.[6] We extract 44,332 discussions of individual galaxies made of 85,314 quotes contained in 18,641 unique papers.

[2] https://huggingface.co/datasets/Smith42/minty-astro-ph

[3] https://www.kaggle.com/datasets/Cornell-University/arxiv

[4] https://developers.openai.com/api/docs/models/gpt-5.5

[5] https://github.com/openai/codex

[6] https://developers.openai.com/api/docs/pricing

Coordinate resolution for grouping discussions Access to the position of the discussed galaxy on the sky allows us to associate discussions across papers and with other catalogs. However, our paper extraction step only captures coordinate information for 3,669 mentions, or 8.3%.

Instead of relying on the coordinates from the extraction step, and to vastly improve coordinate coverage, we use the extracted galaxy identifiers if they follow standard formats such as the NGC, SDSS, and 2MASS identifiers (York et al., 2000; Skrutskie et al., 2006) to match them to an entry on the NED database (Mazzarella and NED Team, 2017). We programmatically query the NED API using galaxy identifiers that match these standard formats. Using this method, we successfully retrieve coordinates for 23,770 out of 25,206 attempted identifier queries, which populates coordinates for 53.6% of mentions.

For the remaining mentions, we task GPT-5.5 agents with finding coordinates by first instructing them to search the original paper, then query the NED API using any compatible galaxy identifier from the paper, not necessarily the formats we attempted. As a fallback, we provide internet access to allow the agents to follow links and references in the paper and retrieve resources such as VizieR catalogs published by the authors (Ochsenbein et al., 2000). We record the internet resources used so that we can audit the coordinate search and we ask the agents to report "unresolved" if they determine the task is impossible.

With agents, we find coordinates for 6,701 records in the source papers, 11,503 through NED, and 1,684 from external internet sources. These results provide coordinates for 19,888 records, or 44.9% of all mentions, leaving 674 records unresolved. Across the programmatic and agentic methods, coordinates are available for 43,546 of 44,332 mentions, or 98.2%.

We then group the 43,546 located galaxy mentions across papers using these sky coordinates. We link positions separated by less than 1" and include neighbors of neighbors. The procedure yields 13,616 entries, each of which we interpret as an individual galaxy discussed in the source papers. However, this relies on the choice of 1" and grouping by sky position does not ensure that these entries correspond to distinct galaxies.

GalaxyWiki In addition to releasing a dataset, we built a web-based wiki intended for astronomers and the general public. For each galaxy page, we generate a two-sentence summary with citations from the available extracted quotes and paper-level summaries using Gemini 3.5 Flash.[7] We programmatically verify that the source quotes and papers used for the summaries are in our dataset. Alongside this summary, each page lists the papers, their summaries, and their quotes, with direct links to the quoted passages where available through the arXiv HTML interface. We allow users to search for galaxies by sky position, UAT topic, name, or text description. To enhance user experience, we retrieve available preview images from the DESI Legacy Imaging Surveys DR10 (Dey et al., 2019), Pan-STARRS (Chambers et al., 2016), and Digitized Sky Survey 2 (McLean et al., 2000) using the coordinates.

3 Results

The extraction yields 44,332 galaxy mention records containing 85,314 quotes from 18,641 papers. The 13,616 grouped entries include 4,187 with multiple mentions and 9,429 (69.2%) with a single mention. Among entries with multiple mentions, the median is three mentions per entry. Every page in the wiki includes a preview image from Legacy Survey DR10 (9,545 pages), Pan-STARRS (3,831 pages), or DSS2 (240 pages).

To assess whether agents accurately copied quotations, we compare each extracted quote with the source paper text. We find character-for-character matches for 66,362 of the 85,314 quotes (77.8%). Normalizing Unicode, spacing, and punctuation yields another 15,971 matches, bringing the total to 82,333 (96.5%). A further 34 match after ignoring case and removing all characters other than letters and numbers. Finally, we find that 2,826 can be split at sentence punctuation into fragments that each match the source after Unicode and whitespace normalization, indicating that agents often combined non-contiguous excerpts into a single quotation. However, textual matching alone does not verify correct attribution or preservation of the source's meaning.

We manually review 120 distinct galaxy mention records. We first review 60 randomly selected records with coordinates, focusing on coordinate verification. We find support for the intended galaxy identification and correct quotations

[7] https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash

in 58 cases. Two records contained quotes that provided little substantive discussion or described a different object. In three otherwise correctly identified records, the galaxy identifiers were truncated or the paper-supplied positional coordinates were rounded, which reduced the precision of the extracted coordinates, with the largest discrepancy reaching 12.2" because an agent decoded a shortened identifier that contained coordinate information. Such errors can separate discussions of the same galaxy into different entries.

We then review another 60 records drawn at random, this time without restricting coordinate availability. During review, all 60 records identify an individual galaxy and have quotes that contain at least three distinct substantive sentences about that galaxy. Two records have passages that only discuss the galaxy jointly with other galaxies, though we determine that these passages still provide useful information for the GalaxyWiki. Four records contain source text in which the MINT-1T parsing pipeline deleted or corrupted LaTeX symbols. This omits details that could be useful to understand the discussed galaxy. Despite these issues, the review supports the conclusion that the pipeline achieves its primary goal, correctly identifying an individual galaxy and extracting its substantive discussion in 118 of 120 cases.

4 Discussion

Our GalaxyWiki is a unique dataset for the field, and has already been used for 1) determining if galaxies discovered by an anomaly detection method have been previously discussed in the literature (Cheng et al., 2026) and 2) enhancing descriptions of galaxy images for training an astronomy-specific vision-language model (Habib et al., 2026).

The GalaxyWiki dataset provides a literature-selected sample of galaxies that have received extended scientific discussion and are therefore of interest to the community. Therefore, one could use the catalog to guide similarity and semantic searches for galaxies with related features (Stein et al., 2021; Koblischke et al., 2026) and support the development of methods for detecting anomalies or identifying galaxies of scientific interest (Lochner and Bassett, 2021; Etsebeth et al., 2024). The database could also support literature surveys in astronomy, complementing semantic retrieval and corpus-wide concept extraction (Iyer et al., 2024; Ting et al., 2025). We can now track which galaxies are discussed and cited most often and how these patterns vary across subfields and years.

The same literature-processing methodology could, in principle, extend beyond galaxies to stars, supernovae, exoplanets, and even theories of dark matter. It could also support similar databases beyond astronomy, drawing on bioRxiv for genes or on ChemRxiv for materials, building on precedents for gene-focused literature synthesis (Skarlinski et al., 2024) and structured extraction of materials information (Dagdelen et al., 2024).

5 Conclusion

We create GalaxyWiki, a database that brings together scientific interpretations of galaxies and discussions of what makes them noteworthy. It allows astronomers and the general public to check whether a galaxy has been discussed in the extracted literature, read those discussions, and continue to the relevant papers. GalaxyWiki illustrates how language models can make previously impractical scientific projects feasible by performing parallelizable and verifiable tasks. Future work will expand the corpus and evaluate mention identification, coordinate resolution, and grouping to build a more complete collection.

Limitations

Even though we can verify that the extracted quotes are in the original papers, we cannot ensure that we have retrieved the mentions of every galaxy in a given paper. It is likely that the language model misses the mentions of some galaxies during the extraction step. In fact, our setup guarantees that some galaxies have been missed since we set a limit of 30 galaxies per paper, which is hit for 39 papers. Notably, in one paper the language model recorded 250 galaxies despite this prompted limit, which we inspect and find minor errors in some extracted sentence counts.

Our coverage is also limited by the source corpus, because we use only the astro-ph subset available through MINT-1T (Awadalla et al., 2024). Papers without an arXiv version are also absent. Expanding the corpus in future work would help produce a more complete dataset.

In our programmatic NED resolution, a pattern intended to capture Abell designations (Abell, 1958) also matched local identifiers like "Galaxy A1", producing false positives. After identifying

the issue, we invalidated 112 of the 156 rows resolved with this pattern.

Acknowledgments

This work was enabled by credits provided through OpenAI's ChatGPT Futures award. We also acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [funding reference number 721697856] and the Data Sciences Institute at the University of Toronto through grant number DSI-DSFY4R1P01. We thank Gabriel Pfaffman and James Garland for valuable discussions and helpful feedback that improved this work. This work used the NASA/IPAC Extragalactic Database and the SIMBAD database.

References

George O. Abell. 1958. The Distribution of Rich Clusters of Galaxies. *The Astrophysical Journal Supplement Series*, 3:211.

Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Kumar Guha, Matt Jordan, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, Ran Xu, Yejin Choi, and Ludwig Schmidt. 2024. MINT-1T: Scaling Open-Source Multimodal Data by 10x: A Multimodal Dataset with One Trillion Tokens. *arXiv e-prints*, arXiv:2406.11271.

K. C. Chambers, E. A. Magnier, N. Metcalfe, H. A. Flewelling, M. E. Huber, C. Z. Waters, L. Denneau, P. W. Draper, D. Farrow, D. P. Finkbeiner, C. Holmberg, J. Koppenhoefer, P. A. Price, A. Rest, R. P. Saglia, E. F. Schlafly, S. J. Smartt, W. Sweeney, R. J. Wainscoat, and 107 others. 2016. The Pan-STARRS1 Surveys. *arXiv e-prints*, arXiv:1612.05560.

Yixuan Cheng, Nolan Koblischke, and Jo Bovy. 2026. Identifying Scientifically Interesting Galaxies with Vision-Language Models. Accepted at the Machine Learning and the Physical Sciences 2026 workshop.

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.

Arjun Dey, David J. Schlegel, Dustin Lang, Robert Blum, Kaylan Burleigh, Xiaohui Fan, Joseph R. Findlay, Doug Finkbeiner, David Herrera, Stéphanie Juneau, Martin Landriau, Michael Levi, Ian McGreer, Aaron Meisner, Adam D. Myers, John Moustakas, Peter Nugent, Anna Patej, Edward F. Schlafly, and 142 others. 2019. Overview of the DESI Legacy Imaging Surveys. *The Astronomical Journal*, 157(5):168.

V. Etsebeth, M. Lochner, M. Walmsley, and M. Grespan. 2024. Astronomaly at scale: searching for anomalies amongst 4 million galaxies. *Monthly Notices of the Royal Astronomical Society*, 529(1):732–747.

Katie Frey and Alberto Accomazzi. 2018. The Unified Astronomy Thesaurus: Semantic Metadata for Astronomy and Astrophysics. *The Astrophysical Journal Supplement Series*, 236(1):24.

Ibrahim Habib, Kevin Li, Trung Nguyen, Gabriel A. Patron, Gaurav Sekar, Nolan Koblischke, Matthieu Le Lain, Shashwat Sourav, Aman Kumar, Kartheik G. Iyer, Joshua S. Speagle, Itai Sfaradi, Michael J. Smith, and Ioana Ciucă. 2026. The Great Chat in the Sky: Aligning Exotic Astronomical Data to Language with AstroBridge. Accepted at the Machine Learning and the Physical Sciences 2026 workshop.

E. Iodice. 2010. Polar Disk Galaxies as New Way to Study Galaxy Formation: the Case of NGC4650A. In *Hunting for the Dark: the Hidden Side of Galaxy Formation*, volume 1240 of *American Institute of Physics Conference Series*, pages 379–382. AIP.

E. Iodice. 2014. Narrow Polar Rings versus Wide Polar Ring/Disk Galaxies. *arXiv e-prints*, arXiv:1402.7325.

Kartheik Iyer. 2026. Lodestone: A living astronomy encyclopedia. `https://lodestone.wiki/`. Accessed 2026-09-20.

Kartheik G. Iyer, Mikaeel Yunus, Charles O'Neill, Christine Ye, Alina Hyk, Kiera McCormick, Ioana Ciucă, John F. Wu, Alberto Accomazzi, Simone Astarita, Rishabh Chakrabarty, Jesse Cranney, Anjalie Field, Tirthankar Ghosal, Michele Ginolfi, Marc Huertas-Company, Maja Jabłońska, Sandor Kruk, Huiling Liu, and 11 others. 2024. pathfinder: A Semantic Framework for Literature Review and Knowledge Discovery in Astronomy. *The Astrophysical Journal Supplement Series*, 275(2):38.

Nolan Koblischke, Liam Parker, Francois Lanusse, Jo Bovy, Irina Espejo, and Shirley Ho. 2026. Semantic Search for 100M+ Galaxy Images Using AI-generated Captions. *The Astrophysical Journal*, 1005(2):170.

L. Kuhn, M. Shubat, and P. Barmby. 2022. Comparing NED and SIMBAD classifications across the contents of nearby galaxies. *Monthly Notices of the Royal Astronomical Society*, 515(1):807–816.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. 2023. PaperQA: Retrieval-Augmented Generative Agent for Scientific Research. *arXiv e-prints*, arXiv:2312.07559.

Jinchu Li, Yuan-Sen Ting, Alberto Accomazzi, Tirthankar Ghosal, and Nesar Ramachandra. 2026. Predicting New Concept-Object Associations in Astronomy by Mining the Literature. *arXiv e-prints*, arXiv:2602.14335.

Chris J. Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alex Szalay, Dan Andreescu, Phil Murray, and Jan Vandenberg. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189.

M. Lochner and B. A. Bassett. 2021. ASTRONOMALY: Personalised active anomaly detection in astronomical data. *Astronomy and Computing*, 36:100481.

T. S. Mankowski, S. J. Slater, and T. F. Slater. 2011. An Interpretive Study of Meanings Citizen Scientists Make When Participating in Galaxy Zoo. In *Earth and Space Science: Making Connections in Education and Public Outreach*, volume 443 of *Astronomical Society of the Pacific Conference Series*, page 150.

J. M. Mazzarella. 2020. Science Discovery With Diverse Multi-Wavelength Datasets Fused in the NASA/IPAC Extragalactic Database (NED). In *Astronomical Data Analysis Software and Systems XXIX*, volume 527 of *Astronomical Society of the Pacific Conference Series*, page 3.

Joseph M. Mazzarella and NED Team. 2017. Evolution of the NASA/IPAC Extragalactic Database (NED) into a Data Mining Discovery Engine. In *Astroinformatics*, volume 325 of *IAU Symposium*, pages 379–384.

B. J. McLean, G. R. Greene, M. G. Lattanzi, and B. Pirenne. 2000. The Status of the Second Generation Digitized Sky Survey and Guide Star Catalog. In *Astronomical Data Analysis Software and Systems IX*, volume 216 of *Astronomical Society of the Pacific Conference Series*, page 145.

F. Ochsenbein, P. Bauer, and J. Marcout. 2000. The VizieR database of astronomical catalogues. *Astronomy and Astrophysics Supplement Series*, 143:23–32.

Yasna Ordenes-Briceño, Iskren Y. Georgiev, Thomas H. Puzia, Paul Goudfrooij, and Magda Arnaboldi. 2016. Compact stellar systems in the polar ring galaxies NGC 4650A and NGC 3808B: Clues to polar disk formation. *Astronomy and Astrophysics*, 585:A156.

Wujun Shao, Rui Zhang, Pengli Ji, Dongwei Fan, Yaohua Hu, Xiaoran Yan, Chenzhou Cui, Yihan Tao, Linying Mi, and Lang Chen. 2024. Astronomical Knowledge Entity Extraction in Astrophysics Journal Articles via Large Language Models. *Research in Astronomy and Astrophysics*, 24(6):065012.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv e-prints*, arXiv:2409.13740.

M. F. Skrutskie, R. M. Cutri, R. Stiening, M. D. Weinberg, S. Schneider, J. M. Carpenter, C. Beichman, R. Capps, T. Chester, J. Elias, J. Huchra, J. Liebert, C. Lonsdale, D. G. Monet, S. Price, P. Seitzer, T. Jarrett, J. D. Kirkpatrick, J. E. Gizis, and 12 others. 2006. The Two Micron All Sky Survey (2MASS). *The Astronomical Journal*, 131(2):1163–1183.

M. Spavone, E. Iodice, M. Arnaboldi, G. Longo, and O. Gerhard. 2011. Chemical abundances of the PRGs UGC 7576 and UGC 9796. I. Testing the formation scenario. *Astronomy and Astrophysics*, 531:A21.

George Stein, Peter Harrington, Jacqueline Blaum, Tomislav Medan, and Zarija Lukic. 2021. Self-supervised similarity search for large scientific datasets. *arXiv e-prints*, arXiv:2110.13151.

Yuan-Sen Ting, Alberto Accomazzi, Tirthankar Ghosal, Tuan Dung Nguyen, Rui Pan, Zechang Sun, and Tijmen de Haan. 2025. AstroMLab 5: Structured Summaries and Concept Extraction for 400,000 Astrophysics Papers. *arXiv e-prints*, arXiv:2511.12353.

Jia-Qi Wang, Jin-Qu Zhang, and Zhi-Qiang Zou. 2026. An Automatic Knowledge Graph Construction Method from Astronomical Literatures via Large Language Models. *Research in Astronomy and Astrophysics*, 26(3):035011.

M. Wenger, F. Ochsenbein, D. Egret, P. Dubois, F. Bonnarel, S. Borde, F. Genova, G. Jasniewicz, S. Laloë, S. Lesteven, and R. Monier. 2000. The SIMBAD astronomical database. The CDS reference database for astronomical objects. *Astronomy and Astrophysics Supplement Series*, 143:9–22.

Donald G. York, J. Adelman, John E. Anderson, Jr., Scott F. Anderson, James Annis, Neta A. Bahcall, J. A. Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, William N. Boroski, Steve Bracker, Charlie Briegel, John W. Briggs, J. Brinkmann, Robert Brunner, Scott Burles, Larry Carey, Michael A. Carr, and 126 others. 2000. The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, 120(3):1579–1587.

Star Light, Star Bright: Can Multimodal Models Parse an Astronomy Visual Question Answering Dataset Right?

J. P. Naiman
School of Information Sciences
University of Illinois, Urbana-Champaign, IL 61820, USA
jnaiman@illinois.edu

Abstract

Figures are essential for scientific communication, and the field of Visual Question Answering (VQA) aims to test the figure understanding capabilities of large multimodal models (LMMs). Here, we investigate the accuracy of several low-cost LMMs on domain-specific VQA tasks for astronomical *images of the sky*. As instrument for this task we build a domain-specific VQA dataset comprising 2001 partially-synthetic figures divided evenly between contour plots, real telescope images drawn from NASA's SkyView, and "fake" sky images which share the axes and angular scales of the real images but draw their colors from a gaussian mixture model, alongside 47357 questions. Each figure is released as pristine and with realistic page-aging degradations mimicking the historical astronomical corpus. Testing three low-cost proprietary LMMs, we find models frequently fail to answer domain-specific questions at high accuracy. An initial comparison with 104 human-verified, astrometrically tagged figures from historical journal publications shows similar median errors in angular extent. All code is available on GitHub: https://github.com/jnaiman/SkyImagesWASP2026 and dataset is available on HuggingFace: https://huggingface.co/datasets/ReadingTimeMachine/astro_sky_image_vqa.

1 Introduction

In academic articles, figures synthesize and interpret scientific information. Their centrality has made figure parsing a main task of pipelines that digitize scientific documents (e.g., document layout analysis, Binmakhashen and Mahmoud, 2019; Ding et al., 2026; Kumar and Lehal, 2025), alongside figure classification (Jobin et al., 2019) and dataset extraction (Davila et al., 2019). With the growing prevalence of large multimodal models (LMMs) in scientific practice, much recent work tests whether these models can parse figures and "understand" their meaning (Hoque et al., 2022; Farahani et al., 2023) – commonly through Visual Question Answering (VQA), in which a model is given an image and questions about the image. Initially focused on natural images (e.g., Kafle and Kanan, 2017), chart-based VQA is now an expanding field (e.g., Huang et al., 2024). Early studies covered simple bar, line and scatter charts (Kahou et al., 2018), while follow-up work added more complex chart types (e.g., funnel, Han et al., 2023) and plots whose relationship to the underlying data is indirect, such as histograms and contour plots (Naiman et al., 2026; Evans et al., 2026). Absent are domain-specific VQA tasks, which are necessary to probe realistic scientific LMM applications (Chang et al., 2022; Huang et al., 2024).

Astronomy relies on *images of the sky*, a figure type that would benefit from domain-specific VQA. Sky images share some complexity with contour plots, mapping light intensity in a wavelength range to a color rather than a spatial scale (Evans et al., 2026), but encode position in right-ascension (RA) and declination (DEC) – a non-Cartesian system that can deviate substantially from a rectangular grid over large sky areas. Historical sky figures (published before ∼1997) add a further complication: many are grayscale rather than color, and scanning for upload to digital archives introduces errors of its own (Grant et al., 2000; Baird, 2007).

To the author's knowledge, there are no VQA datasets centered on images of the sky, for either present-day or historical figures. However, benchmarks focused on astronomy have recently been developed for other AI and machine learning (ML) tasks. AstroVisBench (Joseph et al., 2025) evaluates whether models can write the code that produces astronomical visualizations, and AstroMMBench (Shi et al., 2025) evaluates domain-specific reasoning over curated astronomical imagery via multiple-choice questions. However, neither mea-

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 180–191
November 10, 2026 ©2026 Association for Computational Linguistics

sures LMM free-form answers requiring calculations with a known ground truth.

Our research question is: *can low-cost LMMs perform domain-specific parsing of astronomical sky figures – including distinguishing real and synthetic sky sources – and does this degrade for historical-style figures?* To answer it, we present a VQA dataset focused on images of the sky in both present-day and historical contexts, alongside two visually similar plot types: contour plots drawn from several two-dimensional distributions, and "fake" sky images that keep their RA/DEC axes but draw their colors from a gaussian mixture model (GMM) rather than from a telescope. The data is used to test low-cost proprietary LMMs, comparing accuracy between original ("pristine") figures and those modified by realistic aging processes. Code, images, question sets and LMM answers are released for use by the community.

2 Dataset Construction

2.1 Figure Generation

Single-panel figures are generated following the work of Naiman et al. (2026); Evans et al. (2026) and are rendered with matplotlib (Hunter, 2007). Style, colormap, DPI, aspect ratio, font and font size are randomly selected; titles and axis labels are drawn from a word pool mined from the arχiv, with inline mathematics substituted at a fixed rate (Naiman et al., 2023). Alongside each image the underlying data array, sampled generation parameters, and pixel-coordinate bounding boxes for the plot area, axes, tick labels, title, axis labels, colorbar and colorbar label are saved. Figures are rejected and re-sampled if labels run off the canvas, if bounding boxes overlap beyond a tolerance, or if the plot area falls below a fraction of the figure.

Figure types are sampled evenly between contour plots with colors from various distributions (random, linear, gaussian mixture model) (Evans et al., 2026), images of the sky taken from telescopes, and images of the sky created by a gaussian mixture model. The GMM-sky images serve as a control: real-sky images do not provide figures with realistic axes, angular scale, rendering and color-scale distribution without a telescope signal. Comparing answers between the two families isolates whether errors stem from the domain-specific RA/DEC axis conventions (shared by both) or from the image source (which differs).

For both types of images of the sky, axes are stored as world coordinate system (WCS) with RA/DEC tick labels in several styles and a sampled epoch (e.g., J2000, B1950). A (target, wavelength) pair draws from a set of objects from the astronomical literature (targets described in Naiman 2026), and retrieved from sky surveys within NASA's SkyView (McGlynn et al., 1996) through astroquery (Ginsburg et al., 2019), with WCS handling via astropy (Astropy Collaboration, 2022).

The physical size of an image centered on a target on the sky can change across surveys and applications. To mimic a realistic scale of a sky image shown in an article, when SIMBAD (Wenger et al., 2000) sky-size for images (angular field of view, FOV) is available for a target, we use this to set the range of RA/DEC from which to extract an image. When absent, we make use of the FOV and angular resolution information stored in SkyView for each sky survey and select RA/DEC ranges from 50 times the smallest resolution up to 0.75 times the FOV. GMM-generated sky images follow the same distribution of angular sizes as the real sky images. Real-sky cutouts are cached at 300×300 and square. To create a set of realistic aspect ratios (and to match with contour plot parameters), the grid is sampled (n_x from 50–300, $n_y = n_x/\text{aspect}$). The index map between initial and rescaled pixels is exactly affine, such that arcsec/pixel stays isotropic and the center is preserved. A logarithmic stretch with clipping at the 99th percentile and its inversion is applied to both GMM- and real-sky images – the net effect removes extreme outliers while leaving the bulk of the distribution on its original scale.

Units are not included in colorbars. Only $\sim$20% of retrieved real-sky images declare a form of physical units (e.g., BUNIT – Jy/beam, MJy/sr, K, Rayleighs), and typically several data processing steps are applied between raw sky image data and published real-sky figures (e.g., Starck and Murtagh, 2006). A fuller pipeline of image processing (e.g., Joseph et al., 2025) to more closely align with the domain science (and thus, physical colorbar units) is left as future work.

Our synthetically generated figures trade the visual realism of a human-collected and annotated figure dataset for the detail of access to underlying data and the precision of bounding box locations (Methani et al., 2020; Wang et al., 2024; Masry et al., 2025; Naiman et al., 2026). Our inclusion of telescope-derived images of the sky with realistic angular scales acts as a compromise between a fully-synthetic and a hand-annotated figure dataset.

2.1.1 Page Aging of Figures

The figure generation pipeline described previously produces pristine scientific charts; however, much of the astronomical literature is not "born digital", but instead consists of human-scanned copies of originally paper-only journal articles. Given that the storage and scanning procedures for historical images can degrade the uploaded images, we also build a copy of each figure with several realistic "page aging" image augmentations. We make use of the Augraphy (Groleau et al., 2023) package, which models damage as accumulating in three phases – *ink* (e.g., ink bleed), *paper* (e.g., color, texture) and *post* (e.g., stains, folding). Our application pulls randomly from this set of effects following Guan et al. (2025) to mimic a modified form of the "archive" set of effects from s1mb1o (2025), which is meant to align with expected degradations resulting from scanning historical documents. As many of the historical articles in the field of astronomy were scanned at high resolution by trained archivists (Grant et al., 2000), the uploaded documents retain a lower level of page augmentations – denoted as "archive-light" in this paper. Finally, we transform 2/3 of the images to grayscale to mimic the color palette of historical articles. Not all of the randomly selected original colormaps are suitable to transform to grayscale which adds a further source of error in the aged figures.

An example of a real-sky image (left panel) and aged counterpart (right panel) is shown in Figure 1.

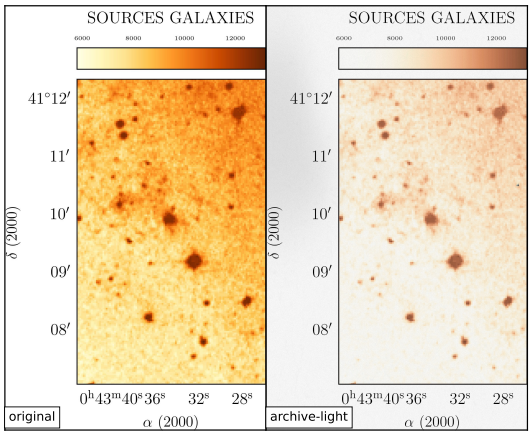


Figure 1: A dataset figure before (left) and after (right) archive-light aging.

2.2 Question Generation

Questions are generated for each figure using the four-level complexity model of Lundgard and Satyanarayan (2022), as adapted to chart-based VQA in other works (Tang et al., 2023; Naiman et al., 2026; Evans et al., 2026). We generate questions at Levels 1–3, spanning what is directly readable off the figure, what requires reading values, and what requires inference beyond a single statistic. Level 4 requires external context or domain expertise and is out of scope for this work. Example questions are listed in Table 1; Appendix B gives further detail and the accessibility motivation for this question scaffolding.

3 Results and Analysis

From our initial dataset of 2001 images, evenly spread between contours, real-sky images, and GMM-sky images and a total of 47357 questions, we randomly select 255 images (85 in each category) for evaluation with three proprietary LMMs, yielding 6035 questions per model. We use each in a low-cost configuration: GPT-5.4-nano (OpenAI, 2026) at default reasoning and verbosity (GPT), Gemini-3.5-flash-lite (Google DeepMind, 2026) with thinking_level = minimal (Gemini), and Claude-Haiku-4.5 (Anthropic, 2025) with thinking disabled (Claude). We restrict this preliminary run to the low-cost tier, where Naiman et al. (2026) found accuracy varied most. These are each provider's lower-cost reasoning setting and are not strictly equivalent across providers. The models selection and question scale assumes a cost-sensitive archiving institution, while holding the 95% Wilson interval (Kahouadji, 2025) on a single accuracy to at most ±10.4 percentage points (pp) with 85 figures per category, and ±7.4 pp for the 170 pooled sky images – enough to resolve the large differences discussed below, but not fine-grained model rankings. Questions are issued one per API call, so each answer is an isolated zero-shot observation. Each model is run over two conditions — the original figures and archive-light. Answers are requested as JSON with entries for raw answer and explanation. Only 48 of the 36210 answers (0.13%) fall outside this pair and are unparsable at scale – 21 that are not valid JSON and 27 explicit `null` refusals; per-model counts are given in Appendix A. Here, we report question types that probe sky-image features – plot type, provenance and angular scale in RA/DEC.

Plot type and Provenance. Figure 2 shows the accuracy of answers to "What is the plot type of the plot in this figure? Please choose the plot type from the following list: [contour, image of the sky]." and,

for sky images, "What is the underlying distribution used to create the data in this figure? Please choose the distribution from the following list: [gaussian mixture model, real image of the sky]."

These questions test two sources of possible confusion in the models – the presence of an image-based figure (contour) and the presence of the specific x/y labels of RA/DEC (GMM-sky). The tests are motivated by work showing scholarly documents can be a carrier for problematic adversarial content aimed at AI-assisted pipelines (Lin, 2026; Nagaraja et al., 2026). GMM-sky images are sampled to the same pixel, angular and color-scale distributions as the real ones, with parameters chosen to span the range from point sources to diffuse emission. We did not pursue statistical indistinguishability as the aim was to remove obvious artifacts while preserving the structural differences the task is meant to probe. A gradient-boosted classifier on low-level image features (e.g., field size) separates real from GMM-sky images with 85% accuracy (Appendix D). The two families are thus distinguishable from image content, so low LMM accuracy reflects model capability rather than an impossible task.

The accuracy shown in Figure 2 is measured as a simple binary (yes/no) for both original and archive-light historical figures. The left, plot type panels of Figure 2 show that when confronted with a contour plot Claude does worse than guessing (15% original, 22% archive-light) and GPT only modestly better (66% and 56%), except Gemini which shows ≈80% accuracy for both original and archive-light figures. Gemini and Claude recover sky-image plot types almost always, but this is to be expected from the over-guessing of this type of figure evident in the contour plot panels; GPT, by contrast, labels only 53% of original sky images correctly. For sky image source, the models clearly favor guessing "real image of the sky" and frequently miss when the image source is a GMM distribution. Only Claude returned unparsable answers here (Appendix A). Across models and conditions, GMM-sky images are identified correctly only 4–14% of the time, against 80–98% for real-sky images. With $n = 85$ per cell, differences of $\lesssim$10 pp lie within the 95% interval of a single cell, but the largest gaps do not: for example, Claude's 15% versus Gemini's 81% on contour plots gives $z = 8.6$ ($p < 10^{-16}$, simple two-proportion test).

Our dataset includes open-ended (no lists provided) versions of several questions. For example,

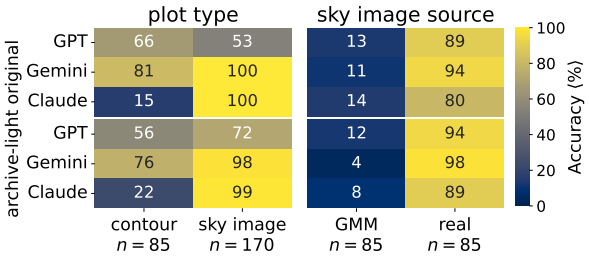


Figure 2: Accuracy of identifying plot type (left) and real vs generated sky (right) for the three low-cost models. Models are offered explicit choices for each.

in the plot type question raw accuracy is near zero for all three models as they answer with strings such as "astronomical survey image" that describe the figure correctly and never match the ground-truth label. Further parsing of open-ended answers is left for future work.

Angular height. Figure 3 shows the accuracy of answers to the question "What is the angular height of the sky region shown in this figure, measured along the declination axis? Give the value in arcminutes." This requires reading the y-axis of the figure and translating the domain-specific measure of angular size (see Figure 1) into another typical unit of angle in astronomy (arcmin, 1/60°).

The accuracy is measured by the log-ratio, $\mathcal{R} = \log_{10}(\hat{y}/y)$, defined for $y > 0$ and $\hat{y} > 0$ and scales the magnitude of a prediction similarly for under or over estimates. The metric cannot be used when LMM and ground truth (GT) answers have different signs which occurs in our dataset for a single GPT, archive-light answer.

Questions about angular scale probe how much detail is expected (angular resolution) and what other targets fall within the field (angular height/width). While Gemini does the best overall (peaked around $\mathcal{R} \approx 0$ in Figure 3), the other models have larger spreads tending toward overestimation with GPT having 53% (60%) of its responses at $> +1$ dex and Claude having 32% (32%) for pristine (arch-light/archive-light) figures. Initial parsing of explanations returned by the models suggests the slight overestimations ($0.5 < \mathcal{R} < 1$) stem from imprecise calculation of the y-axis ranges and rounding of answers. Larger errors ($\mathcal{R} > 1$) often result from misreading of the y-axis tick marks due to domain-specific norms (e.g., the bottom y-tick in Figure 1 is read as $08°00'$ instead of $08'$). Splitting Figure 3 by source shows the GMM-sky images are potentially easier for this axis-reading question: on

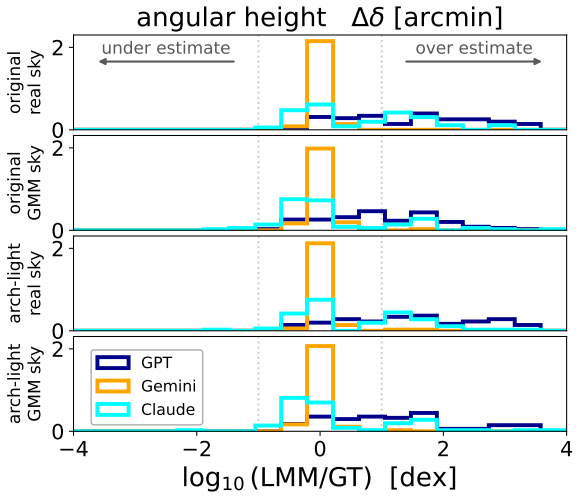


Figure 3: Error distribution of angular height for GMM- and real-sky images measured as log-ratio, $\mathcal{R}$. Zero indicates agreement with ground truth, ± 1 indicates predictions too large or too small by a factor of ten.

the original figures, Claude overestimates by more than a factor of ten significantly more often on real-sky than on GMM-sky figures (41% vs. 24%, $p = 0.01$), and GPT shows the same direction, though not significantly (59% vs. 47%, $p = 0.12$).

Published figures. To test whether results on our figures carry over to published ones, we collect sky images with human-verified astrometry from the AAS Astronomy Image Explorer (AIE) (American Astronomical Society and IOP Publishing, 2014; Naiman, 2020), which indexes figures from the Astrophysical and Astronomical Journals and includes Astronomy Visualization Metadata (AVM) in the WorldWide Telescope convention (Rosenfield et al., 2018). The AVM stores central RA/DEC and pixel scale, but does not include figure bounding box. Following an initial pass in OpenCV for potential image corners (Bradski, 2000; Suzuki and Abe, 1985; Douglas and Peucker, 1973) bounding boxes and figure canvas spans were verified by hand using Label Studio (Tkachenko et al., 2020). Along with automated checks of AVM to SIMBAD's position of the object named in its publication's title (if available) and checks for coordinate duplication or nonphysical field sizes, human verification reduced the AVM-tagged set of figures from 163 to 104.

As AVM tags exclude information for other figure elements (e.g., colorbar), questions are limited to angular scale and passed to our best-performing (Gemini) model. Figure 4 compares angular-width errors (change in RA de-projected from spherical coordinates, see Appendix C). Answers (103, 1 unparsable) are similar across synthetic (real-sky) and published figures: median $\mathcal{R} = -0.08$ dex against -0.05 dex for archive-light, with 55% vs. 51% of answers within a factor of two.

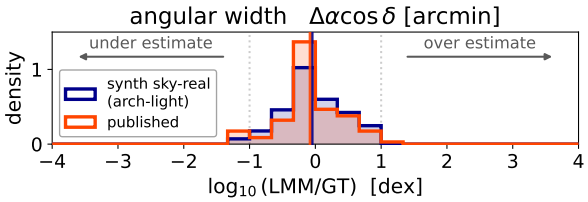


Figure 4: Log-ratio error, $\mathcal{R}$, of Gemini's angular-width answers for real-sky figures (archive-light, blue; $n = 85$) and for 103 human-verified published figures from the Astronomy Image Explorer (orange).

4 Conclusions and Future Work

We have introduced a domain-specific chart VQA dataset for astronomical figure types: contour plots and images of the sky. Sky panels are drawn both from real telescope observations and from GMMs, generated for equal numbers in contour, real- and GMM-sky image categories. We include both pristine images and "aged" images to mimic those in the historical astrophysics corpus.

Across three proprietary low-cost LMMs we find figure-type identification often fails to distinguish between sky and contour plots and, for sky images, between real and GMM images. Models tend to overestimate the angular heights of sky regions.

On a small set of human-verified published figures, Gemini's angular-width answers match its median behavior on our partially synthetic figures. We detect no difference in accuracy between pristine and aged images apart from one cell (GPT labels aged sky images correctly more often – 53% vs. 72%, $p < 0.001$). At $n = 85$ per model/category we can exclude only aging effects larger than $\sim$21 pp at 80% power. Similar questions asked with less domain-specific charts do not display such large inaccuracies (e.g., Naiman et al., 2026). These results motivate the need for the development of further domain-specific VQA datasets (Chang et al., 2022; Huang et al., 2024; Naiman et al., 2026).

Future work could focus on a more domain-accurate image processing pipeline, target other types of astronomy-specific figures (e.g., spectra), and investigate the explanations and open-ended answers stored as part of the LMM responses in our dataset which we did not explore here.

Limitations

The figures are generated, not collected. The arrays inside the real sky panels are genuine survey cutouts, but the figure drawn around them is not a published figure, and lacks conventions real ones carry — domain-standard colormaps, compass roses, beam ellipses, overlaid catalog positions. Our published-figure comparison suggests angular extent transfers in the median, but it covers only one model and two questions (angular width, and angular height in Appendix C). The 104 AVM-tagged figures are also a small, non-random sample of the sky images in the historical literature – those an archive happened to tag with astrometry (Naiman, 2020) – so the comparison should be read as an example of how results on our figures carry over, not as a definitive measure of LMM accuracy on published sky figures.

Matching the synthetic sky fields to the real ones removes the obvious cues to provenance, but does not make the two families indistinguishable and was not intended to. The provenance question we pose is therefore hard rather than impossible, and the accuracies we report should be read against a task that remains somewhat solvable in principle from image content alone (see Appendix D).

We evaluate only the lower-cost configuration of three proprietary models, on a single aging preset and these configurations are not strictly equivalent in reasoning effort across providers. Open-weight models, higher-reasoning configurations and astronomy-tuned models are untested, and model versions and prices change quickly enough that these particular numbers have a short shelf life.

Each condition is a single run over 255 of the 2001 figures. Per-cell uncertainties are up to ± 10.4 pp, so small differences between models, or between pristine and aged figures, should not be overinterpreted.

Our published-figure comparison is small (104 human-verified figures, one model) and limited to the geometric questions that astrometric metadata can ground; colormap, provenance and value questions have no ground truth for figures we did not draw. Its angular extents rely on human-verified plot boxes, and pixel scale cannot be compared at all, because the published astrometry counts scan pixels rather than data pixels. We do not include a human baseline on the real-versus-GMM question, which would require a hand-annotated corpus beyond the scope of this short paper.

The dataset is designed as an evaluation benchmark. We do not test the use of the synthetic images for training or data augmentation, though it is not outside the realm of possibility that this dataset could be used as training or pre-training for a variety of models.

Scoring is exact match after normalization for categorical answers and a log-ratio for numeric ones. As the open-ended plot-type results show in our preliminary investigation, exact match understates a model that is substantively right in unanticipated vocabulary. The questions themselves are generated from templates over known parameters: correct by construction, but not always a sample of what a working astronomer would ask.

Ethics Statement

The sky cutouts used here are public data retrieved from NASA's SkyView service and are redistributed within the dataset in derived form, with full provenance metadata. The published figures used for comparison may remain under their journals' copyright; we release only their metadata, questions and model answers, with a script that retrieves the images from the Astronomy Image Explorer. The synthetic sky panels in this dataset are visually plausible but are not observations, and are labeled as such in every record; we note this explicitly because the same matching procedure that makes them useful as a control would make unlabeled synthetic sky images misleading if separated from their metadata.

More broadly, image falsification is a recognized form of scientific misconduct[1] (a growing issue along with adversarial image and text content injection (Lin, 2026; Nagaraja et al., 2026)), and our synthetic sky images are, by design, plausible enough that the models we test usually label them as real. The procedure that makes them a useful control could be misused to fabricate observations in a manuscript, and our results suggest that low-cost LMMs would not reliably flag them. We mitigate this by labeling every synthetic image in its record and releasing the generation code and parameters so that any image can be traced, and we encourage use of the dataset to develop provenance detectors rather than as a source of figures, though we cannot limit the codebase's use toward generation of synthetic figures passed off as authentic.

[1] https://bps.stanford.edu/home/instances-scientific-misconduct/falsification-images

AI was used to port code from prior work to the current repository and help to provide documentation. AI was also used to write analysis and verification code (e.g., the classifier baseline, statistical checks, and table and cost generation, partially adapted from author-written code), all of which was reviewed by the author. It was additionally used to assist in the literature review (all sources verified by humans), with a draft of the methods section, appendices, and the abstract, and with the writing of the Limitations and Ethics Statement section. It was also used to help shorten the paper to fit page requirements and to help draft revisions responding to reviewer comments; all revised text was checked by the author. AI was trialed for other parts of the writing, but was not found useful in other sections so was abandoned for the main section of the paper.

References

American Astronomical Society and IOP Publishing. 2014. Astronomy image explorer. https://www.astroexplorer.org/. Launched 2014; accessed 2026-10-08.

Anthropic. 2025. Claude Haiku 4.5. https://www.anthropic.com/claude/haiku. Large language model. Released October 2025. API model identifier: `claude-haiku-4-5`.

Astropy Collaboration. 2022. The astropy project: Sustaining and growing a community-oriented open-source project and the latest major release (v5.0) of the core package. *The Astrophysical Journal*, 935(2):167.

Henry S. Baird. 2007. The state of the art of document image degradation modelling. In Bidyut B. Chaudhuri, editor, *Digital Document Processing: Major Directions and Recent Advances*, Advances in Pattern Recognition, pages 261–279. Springer, London.

Galal M. Binmakhashen and Sabri A. Mahmoud. 2019. Document layout analysis: A comprehensive survey. *ACM Comput. Surv.*, 52(6).

Gary Bradski. 2000. The OpenCV library. *Dr. Dobb's Journal of Software Tools*.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. 2022. Mapqa: A dataset for question answering on choropleth maps. In *NeurIPS 2022 First Table Representation Workshop*.

Kenny Davila, Bhargava Urala Kota, Srirangaraj Setlur, Venu Govindaraju, Christopher Tensmeyer, Sumit Shekhar, and Ritwick Chaudhry. 2019. ICDAR 2019 competition on harvesting raw tables from infographics (CHART-Infographics). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1594–1599, Sydney, Australia. IEEE.

Yihao Ding, Soyeon Caren Han, Jean Lee, and Eduard Hovy. 2026. Deep learning based visually rich document content understanding: a survey. *Artificial Intelligence Review*, 59(4):114. Open access.

David H. Douglas and Thomas K. Peucker. 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica*, 10(2):112–122.

Daniel John Evans, Jill P. Naiman, JooYoung Seo, and Stephen J. Downie. 2026. Do you "see" what I "see"? A Multi-panel Visual Question Answer Dataset for Large Language Model Chart Analysis. In *The 30th International Conference on Theory and Practice of Digital Libraries*, The 30th International Conference on Theory and Practice of Digital Libraries.

Ali Mazraeh Farahani, Peyman Adibi, Mohammad Saeed Ehsani, Hans-Peter Hutter, and Alireza Darvishy. 2023. Automatic chart understanding: A review. *IEEE Access*, 11:76202–76221.

Adam Ginsburg, Brigitta M. Sipőcz, C. E. Brasseur, Philip S. Cowperthwaite, Matthew W. Craig, Christoph Deil, James Guillochon, Giannina Guzman, Simon Liedtke, Pey Lian Lim, Kelly E. Lockhart, Michael Mommert, Brett M. Morris, Henrik Norman, Madhura Parikh, Magnus V. Persson, Thomas P. Robitaille, Juan-Carlos Segovia, Leo P. Singer, and 4 others. 2019. astroquery: An astronomical web-querying package in python. *The Astronomical Journal*, 157(3):98.

Google DeepMind. 2026. Gemini 3.1 Pro. https://deepmind.google/models/model-cards/gemini-3-1-pro/. Large language model. Released February 19, 2026. API model identifier: `gemini-3.1-pro`.

Carolyn S. Grant, Alberto Accomazzi, Guenther Eichhorn, Michael J. Kurtz, and Stephen S. Murray. 2000. The NASA Astrophysics Data System: Data holdings. *Astronomy and Astrophysics Supplement Series*, 143(1):111–135.

Alexander Groleau, Kok Wei Chee, Stefan Larson, Samay Maini, and Jonathan Boarman. 2023. Augraphy: A data augmentation library for document images. In *Document Analysis and Recognition – ICDAR 2023*, volume 14191 of *Lecture Notes in Computer Science*, pages 384–401. Springer Nature Switzerland.

Shuhao Guan, Moule Lin, Cheng Xu, Xinyi Liu, Jinman Zhao, Jiexin Fan, Qi Xu, and Derek Greene. 2025. PreP-OCR: A complete pipeline for document image restoration and enhanced OCR accuracy. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*.

Enamul Hoque, Parsa Kavehzadeh, and Ahmed Masry. 2022. Chart question answering: State of the art and future directions. *Computer Graphics Forum*, 41(3):555–572. EuroVis 2022 State-of-the-Art Report.

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. *arXiv preprint*. ArXiv:2403.12027 [cs] version: 2.

John D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

K. V. Jobin, Ajoy Mondal, and C. V. Jawahar. 2019. DocFigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 1, pages 74–79. IEEE.

Sebastian Antony Joseph, Syed Murtaza Husain, Stella S. R. Offner, Stéphanie Juneau, Paul Torrey, Adam S. Bolton, Juan P. Farias, Niall Gaffney, Greg Durrett, and Junyi Jessy Li. 2025. AstroVisBench: A code benchmark for scientific computing and visualization in astronomy. *Preprint*, arXiv:2505.20538. NeurIPS 2025 Datasets and Benchmarks Track.

Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20. Language in Vision.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. 2018. Figureqa: An annotated figure dataset for visual reasoning. *Preprint*, arXiv:1710.07300.

Nabil Kahouadji. 2025. A comprehensive comparison of the wald, wilson, and adjusted wilson confidence intervals for proportions. *arXiv preprint arXiv:2508.10223*.

Jiho Kim, Arjun Srinivasan, Nam Wook Kim, and Yea-Seul Kim. 2023. Exploring Chart Question Answering for Blind and Low Vision Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–15, New York, NY, USA. Association for Computing Machinery.

Atul Kumar and Gurpreet Singh Lehal. 2025. Exploring document layout analysis: A review. In *Innovative Computing and Communications: Proceedings of ICICC 2025*, volume 1438 of *Lecture Notes in Networks and Systems*, pages 367–383, Singapore. Springer.

Zhicheng Lin. 2026. Hidden prompts in manuscripts exploit AI-assisted peer review. *Communications of the ACM*, 69(7):53–56.

Alan Lundgard and Arvind Satyanarayan. 2022. Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content. *IEEE Transactions on Visualization & Computer Graphics (Proc. IEEE VIS)*.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2025. ChartQAPro: A More Diverse and Challenging Benchmark for Chart Question Answering. *arXiv preprint*. ArXiv:2504.05506 [cs] version: 2.

T. McGlynn, K. Scollick, and N. White. 1996. SkyView: The multi-wavelength sky on the internet. In *New Horizons from Multi-Wavelength Sky Surveys*, number 179 in IAU Symposium, page 465. Kluwer Academic Publishers.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. PlotQA: Reasoning over Scientific Plots. pages 1527–1536.

Neha Nagaraja, Lan Zhang, Zhilong Wang, Bo Zhang, and Pawan Patil. 2026. Image-based prompt injection: Hijacking multimodal LLMs through visually embedded adversarial instructions. In *3rd International Conference on Foundation and Large Language Models (FLLM)*, Vienna, Austria.

J. P. Naiman, Morgan G. Cosillo, Peter K. G. Williams, and Alyssa Goodman. 2023. Large synthetic data from the arχiv for ocr post correction of historic scientific articles. In *Linking Theory and Practice of Digital Libraries: 27th International Conference on Theory and Practice of Digital Libraries, TPDL 2023, Zadar, Croatia, September 26–29, 2023, Proceedings*, page 265–274, Berlin, Heidelberg. Springer-Verlag.

Jill Naiman. 2020. The Reading Time Machine: Transforming Astrophysical Literature into Actionable Data. NASA Proposal id.20-ADAP20-225.

Jill P. Naiman. 2026. Astronomical targets from the literature for sky-image figure generation. Submitted to ACM Transactions on AI for Science.

Jill P. Naiman, Daniel J. Evans, and JooYoung Seo. 2026. What Lies Beneath: A Call for Distribution-based Visual Question & Answer Datasets. *arXiv preprint*. ArXiv:2601.22218 [cs].

OpenAI. 2026. ChatGPT (GPT-5.4). `https://openai.com/index/introducing-gpt-5-4/`. Large language model. Released March 5, 2026. API model identifier: `gpt-5.4`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Philip Rosenfield, Jonathan Fay, Ronald K. Gilchrist, Chenzhou Cui, A. David Weigel, Thomas Robitaille, Oderah Justin Otor, Alyssa Goodman, et al. 2018. AAS WorldWide Telescope: A seamless, cross-platform data visualization engine for astronomy research, education, and democratizing data. *The Astrophysical Journal Supplement Series*, 236(1):22.

s1mb1o. 2025. scan-simulator: A synthetic degradation pipeline for scans, photos and aged documents. `https://github.com/s1mb1o/scan-simulator`. Software, MIT license.

JooYoung Seo, Yilin Xia, Bongshin Lee, Sean McCurry, and Yu Jun Yam. 2024. MAIDR: Making Statistical Visualizations Accessible with Multimodal Data Representation. ArXiv:2403.00717 [cs].

Jinghang Shi, Xiaoyu Tang, Yang Huang, Yuyang Li, Xiao Kong, Yanxia Zhang, and Caizhan Yue. 2025. AstroMMBench: A benchmark for evaluating multimodal large language models capabilities in astronomy. *Preprint*, arXiv:2510.00063.

Jean-Luc Starck and Fionn Murtagh. 2006. *Astronomical Image and Data Analysis*, 2 edition. Astronomy and Astrophysics Library. Springer, Berlin, Heidelberg.

Satoshi Suzuki and Keiichi Abe. 1985. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1):32–46.

Benny Tang, Angie Boggust, and Arvind Satyanarayan. 2023. VisText: A Benchmark for Semantically Rich Chart Captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7268–7298, Toronto, Canada. Association for Computational Linguistics.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020. Label Studio: Data labeling software. `https://github.com/HumanSignal/label-studio`. Open source software, version 1.23.2 used.

Lucy Lu Wang, Isabel Cachola, Jonathan Bragg, Evie Yu-Yen Cheng, Chelsea Haupt, Matt Latzke, Bailey Kuehl, Madeleine van Zuylen, Linda Wagner, and Daniel S. Weld. 2021. Improving the accessibility of scientific documents: Current state, user needs, and a system solution to enhance scientific PDF accessibility for blind and low vision users. *arXiv preprint arXiv:2105.00076*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. *Advances in Neural Information Processing Systems*, 37:113569–113697.

M. Wenger, F. Ochsenbein, D. Egret, P. Dubois, F. Bonnarel, S. Borde, F. Genova, G. Jasniewicz, S. Laloë, S. Lesteven, and R. Monier. 2000. The SIMBAD astronomical database. The CDS reference database for astronomical objects. *Astron. Astrophys. Suppl. Ser.*, 143:9–22.

A Answer Parsing

Answers are requested as JSON with entries for the raw answer and an explanation. A small number fall outside that pair and cannot be parsed at scale. Out of 6035 questions per model, this occurred for 12 (22), 4 (3) and 5 (2) answers for GPT, Gemini and Claude respectively, where the first number is for the original images and the bracketed number for archive-light. For Gemini's and Claude's are responses that are not valid JSON: most break on the quote characters of sexagesimal tick labels (e.g., `12'00"`), and four of Claude's answer the provenance question in free text with no JSON at all. 27 of GPT's 34 are valid JSON with an explicit `null` answer, typically for color statistics it judges unrecoverable from the image. For the plot-type and provenance questions of Figure 2 specifically, only Claude returned unparsable answers: 3 for the original images and 1 for archive-light, out of 425 each.

B Question Levels

Questions follow the four-level figure description model of Lundgard and Satyanarayan (2022), adapted to chart-based VQA in Tang et al. (2023); Naiman et al. (2026); Evans et al. (2026). Level 1 covers what is directly readable off the figure (e.g., plot type); Level 2 requires reading values off the axes or colorbar (e.g., the median of the data along an axis, or the angular height of the field); Level 3 requires inference not reducible to a single statistic, such as which distribution family the underlying data came from or multi-step, complex, calculations. Level 4 requires external context or domain expertise and is not generated here.

Beyond quantifying question complexity, this scaffold matters for accessibility: many scientific figures remain unusable for blind and low-vision readers (Wang et al., 2021), and Levels 2 and

3 are the tiers those readers report as most useful (Lundgard and Satyanarayan, 2022; Seo et al., 2024; Kim et al., 2023).

Table 1 gives example questions for our figures; Table 2 lists every question asked of the published figures and where its ground truth comes from.

C Angular Extent of a Sky Figure

Positions on the sky are given in equatorial coordinates: right ascension α, measured eastward along the celestial equator, and declination δ, the angle north or south of it. By convention α is written in hours, minutes and seconds of time, with $24^{\mathrm{h}} = 360°$, so that $1^{\mathrm{h}} = 15°$, $1^{\mathrm{m}} = 15'$ and $1^{\mathrm{s}} = 15''$, while δ is written in degrees, arcminutes and arcseconds. There are two potential issues when reading a sky figure. First, a minute of right ascension is not an arcminute: an axis spanning 2^{m} of α covers $30'$. Second, lines of constant α converge toward the celestial poles, so the true angle subtended by an interval $\Delta\alpha$ shrinks with declination; along a line of constant δ it is $\Delta\alpha \cos\delta$. At $\delta = 60°$, one hour of right ascension spans $7.5°$ on the sky rather than $15°$. Angular height, the extent along the declination axis, needs neither correction, since $\Delta\delta$ is already an angle.

Our ground truth applies the $\cos\delta$ factor at the mid-declination of the displayed field, $w = (\alpha_{\max} - \alpha_{\min}) \cos\bar{\delta}$, whereas for the published figures width is the great-circle length of the plot box's top and bottom edges, measured through the AVM astrometry. The two definitions agree to within 2% for 94 of the 104 published figures (median 0.4%) and differ appreciably only for fields spanning several degrees, by up to 23% for a $16°$-wide field at $\delta \approx 58°$.

Angular height on published figures. Gemini was also asked the angular height of each published figure, worded identically to our figures (Table 2). Figure 5 compares the errors. In the median the two sets agree, as for width: $\mathcal{R} = -0.04$ dex for published figures against -0.05 dex for archive-light. Published figures scatter more, however: 76% of answers fall within a factor of two, against 94% for our figures ($p < 0.001$, two-proportion test), and 10% are too large by more than a factor of ten, against 2%. Since height needs no $\cos\delta$ correction, this extra scatter cannot come from the right ascension conventions above.

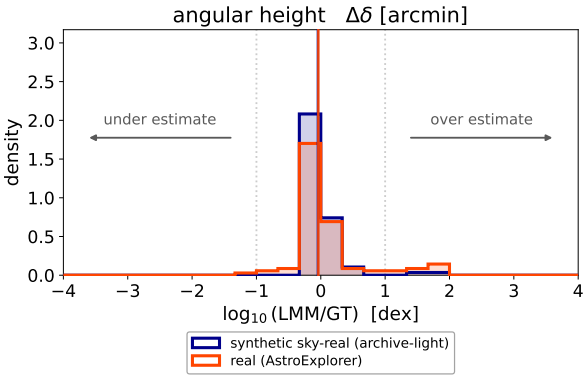


Figure 5: As Figure 4, for angular height: log-ratio error, $\mathcal{R}$, of Gemini's answers for our real-sky figures (archive-light, blue; $n = 85$) and for the 104 human-verified published figures (orange).

D Classifier Baseline for Provenance

To determine if the provenance question can be answered from image content, we train a gradient-boosted classifier (Pedregosa et al., 2011) to separate real-sky from GMM-sky figures using three low-level features of each figure's data array: the field size (the geometric mean of its angular width and height), the concentration of its emission (the share of the background-subtracted signal held by the brightest 5% of pixels), and the sign of the data minimum. Over all 1334 sky figures (667 of each family), 5-fold cross-validation gives 85% accuracy, against 50% for always guessing one family. Most of this comes from concentration, which alone reaches 80%: point-source-dominated real fields are more concentrated than the GMM fields. Field size and the sign of the minimum carry little information on their own, as the matching of the two families intends.

We leave out a fourth feature, the color span (the dynamic range of the data). GMM fields borrow their color ranges from real cutouts, and in this release the same ranges recur across many GMM figures: 667 GMM figures share 56 distinct spans, while real figures have 473. A classifier can exploit that regularity across the dataset – color span alone reaches 81% – but no single figure reveals it, so it says nothing about whether the provenance question is answerable from one image. The code is released with the dataset generation code with updates to pull a color range distribution more closely resembling that of the real-sky images intended in future releases.

Lvl	Question [choices offered]	Applies to	Example Answer
1	What is the plot type of the plot in this figure? [contour, image of the sky]	all plots	image of the sky
1	What is the colormap that was used in this figure?	all plots	rainbow_r
1	What coordinate epoch is given on the axis labels of this figure? Answer "none" if no epoch is stated.	GMM-sky, real-sky	J2000
1	What is the smallest unit of angle shown on the right ascension tick labels of this figure? [hours, minutes, seconds]	GMM-sky, real-sky	minutes
1	What is the minimum right ascension value covered by the right ascension axis of this figure – that is, the lower limit of the axis range, not of the data? Give the value in hours, minutes and seconds [...]	GMM-sky, real-sky	18h46m50.98s
1	What is the style of the contour plot? [image, contour lines, both]	contour	contour lines
2	What is the angular height of the sky region shown in this figure, measured along the declination axis? Give the value in arcminutes.	GMM-sky, real-sky, published	1.58
2	What is the angular width of the sky region shown in this figure, measured along the right ascension axis, as a true angle on the sky (i.e. including the cos(declination) factor)? Give the value in arcminutes.	GMM-sky, real-sky, published	1.88
2	What is the mean value of the data along the color-axis in this figure?	all plots	108.4
3	What is the underlying distribution used to create the data in this figure? [gaussian mixture model, real image of the sky]	GMM-sky, real-sky	real image of the sky
3	What is the underlying distribution used to create the data in this figure along the color-axis? [random, linear, gaussian mixture model]	contour	linear
3	Approximately what angular size does a single pixel of this sky image span? Give the value in arcseconds.	GMM-sky, real-sky	188.0
3	What is the right ascension of the centre of this sky image, in the J2000 (FK5) frame? [...] Give the value in decimal degrees.	published	91.95

Table 1: Example questions for each level (Lvl column), the plot types they are asked of, and an example ground-truth answer. Choices offered are given in brackets; long prompts are abridged ([...]). Plot types: contour, GMM-sky and real-sky are our figures; published are the human-verified Astronomy Image Explorer figures. Example answers come from single figures of each type (vqa_000001, GMM-sky; vqa_000002 and vqa_000004, real-sky; vqa_000007, contour; rsky_000001, published); numeric answers are in the units the question asks for.

E Archive Aging Details

Table 3 lists the archive-light preset implemented: 11 effects, their Augraphy phase, the independent probability with which each fires for a given figure, and the realized firing rate over the 255-figure corpus. Effects marked *soft* run with parameters pulled back from the Augraphy defaults — ink bleed at intensity 0.1–0.3, ink mottling at alpha 0.05–0.12, paper hue restricted to 28–40 at low saturation, stain blending at alpha 0.18, shadows at opacity 0.05–0.2, rotations restricted to $\pm 1^\circ$ rather than $\pm 3^\circ$, and JPEG quality held at 70–95. Folding and SubtleNoise run at their defaults but fire less often.

Relative to the full archive preset of s1mb1o (2025), three effects are dropped entirely — BleedThrough, LightingGradient and PageBorder. A further four effects available in Augraphy are excluded from both archive (s1mb1o, 2025) and our archive-light presets: Scribbles and WaterMark add marks unrelated to the figure, BookBinding bends the page until the plot is no longer readable as a plot, and Letterpress restyles the ink against the colormap. None of these are likely to apply to our figures as they do not typically mimic damage expected in archive scans (Grant et al., 2000).

For each figure we record the sampled plan — the effects that fired, their phase assignment, the grayscale flag, the page background color and the seed — in a manifest released with the dataset, so any aged image can be regenerated without redistributing it. The plan is recorded rather than recomputed at read time, so a manifest stays valid for the images it describes even if the preset is later changed.

F Cost Analysis

Table 4 gives the cost of running each model once over the evaluation, i.e. every question asked a single time per condition. For Gemini and Claude it is computed exactly from the token counts recorded in each API response; one question costs about $0.0007 with Gemini and $0.0025 with Claude. Our code does not recorde GPT's token usage, and its reasoning tokens are billed but not returned, so its cost is scaled from the provider's billed cost for its final runs: $3.34 for the 4970 calls covering

Lvl	Question	Ground truth	Range	Shared
2	What is the angular width of the sky region shown in this figure, measured along the right ascension axis, as a true angle on the sky (i.e. including the cos(declination) factor)? Give the value in arcminutes.	mean length of the plot box's top and bottom edges on the sky	0.33–957′	yes
2	What is the angular height of the sky region shown in this figure, measured along the declination axis? Give the value in arcminutes.	mean length of the plot box's left and right edges on the sky	0.33–1429′	yes
3	What is the right ascension of the centre of this sky image, in the J2000 (FK5) frame? [...] Give the value in decimal degrees.	sky position of the plot-box centre	23–346°	no
3	What is the declination of the centre of this sky image, in the J2000 (FK5) frame? [...] Give the value in decimal degrees.	sky position of the plot-box centre	−67–75°	no
3	Approximately what angular size does a single pixel of this sky image span? Give the value in arcseconds.	not released: scan pixels, not data pixels	–	yes

Table 2: Questions asked of the 104 human-verified published figures (Gemini only). Ground truth is measured through each figure's AVM astrometry over its human-verified plot box (Appendix C); the range is over all 104 figures, in arcminutes (′) or degrees. "Shared" marks questions worded identically to those asked of our figures, so their scores are directly comparable. Questions about other figure elements (colormap, colorbar statistics, provenance) have no ground truth in the metadata and are not asked; pixel scale was asked but is not scored.

Phase	Effect	p	fired
ink	InkBleed *soft*	0.50	55%
ink	InkMottling *soft*	0.30	34%
paper	ColorPaper *soft*	0.80	80%
paper	NoiseTexturize *soft*	0.50	55%
paper	BrightnessTexturize *soft*	0.50	49%
post	Stains *soft*	0.40	38%
post	Folding	0.25	25%
post	ShadowCast *soft*	0.25	24%
post	Geometric *soft*	0.50	49%
post	SubtleNoise	0.50	48%
post	Jpeg *soft*	0.60	59%

Table 3: The archive-light preset: firing probability and the rate realized over 255 figures. Each aged figure is also converted to grayscale with probability 2/3 (realized: 65%).

the last 105 figures of in each condition, or about $0.0007 per question. All costs use non-batch API pricing so larger datasets would likely be less expensive.

Model	Condition	In (M)	Out (M)	Cost ($)
GPT	both (est.)	–	–	8.11
Gemini	original	7.54	0.86	4.42
Gemini	archive-light	7.54	0.77	4.18
Claude	original	10.34	0.96	15.14
Claude	archive-light	10.35	0.95	15.11
Gemini	published	0.66	0.20	0.69

Table 4: Cost of one run of each model: tokens in and out (millions) and cost in USD at list prices per million input / output tokens (GPT $0.20 / $1.25; Gemini $0.30 / $2.50; Claude $1.00 / $5.00), checked 12 October 2026. Each condition is 6035 calls per model; the published-figure run is 520 calls (Gemini only). GPT, whose responses do not record token usage, is scaled from its provider's billed cost for the last 4970 calls (both conditions).

PSGS@2026: From Figures to Citation Graphs: QLoRA Based Multimodal Linking for AstroCLIMB

Prajwal Ghimire[1,2] **and Shiv Ram Saud**[1,2] **and Aashish Mahato**[1,2] **and Sandeep Dawadi**[1,2]
Ramesh Saud and **Sunil Regmi**[1,2] and **Manoj Shakya**[1,2]

[1]Department of Artificial Intelligence, Kathmandu University
[2]Data Science & Artificial Intelligence Lab, Kathmandu University
{prajwalghimire22,saudshivram373,aashishmtho24,sundeepdwd,rajusaud245}@gmail.com
{sunil.regmi,manoj}@ku.edu.np

Abstract

We describe our system for the AstroCLIMB shared task (WASP 2026), which reconstructs the citation graph of astronomy papers from scientific figures and captions alone. The task requires identifying the relationship between pairs of figures and captions as `same_figure`, `same_paper`, `related_papers`, or `unrelated_papers`. Our system finetunes `google/gemma-4-E4B-it` using QLoRA on 8,200 competition training rows with a shared descriptive prompt and greedy decoding. A zero shot baseline achieves 0.4408 macro F1 on a balanced 1,800 pair validation split held out from the competition training file, and the finetuned system substantially improves this citation-relation classification for the challenging `related_papers` class in particular. Our submitted system achieves 0.64282 F1 on all 10,000 official test pairs. These results show that multimodal finetuning can improve citation-relation inference from scientific figures and captions when document metadata is unavailable.

1 Introduction

AstroCLIMB (Astronomy Citation Linking from Illustrations) is a shared task of the 4th Workshop on Artificial Intelligence for Scientific Publications (WASP) at AACL-IJCNLP 2026 (Grezes et al., 2026). Systems predict exactly one of four relations between a pair of objects (two figures, a figure and a caption, or two captions): `same_figure` (both objects from the same figure; figure–caption pairs only), `same_paper` (same DOI, different figures), `related_papers` (one DOI cites the other), or `unrelated_papers` (none of the above). The relation is symmetric; each pair carries a single label, and test pairs carry no DOI or metadata. Therefore, classification must rely on visual–textual content alone.

Our early experiments showed that our strongest zero shot vision language model already performs well on the `same_figure` class while, like the other baselines, largely failing on the citation graph classes, with `related_papers` near zero across every baseline (Section 4). We therefore finetune **Gemma-4-E4B-it** with QLoRA (Dettmers et al., 2023) on 8,200 labeled pairs, optimized with AdamW (Loshchilov and Hutter, 2019), using the exact prompt format also used at inference, with a fixed evaluation protocol (greedy decoding, label only output, macro F1 via scikit-learn (Pedregosa et al., 2018)). The resulting system (Gemma-4-E4B plus the `adapter_best` checkpoint) is our official submission, scoring 0.64282 on all 10,000 test rows.

2 Related Work

The scientific figure–caption literature provides the visual–textual grounding for our task. SciCap introduces scientific-figure caption generation with baselines on graph plots (Hsu et al., 2021), while Multimodal ArXiv introduces ArXivCap and ArXivQA and benchmarks vision-to-text tasks, showing that in-domain training substantially improves scientific-figure comprehension (Li et al., 2024). SciMMIR benchmarks text–figure/table retrieval over 530K pairs (Wu et al., 2024), whereas SciVer evaluates multimodal scientific claim verification on 3,000 expert-annotated examples from 1,113 papers across four reasoning types (Wang et al., 2025). Neither task infers citation relations from scientific figures. For scholarly document relations, SPECTER learns document representations from titles and abstracts using a citation graph triplet objective (Cohan et al., 2020). Citation intent is studied separately by Structural Scaffolds, which incorporates structural information through multitask learning (Cohan et al., 2019), and GraphCite, which combines citation-graph embeddings with

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 192–198
November 10, 2026 ©2026 Association for Computational Linguistics

citation phrases (Berrebbi et al., 2022). Content Based Citation Recommendation instead ranks citation candidates for a query document from textual content without requiring metadata (Bhagavatula et al., 2018). Text-only astronomy models (astroBERT (Grezes et al., 2021); AstroLLaMA (Nguyen et al., 2023)) motivate the domain focus but use no figures. Thus, prior work has separately studied scientific figure–caption modeling, scientific multimodal retrieval and reasoning, and scholarly citation/document relationships. AstroCLIMB combines these dimensions by requiring relation inference directly from scientific figures and captions without document metadata at test time.

3 Methodology

Task. Each (Object A, Object B) pair is classified into one of four classes C. An object is a figure image or a caption text, giving three pair types: TXT-IMG (figure–caption), IMG-IMG (figure–figure), and TXT-TXT (caption–caption). `same_figure` occurs only for TXT-IMG pairs by construction.

Competition data and distribution. The Kaggle competition release (Grezes, 2026) provides 10,000 labeled training pairs and 10,000 test pairs, with object cells containing base64-encoded image bytes or caption text. A separate **sampled development release** contains 1,800 balanced rows with figure images stored as UUID-keyed PNGs `https://www.kaggle.com/datasets/jcxiv42/astroclimb` and is used only for the four few shot demonstrations but never for training or scoring. The two releases contain different rows and are never mixed. The full training file is imbalanced with 10.0% `same_figure` and 30.0% for each other class, while our validation split contains 450 examples per class (seed 42) that leaves 8,200 training rows with 6.7% `same_figure` and 31.1% for each other class. The split is by pair row index with zero index overlap; figures, captions, or source papers may therefore recur across splits, so validation measures in-distribution pair generalization rather than generalization to unseen papers. The primary evaluation metric is macro F1 which is the unweighted mean of per class F1 (Pedregosa et al., 2018).

Multimodal input. Each example pairs the two objects of one row as $x_i = (I_i, T_i)$, where I_i is the ordered list of figure images (two for IMG-IMG, one for TXT-IMG, empty for TXT-TXT) and T_i is the shared descriptive system prompt plus per object caption blocks (truncated to 2,000 characters) and the closing instruction "Answer with one label:". Images appear before their corresponding text and are resized to 448×448. The true label $y_i \in \mathcal{C}$ is appended after the prompt for supervised fine tuning (SFT).

Model architecture. The backbone is **google/gemma-4-E4B-it** (Team et al., 2026a), a dense model with 4.5B effective parameters and a 150M vision encoder, loaded as `AutoModelForMultimodalLM` (falling back to `AutoModelForImageTextToText`). We train LoRA (Hu et al., 2021) adapters on all attention and MLP projections with rank r=64, α=128, dropout 0.05 ($\Delta W = \frac{\alpha}{r} BA$, effective scaling 2) that minimizes causal LM cross entropy with AdamW (Loshchilov and Hutter, 2019). Both training and inference quantize the frozen backbone to 4-bit NF4 with double quantization and bf16 compute (Dettmers et al., 2023).

Training. The model was trained on remaining residual 8,200 rows after removing the balanced 1,800 row validation split (zero index overlap). Training uses the HF `Trainer`, learning rate 10^{-4}, cosine schedule with 5% warmup (204 steps), per device batch 2 with gradient accumulation 8 (effective batch 16), max sequence length 1,536, and a cap of 2,560 steps with early stopping (patience 3, threshold 0.001, evaluated every 200 steps) on a single NVIDIA A100 (80GB). The best validation loss (0.3464) is observed at step 1,000 and the corresponding checkpoint is saved as `adapter_best`. Early stopping terminates the run at step 1,600. Decoding is greedy (`max_new_tokens=32`, `do_sample=False`). The raw generation is mapped to a class by case insensitive regex over the four labels with falling back to `unrelated_papers` when nothing matches. Figure 1 summarizes the training pipeline of the submitted system.

Inference and submission. Testing uses the official 10,000 row `test.csv`. The base model at test time is loaded in 4-bit on 2×T4 GPUs (Kaggle) with `adapter_best` attached unmerged and the predictions are written one-hot in the required column order. The local evaluations report macro F1 and accuracy via scikit-learn (Pedregosa et al., 2018). Figure 2 summarizes the test pipeline.

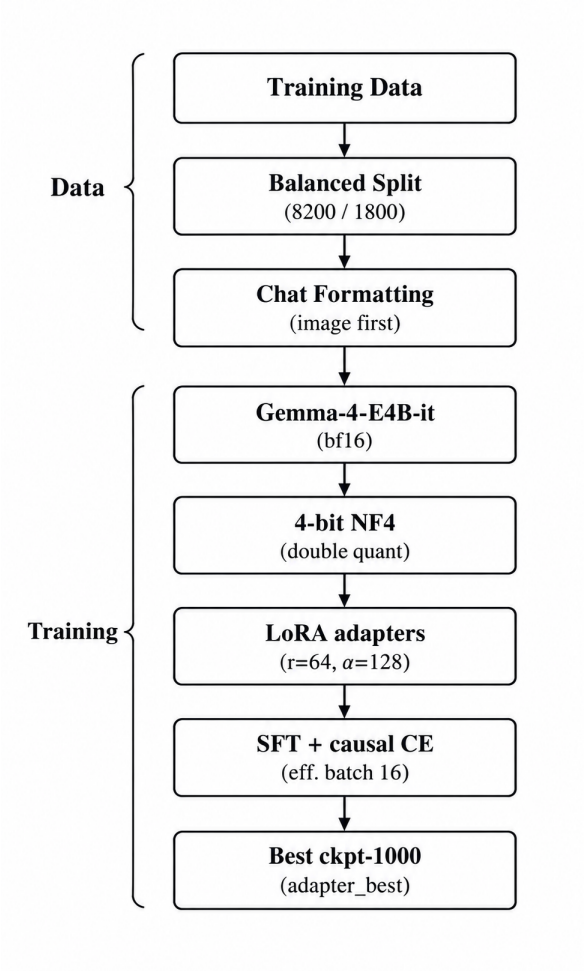


Figure 1: Training pipeline of the submitted system.

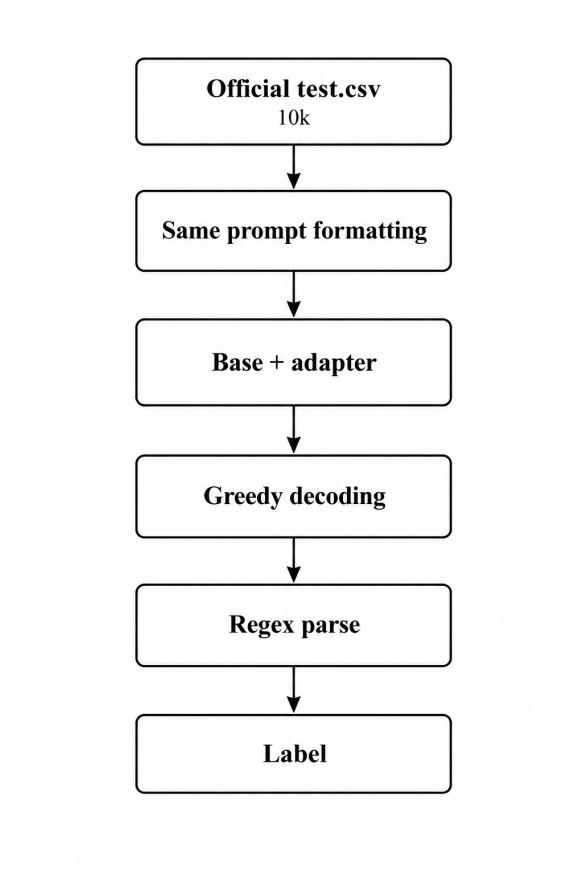


Figure 2: Test/submission pipeline.

4 Results

We evaluated zero shot and few shot vision language model baselines using macro F1 and prediction coverage, and examined the effects of few shot demonstrations and native thinking capabilities, and finetuning on performance.

Zero/few shot baselines. All baselines share one descriptive system prompt and greedy decoding with only the backbone and its native thinking mode varying. Gemma-4-E4B (Team et al., 2026a) leads zero shot at 0.4408 macro F1 (Table 1) whereas Qwen3-VL-8B-Thinking (Bai et al., 2025) collapses to `unrelated_papers` (0.1068), and SmolVLM-500M (Marafioti et al., 2025) scores 0.1723.

The few shot runs include 4 demonstrations (one per class, seed 42, figure–caption pairs preferred) and enable native thinking where the backbone supports it. The decoding is otherwise greedy as in zero shot except that thinking runs use `max_new_tokens=512` instead of 32. The smaller Qwen3-VL-4B and GLM-4.1V-9B variants are used here in place of the Qwen3-VL-8B and GLM-4.6V-Flash variants evaluated zero shot (Team et al., 2026b), so the two tables are not paired backbone-for-backbone. Table 2 compares configurations and results. For Gemma-4-E4B, adding few shot thinking (0.3925) does *not* beat its zero shot score (0.4408) on the identical rows. The Qwen3-VL-4B model covers 397 pairs (0.2497), and GLM-4.1V-9B covers 445 pairs (0.4200). The four demonstrations are drawn from the separate sampled release (disjoint rows); all scores are computed on the validation split.

Model	Think.	Rows	Macro F1	Acc.
Gemma-4-E4B	no	1800	0.4408	0.5206
Qwen3-VL-8B	yes	1800	0.1068	0.2461
SmolVLM-500M	no	1800	0.1723	0.2511
GLM-4.6V-Flash	no	610	0.1020	0.2557

Table 1: Zero shot results on the 1,800-row validation split held out from the training file. Rows list evaluated pairs.

Finetuned system. Our submitted adapter reaches a best validation loss of 0.3464 at check-

Model	Demos	Rows	Macro F1	Acc.
Gemma-4-E4B	4	1800	0.3925	0.505
Qwen3-VL-4B	4	397	0.250	0.330
GLM-4.1V-9B	4	445	0.420	0.524

Table 2: Few shot results with native thinking enabled (demonstrations from the sampled release; evaluation on the same validation split as Table 1). Rows list evaluated pairs and partial row scores are not comparable to full split scores.

point 1,000 on the balanced 1,800 pair split and achieves a macro F1 of 0.5347 (accuracy 0.5483 & per class F1: 0.5705/0.5311/0.3508/0.6864). It substantially improves `related_papers` classification, from near zero F1 in the baselines to 0.3508, while `same_figure` decreases from 0.7760 to 0.5705 with a precision of 0.9945. A comparator SmolVLM-500M LoRA-32 adapter reaches a *lower* validation loss (0.3029) but predicts no `unrelated_papers` on the official test predictions and achieves only 0.23675 on the official leaderboard, compared with the official test F1 of 0.64282 for our system (Table 3).

Model	Val. Loss	Official Test F1
Gemma-4-E4B, LoRA-64	0.3464	**0.64282**
SmolVLM-500M, LoRA-32	0.3029	0.23675

Table 3: Fine-tuned model comparison.

4.1 Per class breakdown & Confusion Analysis

Table 4 reports per class F1 for the leading run per mode (full run list in Tables 1 and 2).

Run	Mode	sf	sp	rp	up
Gemma-4-E4B	ZS	0.7760	0.4108	0.0172	0.5593
Gemma-4-E4B	FS	0.7228	0.2047	0.0505	0.5918
SmolVLM-500M	ZS	0.0128	0.0088	0.3333	0.3341
Gemma-4-E4B	FT	0.5705	0.5311	0.3508	0.6864
GLM-4.1V-9B (445)	FS	0.8653	0.2571	0.0	0.5575
Qwen3-VL-4B (397)	FS	0.4709	0.2381	0.0612	0.2286

Table 4: Per class F1 on the 1,800 row validation split (truncated runs: evaluated subsets, pair counts in parentheses). Leading runs only (sf = `same_figure`, sp = `same_paper`, rp = `related_papers`, up = `unrelated_papers`). ZS = zero shot, FS = few shot, FT = finetuned. Truncated runs are not comparable with full split runs.

The submitted adapter shows strongly asymmetric errors (Figure 3) on the validation split. `same_figure` is highly precise (0.9945) but has low recall (0.4000), whereas `same_paper` has higher recall (0.7600) but lower precision (0.4081). `related_papers` remains weakest (precision 0.5392, recall 0.2600) and `unrelated_papers` is more balanced (precision 0.6170, recall 0.7733).

The primary confusion on validation is `same_figure` → `same_paper`: 256 of 270 missed `same_figure` pairs receive this label. Validation `related_papers` errors concentrate in figure–figure pairs, with recall of 0.0867 and 104 of 150 cases predicted as `same_paper`. The same pattern appears on the official test predictions but is less severe: `same_figure` precision is 0.9386 with recall 0.6420, and 333 of 358 missed `same_figure` cases are classified as `same_paper`. Test precision and recall for the remaining classes are 0.6453 and 0.6463 (`same_paper`) and 0.5573 and 0.3823 (`related_papers`), while figure–figure pairs again show the weakest `related_papers` F1 (0.3166) (Figure 4).

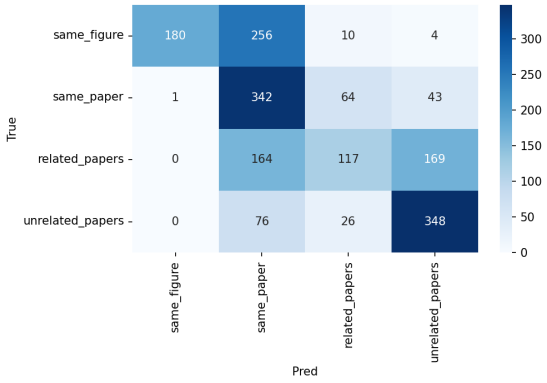


Figure 3: Confusion matrix of the submitted adapter on the balanced 1,800 pair split.

5 Competition Submission Result

Our submission scores all 10,000 test rows and achieves **0.64282**, with predicted class counts 684/3,005/2,058/4,253 (`same_figure`/`same_paper`/`related_papers`/`unrelated_papers`), which shows the use of all four labels rather than collapse onto a single class, and the expected rarity of direct figure–caption matches. No test labels were used in training, model selection, or prompt design. The test set analyses use the organizers' released solution file and were performed after the leaderboard submission. Errors on the official test predictions are asymmetric: `same_figure`

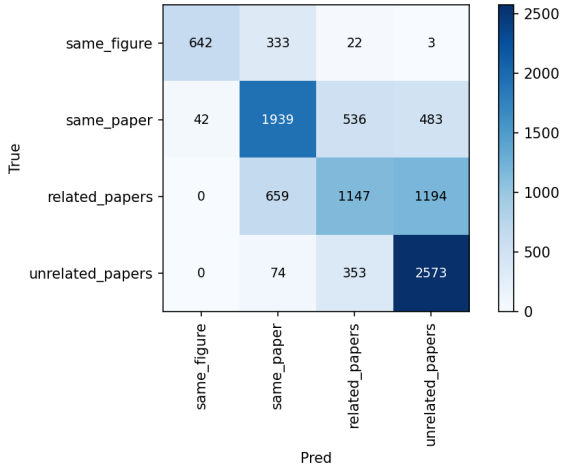


Figure 4: Confusion matrix of the official 10,000 test predictions.

has precision 0.9386 and recall 0.6420, with 333/358 missed pairs predicted as `same_paper`. The `related_papers` class remains weakest overall on test too (precision 0.5573, recall 0.3823, F1 0.4535, up from F1 0.3508 on validation), with figure–figure pairs again the hardest subset.

6 Conclusion

The QLoRA finetuning of Gemma-4-E4B-it on 8,200 competition training rows achieved a best validation loss of 0.3464 and a macro F1 of 0.5347 on the balanced 1,800 row validation split. The resulting submitted system achieved 0.64282 F1 on the official 10,000 row test set. Few shot thinking performed below zero shot for the same backbone (0.3925 vs. 0.4408). The larger thinking models evaluated also score below Gemma-4-E4B's zero shot baseline on their evaluated pairs. Several models cover only partial row subsets and are not directly comparable to full split results. We do not run caption-only or image-only ablations, so we cannot separate multimodal gains from textual topic similarity. The `adapter_best` weights (best validation loss 0.3464) and the official 10,000-row submission are public at https://huggingface.co/ShivRamSaud/gemma4-e4b-qlora-astroclimb-v2-fresh; together with the fixed protocol above and the seed 42 row index split of the public competition file, they document the run (see Appendix A). The training script, preprocessing code, and split row indices are not released; exact reproduction requires reimplementing them from the protocol, prompt, and seed 42 split specification given here. Future work should fine-tune the thinking capable backbones evaluated here as baselines and explore pair type aware training objectives for the hard `related_papers` class.

7 Acknowledgements

We thank the organizers of the AstroCLIMB shared task at WASP 2026 for the competition release and leaderboard evaluation (Grezes, 2026).

A Reproducibility artifacts

Released weights and predictions. The submitted `adapter_best` LoRA adapter (best validation loss 0.3464 at step 1,000) and the official 10,000-row test submission are public at https://huggingface.co/ShivRamSaud/gemma4-e4b-qlora-astroclimb-v2-fresh. The `adapter_best/` folder contains `adapter_model.safetensors`, `adapter_config.json`, the tokenizer and processor configs, and the training arguments. `submission/` holds the scored test predictions.

Not in the repository. The repository contains no training or preprocessing scripts, no prompt file, and no split-index file, and its README is the unfilled PEFT template (no model card). These are instead fixed by the specification in the paper: preprocessing is 2,000-character caption truncation with 448×448 image resizing, the split is seed 42 row indices of the public competition training file (8,200 train / 1,800 validation), decoding is greedy with regex label mapping falling back to `unrelated_papers`, and the exact inference prompt is reproduced below.

A.1 Prompt format

The training, validation, and test use a single fixed prompt contract which is a one descriptive system prompt followed by per object content blocks, with images ordered before text, closed by an explicit answer instruction. The processor renders the turns through its chat template with the generation prompt appended, so training and inference share identical formatting. The exact system prompt is:

```
You are an expert in
astrophysics figures and
captions.  Given Object A
and Object B (each is either
a figure image or a caption
text), classify their relation
into exactly ONE label based
ONLY on what you see/read - no
DOI or metadata is provided.
Classes:  - same_figure:  The
```

```
caption directly describes
the figure in front of you.
Visual elements (axes, labels,
numbers, morphology) are
mentioned verbatim in the
text, or the text reads like
"Figure X shows..." matching
the image. - same_paper:
Same study, different figures.
Similar writing style, same
instruments/datasets/authors
hinted in text, or visual
style (fonts, colors, layout)
is consistent, but NOT a
direct caption-figure match.
- related_papers: Different
papers where one builds on
the other. Overlapping
methods, shared datasets, or
a figure/caption that looks
like a cited prior result,
but style/authors differ. -
unrelated_papers: No clear link.
Different topics, instruments,
scales, or writing/visual style
with no overlap. Base your
decision only on visual and
textual content. Do not assume
same_figure is impossible
for any pair type - judge
from alignment. Output ONLY
the lowercase label (e.g.,
related_papers), no explanation,
no punctuation. Choose
exactly one of: same_figure,
same_paper, related_papers,
unrelated_papers.
```

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Dan Berrebbi, Nicolas Huynh, and Oana Balalau. 2022. Graphcite: Citation intent classification in scientific publications via graph embeddings. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 779–783, New York, NY, USA. Association for Computing Machinery.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 238–251, New Orleans, Louisiana. Association for Computational Linguistics.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3586–3596, Minneapolis, Minnesota. 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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Felix Grezes. 2026. Astroclimb: Astronomy citation linking from illustrations. https://www.kaggle.com/competitions/astroclimb. Kaggle.

Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J. Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S. Grant, Donna M. Thompson, Roman Chyla, Stephen McDonald, Timothy W. Hostetler, Matthew R. Templeton, Kelly E. Lockhart, Nemanja Martinovic, Shinyi Chen, Chris Tanner, and Pavlos Protopapas. 2021. Building astrobert, a language model for astronomy & astrophysics. *Preprint*, arXiv:2112.00590.

Felix Grezes, Kelly Lockhart, Alberto Accomazzi, Atilla Alkan, Ethan Seefried, and Tirthankar Ghosal. 2026. 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.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. Smolvlm: Redefining small and efficient multimodal models. *Preprint*, arXiv:2504.05299.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciucă, Charlie O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Li, Josh Peek, Kartheik Iyer, Tomasz Różański, Pranav Khetarpal, Sharaf Zaman, David Brodrick, Sergio J. Rodríguez Méndez, Thang Bui, Alyssa Goodman, and 5 others. 2023. Astrollama: Towards specialized foundation models in astronomy. *Preprint*, arXiv:2309.06126.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2018. Scikit-learn: Machine learning in python. *Preprint*, arXiv:1201.0490.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026a. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026b. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Chengye Wang, Yifei Shen, Zexi Kuang, Arman Cohan, and Yilun Zhao. 2025. SciVer: Evaluating foundation models for multimodal scientific claim verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8562–8579, Vienna, Austria. Association for Computational Linguistics.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, Noura Al Moubayed, Jie Fu, and Chenghua Lin. 2024. SciMMIR: Benchmarking scientific multi-modal information retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12560–12574, Bangkok, Thailand. Association for Computational Linguistics.

Towards Verifiable Answers: Reinforcement Learning with a Verifier Reduces Hallucinated References in Literature-Based Question Answering

Rushank Goyal
Department of Computer Science
Stanford University
rushankg@stanford.edu

Abstract

Experts and laypeople alike increasingly encounter the scientific literature by posing questions to large language models (LLMs) rather than through database searches, making a model's cited sources the reader's main window into the original literature. Yet LLMs frequently suffer from citation hallucinations. Because source existence is objective and externally checkable, it provides an unusually clean signal for reinforcement learning with verifiable rewards (RLVR). We train LLAMA-3.1-8B-INSTRUCT and LLAMA-3.2-3B-INSTRUCT with GRPO and LoRA against a deterministic verifier that resolves each generated reference against scholarly databases CrossRef and OpenAlex. On 40 common medical questions in the MedAESQA dataset, the hallucination rate decreases from 85.5% for the base $8B$ model to 76.7% ($p < 10^{-4}$) for the best post-trained checkpoint. The post-trained model produces 71% more verified references at unchanged citation density, generates fewer misleading DOIs (down to 14% vs 54% for the base), and favors canonical sources such as the *New England Journal of Medicine*. A six-part audit finds no evidence of reward hacking. Ultimately, this deterministic verifier supplies a practical, objective signal that reduces hallucinated references, offering a novel method to improve citation accuracy and a stepping stone toward fully verifiable LLM answers for literature-based QA.

1 Introduction

When large language models (LLMs) answer questions that require grounding in the scientific literature, they commonly include 'hallucinated' references, i.e. references that point to scholarly articles that do not, in fact, exist as written (Alansari and Luqman, 2026). Such mistakes cause two distinct issues: firstly, an in-text citation to a fabricated reference lends false authority to a claim that may be incorrect; secondly, even if the claim itself is correct, the reader hits a dead end when trying to use the bibliography to navigate to the original literature themselves. Thus, hallucinated references affect a reader's trust in the answer.

Given the increasing number of users, particularly lay users, who use LLMs for literature-based question answering (QA), it is necessary to address the issue of citation hallucination (Chatterji et al., 2025; Mendel et al., 2025). Accurate QA is important for high-stakes tasks such as medical or legal advice, where providing a window into the primary literature is critical, as well as enterprise QA tasks like answering company-specific customer support inquiries. Indeed, any field where QA involves synthesizing accurate answers based on scholarly research, e.g. policymakers asking questions answered by economic analyses (Buchanan et al., 2024), benefits from better citation generation in LLM responses.

Errors are hard to catch on the user's end since a hallucinated reference looks stylistically indistinguishable from a real one to a layperson, with a realistic author list, a plausible title, and a well-formed DOI. This is also why another popular tool for QA finetuning, Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), is also a poor fit for this task, as human annotators cannot reliably catch fabricated references any more than an end user.

The traditional tool for addressing reference accuracy in QA settings is thus retrieval-augmented generation (RAG), which is used by popular general-purpose QA systems like Perplexity and You.com and also specialized medical platforms like OpenEvidence (Low et al., 2025). RAG systems, however, require a realtime inference-time retrieval index and consume more resources and time, while being vulnerable to information poisoning attacks from external sources (Chang et al.,

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 199–209
November 10, 2026 ©2026 Association for Computational Linguistics

2025). In general, too, the high stakes and wide applicability of the problem warrant the exploration of novel techniques to reduce fabricated references.

The central observation we make is that reference or citation existence is *objective* and *automatically checkable* since a reference either resolves to a real record in a scholarly index or it does not. This makes citation existence an unusually clean target for reinforcement learning with verifiable rewards (RLVR), which has recently been a popular technique for post-training language models (Shao et al., 2024).

Our approach in this paper rests on the hypothesis that LLMs, within their large parametric knowledge, do contain memory of correct references, but pretraining regimes provide little incentive to maximize truthfulness—[1]only likelihood. We posit that, when post-trained with an appropriate reward function, LLMs can be trained to cite sources more accurately, increasing readers' trust in QA responses.

We therefore train an open-weights model with Group Relative Policy Optimization (GRPO) against a deterministic citation verifier, asking whether the model can be pushed to surface real bibliographic information from within its parametric knowledge. Specifically:

- We design and train against a deterministic, external citation-existence reward and show statistically significant hallucination reductions by 8.8 pp in the post-trained (PT) model (85.5% base to 76.7% PT, $p < 10^{-4}$).
- We conduct a thorough six-part audit that shows no increase in duplicate citations, no decrease in citation density, and fewer fully fabricated responses, among other results that greatly reduce the likelihood of the result being reward-hacked.
- We characterize how the trained model's citing behavior changes and discuss an important failure mode in which existence is improved at the cost of relevance.

The paper is intended as exploratory work demonstrating that an automated citation verifier provides a viable RL training signal, setting a foundation for future systems that can enforce more complete scholarly metadata and scientific relevance checks for each reference.

[1]This em-dash has been written by a human!

2 Related Work

How bad is citation hallucination? Across models and domains, LLMs fabricate a large fraction of the citations they produce. Walters and Wilder (2023) documented widespread fabrication and error in ChatGPT-generated bibliographies; in medicine, Chelli et al. (2024) measured hallucination rates of 28.6%–91.4% across GPT-3.5, GPT-4, and Bard for systematic-review references, and Mugaanyi et al. (2024) found citation reliability varies sharply by discipline, with fabricated or malformed DOIs a dominant error mode. More recent work in anatomy QA (Ülkir and Paslı, 2026) and legal citation benchmarking (Chen et al., 2026a) continues to show high citation hallucination rates, ranging from 25.2% to 47.5% for the former and as high as 94% for the latter.

RL for factuality. Verifiable-reward RL has reduced hallucination on short factoid QA, where Wei et al. (2026) optimize a ternary truthful/abstain/wrong reward with GRPO, and on long-form generation, where Chen et al. (2026b) train reasoning models to be more factual with a precision/detail/relevance reward. The increased factuality is not, however, a guaranteed outcome of RL-based post-training process, as reasoning-oriented RL can often *increase* hallucination rates (Li and Ng, 2026). The authors attribute this effect to high-variance gradients and entropy-induced randomness. These results all reward truthfulness in short-form QA, whereas we focus on long-form QA, with a desire to optimize truthfulness as well as utility and trustworthiness for the reader using the existence of real bibliographic citations as a proxy.

Reward over-optimization. Gao et al. (2023) show that optimizing a learned proxy reward eventually degrades the true objective. The reward is deterministic rather than learned, so classical reward-model over-optimization does not directly apply to this work. Nonetheless, the training runs do show a plateau and slight regression after a certain point. Possible explanations for this are discussed in §5.

3 Method

3.1 Deterministic verifier

The reward is computed by a deterministic citation verifier, a Python port of CheckIfExist (Abbonato, 2026) with modifications to increase speed. Figure 1 shows the pipeline. For each extracted reference it (i) checks an in-memory cache keyed

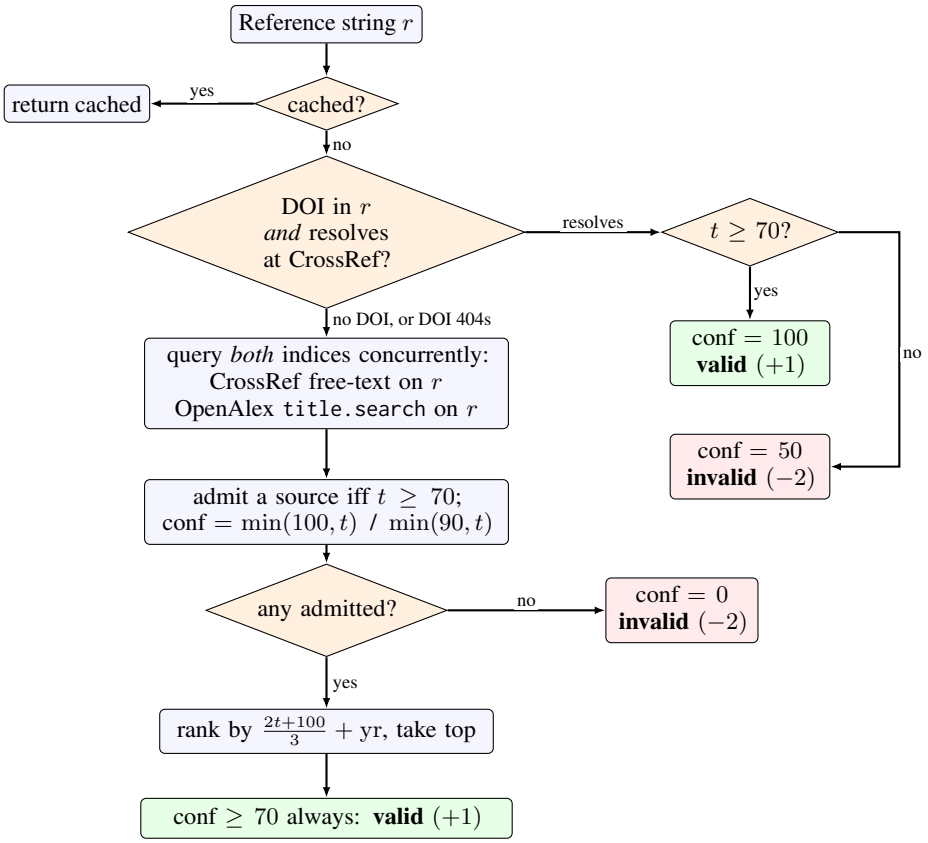


Figure 1: Citation verifier pipeline. t is the title similarity between a retrieved title and the reference string r.

on a hash of the reference fields, so that citations repeated across a prompt's G rollouts incur no repeated API calls; (ii) takes a *DOI fast path*, i.e. if a DOI is present it is resolved against the CrossRef REST API, scoring confidence 100 when the resolved title matches and a capped, sub-threshold confidence of 50 when it does not; and (iii) when no DOI is present (or the DOI lookup fails), dispatches concurrent title searches to CrossRef and OpenAlex.

The best candidate is accepted only if its title similarity exceeds 70. Title similarity (0-100) is the higher of a character-level Levenshtein ratio and a word-level Jaccard overlap, plus a containment boost. Note that title matches in CrossRef and OpenAlex carry confidence ceilings of 100 and 90 respectively—reflecting metadata quality—in line with the code implementation of CheckIfExist. The original verifier queried a third index, Semantic Scholar (S2), which was removed after an ablation on a 47-reference ground-truth corpus showed identical verdicts with and without S2 querying, even though including S2 carried a much higher API rate-limit cost that bottlenecked training speed.

References are extracted from the model output in two passes: a numbered reference-list parser keyed on a "References"/"Bibliography" header, and a parser for inline citations (`[1]`, `[1,2]`, `[1-3]`, etc.) used to compute the per-sentence coverage term $k/|S|$.

3.2 Reward

Let N be the number of extracted references, N_{valid} and N_{invalid} the verified counts, S the set of body sentences, and k the number of sentences with no inline citation. The reward is

$$R = \frac{N_{\text{valid}} - 2\,N_{\text{invalid}}}{N} \;-\; 0.1\,\frac{k}{|S|}, \qquad (1)$$

with $R = -1$ when $N = 0$, giving range $[-2.1, +1.0]$ for the reward function.

Each term targets a specific failure mode. The core term asymmetrically penalizes invalid citations by a factor of two, so a fabrication outweighs a real citation and a positive reward requires mostly real references rather than a lucky majority. Additionally, the normalized sentence term $k/|S|$ discourages confident, uncited prose while dividing by $|S|$ so that long answers are not penalized more than short ones at equal coverage; when every sentence carries a citation it vanishes to zero. The

Setting	Value
Base models	LLaMA-3.1-8B / 3.2-3B-Instruct
Adapter	LoRA r=32, α=64, all proj.
Optimizer	AdamW, lr 5×10^{-6}, bf16
Schedules	linear decay; constant 1.25×10^{-6}
Group size	$G = 4$ rollouts / prompt
Sampling	T=0.8, top-p 0.95, 1500 tok.
Train data	COVID-QA, 1,479 questions
Eval data	MedAESQA, 40 Q $\times$ 4 rollouts
Hardware	Modal A100-80GB, $\sim$60 s/step

Table 1: Training and evaluation configuration.

paper uses a single symbol R throughout to indicate the reward.

3.3 Training and evaluation

We fine-tune LLaMA-3.1-8B-Instruct and LLaMA-3.2-3B-Instruct using Group Relative Policy Optimization[2] (GRPO) (Shao et al., 2024). GRPO was found to be appropriate here since each reward is computed directly using a deterministic verifier rather than a separate learned critic.

Weights are updated using Low-Rank Adaptation (LoRA) (Hu et al., 2022), a parameter-efficient method that freezes the pretrained model and trains small inserted low-rank matrices instead. We use rank-32 LoRA with α=64 and dropout of 0.05, with adapters on all seven attention and MLP projections for a total of $\sim$84M (8B) / $\sim$49M (3B) trainable parameters. Training was run on a single NVIDIA A100 GPU. Table 1 lists the full training details.

Data. Training prompts are drawn from COVID-QA (Möller et al., 2020), which contains 2,019 expert-annotated QA pairs over scholarly articles about COVID-19. We removed yes/no, single-statistic, and date-lookup items that do not call for cited synthesis, retaining a dataset of 1,479 broad-synthesis questions. Only the questions are used: the gold answers are discarded, so the policy receives no supervised target and the entire training signal comes from the verifier. Each prompt pairs the question with a fixed system prompt (see A) that requires a 2–3 paragraph answer, an inline citation marker on every factual claim, 5–10 real sources, and a references section in a parseable format, with an explicit instruction not to fabricate.

[2]Implemented using the TRL and PEFT libraries from HuggingFace.

Procedure. At each step the policy samples $G = 4$ completions for one prompt (T=0.8, top-p 0.95, $\leq$ 1500 new tokens). Every completion is scored independently by the verifier and GRPO forms advantages relative to the group mean. With `per_device_train_batch_size=`G and no gradient accumulation, the number of optimizer steps is equal to the number of questions consumed. Rewards are computed online against the live CrossRef and OpenAlex APIs with an in-memory cache for cost savings.

Runs The main runs cover COVID-QA[0:1000] under a linear decay from 5×10^{-6}. Because reward plateaus after step 750, we trained another run from `checkpoint-750` at a constant LR of 1.25×10^{-6}—which is the learning rate that the decay schedule reaches at step 750—over the 729 remaining, unseen questions COVID-QA[750:1479], to separate a genuine plateau from schedule-induced decay to zero.

Evaluation 40 questions from MedAESQA (Gupta et al., 2025) were used for evaluation, with $G = 4$ rollouts each for a total of 160 rollouts per checkpoint. All statistical comparisons use paired t-tests over the 40 questions, applied to per-question mean reward and per-question hallucination rate. As MedAESQA is an entirely different corpus, with de-identified medical questions from the general public, all gains represent cross-dataset transfer of accurate citation abilities rather than improved in-distribution fit.

4 Results

4.1 GRPO reduces hallucination

Table 2 and Figure 2 give the trajectory of the training runs of the 8B model and both the 3B and 8B models, respectively. Reward rises steeply through step 750, where the 8B model peaks at $R = -1.31$ and hallucination bottoms at 76.7%, an 8.8 pp reduction over baseline ($t = 5.2$, $p < 10^{-4}$) with $\Delta R = +0.26$ (95% CI $[0.16, 0.36]$; $t = 5.1$, $p < 10^{-4}$). The number of valid citations rises from 168 to 287, i.e. the model emits 71% more verified references on the same questions, while references per rollout barely move, going from 7.24 to 7.70.

Continued training after this point does not help and appears to worsen performance across both model sizes. Under linear decay, training from step 750 to 1000 gives $\Delta R = -0.05$ between

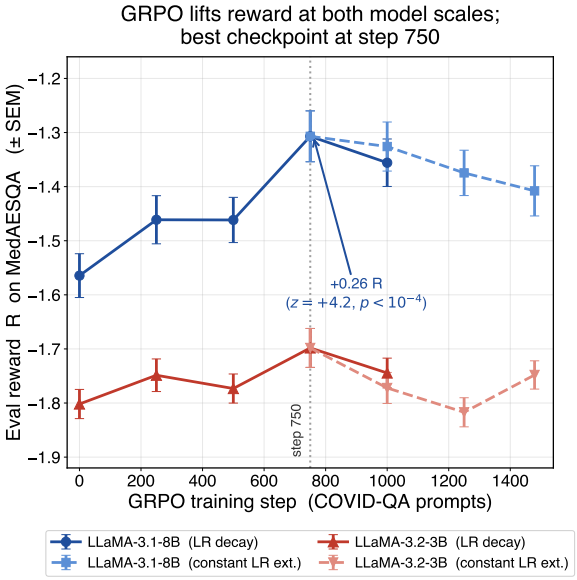


Figure 2: Reward across GRPO training. The 8B and 3B runs trace the same arc and peak at step 750 under both LR schedules, then plateau or mildly regress.

those two points, while the constant-LR continuation run drifts down to $R = -1.41$ by step 1479. The 3B model behaves roughly identically in shape, peaking at step 750 ($R = -1.70$).

Both model sizes exhibit remarkably similar reward trajectories: reward improves rapidly during early GRPO training, peaks at approximately step 750, and subsequently plateaus or mildly regresses across both LR schedules. This effect can be driven by multiple possible reasons: unwanted divergence from the base model after a substantial number of training steps; saturation of the reward signal; or noise from the verifier, which is exactly objective in theory but imperfect in its current implementation.

4.2 Reward gains are broadly distributed

We observe that, at the best (step 750) checkpoint, the model's reward improvement is distributed across the eval set and not driven by a few easy items. Figure 3 plots the per-question change in reward from baseline to step 750: 30 of 40 MedAESQA questions improve and only 10 regress (and even then, the magnitude of regression is much lower than the magnitude of gain). While gain concentrated on a few questions might have implied narrow domain-specific improvement—e.g. for questions that have topic overlap with those in the COVID-QA training set—this broad improvement suggests more general behavioral change.

Checkpoint	R	Halluc.	Refs/roll
Baseline (8B)	−1.564	85.5%	7.24
Step 250	−1.461	81.6%	7.61
Step 500	−1.462	81.0%	7.33
Step 750	**−1.307**	**76.7%**	7.70
Step 1000 (decay)	−1.356	79.2%	7.64
Cont. → 1000	−1.326	77.8%	7.64
Cont. → 1250	−1.375	79.2%	7.52
Cont. → 1479	−1.408	79.7%	7.72

Table 2: GRPO trajectory on MedAESQA with detailed metrics for the main run (8B model).

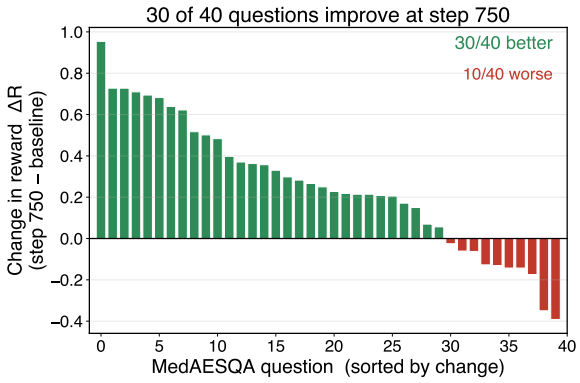


Figure 3: Per-question reward change between baseline and step 750 model (8B). $30/40$ questions improve.

4.3 A six-step audit against reward hacking

To further rule out various sources of reward hacking, we audited the best checkpoint along six axes. No anomalous behavior indicative of reward hacking was observed.

Four of the axes remained stable: reference recycling across questions, within-rollout duplication, mismatches between inline markers and reference lists, and number of sentences with 'stuffed' citations. Specifically, the top 5 most-reused references accounted for 1.6% of citations at both baseline and step 750 with no reference shared across five or more questions, duplicate rates remained near 10%, inline/list ID mismatches were at 0%, and the number of sentences with stuffed (≥ 4) citations was also 0.

The remaining two axes shifted in favor of the post-trained model. First, we checked *which* rollouts earned positive reward (Figure 4). At baseline, the few rollouts with $R > 0$ achieve it by citing sparsely, with a mean of 2.5 references versus 7.2 overall,. At step 750, positive-reward rollouts actually carry mean 7.5 references versus 7.7 overall, i.e. the model earns reward at full citation density. Secondly, we inspect at the shape of the reward distribution (Figure 5). The gain comes

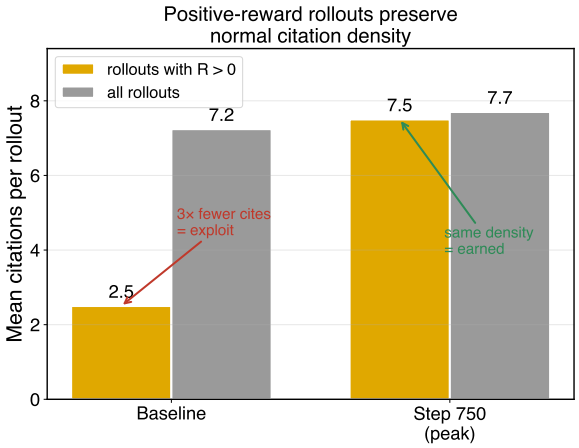


Figure 4: Citation density of positive-reward rollouts. At baseline, $R > 0$ rollouts cite $3\times$ less than average (an exploit); at step 750 they cite at average density, earning reward by citing real papers rather than by abstaining.

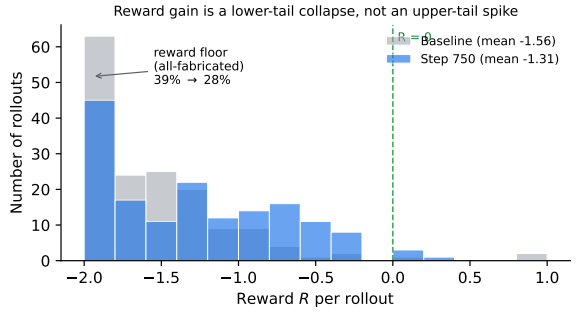


Figure 5: Reward distribution over 160 rollouts. The step-750 gain is a collapse of the lower tail—fewer fully-fabricated answers at the reward floor ($39\% \rightarrow 28\%$)—rather than a spike at the high-reward end.

from shrinking the fraction of fully-fabricated answers at the reward floor (39% to 28%), not from inflating already-good ones, as the proportion of rollouts with $R > 0$ changes minimally, from 1.9% to 2.5%. In other words, RL is fixing the worst answers rather than exploiting gain from those responses that are already the best. Thus, the 8.8 pp reduction represents a genuine improvement in the model's abilities.

4.4 Ruling out eroded caution and advantage noise as causes of performance regression

Before we turn our attention to the learned policy, we share results from tests that check two natural explanations for the post-750 regression. Both explanations are rejected based on the results.

First, if RL eroded the model's epistemic caution, it would hedge less as it regresses, and with lower hedging tendencies it might over-claim or include riskier, i.e. less trustworthy or verified, sources. Instead, hedging—as measured by modal and evidential phrases ("may," "suggests," "further research") per 1K characters—is largely unchanged (see Table 3): the 8B model holds at $\approx$4.1–4.2 across baseline and peak checkpoints, while the 3B peak actually hedges more than its baseline.

Second, GRPO's group-relative advantage could amplify noise if within-group reward variance shrank late in training, destabilizing the computation of the standardized advantage and causing noisy and ultimately counterproductive gradient updates. However, within-group standard deviation (measured on the G=4 eval groups) is flat on MedAESQA and grows on the training distribution ($0.28 \rightarrow 0.49$), and degenerate all-equal-reward groups—which contribute no gradient due to lack of any relative difference among rollouts—become rarer with training (8B $10\% \rightarrow 2\%$; 3B $35\% \rightarrow 15\%$). Neither eroded calibration nor advantage-noise amplification explains the decline.

Checkpoint	Hedges / 1K chars
8B baseline	4.18
8B step 750 (peak)	4.05
3B baseline	4.26
3B step 750 (peak)	**4.73**

Table 3: Epistemic hedging is stable or increases with training, refuting the hypothesis that the performance regression is driven by eroded caution.

4.5 The model learns to cite more carefully

We discuss how the behavior of the post-trained model (i.e. the best step-750 checkpoint) changes compared to the baseline LLM.

An important result is that references carrying an explicit DOI collapse from 54% to 14%. Given the verifier's penalty for a resolving-but-mismatched DOI, a fabricated DOI is a high-variance bet since it could resolve to a completely different paper. Prior surveys have shown that fabricated DOIs are among the most types of hallucinations (Mugaanyi et al., 2024). After the RL training run, the LLM learns not to fabricate DOIs and rely on recall-able title/author matches.

A natural concern is that citing fewer DOIs might be an easy "reward hack" i.e. the observed improvement in reward R is driven entirely by a learned DOI-avoidance behavior. We investigate this further by performing an analysis on rollouts where none of the references carry DOIs ($52/160$

MedAESQA Q37: "if hct is 34, does pt need blood transfusion?"
(✓ = verified real reference, × = hallucinated reference; best rollout per model)

Baseline — $1/3$ verified (67% hallucinated).
A hemoglobin (Hb) concentration below 7 g/dL is typically considered an indication for blood transfusion in patients with acute blood loss [1]. However, a hemoglobin concentration (Hct) of 34% does not necessarily indicate the need for a blood transfusion, as this value alone does not provide information about the patient's overall clinical status or the presence of symptoms.
In fact, some patients with Hb levels as low as 6.5 g/dL may not require a blood transfusion if they are asymptomatic and have no evidence of organ dysfunction [2]. The decision to transfuse should be based on a comprehensive assessment of the patient's clinical condition, including symptoms, vital signs, and laboratory values.
Guidelines from the AABB (formerly the American Association of Blood Banks) and the AABB-AST (American Society for Apheresis) recommend that blood transfusions be considered in patients with a Hb concentration below 7 g/dL or a hematocrit (Hct) below 21% if they have symptoms or signs of inadequate oxygen delivery [3]. Therefore, a Hct of 34% is unlikely to require a blood transfusion unless the patient has significant symptoms or evidence of organ dysfunction.
× [1] Vincent, J. L., et al. (2011). Red blood cell transfusion in critically ill patients. Journal of the American Medical Association, 305(14), 1441–1448. DOI: 10.1001/2011.jama.442
✓ [2] Carson, J. L., et al. (2016). Red blood cell transfusion: A clinical practice guideline from the AABB. Annals of Internal Medicine, 165(8), 557–564. DOI: 10.7326/M16-0933
× [3] AABB. (2017). AABB Clinical Transfusion Guidelines. AABB, Bethesda, MD.

Step 750 — $2/4$ verified (50% hallucinated).
Hemoglobin (Hb) levels are a critical determinant in deciding the need for blood transfusions, not hematocrit (Hct) values alone. A Hct value of 34 is considered low, but it may not necessarily indicate anemia or the need for transfusion [1]. The decision to transfuse should be based on the patient's symptoms, physical examination, and laboratory results, including Hb levels.
According to the American Society of Anesthesiologists (ASA) guidelines, blood transfusions are recommended for patients with Hb levels below 6-8 g/dL, especially if they have symptoms such as fatigue, shortness of breath, or chest pain [2]. However, some studies suggest that even at Hb levels above 6-8 g/dL, patients may still experience impaired oxygen delivery and exercise capacity [3]. Therefore, the decision to transfuse should be individualized and based on the patient's specific clinical context.
A more appropriate approach might be to assess the patient's oxygen delivery and consumption, rather than relying solely on Hct or Hb values. For example, a study by Vincent et al. [4] found that patients with Hb levels between 7-10 g/dL who were not symptomatic did not benefit from transfusion, suggesting that a more nuanced approach to blood transfusion is warranted.
Ultimately, the decision to transfuse should be made by a healthcare provider, taking into account the patient's individual circumstances.
× [1] Vincent JL, et al. Transfusion requirements in critically ill patients. New England Journal of Medicine, 2008.
✓ [2] Practice guidelines for blood component therapy: A report by the American Society of Anesthesiologists Task Force on Blood Component Therapy. Anesthesiology, 1996.
✓ [3] Hébert PC, et al. A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. New England Journal of Medicine, 1999.
× [4] Vincent JL, et al. Transfusion requirements in critically ill patients: a pilot study. Intensive Care Medicine, 2002.

Figure 6: A matched example from the evaluation set showing reduced hallucination in the post-trained model.

for base, $132/160$ for PT). Even in this subset, the hallucination rate falls from 85.3% in the baseline to 75.2% in the post-trained model. This -10.1 pp reduction shows that the performance gain persists independently of DOI presence, suggesting that the model genuinely improves its recall of text-based bibliographic features rather than merely gaming the metric.

We also propose that DOI-avoidance can be considered a behavior improvement, as, when fabricated DOIs are included, lay users are liable to paste them directly into a search engine and end up at a nonexistent or (worse) mismatched paper.

Secondly, the model gravitates more often toward canonical, high-impact sources: the *New England Journal of Medicine* (NEJM), one of the longest-running and most prestigious journals in the medical field, is mentioned 68 times in the baseline's 160 rollouts but 89 times in the same number of rollouts by the post-trained model. Generally, research articles from prominent journals and other high-impact publications are discussed frequently in subsequent literature, suggesting the post-trained model prefers easily recallable sources that it saw more often during pretraining. This aligns with established findings that a language model's ability to accurately recall parametric knowledge is strongly correlated with a document's frequency of occurrence in the pretraining data (Kandpal et al., 2023).

An affinity for more canonical and frequently referenced sources is desirable behavior, since such landmark papers are thoroughly vetted and thus highly likely to be both comprehensive and accu-

rate should the user decide to dive deeper into the bibliography to learn more.

Figure 6 makes this concrete on a sample question from MedAESQA. The baseline produces three references, of which only 1 resolves correctly (67% hallucination rate), while the post-trained model has 2 verified references out of 4 (50% hallucination rate). Even the one that 'resolves correctly' in the baseline's rollout has an incorrect year, as the deterministic verifier focuses on the title.

4.6 Existence is not relevance

The same instinct to retrive canonical papers that drives broad gains can backfire on some questions. We take the example of a different question from MedAESQA, which asks how cannabidiol (CBD) affects liver enzymes. This is a niche, recent topic with sparse indexed literature; here, the baseline model cites an on-point study ("Effects of cannabidiol on liver enzymes in healthy volunteers"), while the post-trained model drifts to the more famous CBD literature on anxiety, epilepsy, and schizophrenia. While those papers are more canonical and more recallable, they are off-topic to the user's questions. This is a limitation of the current verifier, which only rewards existence and not relevance (see §6).

5 Discussion

A verifier can supply what learned rewards lack. Whereas RL against learned, RLHF-style rewards can *increase* hallucination (Lin et al., 2024) or be gamed by exploiting the scorer (Chen et al., 2026b), a deterministic external verifier gives an unspoofable factual signal: a citation either resolves in a scholarly index or it does not. Our main result of an 8.8 pp reduction mirrors for long-form QA what other RLVR-for-truthfulness work has shown for short-answer factoid QA (Wei et al., 2026).

The plateau is an open question. Both scales peak at step 750 and then plateau or mildly regress, and we have separately ruled out reward hacking (§4.3), eroded caution, and advantage-noise amplification (§4.4) as possible factors. Two hypotheses from recent RLVR theory might explain this phenomenon. Cui et al. (2025) show that policy entropy collapses early and trades predictably for reward, implying a performance ceiling once exploration is exhausted. A different paper by Wu et al. (2026) argues that RLVR sharpens probability mass onto what the base model already favors while trimming correct-but-rare outputs. The observations in this paper are consistent with both: firstly, the regression is concentrated in the lower tail of the reward distribution, and once the low-performing responses are improved, the model plausibly hits a 'ceiling' that it cannot cross; secondly, the retreat toward a small set of canonical venues resembles the distribution sharpening that the second hypothesis predicts. The first effect can likely be resolved with a larger model (e.g. LLaMA's 70B variant) that has an order of magnitude more parameters and thus more parametric knowledge to draw upon.

Vanilla GRPO can also suffer from advantage collapse, especially with small groups of 4 rollouts (He et al., 2026). Sampling more rollouts per question ($G = 8$ or 16 or even 32 instead of of $G = 4$) or creating a larger set of possible questions (likely synthetically since we only need questions and not answers) are easy methods to improve performance. On the other hand, adapting the data distribution at train-time (Panaganti et al., 2026) to sample more rollouts from higher difficulty questions or questions that require greater literature search depth could improve reward gain without additional data.

Lastly, the imperfect nature of the verifier likely also impedes improvement at some point (see §6).

We end this section by stressing that the decline is statistically mild, so these are offered as candidate explanations for a possible issue that could arise in similar research, rather than confirmed mechanisms behind a clearly documented phenomenon. From a practical perspective, an early stop at the reward peak is likely ideal.

6 Conclusion

We follow the broader trend towards verifiable rewards for LLM post-training, presenting an exploratory demonstration that a deterministic citation verifier can supply a useful training signal to reduce hallucinated references in long-form, literature-grounded QA.

Specifically, we show that training an open-weights model with GRPO against a deterministic citation verifier measurably reduces hallucinated references in medical QA (-8.8 pp, $p < 10^{-4}$), and a six-part audit finds no evidence of common reward-hacking behaviors. Instead, we see the post-trained model emits 71% more verified references without any change in citation density, emits far fewer DOIs, recalls named landmark trials, and gravitates to canonical sources. The same instinct

can backfire on niche topics, where existence is improved at the cost of relevance, providing a concrete target for future reward design. Ultimately, the deterministic verifier supplies an objective, factual signal that reduces the rate of hallucinated references, providing an interesting, novel method to improve reference accuracy that can serve as a stepping stone toward fully verifiable QA systems.

Limitations

This paper uses a single held-out evaluation set (MedAESQA, 40 questions), which contains data for but one domain of literature-based QA. Additionally, all our training runs are single-seed.

The verifier is imperfect, as only the title of a work is checked and not the journal, year, page numbers, etc. Some imperfections were introduced by our simplifications, while others are more fundamental limits of the databases themselves. We encourage future researchers to improve upon this design. Even with these fixes, the risk of false positives and negatives remains as string matching over large databases is challenging, and certain citations might be to material not indexed in common databases. Regardless of the design, we encourage a thorough analysis of the precision and recall of such verifiers, which we were not able to do for our verifier in the current study.

Even if it were perfect, the verifier checks citation *existence*, not *relevance*: as §4.6 shows, a real reference can be tangential to the claim it supports, so a fluent answer with verifiable-but-irrelevant citations would get high reward while potentially being less useful to the reader. In the worst case, the LLM could actively print wrong statements but attach technically-correct references to 'pass' the verifier. Thus, a highly-accurate 'verifier' for scientific relevance would be an important next step in the evolution of this field.

We believe that beyond existence, truthfulness and relevance themselves can be rewarded using a 'verifier.' Specifically, whether a paper P supports a claim C can be verified to a high degree of confidence. We are confident that future research can reward directly for truthfulness instead of relying on existence as a proxy.

Ethics Statement

Our intention in this research is to improve the reliability of literature-based QA, ensuring that when laypeople or professionals query an LLM, they are pointed toward real, verifiable sources. However, we acknowledge a potential dual risk: namely, by training an LLM to reliably generate real citations, this methodology could theoretically be exploited to automatically generate seemingly credible scientific manuscripts.

To mitigate this, our methodology is strictly designed for focused QA rather than free-form generative writing. Our training pipeline directs the model to generate concise, 2–3 paragraph responses that directly address a specific user query. This design does not support the sustained argumentative synthesis or document-level coherence required for genuine academic authorship. We condemn any use of such methodology for academic misconduct, ghostwriting, or the automated fabrication of scientific literature. This work is intended solely to build trustworthy, verifiable QA assistants that help human readers safely navigate existing knowledge.

Acknowledgments

This research originated as a course project for Stanford University's CS224R. I would like to thank the teaching staff for their instruction and for coordinating with Modal to provide the compute credits that made this work possible. I am also grateful to the anonymous workshop reviewers for their insightful feedback that resulted in an improved final manuscript.

References

Diletta Abbonato. 2026. CheckIfExist: Detecting citation hallucinations in the era of ai-generated content. *Preprint*, arXiv:2602.15871.

Aisha Alansari and Hamzah Luqman. 2026. Large language models hallucination: A comprehensive survey. *Preprint*, arXiv:2510.06265.

Joy Buchanan, Stephen Hill, and Olga Shapoval. 2024. Chatgpt hallucinates non-existent citations: Evidence from economics. *The American Economist*, 69(1):80–87.

Zhiyuan Chang, Mingyang Li, Xiaojun Jia, Junjie Wang, Yuekai Huang, Ziyou Jiang, Yang Liu, and Qing Wang. 2025. One shot dominance: Knowledge poisoning attack on retrieval-augmented generation systems. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18811–18825, Suzhou, China. Association for Computational Linguistics.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use chatgpt.

Working Paper 34255, National Bureau of Economic Research.

Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, and Caroline Ruetsch-Chelli. 2024. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, 26:e53164.

Sijia Chen, Hang Yin, and Shunfan Zhou. 2026a. Legalcitebench: Evaluating citation reliability in legal language models. *Preprint*, arXiv:2605.10186.

Xilun Chen, Ilia Kulikov, Vincent-Pierre Berges, Barlas Oğuz, Rulin Shao, Gargi Ghosh, Jason Weston, and Wen tau Yih. 2026b. Learning to reason for factuality. *Preprint*, arXiv:2508.05618.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *Preprint*, arXiv:2505.22617.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, pages 10835–10866.

Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. 2025. A dataset of medical questions paired with automatically generated answers and evidence-supported references. *Scientific Data*, 12(1):1035.

Xixiang He, Qiyao Sun, Ao Cheng, Xingming Li, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. 2026. Advantage collapse in group relative policy optimization: Diagnosis and mitigation. *Preprint*, arXiv:2605.21125.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. *Preprint*, arXiv:2211.08411.

Junyi Li and Hwee Tou Ng. 2026. Reasoning models hallucinate more: Factuality-aware reinforcement learning for large reasoning models. *Advances in Neural Information Processing Systems*, 38:151064–151085.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. *Advances in Neural Information Processing Systems*, 37:115588–115614.

Yen Sia Low, Michael L Jackson, Rebecca J Hyde, Robert E Brown, Neil M Sanghavi, Julian D Baldwin, C William Pike, Jananee Muralidharan, Gavin Hui, Natasha Alexander, et al. 2025. Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems. *Digital health*, 11:20552076251348850.

Tamir Mendel, Nina Singh, Devin M Mann, Batia Wiesenfeld, and Oded Nov. 2025. Laypeople's use of and attitudes toward large language models and search engines for health queries: survey study. *Journal of medical Internet research*, 27:e64290.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Joseph Mugaanyi, Liuying Cai, Sumei Cheng, Caide Lu, and Jing Huang. 2024. Evaluation of large language model performance and reliability for citations and references in scholarly writing: Cross-disciplinary study. *Journal of Medical Internet Research*, 26:e52935.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Kishan Panaganti, Zhenwen Liang, Wenhao Yu, Haitao Mi, and Dong Yu. 2026. Group distributionally robust optimization-driven reinforcement learning for llm reasoning. *Preprint*, arXiv:2601.19280.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Mehmet Ülkir and Bahattin Paslı. 2026. Reference hallucination, citation reliability, and readability of large language models in anatomy-related question answering. *Clinical Anatomy*.

William H Walters and Esther Isabelle Wilder. 2023. Fabrication and errors in the bibliographic citations generated by chatgpt. *Scientific reports*, 13(1):14045.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Jingxiang Chen, Mohammad Kachuee, Teja Gollapudi, Yiwei Liao, Nicolas Scheffer, Rakesh Wanga, Anuj Kumar, Yu Meng, Wen-Tau Yih, and Xin Luna Dong. 2026. TruthRL: Incentivizing truthful LLMs via reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 134156–134175. PMLR.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. 2026. The invisible leash: Why rlvr may or may not escape its origin. *Preprint*, arXiv:2507.14843.

A System Prompt

Every training and evaluation rollout used the following system prompt, held fixed throughout.

```
You are a medical research assistant answering questions based on the published scientific literature. Your responses must:

1. Be 2–3 paragraphs long.
2. Support every factual claim with an inline citation marker, e.g. [1] or [2, 3].
3. Draw on 5–10 real, peer-reviewed papers or authoritative sources.
4. End with a References section listing each cited source in this exact format:

References
[1] Author(s). Title. Journal or Conference, Year. DOI: xx.xxxx/xxxxx
[2] Author(s). Title. Journal or Conference, Year.

Only cite papers that actually exist. Do not fabricate titles, authors, journals, or DOIs.
```

The user turn contained the COVID-QA or MedAESQA question alone.

Beyond Visual Similarity: Parameter-Efficient Vision–Language Models for Scientific Citation Linking on AstroCLIMB

Ashmani Kumar **Shubham Sharma** **K S Ramalakshmi** **Ankush Chopra**
AI COE / Tredence Analytics
Bengaluru, Karnataka, India
{ashmani.kumar, shubham.sharma, ramalakshmi.ks, ankush.chopra}@tredence.com

Abstract

This paper describes our submission to the AstroCLIMB shared task, which evaluates whether multimodal models can identify citation relationships among astronomy papers from pairs of figures and captions. We explore an empirical progression spanning frozen embedding baselines, dual encoders, cross-encoders, and generative vision–language models (VLMs) adapted via parameter-efficient fine-tuning (LoRA with 4-bit quantization). Among our evaluated configurations, a fine-tuned Qwen3.8-27B achieves a test Macro-F1 of approximately 0.74. An auxiliary specialist adapter trained to re-rank the frequently confused *same paper* and *related papers* classes yields comparable performance within expected run-to-run variance. Across input modalities, Text–Text pairs achieve the highest classification accuracy, followed by Text–Figure pairs, while Figure–Figure pairs prove the most challenging. These findings demonstrate that visual similarity alone remains an insufficient proxy for semantic citation linkage in specialized scientific literature.

Keywords: Multimodal Learning, VLM, Citation Linking, PEFT.

1 Introduction

Scientific figures and captions contain valuable insights, but they are often overlooked by automated research discovery tools. The AstroCLIMB(Grezes et al., 2026) shared task at WASP, IJCNLP–AACL 2026 evaluates whether multimodal models can partially infer citation linkages among astronomy publications directly from isolated pairs of figures and captions. The benchmark tests three configurations of the modality: Figure–Figure (`ii`), Text–Figure (`ti`) and Text–Text (`tt`). Each pair is assigned to one of four relational proximity tiers: *same figure*, *same paper*, *related papers*, or *unrelated papers*.

This task is challenging because astronomy figures from different studies often look similar, while distinguishing figures from the same paper and those from closely related papers requires deeper scientific understanding. We evaluate several models, from AstroBERT-based dual encoders to efficient vision-language models. Our best-performing configurations achieve approximately 0.74 test Macro-F1. A fine-grained analysis across modalities reveals a consistent performance hierarchy (`tt` $>$ `ti` $>$ `ii`), confirming that textual semantics currently provide more reliable citation signals than visual features alone.

2 Related Work

Scientific Document Understanding (SDU): While citation analysis historically relies on text (Beltagy et al., 2019; Devlin et al., 2019), SDU increasingly targets non-textual evidence like charts and plots, which pose unique challenges due to technical axes and visual conventions (Masry et al., 2022). In astronomy, domain-specific models like ASTROBERT (Grezes et al., 2021, 2022) parse literature from the NASA ADS, yet multimodal citation linking directly from isolated artifact pairs remains underexplored, motivating the AstroCLIMB benchmark.

Multimodal Modeling and PEFT: Architectures broadly follow two paradigms: dual encoders mapping features into a shared space (e.g., ViT (Dosovitskiy et al., 2021), SigLIP 1/2 (Zhai et al., 2023; Tschannen et al., 2025)) and generative VLMs combining patch encoders with causal decoders (e.g., PaliGemma (Beyer et al., 2024), Qwen (Wang et al., 2024)). To mitigate computational costs, Parameter-Efficient Fine-Tuning (PEFT) via LoRA (Hu et al., 2021) and 4-bit quantization (Dettmers et al., 2024) allows adapting large backbones on a single workstation GPU. We evaluate whether unified generative VLMs outperform mod-

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 210–215
November 10, 2026 ©2026 Association for Computational Linguistics

ular dual encoders on specialized scientific figures.

3 Dataset and Problem Statement

The AstroCLIMB shared task targets semantic citation relationships between multimodal artifact pairs from NASA's Astrophysics Data System (ADS). Formally, each instance is denoted as (x_1, x_2, τ, y), where x_1 and x_2 are scientific artifacts (figures or text captions), $\tau \in \{\texttt{ii}, \texttt{ti}, \texttt{tt}\}$ specifies the modality configuration (Figure–Figure, Text–Figure, or Text–Text), and $y \in \mathcal{Y}$ is the relationship label. The objective is to infer:

$$\hat{y} = \arg\max_{c \in \mathcal{Y}} P\big(y = c \mid x_1, x_2, \tau\big) \qquad (1)$$

The label space $\mathcal{Y}$ defines four hierarchical proximity tiers: (1) **Same Figure** (composite panels or true figure–caption pairs), (2) **Same Paper** (distinct artifacts from the same paper), (3) **Related Papers** (artifacts from papers connected via citations or topical overlap), and (4) **Unrelated Papers** (independent publications).

Dataset Breakdown: The benchmark comprises labeled training pairs and an unlabelled test set evaluated blindly via Kaggle. We partitioned the labeled data into training and validation sets using stratified sampling to strictly preserve both the class distribution and the proportions of each input modality. The complete breakdown across splits, classes, and modality configurations is detailed in Table 1.

4 Methodology

We evaluate model architectures across four progressive stages on all pair modalities (`ii`, `ti`, `tt`).

4.1 Architectural Progression

The benchmarked systems are organized into four paradigms:

- **Baselines (E0):** Zero-shot cosine similarity over frozen OpenAI text and SigLIP multimodal embeddings.
- **Biencoders and Cross-Encoders (E1):** Dual-tower metric models pairing text (ASTROBERT) and vision (SigLIP-1/2), alongside cross-encoders (PaliGemma).
- **Generative VLMs (E2):** Unified autoregressive decoders like Qwen3-VL-8B, Qwen3-VL-32B, and Qwen3.8-27B (checkpoint: `unsloth/Qwen3.8-27B-unsloth-bnb-4bit`) (Wang et al., 2024)—adapted via 4-bit NF4 and 16-bit LoRA.
- **Specialist and Cascade (E3):** A targeted 2-class adapter for the `same_paper` vs. `related_papers` boundary, alongside a two-level hierarchical cascade.

Table 2 summarizes the evaluated model configurations.

4.2 Training and Adaptation

Biencoder Fusion and Unfreezing: For dual encoders, visual (e_v) and textual (e_t) projections are combined via difference concatenation: $e_{\text{fused}} = [e_v \parallel e_t \parallel |e_v - e_t|]$. While base layers remain frozen, the top four transformer blocks of SigLIP-2 are unfrozen during training.

Parameter-Efficient VLM Adaptation: VLMs were fine-tuned via UNSLOTH (Unsloth, 2024) with LoRA (W_q, W_k, W_v, W_o) on two NVIDIA RTX PRO 6000 Blackwell GPUs (96 GB VRAM each). The training of our best model (Qwen3.8-27B) took approximately 9.5 GPU-hours for 10 epochs, updating 124.4 million out of 27.4 billion parameters (0.45%). We benchmarked 4-bit NF4 quantization (Dettmers et al., 2024) against unquantized 16-bit LoRA and swept adapter ranks ($r \in \{32, 64\}$). We used the AdamW 8-bit optimizer with a learning rate of 1×10^{-4}, $\alpha = 16$, and a batch size of 4 with 8 gradient accumulation steps. At test time, inference validation took approximately 1 hour for 10,000 samples.

Data Augmentation: Training incorporated cross-modal Text–Figure (`ti`) sample upsampling to rebalance pair modalities. Although `ti` pairs constitute 40% of the dataset, initial cross-encoder validation showed it was the weakest modality (Macro-F1 of 0.3604). Therefore, `ti` pairs were deliberately upsampled to improve performance. We also used pair-order permutation ($A \leftrightarrow B$) to enforce symmetric invariance.

4.3 Inference and Post-Processing

Next-Token Logit Classification: To eliminate autoregressive generation latency, classification probabilities for $y \in \{0, 1, 2, 3\}$ were extracted directly via a 4-way softmax over the first-step vocabulary logits of the corresponding label tokens.

Test-Time Augmentation and Calibration: At test time, predictions were symmetrized via pair-order averaging: $P_{\text{sym}} = \frac{1}{2}(P(y \mid A, B) + P(y \mid$

	Class Distribution (Labeled)				Modality Distribution			
Split	**Same Fig.**	**Same Paper**	**Related**	**Unrelated**	**Fig–Fig (`ii`)**	**Text–Fig (`ti`)**	**Text–Text (`tt`)**	**Total**
Train	850	2,550	2,550	2,550	2,550 (30%)	3,400 (40%)	2,550 (30%)	8,500
Validation	150	450	450	450	450 (30%)	600 (40%)	450 (30%)	1,500
Test	—	—	—	—	3,000 (30%)	4,000 (40%)	3,000 (30%)	10,000
Total	**1,000**	**3,000**	**3,000**	**3,000**	**6,000 (30%)**	**8,000 (40%)**	**6,000 (30%)**	**20,000**

Table 1: AstroCLIMB dataset breakdown across target classes and input modalities. Ground-truth test labels were withheld for blind evaluation on the Kaggle leaderboard.

Model Family / Setup	**Paradigm**	**Precision**	**Adaptation Technique**	**Modalities**
OpenAI / SigLIP Embeddings	Metric Baseline	FP16 / FP32	Zero-Shot Cosine Similarity	`tt`, `ti`, `ii`
BERT / AstroBERT + ViT	Biencoder / CE	FP16	Supervised Metric Head	`ti`, `ii`
AstroBERT + SigLIP-1 / 2	Biencoder	FP16	Unfreeze Last 4 Blocks + Aug	`ti`, `ii`
PaliGemma	Cross-Encoder	FP16	End-to-End Supervised Tuning	`tt`, `ti`, `ii`
Qwen3-VL-8B / Qwen3-VL-32B	Generative VLM	4-bit (NF4)	Unsloth LoRA Adapter	`tt`, `ti`, `ii`
Qwen3.8-27B	Generative VLM	4-bit / 16-bit	Unsloth LoRA (Rank Sweeps)	`tt`, `ti`, `ii`
Specialist Adapter	Targeted VLM	4-bit (NF4)	Focused 2-Class Boundary SFT	`tt`, `ti`, `ii`

Table 2: Systematic overview of benchmarked architectures across evaluation stages (E0–E3).

$B, A)$). Finally, validation-fitted additive logit offsets were applied prior to argmax selection to calibrate against class skew.

Threshold Calibration for Embedding Baselines (E0): For the frozen embedding baselines (E0a and E0b), predictions are derived by partitioning continuous cosine similarities $s \in [-1, 1]$ across three strictly ordered decision thresholds $\tau_1 < \tau_2 < \tau_3$:

$$\hat{y}(s) = \begin{cases} \texttt{unrelated_papers} & \text{if } s < \tau_1, \\ \texttt{related_papers} & \text{if } \tau_1 \le s < \tau_2, \\ \texttt{same_paper} & \text{if } \tau_2 \le s < \tau_3, \\ \texttt{same_figure} & \text{if } s \ge \tau_3. \end{cases}$$

Because these frozen metric models lack parameterized classification heads, the cut points are calibrated on the training split to maximize Macro-F1 under this decision rule. We compare two threshold search procedures: (i) a uniform grid search over all ordered candidate triples, and (ii) Nelder–Mead optimization using an unconstrained, order-preserving reparameterization:

$$\tau_1 = x_1, \qquad \tau_2 = \tau_1 + e^{x_2}, \qquad \tau_3 = \tau_2 + e^{x_3},$$

initialized from midpoints between adjacent class-conditional similarity quantiles (the 75th percentile of class k and the 25th percentile of class $k+1$). The optimal thresholds are fixed and evaluated directly on validation and test splits. Subsequent supervised models (E1–E3) bypass this procedure, relying instead on direct softmax logit classification.

5 Experiments

We evaluate our modeling progression across all three pair modalities (`ii`, `ti`, `tt`) on the AstroCLIMB benchmark. Table 3 summarizes the official test set Macro-F1 scores across all development phases. Generative VLM experiments were evaluated using seed 3407, whereas bi-encoder and cross-encoder experiments used seed 42.

5.1 Experimental Setup and Protocol

Models are evaluated on test Macro-F1 across all four classes, alongside modality-stratified breakdowns. Training and evaluation were conducted on NVIDIA RTX PRO 6000 Blackwell GPUs (96 GB VRAM) using the stratified validation split. Deterministic first-token logit extraction and test-time augmentation follow the protocol in Section 4.3.

5.2 Baselines and Discriminative Encoders (E0–E1)

Zero-shot cosine similarity over frozen embeddings (E0) yields the lowest overall performance, confirming that general-purpose metric spaces fail to capture hierarchical citation proximity (Table 3). Supervised bi-encoders and cross-encoders (E1) improve systematically over these baselines as visual towers advance from ViT to SigLIP-2, with further gains from unfreezing upper visual blocks and cross-modal Text–Figure augmentation. However, modular architectures ultimately plateau well below unified VLMs due to representation bottlenecks in late fusion.

ID	Stage	Experiment / Configuration	Test mF1
E0a	Embedding Baseline	OpenAI embedding cosine similarity	0.26608
E0b	Embedding Baseline	SigLIP embedding cosine similarity	0.26619
E1a	Biencoder / CE	AstroBERT + ViT biencoder	0.38688
E1b	Biencoder / CE	AstroBERT + SigLIP-1 biencoder	0.40393
E1c	Biencoder / CE	PaliGemma cross-encoder	0.38836
E1d	Biencoder / CE	AstroBERT + SigLIP-2 biencoder	0.44641
E1e	Biencoder / CE	AstroBERT + SigLIP-2 (unfreeze last 4 blocks)	0.45818
E1f	Biencoder / CE	AstroBERT + SigLIP-2 + Text–Figure sample augmentation	0.46935
E2a	VLM SFT	Qwen3-VL-8B	0.70171
E2b	VLM SFT	Qwen3-VL-32B	0.72534
E2c	VLM SFT	Qwen3.8-27B (4-bit LoRA, $r = 32$)	**0.73925**
E2d	VLM SFT	Qwen3.8-27B (increased LoRA rank, $r = 64$)	0.71072
E2e	VLM SFT	Qwen3.8-27B + data augmentation	0.73889
E2f	VLM SFT	Qwen3.8-27B (16-bit LoRA, $r = 32$)	0.73300
E3a	Specialist	Specialist model (focused 2-class adapter)	**0.73929**
E3b	Cascade	Two-level uncalibrated cascade	0.37401

Table 3: Catalogue of experiments with official test macro-F1 across all development phases.

5.3 Vision–Language Fine-Tuning (E2a–E2f)

Unified autoregressive VLMs substantially outperform modular encoders (Stage E2). Performance scales with model generation and capacity: the newer Qwen3.8-27B backbone consistently outperforms both Qwen3-VL-8B and the nominally larger Qwen3-VL-32B, reflecting stronger foundational pretraining. Across adaptation settings, 4-bit LoRA provides an effective, parameter-efficient baseline. Unquantized 16-bit training, higher adapter rank, and symmetric data augmentation all perform within a narrow score band; given our limited dataset size, these minor variations reflect optimization sensitivity and sample variance rather than fundamental scaling limits.

5.4 Specialist Modeling and Cascading (E3a–E3b)

To address the primary confusion between `same_paper` and `related_papers`, we trained an auxiliary 2-class specialist adapter (E3a). However, introducing this specialist did not meaningfully improve performance over the monolithic generalist baseline, yielding virtually identical results within expected run-to-run variance. This suggests that isolated artifact pairs may lack sufficient document-level context to resolve this subtle boundary, regardless of focused re-ranking. In contrast, an uncalibrated two-level cascade (E3b) degrades performance severely, falling to a macro-F1 of 0.374. This cascading failure happens because the first stage (differentiating `related_papers` vs `unrelated_papers`) achieved a weak macro-F1 of 0.55, while the second stage achieved ~ 0.75. The poor initial routing meant candidate pairs were incorrectly discarded early, causing these early-stage routing errors to propagate irreversibly.

6 Results and Analysis

6.1 Architectural Comparisons and Adaptation Ablations

Across the evaluated paradigms (Table 3), zero-shot embedding similarity performs worst, while supervised bi-encoders and cross-encoders markedly improve results. However, neither matches the performance of generative VLMs. Within the Qwen family, the newer Qwen3.8-27B consistently outperforms earlier backbones, including both the 8B model and the nominally larger Qwen3-VL-32B, benefiting from stronger foundational pretraining.

When adapting Qwen3.8-27B, ablations over LoRA rank and precision (4-bit vs. 16-bit) remained within a narrow score band. Given the limited size of the dataset, we cannot conclusively determine whether scaling rank or precision systematically improves performance or simply reflects optimization noise. Under these data-constrained conditions, standard 4-bit LoRA provides an effective, memory-efficient baseline.

6.2 Modality-Level Breakdown

Validation results follow a strict hierarchy: `tt` $>$ `ti` $>$ `ii` (Table 4). Text–Text (`tt`) pairs benefit from explicit domain terminology (e.g., coordinates and instruments), while Text–Figure (`ti`) pairs effectively ground text into visual elements.

Modality	Val mF1	Val Acc.	n
Overall	0.7439	0.7167	1,500
Text–Text (`tt`)	0.7798	0.7844	450
Text–Figure (`ti`)	0.7014	0.7067	600
Figure–Figure (`ii`)	0.6635	0.6622	450

Table 4: Modality-stratified validation metrics for Qwen3.8-27B.

Conversely, Figure–Figure (`ii`) pairs are the most difficult because independent studies often share standardized visualization conventions (e.g., contours, spectra, and color-magnitude plots) that create spurious visual similarity.

6.3 Class-Level Performance and Error Analysis

Validation confusion matrices show that errors are heavily concentrated along specific boundaries. While `same_figure` (via layout continuity) and `unrelated_papers` (via lexical divergence) are classified reliably, distinguishing `same_paper` from `related_papers` causes the vast majority of misclassifications. Papers that cite each other or target identical astronomical objects share nearly identical vocabulary and plotting styles, making isolated artifact pairs fundamentally ambiguous without broader document context.

7 Conclusion and Future Work

We investigated an empirical progression of multimodal architectures for scientific citation linking on AstroCLIMB. While metric dual encoders are constrained by cross-modal representation gaps, parameter-efficient fine-tuning of generative vision–language models achieves competitive performance, reaching approximately 0.74 test Macro-F1. Our experiments demonstrate that model performance is not primarily limited by architectural capacity, but rather by visual ambiguity in Figure–Figure pairs and the subtle semantic boundary separating artifacts from the same paper versus related papers.

Future work includes leveraging full-paper metadata, domain-specific vision encoder pretraining, and exploring dedicated decision models (e.g., KEV/JEV) for routing, enabling more reliable model selection and reducing early routing errors.

8 Limitations

Lack of Document Context: Looking at artifact pairs alone is not enough. Without surrounding text, methodology, or author metadata, telling apart *same paper* from *related papers* is very difficult—especially when two papers study the same celestial object or use the same data pipeline.

Visual Domain Gap on Figure–Figure (`ii`) Pairs: Standard vision models are not trained on astronomy plots. On `ii` pairs, the vision encoder often focuses on shared plot styles (such as colorbars, axes, and line layouts) instead of the actual science, causing false matches between unrelated papers.

Simple Decision Rule and Cascade Failures: Our system uses a simple argmax rule over logits rather than an adaptive decision model. As our cascade test (E3b) showed, hard multi-stage decisions fail easily: errors in the first step permanently discard correct pairs. A better decision model must adjust confidence thresholds for each modality, since performance varies widely ($tt > ti > ii$).

Inference Speed and Scalability: Although 4-bit LoRA allows training on a single GPU, running a 27B model is still slow and resource-heavy. This makes it difficult to scale the system across millions of papers in the NASA ADS.

Lack of Explanations: Taking predictions directly from the first token makes inference fast, but the model cannot output reasoning steps or explain why two items are linked.

Domain Focus: Our findings are specific to astronomy publications. They may not directly apply to fields like biomedicine or computer science, which use different types of figures and citation styles.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *Preprint*, arXiv:1903.10676.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, and 16 others. 2024. Paligemma: A versatile 3b vlm for transfer.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *NeurIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Felix Grezes, Thomas Allen, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J. Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S. Grant, Donna M. Thompson, Timothy W. Hostetler, Matthew R. Templeton, Kelly E. Lockhart, Shinyi Chen, Jennifer Koch, Taylor Jacovich, and Pavlos Protopapas. 2022. Improving astrobert using semantic textual similarity.

Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J. Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S. Grant, Donna M. Thompson, Roman Chyla, Stephen McDonald, Timothy W. Hostetler, Matthew R. Templeton, Kelly E. Lockhart, Nemanja Martinovic, Shinyi Chen, Chris Tanner, and Pavlos Protopapas. 2021. Building astrobert, a language model for astronomy astrophysics.

Felix Grezes, Kelly Lockhart, Alberto Accomazzi, Atilla Alkan, Ethan Seefried, and Tirthankar Ghosal. 2026. 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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features.

Unsloth. 2024. Unsloth: Fast finetuning for large language models. `https://github.com/unslothai/unsloth`.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11985.

Multimodal Feature Integration for Pairwise Relationship Classification in Astronomy Literatures

Raj Mane[1], Ananya Kalia[1], Aashna Shetty[1], Rutal Mahajan[2], Preksha Pareek[1]

[1]Department of Computer Engineering,
Thakur College of Engineering and Technology, Mumbai, India
[2]Department of Computer Science,
Utica University, USA

Abstract

In scientific articles, figures often convey experiments and observations, as well as physical intuition, to the readers. Yet, surface metadata alone can be insufficient to unpack how two scientific objects relate to each other. We explore this with AstroCLIMB, which infers relationships between pairs of figures and metafigures based on four classes: same figure, same paper, related papers and unrelated papers. We design a multimodal pairwise pipeline that consolidates the textual visual semantic, and domain-specific signal sources from object pairs. TF-IDF and SBERT text and image-based text-image feature pairings, combined with lexical overlap, entity, and instrument overlap, numerical overlap, and interaction features conditioned on methods, comprise the feature set fed into a weighted ensemble of LightGBM, XGBoost, and CatBoost trained with 313 engineered features on five-fold stratified cross-validation, optimizing for class-specific thresholds to produce a Macro-F1 of 0.5016 on out-of-fold data. Results show that yet stronger textual and multimodal counterpart representations exhibit complementary yet useful pairwise signals, although establishing related papers still remains a significant challenge.

1 Introduction

Scientific publications are multimodal documents, which include prose, figures, tables, and captions that communicate significant information (Kashyap et al., 2023). Figures communicate observations, model results, measurements, and comparisons and can be central to any scientific work.

Recently, there have been some work on scientific figure captioning, figure extraction, multimodal scientific understanding, scientific information retrieval, and citation prediction. AstroCLIMB considers a different task: instead of generating a caption, retrieving a document, or answering a question on a single figure, it classifies the type of relationship between a pair of scientific objects. AstroCLIMB offers four symmetric relationship types namely same figure, same paper, related papers, and unrelated papers, and attempts to recover, in a limited fashion, the citation graph of astronomy publications using figures and captions rather than relying solely on standard metadata. The method uses several visual textual semantic, and domain-specific signals in a single feature-based classification method.

It then generates pairwise features based on overlap in lexical units, TF-IDF and SBERT similarity, CLIP-based text and image representations, overlap in astronomical entity and instrument concepts, numerical similarity, length of objects, sentence-based features, and modals. The trained classification model is a weighted meta ensemble of LightGBM, XGBoost, and CatBoost classifiers, exploiting 313 engineered features among the evidence sources. The model is assessed by five-fold stratified cross-validation, with an out-of-fold Macro-F1 of 0.5016 after improvement of the decision-thresholds for each class.

2 Related work

2.1 Scientific Figure Understanding

Figure captioning for scientific figures has demonstrated that there is information in figures that might not be captured in author captions. SciCap established a large-scale benchmark for scientific figure captioning and demonstrated both the potential and the challenges of generating an informative scientific caption (Hsu et al., 2021). SciCap+ demonstrated that figure-mention paragraphs and OCR information can be used for captioning, and so the importance of surrounding text (Yang et al., 2024). Summarization-based captioning in the same way uses the paragraphs that mention figures (Huang et al., 2023). Scientific figure captioning with large multimodal models was then

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 216–223
November 10, 2026 ©2026 Association for Computational Linguistics

explored and shown to be a challenging problem (Hsu et al., 2026).

Another task that we will need is figure extraction from scientific documents. Figure extraction is needed for caption generation. FigEx has investigated extraction and alignment of subfigures and subcaptions (Song et al., 2025) while SciFig offers large-scale data for scientific figure classification and understanding (Anonymous, 2021). These are related to AstroCLIMB as figure extraction and multimodal understanding is required for relation discovery, but in fact their main focus is on extraction, captioning, or classification of figures, rather than predicting the relationship between two scientific figures.

2.2 Scientific Multimodal Learning

Multimodal ArXiv offers scientific large-scale image-text data and demonstrates how domain-adaptive training can cause scientific understanding, but there are still differences in capturing subtle scientific semantics (Li et al., 2024). SciMMIR benchmarks sci-multimodal scientific retrieval and compare vision language models under the zero-shot and fine-tuned settings (Wu et al., 2024). SciVQA tests reasoning on scientific figures, the results show that there is still room to improve on visual reasoning tasks, as the current strong multimodal systems do not yet outperform humans on this task (Borisova et al., 2025). The experiments above emphasize the importance of representations that are adapted to scientific content, rather than solely relying on general web training of vision-language models.

2.3 Scientific Similarity and Citation Analysis

SPECTER Learning document representations using citation information and handling tasks such as citation prediction and classification (Cohan et al., 2020). Scientific RepEval: a system for evaluating representations of scientific documents over multiple tasks (Singh et al., 2023). Multi-vector methods learn fine-grained document similarities and aspect-based similarity learning directly from citation sections (Mysore et al., 2022; Ostendorff et al., 2020). Citation recommendation methods use document contents or citation context to recommend relevant scientific papers (Bhagavatula et al., 2018; Duma et al., 2016). These techniques demonstrate the importance of scientific relations and domain-specific knowledge. Yet, most of these methods are based on text and never deal with classifying relationships between scientific figures directly by integrating both visual and text methods.

2.4 Vision-Language Models

Recently, a lot of state-of-the-art approaches have been proposed that seamlessly connect visual and textual ways. CLIP learns text and image representations aligned with large scale image-text data (Radford et al., 2021). SigLIP learns image and text using a sigmoid loss for the image-text relation (Zhai et al., 2023). BLIP-2 connects a frozen visual encoder with a language model through a Querying Transformer (Li et al., 2023), LLaVA combines a vision encoder with a large language model to follow instructions from visual input (Liu et al., 2023). ColPali learns visual document representation and late interaction for document retrieval (Faysse et al., 2024). These approaches gave us some clues on multimodal learning but ultimately AstroCLIMB is task-specific, which is to classify the relation types between astronomy figure-text pairs.

3 Methodology

The proposed system performs this multimodal relationship classification of between pairs of astronomy figures and their respective captions. The system flow is illustrated in Figure 1.

3.1 Extraction and Preprocessing

The data is a collection of astronomy figures subpanels captions, and text data from the NASA ADS and AstroCLIMB collection (Grezes et al., 2026). Textual data is stored as UTF-8 string and subject to simple normalization as well as whitespace handling before text feature extraction.

For each object, we obtain characters words sentences, references to figures, tables, and equations. Astronomy-specific features are obtained from catalogue and object identifier regular-expression patterns, which include NGC IC M, HD HIP Kepler, Wasp, and other common catalogue and object identification conventions, and for instruments and surveys, by matching text to a list of instrument and survey names, including JWST, Hubble, Chandra, ALMA, TESS, Gaia, and SDSS. Identifiers and instrument names are normalized for consistent matching. For each pair, we then extract shared identifiers and shared instrument labels as well as their Jaccard similarities, and the number of numerical entries in common. These features work as explicit features for the presences of common objects

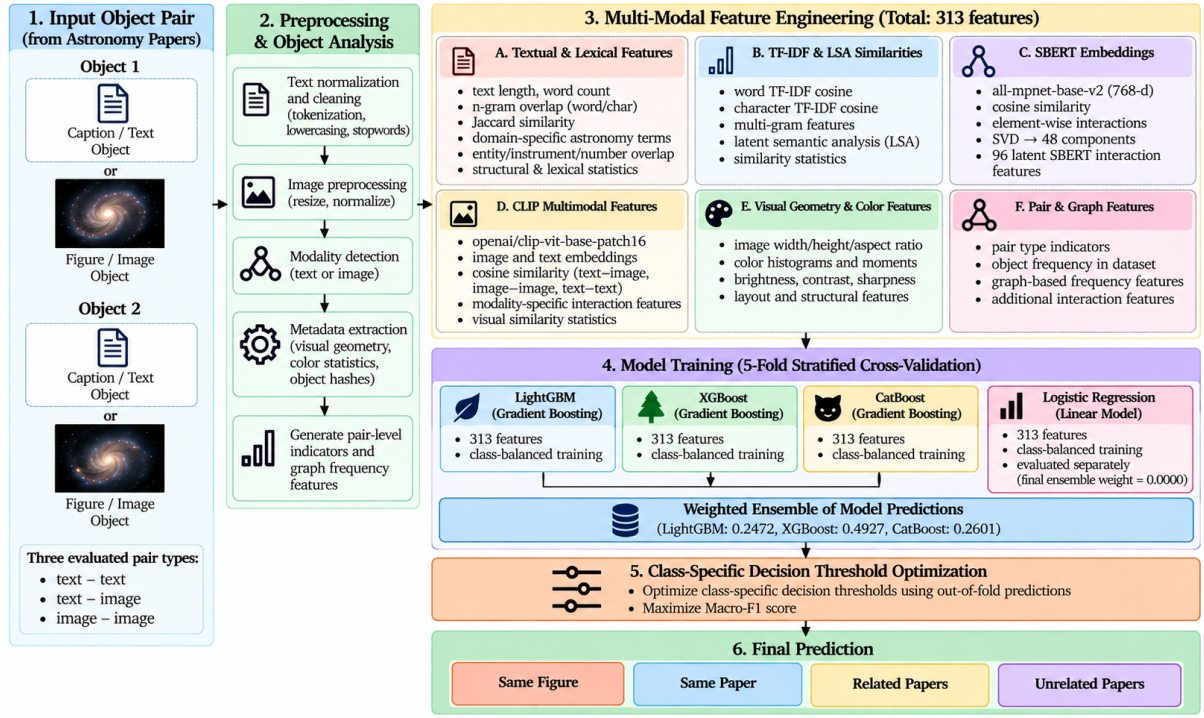


Figure 1: Implemented AstroCLIMB pipeline showing pairwise feature extraction, modality-aware interactions, weighted ensemble classification using LightGBM, XGBoost, and CatBoost, and class-specific threshold optimization.

and common observational instrumentation. We also extract object-hash frequency features from the available train and test pairs. For each pair, we extract the minimum and maximum of the two object-hash frequencies to provide an idea of the frequency of the two objects being paired.

We augment the lexical features with pretrained semantic encoders. For text, SBERT gives us 768-dimension text embeddings with the `all-mpnet-base-v2` checkpoint. For visual and cross-modal encodings, we use CLIP with the `openai/clip-vit-base-patch16` checkpoint, which gives us 512-dimension projected embeddings. The TF-IDF features encode the lexical similarity, while the SBERT embeddings measure semantic similarity. The TF-IDF vocabulary is fitted on the concatenated train and test corpus without the class labels.

Before CLIP text encoding each string is limited to 350 characters. The CLIP text encoder defaults to truncating the input to 77 tokens (including any special tokens required by the encoder), due to the context-length restriction of the encoder. As a result, information provided later in longer captions may not be encoded by CLIP. Both CLIP encoders are used in their original, pretrained state (i.e. without fine-tuning on the AstroCLIMB dataset). In the feature-extraction run reported here, CLIP text embeddings were computed for 14,581 text instances, while CLIP image embeddings were computed for 14,598 images. All images were successfully decoded in this run.

For images, we use CLIP image features and compare them with the paired object representations. This is image-image or cross-modality text-image matching.

Lastly, we explicitly encode the modality (text or image) of each object, which means that the pairwise feature construction stage can identify whether a pair is text-text, text-image, or image-image and include this information when constructing interaction features for the final classification.

3.2 Pairwise Feature Construction

For each pair of objects, the system generates a 313-dimensional feature vector of combined textual, semantic, visual, structural, and modality-aware features. The set of features contains the TF-IDF and SBERT similarities, CLIP-based text and image features, lexical overlap, features specific to astronomy, and the interactions of the compressed embeddings, aiming to combine complementary signals between the paired objects.

Object-level statistics object length, number of

words, number of sentences; Structural markers references to figure, table, and equation; Domain-oriented features number of astronomical objects and instruments in each object, and shared number of astronomical objects, shared astronomical instruments, and shared numerical values between the two objects.

Lexical overlap is measured with word-level Jaccard similarity, word overlap count, character 3-gram Jaccard similarity, and the fraction of common words compared to the shorter object. Entity, instrument, and numerical overlaps are captured using both shared-count and Jaccard-based features.

The pipeline explicitly encodes whether each object is an image and whether the pair is a mixed-modality pair. These semantic similarities are combined with modality indicators to create interaction features: TF-IDF, SBERT, and CLIP similarities are multiplied by indicators for each of the four possible modality configurations (text-text, text-image, image-text, image-image). The evaluated dataset contains text-text, text-image, and image-image pairs; no image-text pairs are present in the evaluated instances. Additional difference features, including SBERT minus TF-IDF, CLIP minus SBERT, and CLIP minus TF-IDF, are included to capture disagreement between the different representation spaces. The average and maximum text similarity signals are included as aggregate semantic features.

3.3 Ensemble Classification

The 313-dimensional pairwise feature vectors act as input to an ensemble of LightGBM, XGBoost, and CatBoost classifiers. A linear classifier is also tested, but is assigned zero weight when forming the ensemble. The training set includes 10,000 labeled instances sorted into four classes of relationship: 1,000 same-figure 3 000 same-paper 3 000 related-paper, and 3,000 unrelated-paper.

The optimal parameters used for the LightGBM classifier are a gradient-boosted decision tree with a learning rate of 0.03, 127 leaves, maximum depth of 9, row subsample of 0.8, feature subsample of 0.7, and up to 3,500 estimators. For XGBoost the parameters are a learning rate of 0.03, maximum depth of 8, minimum child weight of 15, row subsample of 0.8, feature subsample of 0.7, and up to 3,500 estimators. For CatBoost these are a learning rate of 0.03, depth 7 3 500 maximum iterations, and an L2 leaf reg regularization coefficient of 4.0. Early stopping with a patience of 150 rounds is set for all three tree-based classifiers. Weights for class imbalance are used in CatBoost, while sample weights are applied for LightGBM and XGBoost.

Table 1: Macro-F1 scores of individual classifiers and the weighted ensemble before and after threshold optimization.

Model / Configuration	**Macro-F1**
LightGBM (mean fold score)	0.48358
XGBoost (mean fold score)	0.49150
CatBoost (mean fold score)	0.47694
Linear classifier (mean fold score)	0.48082
Weighted ensemble (raw OOF)	0.48846
Weighted ensemble (threshold-optimized OOF)	**0.50163**

For scoring, 5-fold stratified cross validation is used to estimate out of fold performance with random seed 42. The predictions of the three models used contribute to the ensemble and are combined with weights chosen from out of fold predictions. Class specific decision multipliers are then multiplied into the ensemble scores to optimize about Macro-F1. In Table 1 ,the final macro-F1 after threshold improvement is presented as 0.50163 versus 0.48846 pre-tuning.

The final level evaluation is based on the out-of-fold predictions from all 5 folds. This way, each example in the train set has a prediction from a model that did not see that example during training.

4 Experiments and Results

The updated system is evaluated using five-fold stratified cross-validation on 10,000 labeled training instances. The dataset contains four relationship classes: same figure (1,000 instances), same paper (3,000 instances), related papers (3,000 instances), and unrelated papers (3,000 instances). Each fold evaluates the models on held-out instances that are excluded from the corresponding training phase.

Four classifiers are evaluated: LightGBM, XGBoost, CatBoost, and a linear classifier. Their out-of-fold predictions are used to determine the ensemble weights. The final ensemble assigns weights of 0.2472 to LightGBM, 0.4927 to XGBoost, and 0.2601 to CatBoost. The linear classifier receives zero weight in the final ensemble. Thus, the final predictions combine the outputs of the three contributing tree-based models.

The individual-classifier results show the model level comparison when using the same feature space. In Table 2, XGBoost scored the highest

Table 2: Per-class performance of the threshold-optimized ensemble on out-of-fold predictions.

Class	Prec.	Recall	F1
Same Figure	0.4612	0.7310	0.5656
Same Paper	0.6123	0.4170	0.4961
Related Papers	0.4071	0.4153	0.4112
Unrelated Papers	0.5086	0.5613	0.5337
Macro Avg.	0.4973	0.5312	0.5016
Weighted Avg.	0.5045	0.4912	0.4888

mean five-fold validation Macro-F1 (0.49150), followed by LightGBM (0.48358), the linear classifier (0.48082), and CatBoost (0.47694). The weighted ensemble resulted in an OOF Macro-F1 of 0.48846, which increased to 0.50163 following class-specific threshold tuning. This class-specific threshold improvement results helps spell out the effect of classifier choice and ensemble calibration; they do not yet show the effects of the individual feature ways.

The experiments above report performance for the ensemble with five-fold out-of-fold predictions. The raw weighted ensemble obtains a Macro-F1 of 0.48846, rising to 0.50163 after class-specific decision-multiplier tuning. These results provide performance benchmarks for the specified system; other than the projected TF-IDF-only and CLIP-only baselines, we cannot compare majority-class predictions under the same validation regime to measure the additive contribution of the whole multimodal feature set. For this reason we consider these results an experimental validation of the specified pipeline and not as a controlled experiment demonstrating the individual benefit of each modality.

Class-specific threshold tuning results before class-specific decision thresholds tuning, the ensembleout-of-fold Macro-F1 is 0.48846. When using class-specific decision multipliers, the threshold-optimized Macro-F1 is 0.50163. The absolute gain is 0.01317. The results show that class-specific decision thresholds improvement can improve the class balance of the ensemble for the four classes.

The threshold-optimized results show that the same-figure class obtains the highest F1 score at 0.5656, followed by unrelated papers at 0.5337 and same paper at 0.4961. Related papers remain challenging, with an F1 score of 0.4112. The same-figure class has the highest recall (0.7310), whereas the same-paper class has the highest precision (0.6123). This illustrates the different precision-recall trade-offs across the relationship categories.

The threshold-optimized classification report gives an overall accuracy of 0.4912. The Macro-F1 score is 0.50163, while the weighted-average F1 score is 0.4888. Macro-F1 is used as the primary metric because it gives equal importance to each relationship class despite the difference in class support.

The final submission contains 10,000 test predictions, with one predicted class assigned to each instance. The predicted distribution is 1,309 same-figure instances (13.09%), 1,657 same-paper instances (16.57%), 3,474 related-paper instances (34.74%), and 3,560 unrelated-paper instances (35.60%). The submission file and the corresponding soft predictions were generated by the final pipeline.

In addition to cross-validation evaluation, the proposed system was evaluated in the AstroCLIMB Kaggle competition. At the time of recording the leaderboard results, the submission ranked 10th with a score of 0.48121. This competition result is reported separately from the out-of-fold Macro-F1 scores because the two results arise from different evaluation settings. The competition leaderboard is available at https://www.kaggle.com/competitions/astroclimb/leaderboard.

5 Error Analysis and Discussion

Class-by-class results reveal that the four relation classes display varying precision-recall behavior. The same-figure class has the highest F1 of 0.5656, but it also has a recall of 0.7310. This shows the threshold-optimized model is capturing many same-figure pairs, but with a precision of 0.4612, some of the pixels predicted to belong to this class are actually false positives.

The most precise class is the same-paper class at 0.6123, but the class recall is low at 0.4170. This shows that predictions made for this class are relatively stringent, although we are assigning a fair number of actual same-paper relations to other classes. It may be more difficult to differentiate same-paper versus other links without specific evidence of shared source documents.

Related papers is also an ambiguous class with an F1 score of 0.4112. Truly related papers can mention similar scientific terms, astronomical objects instruments observations, and research goals without being the same paper (or cited by the same paper). Because of this, using lexical or seman-

tic similarity on its own may not be sufficient to tell related-paper relationships apart from same- or unrelated-paper relationships.

F1 for the unrelated-papers class is 0.5337, with precision 0.5086 and recall 0.5613. Of course, two papers from different sources may still talk about the same objects, instruments, measurements, and scientific things. So high semantic similarity does not mean that two objects must be related to each other.

The final ensemble combines LightGBM, XGBoost, and CatBoost, with XGBoost receiving the largest weight. By combining their predictions, the system uses the complementary decision patterns of the three classifiers. Class-specific decision multipliers further improve the out-of-fold Macro-F1 from 0.48846 to 0.50163. Yet, since these multipliers are optimized using out-of-fold predictions, the threshold-optimized score should be interpreted in this improvement procedure, rather than as an independent estimate of generalization performance.

Overall, the findings indicate that integrating lexical semantic visual, structural, and astronomy-specific features can aid in multimodal relationship classification. Yet, the observed differences in precision and recall across classes show that accurately identifying the correct relationship involves more than simply assessing similarity between paired objects.

The system uses lexical, semantic, visual, structural, and astronomy-oriented features. These groups of features complement each other though we do not attempt to measure the impact of each group of features in isolation by 'ablating' features in a controlled manner. As such, these results should not be used so that each group of features contributes to the improvements in Macro-F1.

Object length, number of words, and number of sentences are cheap content-based indicators that could signal whether an object has lots of text and might this way help differentiate between text-rich objects and short captions, images, or other objects. But these statistics could also be a function of object modality or formatting conventions, or properties of dataset construction rather than the relationship of interest. We can't claim that they encode the semantics of the relationship without a rigorous experimental test.

6 Conclusion

This work investigates multimodal relationship classification for astronomy figures and text using the AstroCLIMB benchmark. The proposed system combines lexical, semantic, visual, structural, and domain-oriented pairwise features, including TF-IDF and SBERT similarities, CLIP-based representations, modality-aware interaction features, astronomical entity and instrument overlap, numerical overlap, and object-level statistics.

The final system uses 313 engineered features and combines LightGBM, XGBoost, and CatBoost predictions through a weighted ensemble. Under five-fold stratified cross-validation, the ensemble achieves an out-of-fold Macro-F1 of 0.48846 before class-specific threshold optimization and 0.50163 after optimization. The per-class results show that same-figure relationships are identified with the highest F1 score, while related-paper relationships remain more difficult to classify.

These findings demonstrate the potential of combining complementary textual and multimodal signals for relationship classification. Future work can investigate OCR-enhanced figure representations, richer document structure, improved figure-caption alignment, and end-to-end vision-language models to better distinguish semantic similarity from direct relationships between scientific objects.

7 Limitations

Our system has several limitations. First, distinguishing related-paper relationships remains challenging, with an F1 score of 0.4112 for this class. Scientific papers from the same research area may share terminology, astronomical entities, instruments, observations, and semantic content even when their relationship is indirect.

Second, visual information is primarily represented through pretrained CLIP embeddings rather than explicit interpretation of plot components. The current pipeline does not explicitly perform OCR over axis labels, legends, tick values, or annotations. Consequently, visually similar figures with different numerical or textual content may remain difficult to distinguish.

Third, the system relies on engineered pairwise features and a weighted ensemble of LightGBM, XGBoost, and CatBoost. Although this approach combines several classifiers and different sources of information, it may not capture all fine-grained

relationships between figure structure and surrounding scientific text.

Finally, the reported threshold-optimized Macro-F1 is calculated from out-of-fold predictions after class-specific decision multipliers have been optimized. This optimization may make the reported score optimistic as an estimate of generalization performance. Independent evaluation of the selected thresholds would provide a more reliable assessment. The pipeline also depends on the quality and completeness of the available textual and visual data; missing, incomplete, or weakly informative captions can reduce the usefulness of lexical and semantic features.

Future work can investigate OCR-enhanced visual representations, richer document structure, improved figure-caption alignment, and end-to-end vision-language fine-tuning.

8 Acknowledgment

This paper was prepared for the AstroCLIMB shared task at WASP 2026. We thank the organizer and reviewers. The implementation and benchmark results described here are based on the project model and evaluation information supplied for the research study.

References

Anonymous. 2021. Scifig: A scientific figure dataset for figure understanding. OpenReview submission.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 238–251.

Elena Borisova, Nina 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*, pages 182–210.

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.

Daniel Duma, Maria Liakata, Amanda Clare, James Ravenscroft, and Ewan Klein. 2016. Applying core scientific concepts to context-based citation recommendation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 1737–1742.

Mickael Faysse, Hugues Sibille, Tong Wu, Behrooz Omrani, Gregory Viaud, Celine Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Felix Grezes, Kelly Lockhart, Alberto Accomazzi, Atilla Alkan, Ethan Seefried, and Tirthankar Ghosal. 2026. 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.

Ting-Yun 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.

Ting-Yun Edward Hsu, Yu-Lin Hsu, Saksham Rohatgi, Chia-Yu Huang, H. Y. S. Ng, Ryan Rossi, Seung Kim, Tong Yu, Lun-Wei Ku, C. Lee Giles, and Ting-Hao Kenneth Huang. 2026. Do large multimodal models solve caption generation for scientific figures? lessons learned from scicap challenge 2023. *Transactions of the Association for Computational Linguistics*, 14:233–252.

Chieh-Yang Huang, Ting-Yao Hsu, Ryan Rossi, Ani Nenkova, Sungchul Kim, Gromit Yeuk-Yin Chan, Eunyee Koh, Clyde Lee Giles, and Ting-Hao Huang. 2023. Summaries as captions: Generating figure captions for scientific documents with automated text summarization. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 80–92.

Abhishek Kashyap, Zhen Yang, and Min-Yen Kan. 2023. Scientific document processing: Challenges for modern learning methods. *International Journal on Digital Libraries*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Lei Li, Yufei Wang, Ruochen Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 14369–14387.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4453–4470.

Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, and Georg Rehm. 2020. Aspect-based document similarity for research papers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6194–6206.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. Scirepeval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566.

Jinyu Song, Abhishek Das, Guanhua Cui, and Yufeng Huang. 2025. Figex: Aligned extraction of scientific figures and captions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16558–16571.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, Noura Al Moubayed, Jie Fu, and Chenghua Lin. 2024. Scimmir: Benchmarking scientific multi-modal information retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12560–12574.

Zhen Yang, Raj Dabre, Hiroki Tanaka, and Naoaki Okazaki. 2024. Scicap+: A knowledge augmented dataset to study the challenges of scientific figure captioning. *Journal of Natural Language Processing*, 31(3):1140–1165.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Scientific Table Generation: Structured Information Extraction from Scientific Literature

Necva Bölücü, Huichen Yang*, Roelien C. Timmer, Stephen Wan,
Brian Jin, Changhyun Lee, Andreas Duenser, Jessica Irons
CSIRO, Australia
{necva.bolucu, huichen.yang, roelien.timmer, stephen.wan,
brian.jin, aiden.lee, andreas.duenser, jessica.irons}@csiro.au

Abstract

We introduce *Scientific Table Generation (STG)*, the task of extracting information from scientific documents to populate a table. We argue that STG can serve as a unifying task for structured knowledge extraction from scientific literature. To support this goal, we present the first multi-disciplinary benchmark for STG with predefined schemas and multi-record extraction, covering six scientific domains and comprising 20 datasets unified under our proposed formulation. We detail the benchmark's construction, including dataset composition, and introduce an *STG evaluation protocol*. We analyse the results of zero-shot large language model (LLM) and vision-language model (VLM) pipelines with typical document processing AI methods for the STG task. In doing so, we identify key challenges of the STG task and propose a set of open challenges, including multimodality, table and figure understanding, normalisation of data, and treatment of reported quantities. All code and data accompanying this work are publicly available at[1]: https://github.com/bolucunecva/table_generation

1 Introduction

The exponential growth of scientific literature has made it increasingly challenging for researchers to extract and organise key findings efficiently (Cai et al., 2024; Reddy and Shojaee, 2025). Manual methods, once manageable, are now overwhelmed by the sheer volume of publications (Borah et al., 2017; Cai et al., 2024). This has created a niche for automated tools (e.g., Elicit[2] and SciLire (Bölücü et al., 2026)), that can extract and structure information from scientific texts.

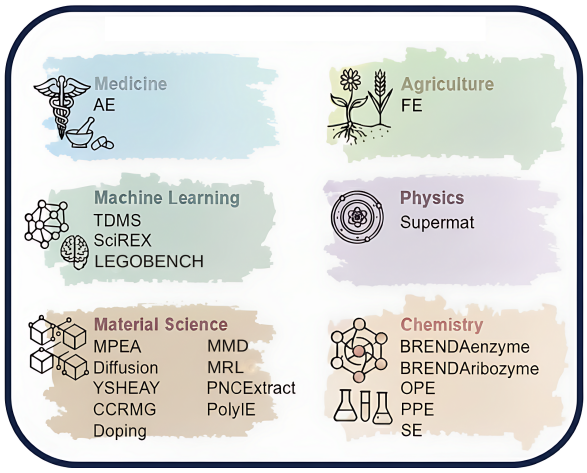


Figure 1: Benchmark covers six scientific domains: *Medicine*, *Agriculture*, *Machine Learning*, *Physics*, *Material Science*, and *Chemistry*.

Information extraction (IE) provides a foundational capability for this transformation by converting knowledge from unstructured sources into structured data (e.g., scientific publications, experimental notes). Such derived data can accelerate knowledge discovery (Horawalavithana et al., 2022), support predictive modelling (Li et al., 2022), and help in the development of AI-powered scientific assistants that can help researchers in summarising and reasoning about large scientific corpora (Lu et al., 2024; Su et al., 2024; Kulkarni et al., 2025).

While many recognise the need for IE solutions for scientific literature, there has not been a strong consensus in the problem framing. This has led to the exploration of this topic under various names within the Natural Language Processing (NLP) community, causing fragmentation and inconsistency across studies. Different terms have been used, such as joint named entity recognition and relation extraction (Townsend et al., 2021; Huguet Cabot and Navigli, 2021), n-ary relation

* This work was done when the author was affiliated with CSIRO.

[1]This repository will save researchers significant time. Instead of manually searching for 20 different datasets, researchers can use our repository to download all datasets directly, as a download link is provided for each paper. In addition, researchers can use our evaluation framework, which streamlines the process and improves overall efficiency.

[2]https://elicit.com

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 224–246
November 10, 2026 ©2026 Association for Computational Linguistics

extraction (Jia et al., 2019; Khalighinejad et al., 2024), Unified Text-to-Schema (Song et al., 2023), structured IE (Dunn et al., 2022), structured information inference (Xie et al., 2023b), ad-hoc schema-based IE (Ghosh et al., 2024), and instruct-and-extract (Jiao et al., 2023). These studies introduce separate datasets, differing in format and evaluation strategies, making it difficult to compare approaches or build systems applicable to all tasks. Moreover, most focus on specialised scientific domains, e.g., materials science (Ghosh et al., 2024; Cheung et al., 2024) and enzymes (Jiang et al., 2025).

We note that the existing common goal in these works remains the same: to extract structured data aligned with a predefined schema, a format into which the extracted information is organised. This is considered a comprehensive-level task (identifying, extracting and understanding the core content of documents) (Krathwohl, 2002) in scientific knowledge benchmarks (Feng et al., 2024; Cai et al., 2025).

In this work, we introduce the term **Scientific Table Generation** (STG), which formalises the task as extracting key values from unstructured scientific documents and organising them into a predefined tabular schema. Noting that scientists often use tables as a data format, we adopt a table-centric approach, as it offers a clear, consistent, and interpretable representation of scientific findings, enabling easier integration and downstream applications (Jiang et al., 2025; Wei et al., 2025).

To support research for this unified task, we curate the first *multi-disciplinary* benchmark (see Figure 1), standardising datasets introduced in prior work into a common table-centric format. For our selection criteria, we focused on (1) responsible AI datasets compliant with FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Wilkinson et al., 2016); (2) document-level extraction of records (records filled with information dispersed throughout a document); and (3) records with more than two fields.

Our assembled benchmark spans 6 scientific domains[3] and includes 20 datasets, reflecting the broad applicability and diversity across scientific domains. In addition, we document an evaluation framework for STG to help evaluate systems across different scientific domains. In this paper, we dive into the nature of the STG task, identifying specific challenges, e.g., inconsistencies in value formatting (e.g., units of measurement) and entity disambiguation. These insights lead us to propose STG open research topics.

The overarching goal of this paper is to establish a robust benchmark for STG, enabling consistent and rigorous evaluation of methodologies developed for this task. Correspondingly, our paper focuses on three primary contributions. First, we present the unified task formalism of STG (§ 3) and describe the benchmark (§ 4), whereby each dataset is reformatted to align with a common STG schema and includes: (i) *links* to documents in PDF format (e.g., DOI, PubMed, or ArXiv IDs); and (ii) ground truth tables aligned with the schema. Second, to facilitate comparisons, we also introduce an evaluation protocol (§ 5) that is customised for the STG task, such as mechanisms to handle challenges in scientific IE, such as varying units, format variations, and multiple records per document. Finally, our third contribution is a characterisation of the technical challenges of the STG task, as revealed by applying baseline LLM and VLM pipelines to the benchmark data. We analyse the inference results to gauge the relative differences across scientific domains and insights about the complexities of the task.

2 Related Work

Scientific IE STG is part of a broader set of IE tasks from scientific documents, such as extracting keyphrases and relations from publications (Augenstein et al., 2017), measuring science evolution via citation and entity extraction (Jurgens et al., 2018), classifying scientific discourse sentiments (Teufel et al., 2009), and automatically constructing scientific knowledge graphs (Ammar et al., 2018). These examples represent a small sample, with STG contributing by focusing on structured extraction and reconstruction of scientific tables reporting experimental results. ArxivDIGESTables (Newman et al., 2024) studies scientific table generation for literature-review synthesis, where both the schema and the values are generated with one row per paper

In our work, we focus on table-centric document IE, a more general problem that subsumes n-ary relation extraction. To focus on the more challenging general problem, we move on from the well-explored space of binary relation extraction

[3]Material Science appears as the dominant class, but is an interdisciplinary field spanning physics, chemistry, biology, and engineering.

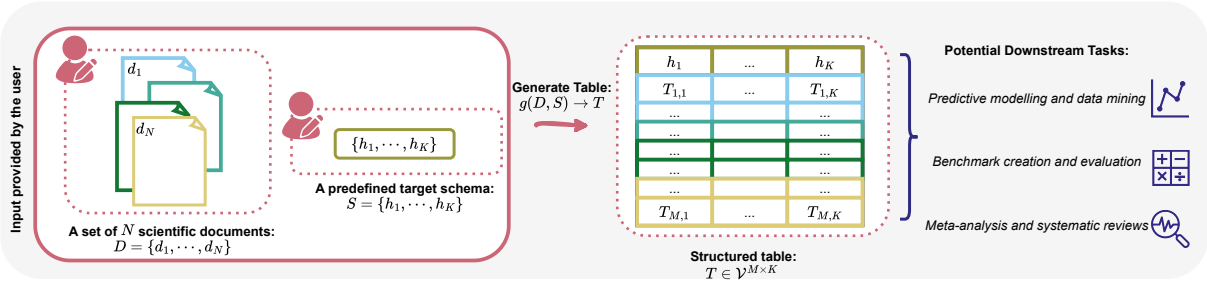


Figure 2: Visualisation of the Scientific Table Generation (STG) Task. First, the user provides N scientific documents $D = \{d_1, d_2, \cdots, d_N\}$ and a predefined target schema $\{h_1, \cdots, h_K\}$. Second, a structured table is generated: $g(D, S) \rightarrow T \in \mathcal{V}^{M \times K}$. The generated tables can be used for many downstream tasks, such as *predictive modelling and data mining*, *benchmark creation and evaluation*, and *meta-analysis and systematic reviews*.

and entity extraction, and specialised n-ary relation extraction instances that have been well explored in prior NLP work. That is, in favour of multi-field records (the rows in the table), we excluded binary relation datasets (e.g. SciERC (Luan et al., 2018) and DocRED (Yao et al., 2019)), abstract-level datasets (e.g. SemEval-2018 (Gábor et al., 2018), BioRED (Luo et al., 2022), and DWIE (Zaporojets et al., 2021)). We also exclude datasets targeting unrelated tasks such as classification or regression (e.g., Darwin (Xie et al., 2023a)).

Scientific Knowledge Benchmarks The presented benchmark is not the first benchmark proposed for evaluating scientific knowledge tasks. Other benchmarks include SciAssess (Cai et al., 2025), SciKnowEval (Feng et al., 2024), SciFact (Wadden et al., 2020), and SciBench (Wang et al., 2024a). SciAssess evaluates LLM performance on scientific literature analysis tasks, subdivided into *memorisation*, *comprehension*, and *analysis&reasoning*. SciKnowEval is similar, targeting scientific LLM tasks across difficulty levels: *studying extensively*, *inquiring earnestly*, *thinking profoundly*, *discerning clearly*, and *practising assiduously* (Feng et al., 2024). SciBench evaluates reasoning capabilities for complex scientific problems (Wang et al., 2024a), while SciFact focuses on verifying scientific claims based on abstracts (Wadden et al., 2020). In contrast, the *STG* benchmark uniquely targets the extraction of tables, which is not addressed by existing benchmarks.

3 Scientific Table Generation

Scientific Table Generation (STG) is the task of automatically generating a structured table of reported data from unstructured scientific publications conforming to a predefined schema, a structure that defines the format into which the extracted information is organised (Borg et al., 2020; Ghosh et al., 2024). The goal is to extract relevant information across different sections of a scientific document, representing these in tabular form, where each row is a scientific record, and each column corresponds to a schema attribute.

We formally describe the STG task as follows. Let $D = \{d_1, d_2, \cdots, d_N\}$ denote a collection of N scientific documents and $S = \{h_1, h_2, \cdots, h_K\}$ be a predefined target schema, where each h_j represents an attribute (or column) in the schema with a total of K columns. The STG task is the IE transformation, $g(D, S) \rightarrow T$, where T is a table of data of the form $T \in \mathcal{V}^{M \times K}$, where $M = \sum_{i=1}^{N} R_i$, where R_i is the total number of scientific records (i.e., rows) per document $d_i \in D$ that contributes to T.

Furthermore, each cell $T_{m,j}$ for row m (i.e. a single record) and column j, contains a string value, $V \in \mathcal{V}$. This value is a string that can be interpreted as a qualitative text answer, a reported quantity, or a category. Figure 2 shows a visualisation of the STG task, illustrating how information from scientific documents is mapped into a structured table consistent with the target schema. This formulation highlights two core challenges: (i) detecting and aggregating relevant evidence across potentially long and complex documents, and (ii) organising the extracted information into structured records that conform to a fixed schema.

4 Benchmark Construction

We introduce the benchmark, a suite of repurposed datasets initially proposed for varying related NLP tasks (e.g., IE, Q&A, . . .), covering 6 scientific domains: *Medicine*, *Agriculture*, *Machine Learning*, *Physics*, *Material Science*, and *Chemistry*.

Domain	Dataset	# Documents	Avg. Pages	# Records	Records / Document	Schema Size	Reference
Machine Learning	TDMS	332	10.52	904	2.72	4	Hou et al. (2019)
	SciREX	372	11.83	1,897	5.34	5	Jain et al. (2020)
	LEGOBENCH	6,995	14.63	145,419	20.79	5	Singh et al. (2024)
Material Science	MPEA	264	8.70	1,544	5.85	17	Borg et al. (2020)
	Diffusion	93	14.02	3,533	37.99	18	Zhang et al. (2010)
	YSHEAY	219	15.65	837	3.82	3	Polak and Morgan (2024)
	CCRMG	24	11.96	43	1.79	3	Polak and Morgan (2024)
	Doping	66	6.66	544	8.24	3	Dunn et al. (2022)
	MMD	3	6.00	140	46.67	16	Xie et al. (2023a)
	MRL	100	5.81	993	9.93	19	Xie et al. (2023a)
	PNCExtract	155	8.92	838	5.41	6	Khalighinejad et al. (2024)
	PolyIE	76	8.58	2,337	30.75	3	Cheung et al. (2024)
Chemistry	BRENDA-enzyme	155	12.45	4,210	27.16	13	Jiang et al. (2025)
	BRENDA-ribozyme	163	11.05	1,756	10.77	17	Jiang et al. (2025)
	OPE	104	7.55	255	2.45	7	Cai et al. (2025)
	PPE	109	6.97	265	2.43	6	Cai et al. (2025)
	SE	96	8.01	2,363	24.61	3	Cai et al. (2025)
Medicine	AE	40	4.30	406	10.15	3	Cai et al. (2025)
Agriculture	FE	326	11.65	326	1.00	15	Bölücü et al. (2025)
Physics	SuperMat	142	8.55	1,301	9.16	4	Foppiano et al. (2021)

Table 1: Statistics of 20 datasets across six domains. The schema of each dataset is given in the Appendix A in Table 3.

Dataset Selection Criteria This benchmark is built on datasets that meet several criteria: (i) document-level datasets (excluding sentence- or abstract-level IE data sets) and full PDFs can be obtained; (ii) adherence to the FAIR principles (Wilkinson et al., 2016) to support reproducibility and community use; and (iii) have n-ary relations that reflect the complexity of scientific knowledge.[4]

These datasets offer a blend of coarse- and fine-grained information, originating from either publication in PDF format or a structured document representation (e.g., XML) of a publication. For datasets that have only released structured representations (e.g., SciREX (Jain et al., 2020), MRL (Xie et al., 2023a)), we collect the corresponding PDFs via automatic scripts.

Design Choice For STG, we stipulate that the task should start with the PDF version of a publication. That is, a key design choice in building this benchmark is that we *do not force* a conversion of the PDFs into plain text, as this presumes some level of preprocessing that may become outdated in time. For example, as multimodal approaches become more advanced, we may *eventually* see a shift away from PDF parsing as a preprocessing step. To facilitate this, we distribute links to the PDF publication (e.g., DOI, ArXiv IDs). This decision has benefits: (i) the input can include non-textual elements (e.g., tables and figures) that are known to contain key information relevant to information reuse; and (ii) enables researchers to start with the original PDF, thereby encouraging transparency of approaches and reproducibility.[5]

Overview and Scope The final benchmark includes 20 datasets that meet our selection criteria. They cover a wide range of scientific concepts, for example, leaderboard construction (e.g., TDMS (Hou et al., 2019), LegoBench (Singh et al., 2024)), enzyme kinetics (e.g., BRENDA (Jiang et al., 2025)), and polymer collar cells (e.g., PolyIE (Cheung et al., 2024)), to name but a few research topics. The collection includes a variety of data types, covering both qualitative and quantitative information at various levels of detail. Table 1 provides an overview of the datasets included in the benchmark. Additional details on the schema structure, document sources, and domain can be found in Appendix A.

5 An Evaluation Protocol for STG

Evaluating IE from scientific documents presents unique challenges. Ideally, one would evaluate with human experts routinely (Dunn et al., 2022; Asai et al., 2024; Wang et al., 2024b); however, this is time-consuming and expensive. We note that scientists often judge the correctness of records as a

[4]Here we use the term n-ary relations to refer to a set of arguments, some of which might be subsumed by a common predicate while other arguments are merely associated as part of a discipline-specific IE task.

[5]Note that, due to copyright reasons, we do not distribute the PDFs as part of this dataset. Rather, we provide links and code for others to download the PDFs themselves.

Domain	Dataset	GPT-OSS				o1-mini				LLaMA 3.3				Nanonets (vision)			
		ChrF	Precision	Recall	F_1	ChrF	Precision	Recall	F_1	ChrF	Precision	Recall	F_1	ChrF	Precision	Recall	F_1
Machine Learning	TDMS	10.91	0.74	16.28	1.41	12.19	2.25	21.03	**4.07**	11.46	1.99	18.38	3.59	9.01	0.36	8.49	0.69
	SciREX	6.45	0.60	10.30	1.14	6.75	1.09	8.44	1.93	6.95	1.48	8.93	**2.54**	3.01	0.01	0.28	0.03
	LEGOBENCH	6.58	1.41	8.76	2.43	12.10	2.70	14.50	**4.55**	11.10	2.42	13.18	4.10	8.52	1.81	12.62	3.17
Material Science	MPEA	1.69	7.05	41.36	12.04	2.78	18.47	42.03	**25.66**	2.00	10.20	37.83	16.07	1.00	0.03	0.86	0.06
	Diffusion	0.74	1.12	19.06	2.12	0.94	22.67	13.20	**16.68**	0.88	10.52	10.16	10.34	0.44	1.06	10.46	1.92
	YSHEAY	13.19	0.57	23.58	1.11	16.31	2.86	44.29	**5.37**	14.25	1.82	34.37	3.47	3.99	1.40	2.51	1.80
	CCRMG	10.56	0.63	19.42	1.22	11.85	0.85	20.16	**1.63**	10.49	0.70	18.56	1.35	5.14	0.32	5.12	0.60
	Doping	10.19	4.83	20.15	7.79	14.96	23.30	19.16	**21.03**	12.52	13.38	20.83	16.29	2.17	1.21	2.77	1.69
	MMD	0.89	1.85	1.74	1.79	0.89	4.83	1.07	1.75	1.09	8.77	3.57	**5.08**	1.07	5.00	0.36	0.67
	MRL	0.57	1.28	3.00	1.79	0.73	2.86	1.25	1.74	0.72	2.44	2.82	**2.62**	0.71	0.38	2.22	0.65
	PNCExtract	8.74	6.52	44.15	11.37	8.64	23.96	39.14	**29.72**	9.40	11.73	27.96	16.53	5.22	0.42	13.17	0.81
	PolyIE	14.66	3.18	15.12	5.26	18.53	15.78	32.83	**21.31**	19.02	11.74	27.00	16.36	8.45	0.58	6.08	1.06
Chemistry	BRENDA-enzyme	2.39	11.01	27.64	15.75	2.55	24.30	26.67	**25.43**	2.53	11.07	30.72	16.28	1.21	0.75	8.22	1.38
	BRENDA-ribozyme	1.35	1.89	7.81	3.04	1.27	6.09	8.02	**6.92**	1.37	4.72	11.39	6.68	0.71	0.04	1.99	0.08
	OPE	3.90	2.06	34.40	3.88	4.75	7.08	38.60	**11.96**	4.76	4.66	37.70	8.29	2.99	0.07	10.81	0.14
	PPE	9.61	4.29	54.03	7.94	10.41	13.01	59.62	**21.37**	10.75	8.86	61.32	15.48	2.90	0.11	1.64	0.21
	SE	8.04	12.38	31.18	17.72	11.22	31.51	36.93	**34.01**	9.74	18.40	30.41	22.93	3.52	1.02	12.43	1.88
Medicine	AE	6.51	4.80	7.85	5.96	5.35	7.76	4.30	5.54	6.49	7.34	6.84	**7.08**	8.02	0.88	24.30	1.70
Agriculture	FE	2.22	0.85	36.58	1.67	2.52	1.89	28.45	**3.54**	2.50	1.49	28.20	2.83	1.40	1.63	5.46	2.52
Physics	SuperMat	6.53	11.91	26.84	16.50	10.05	33.28	29.72	**31.40**	9.04	12.49	19.23	15.15	2.94	1.70	8.46	2.83
	Avg.	6.29	3.95	22.46	6.10	7.74	12.33	24.47	**13.78**	7.35	7.31	22.47	9.65	3.62	0.94	6.91	1.19

Table 2: Performance comparison of models. All scores are reported on a 0–100 scale, with the best F_1 score for each dataset highlighted in **boldfaced**.

whole, rather than a component-based evaluation that focuses on cell contents in isolation. Existing evaluation procedures, moreover, generally do not account for correctness at the record or tuple level (Xie et al., 2023a; Ghosh et al., 2024; Cheung et al., 2024; Feng et al., 2024; Jiang et al., 2025). A notable exception is PNCExtract (Khalighinejad et al., 2024), which aligns predicted and reference records via maximum-weight bipartite matching before scoring. Our protocol builds on this idea and generalises record-level evaluation to the heterogeneous schemas of all datasets in the benchmark. We propose an evaluation protocol that assesses output with a record-level component.

We align predicted and reference records using the Hungarian Maximum Matching algorithm (Kuhn, 1955), computing semantic similarity of record contents. This ensures that each generated record matches the most appropriate reference. After alignment of records, we measure cell-level correctness using two kinds of metrics: **string-based** and **quantity-based**.

String-based metrics have traditionally been used in this space (for an overview of typical usage, see Timmer et al. (2025)). Here, we use precision, recall, and F_1 to quantify the correctness of the extracted cell values. We also use a character-based overlap metric, ChrF (Popović, 2015) to measure the degree to which the generated answer differs from the ground-truth answer. ChrF is language- and tokenisation-independent, making it a good fit for scientific data extraction.

Quantity-based metrics have been used in particular disciplines (e.g., enzyme work (Boorla and Maranas, 2025)), but we note their general applicability in our unified STG task. Here, we use these metrics to measure the distance of extracted numerical values to the ground truth values, revealing insights about the appropriateness of the *form* of the value, for example, with respect to unit consistency and scale. For numerical fields, we compute coverage, which reflects the fraction of entries with a non-null prediction. We also calculate mean absolute error (MAE) to measure the average deviation of these predictions, and the proportion of predictions within one order-of-magnitude ($p1mag$) to capture scale-level accuracy. All metrics are computed only on available predictions.

Formulas for all metrics with record alignment are provided in Appendix B.

6 Gauging STG Task Difficulty

6.1 Method

To gauge the problem difficulty, we employed typical LLM approaches, text-based and vision-based LLMs. Our goal is to use these approaches to characterise the challenges of the STG task through analyses of the limitations in applying generative AI to this task.

For the former, we used both closed-source and open-source models. For text-based generation, we used examples of locally-hostable models, GPT-OSS 120B (OpenAI et al., 2025), LLaMA-3.3-70B-Instruct (Grattafiori et al., 2024), as well as cloud-hosted models, such as o1-mini (Jaech et al., 2024).[6] These models operate on text

[6]We used the `gpt-o1-mini-2024-05-13` model via the

parsed from PDFs. For vision-based generation, we use Nanonets-OCR-s (Mandal et al., 2025)[7], a lightweight 3B vision-language model based on Qwen2.5-VL-3B-Instruct, designed for semantic understanding of document images.

For the LLM models, we follow the IE pipeline of Ghosh et al. (2024)[8], which integrates LangChain with GROBID (Lopez, 2009) for PDF parsing.[9] We note that GROBID is a widely used tool for PDF parsing, especially for scientific literature (Ghosh et al., 2024; Timmer et al., 2025).[10]

For the vision-language model, we provide each page as an image. All approaches use the same instruction to extract relevant information, given the respective inputs.[11]

6.2 Analysis

Table 2 summarises the overall performance of the sampled state-of-the-art approaches on the benchmark datasets. Note that the scores are calculated using our introduced *tuple-based* protocol, which first aligns predicted records with ground truth records. With this initial filtering, our proposed evaluation is stricter than cell-based evaluation approaches that perform cell-based evaluation (for an overview of such evaluation strategies, see Timmer et al. (2025)). This strictness reflects the constraint that the correctness of cell values must be judged in the context of the record, aligning with scientists' expectations of correctness. Our immediate observation is that the absolute values for the scores are quite low. ChrF, which can be interpreted as a percentage of matched character sequences, has a low average score of 7.74 (o1-mini). Similarly, the average F_1 is similarly low (e.g., o1-mini:13.78). This is true regardless of which approach we use, indicating the overall task difficulty and showing that current AI methods do not currently do well with the record-level focus of this task.

PDF Conversion Issues A limitation of the system lies in parsing PDFs. While GROBID is commonly used in the scientific domain (Ghosh et al., 2024; Timmer et al., 2025), it struggles with scanned documents, unusual section headers and embedded figures, often resulting in missing or misaligned content. Other OCR-based tools, such as Mathpix[12] and PyMuPDF[13], have shown potential to handle issues (Jiang et al., 2025). Here, we evaluated other OCR-based PDF conversion techniques but found GROBID to perform the conversion best.[14] We note that even web-based access to LLMs that allow PDF uploads, such as GPT-5 (OpenAI et al., 2025), rely on similar underlying logic for parsing and are subject to the same limitations.

LLM and VLM Context Limitations A key limitation of LLMs lies in their output token constraints (o1-mini: 4096 tokens)[15]. Datasets that have documents with multiple records (e.g., LegoBench or BRENDA-enzyme) often exceed this limit. While chunking mitigates this issue partially, it introduces new problems, as information is often spread throughout a paper, and splitting the document leads to incomplete or inconsistent outputs.

We observe that the evaluated VLM baseline (page-level extraction) performs worse than the evaluated text-based LLMs (with PDF parsing preprocessing). However, this comparison is not fully controlled, and two factors besides input modality may contribute to the gap. First, although Nanonets is designed for document understanding, it processes inputs at the page level. In our setting, accurate table generation often requires aggregating evidence spread across multiple pages, so page-level processing is a likely source of the gap. Second, Nanonets (3B parameters) is substantially smaller than several of the text models (e.g., GPT-OSS and LLaMA 3.3), and model scale may account for part of the difference. Our results are therefore consistent with the hypothesis that methods operating over the full document, abstracting away from presentation artefacts such as page boundaries, are better suited to STG, as suggested by the stronger performance of o1-mini, GPT-OSS and LLaMA 3.3. Disentangling the effects of modality, context scope and model scale would require controlled comparisons, for example a multi-page VLM of comparable size, or text models restricted to page-

Azure OpenAI API with options for data privacy protection.

[7] https://huggingface.co/nanonets/Nanonets-OCR-s

[8] https://github.com/SatanuG/ad_hoc_information_extraction

[9] Details of experimental settings are provided in Appendix C.1.

[10] We also tested alternative PDF parsing approaches, including vision-language models, but found that they did not perform as well as GROBID (Appendix C.2).

[11] See Appendix C for details of the prompt.

[12] https://mathpix.com

[13] https://pymupdf.readthedocs.io

[14] See Appendix D for details.

[15] The context windows and output budgets of recent models are steadily increasing, which may reduce the need for chunking.

level input.

Multi-record Issues LLMs can also overgenerate records per document. For instance, in the FE dataset, where each document corresponds to a single gold record (see Table 1), the o1-mini generated 1,738 records for just 326 documents. Although this boosts recall, it lowers precision and introduces evaluation misalignment (Ghosh et al., 2024). As noted above, since VLMs such as Nanonets operate on single pages, they lack the document-level context needed for accurate multi-record extraction.

Information Specificity and Source Fidelity We observe that systems sometimes extract information with extra detail that is not present in the gold records. For example, in the PPE dataset, the system extracts "PTzBT-14HD" instead of "PTzBT-14HD (Mn = 73 kDa)". Similarly, LLMs were observed to erroneously append "%" to a numeric value. Providing few-shot exemplars in the prompt that follow the expected schema may help to reduce such formatting issues (Ghosh et al., 2024; Bölücü et al., 2025). Another solution is to train models on task-specific data to ensure consistent adherence to the schema (Sundar et al., 2024; Dagdelen et al., 2024; Wei et al., 2025).

Schema Size and Performance Figure 3 shows the relationship between dataset schema size and F_1 score for the o1-mini model across 20 scientific datasets. The analysis indicates a negligible correlation (slope= -0.29, p= 0.50) between the number of columns (schema size) and performance. This suggests that the difficulty of the IE task is due to the inherent nature of the datasets and challenges related to the discipline-specific science. Datasets from similar domains do seem to cluster together, with Material Science and Chemistry datasets generally performing well, while Machine Learning datasets tend to have lower F_1 scores.

Differences Across Domains Figure 4 shows radar plots of model performance with string-based metrics averaged at the domain level. Due to space constraints, we show graphs for two scientific domains, *Chemistry* and *Material Science*, for which we have multiple datasets.[16] The results reveal a consistent pattern: all models achieve higher recall but lower precision. This is consistent with our observation that the evaluated models tend to *overgenerate* records, which explains the tendency for

[16]For graphs of all domains, see Appendix C.3.

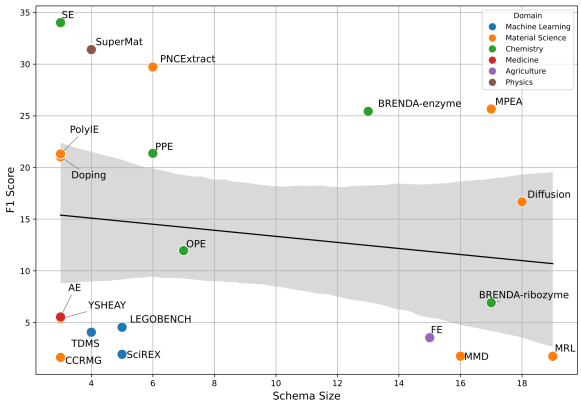
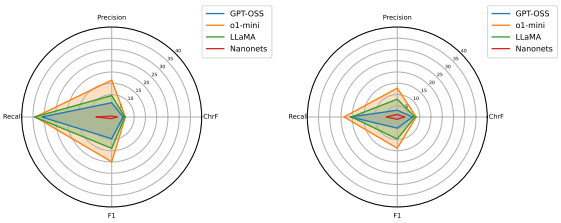


Figure 3: Correlation between F_1 score and schema size for the o1-mini model across 20 datasets.

recall to be high. Recall is also influenced by the density of records in each dataset. Datasets with fewer records per document tend to have higher recall, as models generate more candidate records.



(a) Chemistry (b) Material Science

Figure 4: Radar plots showing the performance of GPT-OSS, o1-mini, LLaMA 3.3, and Nanonets (vision) across domains. Metrics include ChrF, Precision, Recall, and F_1, averaged over all datasets in the domains.

Quantity-based Performance Assessments Our analysis of numerical values shows that this field is challenging. We note that the results show that extraction performance varies widely across domains and datasets. Coverage often remains low in complex material science and agriculture datasets, reflecting the difficulty of predicting all reference values. MAE values can be high, particularly in datasets with large numeric ranges (e.g., LEGOBENCH or BRENDA-enzyme), indicating substantial deviations from reference values when predictions exist. We suspect that this might be due to the extraction of values in different units of measurement The $p1mag$ metric, which indicates the percentage of values within one order of magnitude, is frequently high, suggesting that even when exact numerical predictions are inaccurate, they often remain within an order-of-magnitude. Certain

Chemistry datasets, such as BRENDA-ribozyme and PPE, show better alignment across metrics, likely reflecting more standardised reporting of numerical attributes. These results underscore the challenges of extracting precise numerical data across diverse scientific domains and highlight the need for improved normalisation, alignment, and coverage strategies in future work. For the full results, see Table 5 in Appendix C.4.

7 STG Challenges

Generating scientific tables from literature is challenging due to diverse data modalities, terminologies, and reporting styles. At a high-level we note that the task is diverse. For example, scientific domains differ in required output granularity: some datasets expect a single summary record per paper (e.g., FE), while others need many fine-grained records (e.g., MPEA, Diffusion).

As with any dataset, there are data issues. We also observe empty or missing values in annotations, sometimes deliberately left blank, other times due to annotation errors. In some cases, models correctly extract values that do not appear in the gold data, highlighting flaws in reference annotations (i.e., representation inconsistencies).

In addition to the issues highlighted in the analysis of Section 6, we highlight key open challenges that may accelerate the process in automated table generation.[17]

Representing Numerical Values Evaluation of text variants remains an open problem as exact-match metrics penalise these semantically equivalent but superficially different results. We frequently observe inconsistencies between predicted and gold values for quantitative attributes (e.g., temperature, kcat).

For example, we observed **unit mismatches** where values are expressed using different units than the gold data (e.g., "10 kbar" instead of "1 GPa"), leading to false negatives under exact-match evaluation unless unit normalisation is applied. The other observation is **representation inconsistencies** where there are variations in formatting (e.g., "2–2.5 GPa" versus "2 to 2.5 GPa") which result in semantically equivalent values but still fail exact-match checks. This highlights a common surface-form mismatch where the content is correct, but the formatting differs. These kinds of mismatches can be resolved through normalisation. Other examples include alternative ways to report numerical precision. (e.g., "80 wt%" vs. "0.8"; "2.5" vs. "2.50"; "1.69e-11" vs. "0.0000000000169"). Alternations are numerically equivalent but expressed differently. This highlights the need for normalisation to ensure an accurate evaluation.

Unit and Quantity Standardisation Scientific texts frequently report quantitative values in diverse units, making it challenging to compare and integrate data from different studies.

For example, in enzyme kinetics, the Michaelis constant (K_m) can be reported in various units, such as micromolar (μM), molar (M), or millimolar (mM), each of which requires unit conversion for consistency. Moreover, human-curated databases like BRENDA (Chang et al., 2021) and ChEMBL (Papadatos et al., 2015) standardise units by normalising values across different measurements. Such unit standardisation ensures that data can be easily compared and integrated for downstream analyses and building knowledge-bases (e.g., BRENDA). Using an existing library for normalising units of measurement (the pint library) demonstrated an improvement in performance for the BRENDA-enzyme dataset (F_1 34.2 $\rightarrow$ 38.7), demonstrating the potential for further research to improve STG performance more widely.[18]

Entity Normalisation In scientific literature, the same entity is often represented in multiple forms. For example, "TiO_2", "titania", "AO_2 (A = Ti)", and "titanium dioxide" all refer to the same specific stoichiometry: "TiO_2". To ensure consistency and improve searchability, it is important to normalise these entities into a standard, canonical format. This enables, for example, a query for "titania" to also retrieve papers that mention "titanium dioxide".

In preliminary trials to include some known value normalisation techniques, we observed a performance gain (F_1 increasing: $3.59 \rightarrow 18.56$) for datasets from machine learning. This suggests an opportunity for similar techniques to be developed for other subsets of the benchmark dataset.

Entity Linking In scientific literature, entities such as genes, proteins, and chemicals are often referenced using different names and formats, which can make it difficult to integrate and compare information across studies. The task of entity linking

[17] Descriptions of preliminary attempts to tackle these challenges are described in Appendix D.

[18] https://github.com/hgrecco/pint

addresses this challenge by mapping these entities to their corresponding entries in external structured knowledge bases (e.g., Gene Ontology (GO) (Ashburner et al., 2000), UniProt (Bairoch et al., 2005), and PubChem (Wang et al., 2009)). This ensures that multiple references to the same entity are consistently identified, improving data interoperability and enabling more accurate analysis of scientific findings across sources. In the literature, this task is also referred to as table entity linking (TEL), which specifically aims to map entity mentions in tables to their corresponding entities in a knowledge base (Lou et al., 2023; Wu et al., 2024; Wang et al., 2024c). We leave explorations of entity linking to future work, noting that this is a challenging scenario in which the linked category may not occur verbatim in the source text.

Table and Figure Understanding Scientific papers often present key findings in tables and figures, which require different strategies for extraction. Table understanding involves detecting regions, recognising structures (e.g., headers, merged cells, hierarchies), and parsing the content into a machine-readable format. Table designs vary considerably across domains, from simple grids to complex, nested layouts with implicit row or column relationships (Sun and Liu, 2015; Moosavi et al., 2021; Darmanović et al., 2023).

Figure understanding, on the other hand, focuses on interpreting visual elements such as bar charts, scatter plots, and line graphs. While some figures explicitly encode data (e.g., ChartQA (Masry et al., 2022; Kim et al., 2025)), others convey insights more implicitly through visual trends and relative positions (e.g., Diffusion and MMD datasets). Effectively utilising tables and figures for tasks requires a deep understanding of both visual structure and contextual semantics across modalities. It shows that integrating multimodality helps to extract more accurate information from documents (Khalighinejad et al., 2025). Again, we leave a treatment of tables and figures to future work.

8 Conclusion

We introduce *Scientific Table Generation* (STG), a new task and benchmark designed to unify diverse structured information extraction efforts within a table-centric framework. The benchmark combines 20 datasets from six scientific domains, providing a rich and diverse testbed for evaluating systems that aim to structure scientific content based on predefined schemas. Our formalisation of the STG task, curation of a benchmark dataset, and introduction of a record-based evaluation protocol highlight the difficulty of the task and the limitations of standard zero-shot pipelines for this task.

Our analysis of the output of typical LLM and VLM approaches to document processing reveals several ongoing challenges in the field, including inconsistencies in how values are reported, variations in entity mentions, lack of standard units in extractions, PDF conversion issues, and gaps between text, tables, and figures. All emerge as promising areas for further research.

To promote progress in this area, we make our benchmark and codebase publicly available and invite the research community to contribute new datasets across additional scientific domains. This collaborative approach will help expand STG and enable broader development of models that not only extract information but also structure and normalise it in ways that align with real-world scientific workflows.

Limitations

The study provides a unified and valuable benchmark for the scientific table generation (STG) task. It covers six domains (machine learning, material science, chemistry, agriculture, medicine, and physics). However, there are several limitations. First, the current benchmark is constrained to these six domains; expanding to additional areas such as engineering or environmental science could reveal new domain-specific challenges and improve generalisability.

Second, although the benchmark integrates 20 datasets under a common task formulation, each dataset has its own specialised schema and varies in granularity (coarse- or fine-grained). This variation limits the comparability of results across datasets. Moreover, the dataset sizes are relatively small, which may limit the ability to fine-tune models specifically for each dataset.

Third, we propose subtasks and provide simple baseline solutions to highlight their utility. However, it requires more research to systematically evaluate the role of these subtasks and to explore their integration into multimodal pipelines.

Finally, we do not re-implement or run the methods proposed in the original studies from which the datasets are drawn. As such, our focus remains on

providing a unified benchmark and evaluating general LLM and VLM-based baselines, rather than directly comparing against domain-specific models reported in prior work.

Ethics Statement

Our benchmark is built on existing public datasets adopted from existing repositories. Therefore, we do not foresee any serious or harmful issues related to their content. To ensure full compliance with copyright laws, we provide only the necessary identifiers (e.g., DOI, PubMed ID) for the scientific documents, but we do not distribute the actual documents. Researchers are required to download the relevant files independently. The repository includes detailed instructions to guide users on how to properly place the downloaded documents.

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, and 4 others. 2018. Construction of the Literature Graph in Semantic Scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, and 1 others. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, and 1 others. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Amos Bairoch, Rolf Apweiler, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, and 1 others. 2005. The universal protein resource (UniProt). *Nucleic acids research*, 33(suppl_1):D154–D159.

Necva Bölücü, Jessica Irons, Changhyun Lee, Brian Jin, Maciej Rybinski, Huichen Yang, Andreas Duenser, and Stephen Wan. 2026. Using a human-AI teaming approach to create and curate scientific datasets with the SciLire system. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 428–444, Rabat, Marocco. Association for Computational Linguistics.

Necva Bölücü, Jordan Pennells, Huichen Yang, Maciej Rybinski, and Stephen Wan. 2025. An Evaluation of Large Language Models for Supplementing a Food Extrusion Dataset. *Foods*, 14(8):1355.

Veda Sheersh Boorla and Costas D. Maranas. 2025. CatPred: a comprehensive framework for deep learning in vitro enzyme kinetic parameters. *Nature Communications*, 16(1).

Rohit Borah, Andrew W Brown, Patrice L Capers, and Kathryn A Kaiser. 2017. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ open*, 7(2):e012545.

Christopher KH Borg, Carolina Frey, Jasper Moh, Tresa M Pollock, Stéphane Gorsse, Daniel B Miracle, Oleg N Senkov, Bryce Meredig, and James E Saal. 2020. Expanded dataset of mechanical properties and observed phases of multi-principal element alloys. *Scientific Data*, 7(1):430.

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.

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Wang Changxin, Zhifeng Gao, Hongshuai Wang, Li Yongge, Mujie Lin, Shuwen Yang, Jiankun Wang, Mingjun Xu, Jin Huang, Xi Fang, Jiaxi Zhuang, Yuqi Yin, Yaqi Li, Changhong Chen, and 4 others. 2025. SciAssess: Benchmarking LLM Proficiency in Scientific Literature Analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2335–2357, Albuquerque, New Mexico. Association for Computational Linguistics.

Hengxing Cai, Xiaochen Cai, Shuwen Yang, Jiankun Wang, Lin Yao, Zhifeng Gao, Junhan Chang, Sihang Li, Mingjun Xu, Changxin Wang, and 1 others. 2024. Uni-SMART: Universal Science Multimodal Analysis and Research Transformer. *arXiv preprint arXiv:2403.10301*.

Antje Chang, Lisa Jeske, Sandra Ulbrich, Julia Hofmann, Julia Koblitz, Ida Schomburg, Meina

Neumann-Schaal, Dieter Jahn, and Dietmar Schomburg. 2021. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic acids research*, 49(D1):D498–D508.

Hong Chen, Hiroya Takamura, and Hideki Nakayama. 2021. SciXGen: A scientific paper dataset for context-aware text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1483–1492, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jerry Cheung, Yuchen Zhuang, Yinghao Li, Pranav Shetty, Wantian Zhao, Sanjeev Grampurohit, Rampi Ramprasad, and Chao Zhang. 2024. POLYIE: A Dataset of Information Extraction from Polymer Material Scientific Literature. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2370–2385, Mexico City, Mexico. 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.

Filip Darmanović, Allan Hanbury, and Markus Zlabinger. 2023. SCI-3000: A Dataset for Figure, Table and Caption Extraction from Scientific PDFs. In *International Conference on Document Analysis and Recognition*, pages 234–251. Springer.

Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. 2022. Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238*.

Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. 2024. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*.

Luca Foppiano, Sae Dieb, Akira Suzuki, Pedro Baptista de Castro, Suguru Iwasaki, Azusa Uzuki, Miren Garbine Esparza Echevarria, Yan Meng, Kensei Terashima, Laurent Romary, Yoshihiko Takano, and Masashi Ishii. 2021. SuperMat: construction of a linked annotated dataset from superconductors-related publications. *Science and Technology of Advanced Materials: Methods*, 1(1):34–44.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. 2025. JSONSchemaBench: A Rigorous Benchmark of Structured Outputs for Language Models. *Preprint*, arXiv:2501.10868.

Satanu Ghosh, Neal Brodnik, Carolina Frey, Collin Holgate, Tresa Pollock, Samantha Daly, and Samuel Carton. 2024. Toward Reliable Ad-hoc Scientific Information Extraction: A Case Study on Two Materials Dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15109–15123, Bangkok, Thailand. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sameera Horawalavithana, Ellyn Ayton, Shivam Sharma, Scott Howland, Megha Subramanian, Scott Vasquez, Robin Cosbey, Maria Glenski, and Svitlana Volkova. 2022. Foundation Models of Scientific Knowledge for Chemistry: Opportunities, Challenges and Lessons Learned. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 160–172, virtual+Dublin. Association for Computational Linguistics.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5203–5213, Florence, Italy. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation Extraction By End-to-end Language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

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.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-Level N-ary Relation Extraction with Multiscale Representation Learning. In *Proceedings of*

the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3693–3704, Minneapolis, Minnesota. Association for Computational Linguistics.

Jinling Jiang, Jie Hu, Siwei Xie, Menghao Guo, Yuhang Dong, Shuai Fu, Xianyue Jiang, Zhenlei Yue, Junchao Shi, Xiaoyu Zhang, and 1 others. 2025. Enzyme Co-Scientist: Harnessing Large Language Models for Enzyme Kinetic Data Extraction from Literature. *bioRxiv*, pages 2025–03.

Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. 2023. Instruct and Extract: Instruction Tuning for On-Demand Information Extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10030–10051, Singapore. Association for Computational Linguistics.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Ilknur Karadeniz and Arzucan Özgür. 2019. Linking entities through an ontology using word embeddings and syntactic re-ranking. *BMC bioinformatics*, 20(1):156.

Ghazal Khalighinejad, Defne Circi, L. Brinson, and Bhuwan Dhingra. 2024. Extracting Polymer Nanocomposite Samples from Full-Length Documents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13163–13175, Bangkok, Thailand. Association for Computational Linguistics.

Ghazal Khalighinejad, Sharon Scott, Ollie Liu, Kelly L. Anderson, Rickard Stureborg, Aman Tyagi, and Bhuwan Dhingra. 2025. MatViX: Multimodal Information Extraction from Visually Rich Articles. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3636–3655, Albuquerque, New Mexico. Association for Computational Linguistics.

Wonjoong Kim, Sangwu Park, Yeonjun In, Seokwon Han, and Chanyoung Park. 2025. SIMPLOT: Enhancing Chart Question Answering by Distilling Essentials. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 573–593, Albuquerque, New Mexico. Association for Computational Linguistics.

David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Adithya Kulkarni, Fatimah Alotaibi, Xinyue Zeng, Longfeng Wu, Tong Zeng, Barry Menglong Yao, Minqian Liu, Shuaicheng Zhang, Lifu Huang, and Dawei Zhou. 2025. Scientific Hypothesis Generation and Validation: Methods, Datasets, and Future Directions. *Preprint*, arXiv:2505.04651.

Feiran Li, Le Yuan, Hongzhong Lu, Gang Li, Yu Chen, Martin KM Engqvist, Eduard J Kerkhoven, and Jens Nielsen. 2022. Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis*, 5(8):662–672.

Patrice Lopez. 2009. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27-October 2, 2009. Proceedings 13*, pages 473–474. Springer.

Yuze Lou, Bailey Kuehl, Erin Bransom, Sergey Feldman, Aakanksha Naik, and Doug Downey. 2023. S2abEL: A Dataset for Entity Linking from Scientific Tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3089–3101.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Xinyuan Lu, Liangming Pan, Yubo Ma, Preslav Nakov, and Min-Yen Kan. 2025. TART: An Open-Source Tool-Augmented Framework for Explainable Table-based Reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4323–4339, Albuquerque, New Mexico. 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.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. BioRED: A Rich Biomedical Relation Extraction Dataset. *Briefing in Bioinformatics*.

Souvik Mandal, Ashish Talewar, Paras Ahuja, and Prathamesh Juvatkar. 2025. Nanonets-OCR-S: A model for transforming documents into structured markdown with intelligent content recognition and semantic tagging.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

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

Ahmed Nassar, Nikolaos Livathinos, Maksym Lysak, and Peter Staar. 2022. TableFormer: Table Structure Understanding With Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4614–4623.

Benjamin Newman, Yoonjoo Lee, Aakanksha Naik, Pao Siangliulue, Raymond Fok, Juho Kim, Daniel S Weld, Joseph Chee Chang, and Kyle Lo. 2024. ArxivDIGESTables: Synthesizing Scientific Literature into Tables using Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9612–9631, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b gpt-oss-20b Model Card. *Preprint*, arXiv:2508.10925.

George Papadatos, Anna Gaulton, Anne Hersey, and John P Overington. 2015. Activity, assay and target data curation and quality in the ChEMBL database. *Journal of computer-aided molecular design*, 29:885–896.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter W J Staar. 2022. DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis.

Maciej P Polak and Dane Morgan. 2024. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Chandan K Reddy and Parshin Shojaee. 2025. Towards scientific discovery with generative ai: Progress, opportunities, and challenges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28601–28609.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. In *Advances in Neural Information Processing Systems*, volume 37, pages 18695–18728. Curran Associates, Inc.

Shruti Singh, Shoaib Alam, Husain Malwat, and Mayank Singh. 2024. LEGOBench: Scientific Leaderboard Generation Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14598–14613, Miami, Florida, USA. Association for Computational Linguistics.

Yu Song, Santiago Miret, and Bang Liu. 2023. MatSciNLP: Evaluating Scientific Language Models on Materials Science Language Tasks Using Text-to-Schema Modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3621–3639, Toronto, Canada. Association for Computational Linguistics.

Haoyang Su, Renqi Chen, Shixiang Tang, Xinzhe Zheng, Jingzhe Li, Zhenfei Yin, Wanli Ouyang, and Nanqing Dong. 2024. Two Heads Are Better Than One: A Multi-Agent System Has the Potential to Improve Scientific Idea Generation. *arXiv preprint arXiv:2410.09403*.

Lijun Sun and Yao Liu. 2015. Review of Research on Table Extraction in Scientific and Technical Literature. *International Journal of Knowledge and Language Processing*, 6:48–62.

Anirudh Sundar, Christopher Richardson, and Larry Heck. 2024. gtbls: Generating tables from text by conditional question answering. *arXiv preprint arXiv:2403.14457*.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical Entity Representations with Synonym Marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

Deep Search Team. 2024. Docling Technical Report. Technical report.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore. Association for Computational Linguistics.

Roelien C. Timmer, Yufang Hou, and Stephen Wan. 2025. A Position Paper on the Automatic Generation of Machine Learning Leaderboards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30749–30772, Suzhou, China. Association for Computational Linguistics.

Benjamin Townsend, Eamon Ito-Fisher, Lily Zhang, and Madison May. 2021. Doc2dict: Information extraction as text generation. *arXiv preprint arXiv:2105.07510*.

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.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024a. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 50622–50649. PMLR.

Xingbo Wang, Samantha L Huey, Rui Sheng, Saurabh Mehta, and Fei Wang. 2024b. SciDaSynth: Interactive Structured Knowledge Extraction and Synthesis from Scientific Literature with Large Language Model. *arXiv preprint arXiv:2404.13765*.

Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, and Stephen H Bryant. 2009. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research*, 37(suppl_2):W623–W633.

Yuanzheng Wang, Yixing Fan, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. 2024c. RoCEL: Advancing Table Entity Linking through Distinctive Row and Column Contexts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15284–15298, Miami, Florida, USA. Association for Computational Linguistics.

Galen Wei, Xinchun Ran, Runeem AI-Abssi, and Zhongyue Yang. 2025. Finding the dark matter: Large language model-based enzyme kinetic data extractor and its validation. *Protein Science*, 34(9):e70251.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, and 1 others. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Tianxing Wu, Lin Li, Huan Gao, Guilin Qi, Yuxiang Wang, and Yuehua Li. 2024. Npel: Neural paired entity linking in web tables. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Tong Xie, Yuwei Wan, Wei Huang, Zhenyu Yin, Yixuan Liu, Shaozhou Wang, Qingyuan Linghu, Chunyu Kit, Clara Grazian, Wenjie Zhang, and 1 others. 2023a. Darwin series: Domain specific large language models for natural science. *arXiv preprint arXiv:2308.13565*.

Tong Xie, Yuwei Wan, Wei Huang, Yufei Zhou, Yixuan Liu, Qingyuan Linghu, Shaozhou Wang, Chunyu Kit, Clara Grazian, Wenjie Zhang, and 1 others. 2023b. Large language models as master key: unlocking the secrets of materials science with GPT. *arXiv preprint arXiv:2304.02213*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. DWIE: An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, 58(4):102563.

Youxue Zhang, Huaiwei Ni, and Yang Chen. 2010. Diffusion data in silicate melts. *Reviews in Mineralogy and Geochemistry*, 72(1):311–408.

A Benchmark Insights

We present the details of 6 scientific domains with the 20 datasets used in the benchmark. We summarise the source, content focus, and key characteristics of each dataset.

A.1 Datasets

Machine Learning The machine learning domain focuses on developing and evaluating models across various tasks, often through benchmark datasets (Hou et al., 2019; Singh et al., 2024). Consequently, leaderboard generation becomes important for extracting relevant information and comparing model performance effectively, as well as for table generation from scientific documents. In this domain, we collect different-scale datasets focused on leaderboard generation.

TDMS is a scientific leaderboard dataset that contains information on tasks, datasets, evaluation metrics, and scores from NLP papers. It targets the extraction of the best numerical results from experimental sections.

SciREX supports scientific information extraction, specifically relation extraction and entity linking, based on material, method, and evaluation metrics.

LegoBench is a dataset designed for leaderboard generation, reformulated as a question-answering

Dataset	Schema	Source
TDMS	Task, Dataset, Metric, Score	ACL Anthology ID
SciREX	Material, Method, Metric, Task, Score	Semantic Scholar ID
LEGOBENCH	Task, Dataset, Method, Metric, Score	ArXiv ID
MPEA	high entropy alloy formula, microstructure, processing method, BCC/FCC/other, grain size, experimental density, hardness, type of test, test temperature, yield strength, ultimate tensile strength, elongation, elongation plastic, experimental young modulus, oxygen content, nitrogen content, carbon content	DOI
Diffusion	melt, diffusing species, type of diffusion, type of experiment, test temperature, pressure, diffusivity, SiO2, TiO2, Al2O3, FeOt, MnO, MgO, CaO, Na2O, K2O, P2O5, H2Ot	DOI
YSHEAY	Material, Value, Unit	DOI
CCRMG	Material, Value, Unit	DOI
Doping	BASEMAT, DOPANT, DOPMODQ	DOI
MMD	Formula, Formula units per cell, Atomic sites per cell, Crystal system, Formation energy (eV/atom), Energy relative to convex hull (eV/atom), Structure search, Averaged magnetic moment (μB/atom), Magnetic polarization, Js</(T), Magnetic easy axis, Magnetic anisotropy constants:Ka-c, Magnetic anisotropy constants:Kb-c (MJ/m3), Magnetic anisotropy constants:Kb-a (MJ/m3), Magnetic anisotropy constants:Kd-a (MJ/m3), Curie temperature, TC(K), Methods	DOI
MRL	Electrical resistivity (Ωcm), Seebeck coefficient (μCV/K), ZT (thermoelectric figure of merit), formula, synthesis, form, temperature, Thermal conductivity (W/mK), Average atomic mass (g/mol), Scarcity (wt fraction/abundance), HHIP (Herfindahl-Hirschman Index for elemental production), HHIR (Herfindahl-Hirschman Index for elemental reserves), Unit cell volume (Å^3), Average atomic volume (Å^3), Atoms per unit cell, Electrical conductivity (Ω^{-1} cm^{-1}), Power Factor (W/(K^2m)), S^2 (μV/K)^2, Electronic fraction of thermal conductivity (κ_e/κ)	DOI
PolyIE	Chemical Name, Property Name, Property Value	DOI
PNCExtract	Matrix Chemical Name, Matrix Abbreviation, Filler Chemical Name, Filler Abbreviation, Filler Composition Mass, Filler Composition Volume	DOI
BRENDA-enzyme	Enzyme, Organism, Substrate, km, unit_km, kcat, unit_kcat, km_kcat, unit_km_kcat, Commentary[Temp], Commentary[pH], Commentary[Mutant], Commentary[Cosubstrate]	PubMed
BRENDA-ribozyme	Ribozyme, Organism, Variable_reactant, Kobs, Unit_Kobs, km, Unit_Km, kcat, Unit_Kcat, km_kcat, Unit_Kcat/Km, Kcleav, Unit_Kcleav, Commentary[Temp], Commentary[pH], Commentary[Mutant], Commentary[Others]	PubMed
OPE	Host, Dopant, Td [$^\circ$ C]/Tg [$^\circ$ C]/ET [eV], Von [V], EQE, CE, PE, CIE [x, y]	DOI
PPE	Nickname, PCE_max(%), PCE_ave(%), Voc (V), Jsc (mA cm^2), FF	DOI
SE	solute_name, solvent_name, temperature, pressure, solubility	DOI
AE	Compound, Affinities	DOI
FE	extruder product type, extruder manufacturer company, extruder model name, extruder manufacturer country, scale, screw type, screw dimension, length to diameter ratio, die dimensions, ingredients, moisture in extrusion trials, input variables, response variables, feed characterisation, quality characterisation	DOI
SuperMat	Material, tcValue, Pressure, me_method	DOI

Table 3: Overview of 20 datasets across six domains. Each dataset is linked to a document source and schema.

(QA) task over documents, focusing on extracting structured records from full-text papers. We converted the QA dataset into the STG formalised dataset.

Material Science Materials science explores a wide variety of substances, including polymers, ceramics, and metals. Each of these has distinct physical, chemical, and mechanical properties, making them important for applications in industries such as aerospace, electronics, and construction. Scientific literature in this domain often reports not only material compositions and target properties, but also detailed experimental and contextual information. For example, the MPEA dataset includes parameters such as melt composition, pressure, temperature, and diffusion type. Materials science is a well-studied domain with many datasets covering

detailed information found in the literature (Borg et al., 2020; Khalighinejad et al., 2025).

MPEA is a dataset focused on multi-principal element alloys (MPEAs), also known as high-entropy alloys (HEAs). It contains experimental key properties (e.g. strength, hardness, and Young's modulus).

Diffusion is a dataset that covers the diffusion (transport of elements) in silicate melts and glasses, which is a key process in both geology and materials science. It captures experimental parameters such as temperature, pressure, and diffusion coefficients. This dataset is built on a summary table presented by Zhang et al. (2010). Instead of extracting information from this paper, we gathered the reference papers to build a dataset for this task similar to Ghosh et al. (2024). Similar to this dataset, there are more possible datasets that we have not included.

YSHEAY and CCRMG cover the yield strength of HEAs and the critical cooling rates of metallic glasses, respectively. Both YSHEAY and CCRMG datasets include material names, measured property values, and units. We formulate datasets and gather referenced papers given in the datasets for our benchmark. The datasets extract the sentences from the scientific documents in which the information is presented to simplify the task. We use the full document, similar to the rest of the other datasets.

MRL is derived from a review on thermoelectric materials and includes experimental results and properties (e.g., Electrical resistivity, Seebeck coefficient, thermoelectric figure of merit).

Doping is a dataset for materials chemistry that covers dopant (also called a doping agent), which is a small amount of a substance added to a material to alter its physical properties, such as electrical or optical properties, with its host materials.

MMD is a dataset focused on physical and magnetic properties of materials relevant to magnetism research and materials design including physical and magnetic properties of materials relevant to magnetism research and materials design.

PolyIE is a dataset for polymer scientific literature covering four different material science domains: polymer solar cells, ring-opening polymerization, polymer membranes, and lithiumion batteries. We could not access all documents in the dataset since the DOI numbers are not provided.

PNCExtract is a dataset focused on the extraction of polymer nanocomposites (PNCs) samples from science articles. The variety in PNCs comes from different matrix and filler combinations that modify the properties.

Chemistry The field of chemistry involves a vast array of complex molecular structures, chemical reactions, and properties, alongside a substantial amount of data presented in formulas, reaction equations, and diagrams. Effectively processing and interpreting these components is a significant challenge for language models. Tasks in the chemical domain demand precise understanding of molecular compositions, reaction mechanisms, and material properties.

BRENDA is a database on functional and molecular information of enzymes, based on primary literature. Jiang et al. (2025) construct 2 datasets (enzyme and ribozyme) for enzyme and ribozyme kinetics data collected from scientific literature.

OPE is a dataset that covers OLED molecules and their optical properties. The dataset is built on summary tables from a limited number of scientific documents. Instead of that, we collect reference studies instead of using given documents to make it similar to other datasets in this study and remove (a text-based format for representing chemical structures as a line notation) from the schema since this attribute is not provided in the scientific documents.

PPE is a dataset that contains values like power conversion efficiency (PCE) and open-circuit voltage (VOC) for polymer molecules extracted from scientific documents.

Solubility Extraction (SE) is a dataset Organic electrolytes, extensively used in battery technologies, comprise organic solvents, lithium salts, and additives. Understanding solubility in organic electrolytes is crucial as it impacts the efficiency of electrolytic processes, product selectivity, and equipment design.

Medicine Medicine often involves complex interactions between biological molecules, experimental methods, and treatment conditions. Literature in this domain typically reports experimental results, molecular data, and pharmacological properties essential for therapeutic development.

Affinity Extraction (AE) extract an affinity table containing molecules' tags, SMILES, and their affinities to different targets in bioassays. It tests several key capabilities of LLMs, including understanding complex and domain-specific language, as well as molecules and tables. Affinity data extraction requires not just surface-level text processing

but also a deeper analysis to match different modalities.

Agriculture Agricultural research often involves complex interactions between biological materials, machinery, and environmental conditions. Literature in this domain typically reports experimental setups, crop properties, and processing parameters relevant to food production and sustainability.
Food Extrusion (FE) is a qualitative literature collection that consolidates and standardises research on food extrusion, including details about the extruder and other variables and characteristics explored in the articles

Physics Superconductors Material (SuperMat) is a manually linked annotated dataset of superconductor-related materials and properties.

A.2 Benchmark Curation and Cleaning

Source document recovery. We collect the original source documents for every dataset and rebuild each dataset from those documents, rather than reusing the released annotations as they are. However, for some datasets, we cannot find all the original documents. Some papers are not publicly available. Others are listed without a persistent identifier (e.g., DOI or ArXiv ID) or with too little bibliographic information to locate them reliably.[19] As a result, the number of datasets presented here may not exactly match the original resources. Finally, we do not follow the original dataset splits (train/dev/test), so we do not perform comparisons of baseline results with the methods proposed in the original papers.

Schema normalisation We convert the datasets into a unified table-centric format. We remove certain attributes when the corresponding values cannot be reliably extracted from the PDFs. One example is *SMILES* strings, which are rarely stated explicitly in the text and are usually derived from external databases. In contrast, we retain attributes whose values are expressed in the document but annotated in a normalised or canonical form (e.g., *Task* in TDMS and SciREX datasets), and we treat entity normalisation as one of the STG challenges.

A.3 Resources and Downloads

TDMS https://github.com/IBM/science-result-extractor

SciREX https://github.com/allenai/SciREX

LEGOBENCH https://github.com/lingo-iitgn/LEGOBench

MPEA https://github.com/CitrineInformatics/MPEA_dataset

Diffusion http://www.minsocam.org/MSA/RIM/rim72.html

YSHEAY and CCRMG https://figshare.com/articles/dataset/Data_for_Flexible_Model-Agnostic_Method_for_Materials_Data_Extraction_from_Text_Using_General_Purpose_Language_Models_/21861948

Doping https://github.com/lbnlp/NERRE

MRL and MMD https://github.com/MasterAI-EAM/Darwin

PolyIE https://github.com/jerry3027/PolyIE

PNCExtract https://github.com/ghazalkhalighinejad/PNCExtract

BRENDA (enzyme and ribozyme) https://github.com/JackKuo666/LLM-BioDataExtractor

OPE, PPE, SE, and AE https://github.com/sci-assess/SciAssess

FE https://data.csiro.au/collection/csiro%3A64834v1

SuperMat https://github.com/lfoppiano/SuperMat

B Evaluation metrics

Record Alignment The output table $T \in \mathcal{V}^{M \times K}$ is partitioned into N document-level subtables, where subtable $T^{(i)} \in \mathcal{V}^{R_i \times K}$ contains the R_i records contributed by document d_i. Correspondingly, the predicted table $\hat{T}$ is partitioned into subtables $\hat{T}^{(i)} \in \mathcal{V}^{\hat{R}_i \times K}$, where $\hat{R}_i$ may differ from R_i. Alignment is performed independently within each subtable pair $(T^{(i)}, \hat{T}^{(i)})$.

For each record, we concatenate its K attribute values into a single string and encode it with a sentence encoder ϕ. The pairwise semantic similarity matrix $\mathbf{S}^{(i)} \in \mathbb{R}^{R_i \times \hat{R}_i}$ is then computed as:

[19] Records whose source document could not be retrieved are excluded from the benchmark.

$$S^{(i)}_{m,\hat{m}} = \cos(\phi(g_m),\ \phi(\hat{g}_{\hat{m}})) \tag{1}$$

where $\phi(g_m)$ denotes the sentence embedding of the concatenated attribute values of record g_m, $\cos(\cdot,\cdot)$ denotes cosine similarity, and ϕ is implemented using `all-MiniLM-L6-v2`[20] (**?**).

Since $R_i \neq \hat{R}_i$ in general, $\mathbf{S}^{(i)}$ is a rectangular matrix. The Hungarian algorithm (Kuhn, 1955) finds a maximum-weight partial matching:

$$\pi_i^* = \arg\max_{\pi \in \Pi_i} \sum_{m=1}^{\min(R_i, \hat{R}_i)} S^{(i)}_{m,\pi(m)} \tag{2}$$

where Π_i is the set of all injective mappings from the smaller index set into the larger. This yields $\min(R_i, \hat{R}_i)$ matched pairs $\mathcal{M}_i = \{(g_m, \hat{g}_{\pi_i^*(m)})\}$, with the remaining $|R_i - \hat{R}_i|$ records left unmatched. Unmatched predicted records penalise precision; unmatched ground-truth records penalise recall. The global alignment set across all documents is $\mathcal{A} = \bigcup_i \mathcal{M}_i$.

Cell=level Scoring **String-based metrics** have traditionally been used in this space (for an overview of typical usage, see Timmer et al. (2025)). Here, we use precision, recall, and F_1 to quantify the correctness of the extracted cell values. We also use a character-based overlap metric, ChrF (Popović, 2015) to measure the degree to which the generated answer differs from the ground-truth answer. ChrF is language- and tokenisation-independent, making it a good fit for scientific data extraction.

Each record r is treated as a *bag of cells*: a matched pair $(r_j, \hat{r}_i) \in \mathcal{A}$ contributes $|\hat{r}_i \cap r_j|$ correctly predicted cells to the numerator, where a cell is counted as correct if its value exactly matches the ground-truth cell in the aligned column. Precision normalises by the total number of predicted cells across *all* predicted records (including unmatched ones), ensuring that hallucinated extra records are penalised:

$$\text{Precision} = \frac{\sum_{(i,j)\in\mathcal{A}} |\hat{r}_i \cap r_j|}{\sum_{\hat{r}\in\hat{\mathcal{R}}} |\hat{r}|}, \tag{3}$$

$$\text{Recall} = \frac{\sum_{(i,j)\in\mathcal{A}} |\hat{r}_i \cap r_j|}{\sum_{r\in\mathcal{R}} |r|}, \tag{4}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{5}$$

where $\hat{\mathcal{R}}$ and $\mathcal{R}$ denote the full sets of predicted and ground-truth records, respectively, and $|r|$ is the number of cells in record r.

We also report the micro-averaged ChrF score over aligned records:

$$\text{ChrF}_{\text{micro}} = \frac{\sum_{(i,j)\in\mathcal{A}} \sum_{c\,\in\,\text{cells}(r_j)} \text{ChrF}(\hat{c}_i,\ c)}{\sum_{(i,j)\in\mathcal{A}} |\text{cells}(r_j)|}, \tag{6}$$

where $\text{cells}(r_j)$ is the ordered set of cells in reference record r_j, $\hat{c}_i$ is the predicted cell in the corresponding column of $\hat{r}_i$, and $\text{ChrF}(\hat{c}_i, c)$ is the character-level F-score between the two cell strings. The denominator counts the total number of reference cells across all aligned pairs.

Quantity-based metrics have been used in particular disciplines (e.g., enzyme work (Boorla and Maranas, 2025)), but we note their general applicability in our unified STG task. Here, we use these metrics to measure the distance of extracted numerical values to the ground truth values, revealing insights about the appropriateness of the *form* of the value, for example, with respect to unit consistency and scale.

After record alignment, let N_f be the total number of numeric reference entries in column f across all aligned pairs $(\hat{r}_i, r_j) \in \mathcal{A}$, and let $A_f \subseteq \{1, \ldots, N_f\}$ be the index set of those entries for which a non-null numeric prediction is available. All metrics are computed only over A_f; non-numeric cells and missing predictions are excluded.

Coverage quantifies the proportion of numeric reference entries in column f for which a non-null prediction is available:

$$\text{Coverage}_f = \frac{|A_f|}{N_f} \times 100. \tag{7}$$

[20] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 Last visited: 01-04-2026

Mean Absolute Error (MAE) computes the average absolute deviation of available predictions from the reference values in column f:

$$\mathrm{MAE}_f = \frac{1}{|A_f|} \sum_{i \in A_f} |v_{f,i} - \hat{v}_{f,i}|, \tag{8}$$

where $v_{f,i}$ and $\hat{v}_{f,i}$ are the reference and predicted values for entry i in the aligned pair, respectively.

Within One Order-of-Magnitude ($p1\mathrm{mag}$) is the fraction of available predictions in column f that lie within one order of magnitude of the reference value:

$$p1\mathrm{mag}_f = \frac{100}{|A_f^+|} \sum_{i \in A_f^+} \mathbf{1}\Big(\big|\log_{10} \hat{v}_{f,i} - \log_{10} v_{f,i}\big| \leq 1\Big), \tag{9}$$

where $A_f^+ = \{i \in A_f \mid v_{f,i} > 0 \text{ and } \hat{v}_{f,i} > 0\}$ is the subset of available predictions with positive values required for the logarithm, and entries with zero or negative values are excluded from this metric only.

C Detailed Experimental Settings and Results

C.1 Experimental Settings

We structure outputs in JSON and guide decoding using LangChain's `create_structured_output_runnable` with a Pydantic schema. This approach enforces consistent formatting and reduces errors (Geng et al., 2025).

We conduct the experiments using a zero-shot condition and generalise the simple prompt for use across all domain-specific datasets in our study. We set the LangChain chunk size to 2000 tokens with a 1% overlap ratio following Ghosh et al. (2024). To ensure stable and reproducible output, we set the temperature parameter to 0.

The prompt used in table generation:

```
You are a helpful assistant capable of extracting information from text. Only extract relevant information from the text. Extract nothing if no important information can be found in the text.
Input: {{Paper (text)}}
Schema: {{JSON schema}}
```

For vision-based generation, we follow the prompt used in the model.

Prompt:

```
Extract the text from the above document as if you were reading it naturally. Return the tables in html format. Return the equations in LaTeX representation. If there is an image in the document and an image caption is not present, add a small description of the image inside the <img></img> tag; otherwise, add the image caption inside <img></img>. Watermarks should be wrapped in brackets. Ex: <watermark>OFFICIAL COPY</watermark>. Page numbers should be wrapped in brackets. Ex: <page_number>14</page_number> or <page_number>9/22</page_number>.
Input: {{Image url}}
```

Computing Costs and Resources

In this study, the approximate cost of using o1-mini was $672 USD. Inference with GPT-OSS 120B, LLaMA-3.3-70B and Nanonets was conducted locally on two Nvidia H100 GPUs.

C.2 Different Parsers

We explored alternative PDF parsing approaches using OCR-based tools like Nanonets-OCR-s (Mandal et al., 2025), Mathpix and PyMuPDF, as some have reported success with these tools (Jiang et al., 2025). Despite known issues, GROBID remains the most reliable option for the benchmark.

To investigate the impact of parsers on table generation performance, we experiment with Docling (Pfitzmann et al., 2022; Team, 2024; Nassar et al., 2022)[21] and Nanonets-OCR-s (Mandal et al., 2025)[22] , the latter being an OCR-based model, in addition to GROBID. Since we cannot do intrinsic evaluation by comparing the extracted text from parsers with the ground truth text, we use the STG task as an extrinsic evaluation setting, where PDF parsing quality directly affects downstream performance. The experiments are conducted using LLaMA 3.3 as the generation model. The results, shown in Table 4, indicate that GROBID achieves the strongest overall performance, although the gap between parsers is relatively small.

[21] https://huggingface.co/ds4sd/docling-models

[22] https://huggingface.co/nanonets/Nanonets-OCR-s

Domain	Dataset	GROBID				nanonets/Nanonets-OCR-s				ds4sd/docling-models			
		ChrF	Precision	Recall	F_1	ChrF	Precision	Recall	F_1	ChrF	Precision	Recall	F_1
Machine Learning	TDMS	11.46	1.99	18.38	**3.59**	11.49	1.11	19.39	2.09	11.30	1.20	20.03	2.26
	SciREX	6.95	1.48	8.93	**2.54**	6.13	1.00	10.43	1.83	6.34	0.89	9.70	1.63
	LEGOBENCH	11.10	2.42	13.18	**4.10**	10.45	2.32	12.56	3.92	10.86	2.34	14.12	4.01
Material Science	MPEA	2.00	10.20	37.83	**16.07**	1.85	6.90	39.15	11.74	1.90	7.45	38.87	12.50
	Diffusion	0.88	10.52	10.16	**10.34**	0.90	7.66	13.84	9.86	0.83	8.19	13.43	10.18
	YSHEAY	14.25	1.82	34.37	**3.47**	14.85	1.00	31.50	1.95	15.22	1.03	30.63	1.99
	CCRMG	10.49	0.70	18.56	**1.35**	11.10	0.68	17.16	1.31	11.32	0.61	16.18	1.18
	Doping	12.52	13.38	20.83	**16.29**	12.58	9.99	22.61	13.86	12.21	11.56	23.23	15.44
	MMD	1.09	8.77	3.57	**5.08**	0.89	4.94	3.04	3.76	0.87	5.75	2.54	3.53
	MRL	0.72	2.44	2.82	**2.62**	0.67	2.00	3.50	2.54	0.70	2.21	3.02	2.55
	PNCExtract	9.40	11.73	27.96	**16.53**	9.37	9.68	34.49	15.12	9.19	9.74	32.20	14.96
	PolyIE	19.02	11.74	27.00	**16.36**	18.95	9.94	28.30	14.71	18.22	10.16	27.39	14.82
Chemistry	BRENDA-enzyme	2.53	11.07	30.72	16.28	2.67	11.81	33.09	**17.41**	2.55	11.90	31.97	17.34
	BRENDA-ribozyme	1.37	4.72	11.39	**6.68**	1.45	4.04	12.51	6.11	1.31	4.25	11.36	6.19
	OPE	4.76	4.66	37.70	**8.29**	4.43	3.29	34.79	6.02	4.68	3.90	38.38	7.08
	PPE	10.75	8.86	61.32	**15.48**	10.83	6.84	60.57	12.29	10.81	7.29	60.94	13.02
	SE	9.74	18.40	30.41	**22.93**	9.39	14.91	33.80	20.69	9.56	15.27	35.40	21.34
Medicine	AE	6.49	7.34	6.84	7.08	10.19	11.44	20.13	**14.59**	7.41	6.02	9.11	7.25
Agriculture	FE	2.50	1.49	28.20	**2.83**	2.31	1.06	29.10	2.04	2.23	0.98	27.69	1.89
Physics	SuperMat	9.04	12.49	19.23	**15.15**	8.68	9.11	15.07	11.36	9.24	3.06	5.06	3.81
	Avg.	7.35	7.31	22.47	**9.65**	7.46	5.99	23.75	8.66	7.34	5.69	22.56	8.15

Table 4: Performance comparison of parsers using LLaMA 3.3. All scores are reported on a 0–100 scale, with the best F_1 score for each dataset highlighted in **boldfaced**.

C.3 Domain Level Results

Figure 5 shows radar plots of model performance across six scientific domains, with metrics averaged at the domain level. The results reveal a consistent pattern: all models achieve higher recall but lower precision, showing that while many candidate records are generated, their accuracy remains limited. Recall is also influenced by the density of records in each dataset. Datasets with fewer records per document tend to have higher recall, as models generate more candidate records.

C.4 Quantity-based results

Quantity-based results are presented in Table 5.

D Explorations on STG Open Challenges

D.1 Entity Normalisation

Entity normalisation is the task of mapping semantically equivalent entity mentions in the generated table to a canonical form. In our setting, mentions appear at the *cell level* of the generated tables. Formally, given a set of cell-level entity mentions $E = e_1, e_2, \cdots, e_n$, the goal is to normalise each e_i to a unique canonical entity $c_j \in C$, where C is the set of standard forms.

Our benchmark datasets include normalised entities to support this task (e.g, SciREX, TDMS, FE). Unsupervised methods such as string similarity metrics (e.g., Levenshtein distance or token-based ratios), string matching libraries (e.g., `TheFuzz`[23]), semantic similarity methods (e.g., embedding-based approaches), or entity ranking algorithms (Sung et al., 2020) can provide effective baseline solutions. These approaches rely on different types of similarity (surface-form or semantic) and can be useful depending on how the mention and the canonical entity vary (e.g., "TiO_2" vs. "Titanium dioxide" or "polymer chain" vs. "macromolecule").

Method: We apply a semantic similarity approach for entity normalisation. We use `PapersWithCode`[24], a comprehensive resource on leaderboard entities, to get normalised forms of entities in the TDMS dataset, which itself adopts PapersWithCode for normalisation. Embeddings are computed for both the extracted values and the candidate normalised entities using the pretrained model[25]. Cosine similarity is then applied to map each extracted value to its closest normalised form. This normalisation post-processing results in a substantial performance gain, with the F_1 score increasing from 3.59 to 18.56.

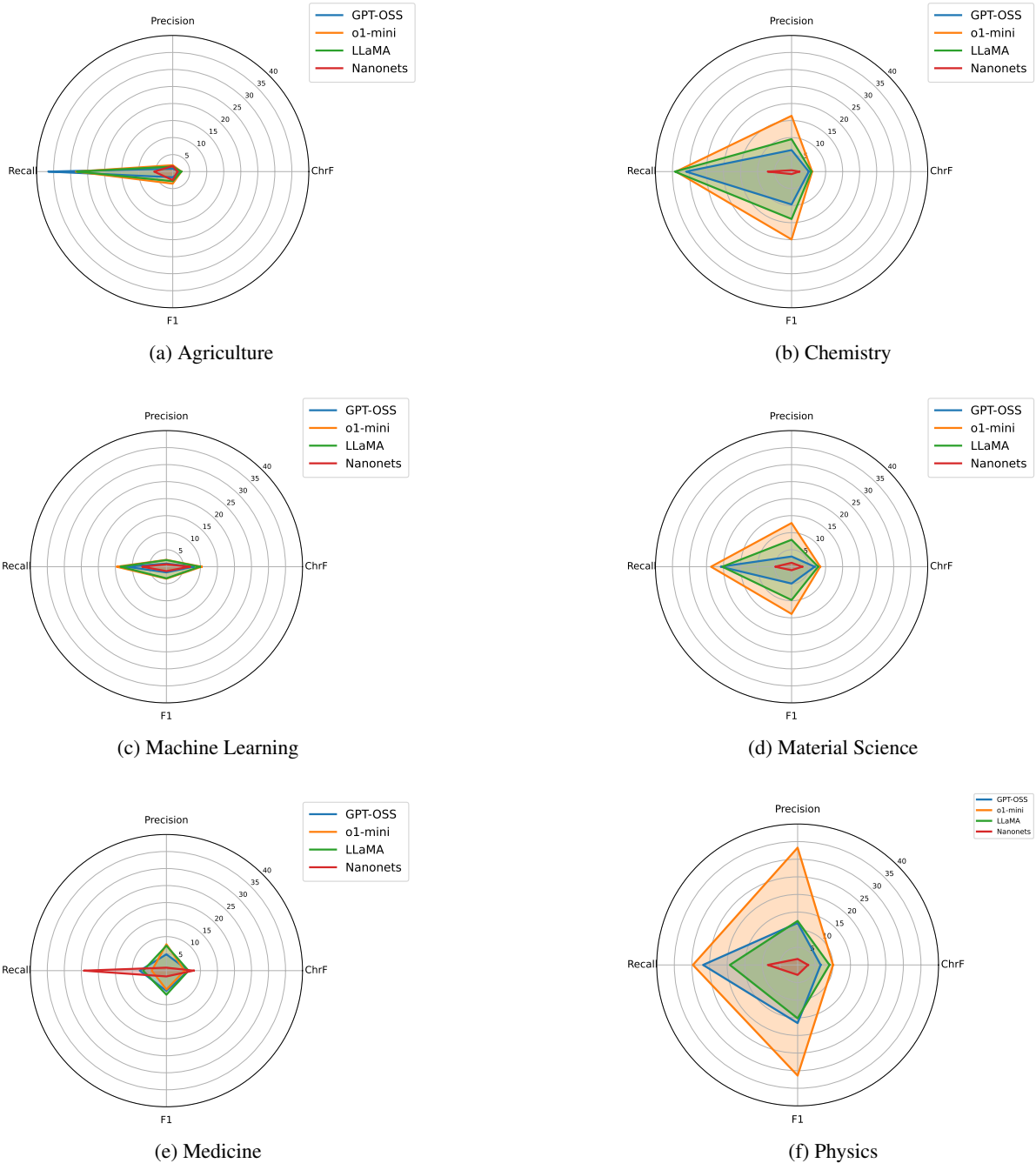
D.2 Unit and Quantity Standardisation

Unit & quantity standardisation is the task of converting quantitative values into a common unit and

[23] https://github.com/seatgeek/thefuzz

[24] https://github.com/paperswithcode/paperswithcode-data

[25] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2



(a) Agriculture (b) Chemistry

(c) Machine Learning (d) Material Science

(e) Medicine (f) Physics

Figure 5: Radar plots showing the performance of GPT-OSS, o1-mini, LLaMA 3.3, and Nanonets (vision) across domains. Metrics include ChrF, Precision, Recall, and F_1, averaged over all datasets in each domain.

standardised numeric format. Given a set of quantity pairs $Q = (v_1, u_1), \cdots, (v_n, u_n)$ extracted from document d, where v_i is a numeric value and u_i is its associated unit (stored in separate columns in the generated table), the goal is to convert each (v_i, u_i) into a standardised pair $(v_i^{'}, u_i^{'})$ using unit conversion rules.

Our benchmark datasets provide raw-value-unit attributes to support this task (BRENDA-enzyme and BRENDA-ribozyme). Simple methods include rule-based conversion systems (Jiang et al., 2025), curated unit mappings, or libraries like `pint`[26].

Method: We apply unit normalisation using `pint` library on the BRENDA-enzyme dataset, where units are annotated as a separate attribute (see Table 3). In this dataset, K_m values are standardised to millimolar (mM) and k_{cat} values to per minute (min^{-1}). The `pint` library enables convert-

[26] https://github.com/hgrecco/pint

Domain	Dataset	Attribute	Coverage (↑)	MAE (↓)	$p\mathbf{1}mag$ (↑)
Machine Learning	TDMS	Score	65.76	30.32	81.56
	SciREX	Score	63.30	496.51	78.82
	LEGOBENCH	Score	70.10	703213.97	76.71
Material Science	MPEA	grain size	40.89	52.14	78.26
		experimental density	11.68	571.10	88.54
		test temperature	37.46	453.88	53.48
		yield strength	67.37	241.70	94.39
		ultimate tensile strength	54.62	237.17	93.31
		elongation	37.92	12.94	81.86
		elongation plastic	40.43	2.49	87.72
	Diffusion	test temperature	75.94	331.87	98.32
		pressure	50.94	1332.31	28.35
		diffusivity	43.20	11.72	17.09
		SiO2	21.45	9.75	90.15
		TiO2	18.83	0.39	84.14
		Al2O3	16.39	2.40	97.51
		FeOt	18.58	2.24	84.42
		MnO	20.25	0.01	99.24
		MgO	17.75	1.07	95.77
		CaO	17.07	2.04	95.41
		Na2O	16.59	0.98	98.64
		K2O	16.02	0.31	99.46
		P2O5	15.62	0.01	100.00
		H2Ot	36.02	1.40	85.96
	YSHEAY	Value	91.36	246.60	86.11
	CCRMG	Value	71.08	515.17	74.58
	Doping	—	—	—	—
	MMD	Formula units per cell	26.92	1.57	100.00
		Atomic sites per cell	11.54	2.67	100.00
		Formation energy (eV/atom)	23.08	16.41	0.00
		Energy relative to convex hull (eV/atom)	30.77	20.35	0.00
		Averaged magnetic moment (μB/atom)	30.77	0.72	87.50
		Magnetic polarization, Js</(T)	11.54	2.67	66.67
		Magnetic anisotropy constants:Kb-c (MJ/m3)	50.00	1.14	50.00
	MRL	Electrical resistivity (Ωcm)	20.98	38.09	7.81
		Seebeck coefficient (μCV/K)	29.51	147.99	72.22
		ZT (thermoelectric figure of merit)	46.56	0.63	83.80
		temperature	71.15	272.25	95.39
		Thermal conductivity (W/mK)	34.38	13.80	83.12
		Average atomic mass (g/mol)	0.33	0.24	100.00
		Atoms per unit cell	3.28	4.73	100.00
	PNCExtract	Filler Composition Mass	81.06	7.88	9.03
		Filler Composition Volume	77.10	6.43	19.39
	PolyIE	Property Value	91.73	43.91	76.13
Chemistry	BRENDA-enzyme	km	57.83	61763.09	43.92
		kcat	56.21	4814.92	71.45
		km_kcat	53.85	633623.38	35.93
		Commentary[Temp]	30.85	4.12	99.49
		Commentary[pH]	26.16	0.41	99.87
	BRENDA-ribozyme	km	57.98	183.64	86.47
		kcat	58.94	5.61	83.61
		km_kcat	29.62	122445.26	74.31
		Commentary[Temp]	36.21	4.50	99.29
		Commentary[pH]	44.39	0.38	100.00
	OPE	Von [V]	66.87	1660.93	76.58
vim	PPE	PCE_max(%)	89.81	0.28	100.00
		PCE_ave(%)	69.62	0.28	100.00
		Voc (V)	79.25	0.06	99.52
	SE	temperature	80.81	34.12	99.02
		pressure	74.03	160.54	94.42
		solubility	64.18	643.14	77.95
Medicine	AE	Affinities	72.83	33849.81	54.48
Agriculture	FE	screw dimension	23.91	40.12	90.91
Physics	SuperMat	tcValue	66.74	9.14	92.64
		Pressure	82.08	2.99	89.66

Table 5: Quantitative evaluation of numerical attributes using o1-mini predictions. **Coverage** indicates the proportion of numerical values predicted, **MAE** denotes Mean Absolute Error, and $p\mathbf{1}mag$ denotes Within One Order-of-Magnitude. Note that the Doping dataset has no numerical attributes.

ing units (e.g., K_m: 15 $\mu M \rightarrow 0.015\ mM$). Evaluation on these quantitative attributes (K_m and k_{cat}) demonstrates an improvement in performance, with the F_1 score increasing from 34.2 to 38.7 after unit normalisation.

D.3 Entity Linking

Entity linking is the task of mapping entity mentions in the generated table to entries in an external knowledge base. Given a set of mentions $M = m_1, m_2, \cdots, m_n$ and a knowledge base $\mathcal{K}$, the goal is to identify a mapping function $f : M \rightarrow \mathcal{K}$ such that $f(m_i)$ returns the correct canonical entity from $\mathcal{K}$ for each mention m_i.

Our benchmark includes machine learning datasets (TDMS, SciREX and LEGOBENCH), which contain *task*, *dataset*, and *metric* attributes presented in their canonical forms and can be evaluated for this subtask. One of the solutions for entity linking includes computing semantic similarity between mentions and candidate entries using pretrained embedding models (Karadeniz and Özgür, 2019; Yamada et al., 2020).

Method: We apply the same normalisation method for entity linking on the TDMS dataset.

D.4 Table and Figure Understanding

Table and figure understanding is the task of extracting structured information from tables and figures in scientific documents. Given a visual input x (a table or figure) embedded in a document d_i, the goal is to parse its structure and extract content to enrich the textual representation of d_i. For tables, this typically involves recovering explicit information such as numerical values, labels, and table structures. For figures, the extracted information may be explicit (e.g., axis labels, data points) or implicit (e.g., observed trends, comparisons, or relationships). Understanding tables and figures is also important for tasks such as reasoning (Lu et al., 2025), text generation (Moosavi et al., 2021; Chen et al., 2021), interpretation (Roberts et al., 2024), and question answering (Borisova et al., 2025), as it enables models to integrate visual and textual information for comprehensive scientific analysis.

All datasets in our benchmark are suitable for the table understanding subtask. For the figure understanding task, material science datasets (e.g., Diffusion, MPEA, and MRL) can be used since they require extensive figure understanding for the STG task. However, our benchmark does not provide ground-truth annotations for table or figure understanding. Therefore, evaluation of these subtasks can only be performed extrinsically, via their impact on STG.

D.5 Quantitative Value Evaluation

Quantitative value evaluation is the task of assessing the correctness of extracted numerical values, accounting for unit normalisation, numeric format, and acceptable tolerances. Given a predicted value $\hat{v}$ and reference value v, the goal is to determine if $\hat{v}$ matches v within a defined equivalence function $f_{\text{eq}}(\hat{v}, v) \in 0, 1$ that incorporates value equivalence, unit consistency, and numeric precision.

Method: We evaluate numerical correctness using quantity-based metrics defined in the evaluation framework (see Section 5), and the results are presented in Table 5.

Recovering Method Parameters from Protocol Prose: Task Definition, Audited Extraction, and the Comparability It Licenses

Fabio Rovai
The Tesseract Academy, London, United Kingdom
fabio@thetesseractacademy.com

Abstract

Repositories usually report metadata quality as completeness, the share of fields filled. We measure instead whether two deposited analyses record enough to be compared. Over all 167,331 pairs of gas chromatography analyses in Metabolomics Workbench, 120 pairs meet the necessary conditions for comparing retention indices and two meet those for comparing retention values; one field, the retention index reference series, recorded for 19 of 579 analyses, removes 99.9% of the pairs that share a stationary phase. The oven temperature programme is usually written in the protocol prose instead, and the series sometimes is, so we define their recovery as an extraction task with ordered output and explicit abstention. Audited by two blind LLM raters whose error detection we measure on planted errors, the programme extractor is 65% precise on 40 random MetaboLights recoveries with no fabricated value, recovers in full 49% of the programmes stated in a separate random sample of 40 protocols, and the series extractor is correct on 57 of 66. On the census corpus itself, recovered series raise the pairs that can compare indices from 120 to 171, or from 78 to 92 when the two series must be identical. Of 239 populated Workbench programme fields, 130 hold only a range, a single temperature or a pointer to a paper, yet a completeness audit counts all of them as filled. Every released value carries its source sentence and is attributed to the extracting pipeline, not to the depositor.

1 Introduction

The standard quality report for a scientific data repository is a completeness audit. Fields are enumerated, populated values are counted, a percentage is published, and improvement programmes target the percentage. Minimum information checklists are written in the same currency (Sumner et al., 2007), and compliance with them has been measured across public metabolomics repositories in exactly these terms (Spicer et al., 2017).

A populated field, however, is not a usable field. In a public reference spectral archive of 137,030 spectra (Horai et al., 2010), the collision energy of a mass spectrum is recorded on 84.2% of records and written in 1,023 distinct ways, so joining on it exactly retains 9.3% of the comparisons that a normalised join retains (Rovai, 2026a). Completeness reports 84.2%. The figure a consumer can act on is 9.3%.

What a consumer needs is not whether a field is filled but whether two records can be used together, and for what. That property is relational rather than intrinsic, graded rather than binary, and computable from the record alone. Section 3 makes it precise as four cumulative conditions over pairs of analyses, first reported with a census in Rovai (2026a): of the 167,331 pairs of gas chromatography analyses in the Metabolomics Workbench corpus (Sud et al., 2016), two meet the condition for comparing measured values. Among pairs that share a stationary phase, the collapse is concentrated in a single field, the retention index reference series, which 19 of 579 analyses record.

The contribution of this paper begins where that diagnosis ends. The field is missing from the structured record and present in the text. In the MetaboLights corpus (Haug et al., 2013), none of 355 gas chromatography studies declares a protocol parameter for the oven temperature programme or for the reference series, while their protocol prose usually describes the programme and sometimes the series. The repair for a metadata deficit that no curation campaign has closed is therefore an information extraction problem over documents the repository already stores, and it is measurable in the units that matter, which are comparisons licensed rather than fields filled.

We make five contributions. We define recovery of two chromatographic method parameters from protocol prose as an extraction task with an ordered

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 247–261
November 10, 2026 ©2026 Association for Computational Linguistics

output and a stated correctness criterion (Section 4). We build an extractor whose governing difficulty is exclusion rather than recognition, because the decoy values sit in the same clauses as the targets at similar magnitudes (Section 5). We audit both extractors by reading, separating precision, structural agreement and fabrication, with two blind LLM raters whose error detection we measure on planted errors (Section 6). We close the loop on the census corpus itself, reproducing the census as a control, adjudicating every recovery that changes it, and measuring the result under three readings of the reference-series condition, together with the cost of transferring an extractor between repositories (Section 7). And we release the recovered values with the sentence each came from, attributed in PROV-O to the extracting pipeline, which makes the release an adjudicable evaluation resource (Section 9).

2 Related work

Extraction as archive infrastructure. This workshop has repeatedly addressed the layer this paper works on. Research-data metadata has been generated from citation contexts as a defined task with a constructed dataset (Watanabe et al., 2025), articles have been linked to the data publications that describe them (McGillivray et al., 2025), URL citations have been classified to promote research artefacts (Tsunokake and Matsubara, 2022), property values have been extracted from tables (Phi et al., 2025), leaderboards assembled by linking table entities (Yang et al., 2022), and concepts extracted from four hundred thousand papers (Ting et al., 2025). Extraction has been placed in an archive's ingest path (Susemiehl and Mazzarella, 2025), reproducibility signals read off articles (Akella et al., 2022), and quantities (Bölücü et al., 2023) and bibliographic structure (Joshi et al., 2023) extracted from scientific text. Two features of that work carry over to ours: the deliverable is typically a corpus or a service component rather than a model, and evaluation often has to be built alongside the extractor, because for many of these targets no reference annotation exists in advance.

Defining a target and building its resource. Our framing follows the same practice. The PICO corpus was constructed specifically to support automatic data extraction from biomedical literature (Mutinda et al., 2022), and MuLMS annotates materials-science text in layers for information extraction (Schrader et al., 2023). Both make the bet we make, that the bottleneck is a precisely defined target with adjudicated examples rather than model capacity.

Method and procedure extraction. Reading method detail out of scientific prose is established practice in chemistry and materials science. ChemDataExtractor recovers chemical identifiers, properties and spectra from articles and tables (Swain and Cole, 2016), synthesis procedures have been converted into ordered action sequences (Vaucher et al., 2020), and a text-mined corpus of inorganic synthesis recipes was built from published methods sections (Kononova et al., 2019). Our task differs in two ways that matter for a repository: the output is an ordered numeric programme, so the right numbers in the wrong order are wrong, and the target is a parameter of the instrument behind a deposited measurement, so the extracted value has a determinate consumer, the comparison the repository's schema cannot license.

Why completeness is the wrong target. The checklists and compliance studies of Section 1 (Sumner et al., 2007; Spicer et al., 2017) count fields filled. The community-level frame is the FAIR principles (Wilkinson et al., 2016), whose interoperability criterion is where our measurement bites: a field that is present, populated and written in 1,023 ways is findable and accessible without being interoperable, and a completeness percentage cannot tell the difference.

Graded comparability. The notion we use is imported rather than invented. Psychometrics tests configural, metric and scalar invariance as separate cumulative claims, each licensing a specific comparison (Meredith, 1993). The same problem is solved under other names elsewhere, including calibration transfer in chemometrics (Wang and Kowalski, 1992) and retrospective harmonisation in epidemiology (Fortier et al., 2017). We adopt the psychometric structure because it alone treats comparability as graded and separately testable; its factor-analytic machinery is not needed, since for an instrument record the predicates are decidable from the metadata.

3 Where comparability is lost

A ladder over pairs. We instantiate the psychometric levels for gas chromatography as four cumulative predicates over an unordered pair of analyses

(a, b), following Rovai (2026a). Writing m for the recorded modality, ϕ for the stationary phase family, σ for the retention index reference series, π for the thermal programme, $\perp$ for a value the record does not carry, and a subscript for which of the two analyses a value belongs to:

$$
\begin{aligned}
L_1 &\iff m_a = m_b \\
L_2 &\iff L_1 \wedge \phi_a = \phi_b \neq \perp \\
L_3 &\iff L_2 \wedge \sigma_a, \sigma_b \neq \perp \\
L_4 &\iff L_3 \wedge \pi_a = \pi_b \neq \perp
\end{aligned}
$$

The rung of a pair is $\max\{i : L_i(a, b)\}$. The first rung says the records are about the same kind of measurement, the second is what comparing elution orders needs, the third what comparing retention indices needs, and the fourth what comparing retention values directly needs. Assignment uses only what the record states. There is no imputation, and at this stage no inference from publication text, which is exactly what separates this section from the rest of the paper.

What the predicates assert. Each predicate is a necessary condition for the comparison it names, read off the record, and none is sufficient. L_2 asks for the same phase family because selectivity, and with it elution order, is a property of the phase (Rohrschneider, 1966; McReynolds, 1970); it does not test column length, film thickness or carrier flow. L_3 asks for a reference series on both sides because an index is defined only relative to its series (Kováts, 1958; van den Dool and Kratz, 1963), and as written it asks for presence, not identity. That is permissive in a way that matters. The FiehnLib library indexes retention against fatty acid methyl ester (FAME) markers from C8 to C30 rather than against n-alkanes, and its authors interconvert the two scales with a fitted fifth-degree polynomial rather than treating them as one (Kind et al., 2009); an index computed against FAMEs and one computed against alkanes are comparable only through such a transformation. L_4 adds programme equality, which puts both runs in the same interpolation regime, logarithmic for isothermal operation and linear for temperature-programmed operation, but which regime was applied is recorded for one analysis in the corpus. The ladder is therefore an operational approximation: a pair that fails a rung does not, on its record, support its comparison, and a pair that passes may still not, and Section 7 reports how the result changes when L_3 also requires the two series to be of the same kind, and then to be the same series.

Rung	Pairs at or above	Share
L_1 configural	167,331	100%
L_2 metric, scale	98,479	58.85%
L_3 metric, calibrated	120	0.072%
L_4 scalar	2	0.001%

Table 1: Comparability census over all pairs of gas chromatography analyses in the Metabolomics Workbench corpus (Rovai, 2026a). The step from L_2 to L_3 removes 99.9% of what survived the first step, and its cause is one unrecorded field.

The census. We classified every unordered pair of the 579 gas chromatography analyses in the Metabolomics Workbench corpus, giving 167,331 pairs. Because every pair is classified rather than sampled, the counts in Table 1 are a census and carry no sampling error; the uncertainty in them is definitional, in ϕ, σ and π.

Two pairs in one hundred and sixty-seven thousand meet the necessary conditions for comparing values. Programme equality is tested on the ordered numbers in a free-text field, which can equate programmes whose numbers play different roles and separate identical programmes written in different units; both pairs it finds carry character-identical programme text.

One field does the damage. The losses are unequal. Differing stationary phases remove 68,852 pairs, differing thermal programmes remove 118, and the absence of a retention index reference series on at least one side of the pair removes 98,359. That single field is stated for 19 of 579 analyses, and the convention by which the index was computed is stated for one. The distinction is not pedantry. Kováts interpolates logarithmically and applies to isothermal operation (Kováts, 1958); van den Dool and Kratz interpolate linearly and apply to temperature-programmed operation (van den Dool and Kratz, 1963). Which one is correct depends on how the oven ran, and a bare index does not say, so the reference series and the programme are jointly the calibration.

4 Task definition

The two fields that govern the census are missing from the structured record in different ways in the two repositories we use (Table 2). MetaboLights offers no slot for either: none of its 355 gas chro-

Workbench, 579 analyses	Programme	Series
Key absent, empty or nullish	340	560
Key populated	239	19
yields a structured value	62	15

Table 2: The state of the two fields in Metabolomics Workbench. A programme is structured when the transferred extractor of Section 7 recovers one from the field; a series when the value names both its kind and its carbon span. Of the other four populated series values, two name only a kind and two hold an index range (1000–3100) rather than a series. MetaboLights has no slot for either field in any of its 355 gas chromatography studies.

matography studies declares a protocol parameter for the reference series or for an oven programme, and two declare a parameter named only *Temperature*. Metabolomics Workbench has a key for each, and both keys are mostly absent or insufficient. The reference-series key is absent from 560 of 579 analyses, and of the 239 analyses whose programme field is populated, 130 hold only a range, a single temperature or a pointer to a paper (Section 7). In MetaboLights the programme, and sometimes the series, is written in the protocol prose instead. We therefore define their recovery as an extraction task. A system receives the free-text protocol of one deposited study and returns, for each of two targets, either a structured value or an explicit abstention.

Target 1: the thermal programme. The output is an ordered sequence of steps. Each step carries an oven temperature in degrees Celsius, and optionally a ramp rate in degrees per minute leading to it and a hold duration in minutes at it. Order is part of the output, because the same temperatures in a different sequence describe a different experiment. A returned programme is correct when its steps, their order, and the rates and holds attached to them agree with the protocol as it is written.

Target 2: the retention index reference series. The output is a normalised carbon series, that is, the homologous series against which a reported retention index was computed, with its kind (n-alkane or FAME) and its span, so that two studies naming it differently become comparable. Correctness is agreement in kind and span with the series the protocol states.

Abstention is a first-class output. A protocol that does not state a programme must yield no programme. This matters more than it usually does, because the consumer of the value is a comparability predicate: a fabricated programme does not merely add noise, it promotes a pair to a rung it has not earned and licenses a comparison that is invalid. An omission leaves a visible gap; an invention is indistinguishable from a recorded measurement. We therefore treat the two error directions as unequal throughout, and prefer an extractor that drops a hold time to one that invents one.

Why this is a task and not a cleanup. The two targets are instances of a pattern we expect to generalise: the field that governs whether two records can be combined is frequently the field the depositing schema has no slot for, or a slot nobody fills, while the prose that accompanies the deposit carries it. The task is therefore defined by the consumer rather than by the text. Given a comparability predicate, the extraction targets are whichever of its arguments the record leaves as $\perp$, and success is measured as comparisons licensed rather than as fields filled.

The task is small enough to run as a shared task of the shape this workshop has carried before (Grezes et al., 2022, 2025): the inputs are public, the outputs short and structured, the adjudication protocol is that of Section 6, and the adjudicated items we release are a starting gold standard.

5 Corpus and method

Corpus. We use the 355 gas chromatography studies of the MetaboLights corpus, a public, general-purpose metabolomics repository whose deposits pair machine-readable assay tables with free-text protocol descriptions. The census of Section 3 is over Metabolomics Workbench, and the extractors were developed on MetaboLights. Section 7 then applies them to the Workbench records themselves, so that the effect of recovery on the census is measured on the census corpus and not inferred across repositories.

Extractor. The programme is read as an ordered token stream rather than by independent pattern matches. The walk proceeds left to right through the protocol, emitting temperatures, ramp rates and durations as they are encountered and binding each rate and duration to the temperature it governs by position in the stream. Reading in order is what makes the binding sound. Independent pattern matches carry no relative order between them, so on a multi-ramp programme a rate can bind

Protocol text (study MTBLS3, verbatim, as stored with the recovered value)

column temperature was 110 °C; this was held for 2 min and then increased by 30 °C/min to 310 °C. The column effluent was introduced into a Trace DSQ quadrupole mass spectrometer (Thermo Electron; **transfer line temperature = 320 °C**, **ion source temperature = 240 °C**, *electron beam = 70 eV)*

Step	Ramp	Target	Hold
0		110 °C	2 min
1	30 °C/min	310 °C	

Figure 1: A recovered programme and the sentence it came from, as released. The two temperatures in bold are decoys inside the same evidence window, one above the oven's final temperature and one inside its range, and both must be rejected. Storing the evidence is what makes a single extraction adjudicable by reading rather than by rerunning.

to the following clause's target, and two named methods can be spliced into a single programme that no instrument ran. The procedure is given as pseudocode in Appendix A, so that it can be reimplemented from this paper alone.

Exclusion is the hard part. Recognition of a temperature is trivial and almost all of the engineering is in refusing the wrong ones. Three failure classes were diagnosed and each has a specific mitigation, which we state because the mitigations are the method.

Decoy capture. Oven temperatures share their clauses, at similar magnitudes, with inlet, injector, transfer line, interface, ion source, detector, quadrupole, septum, desorption and incubation temperatures, among others (Appendix A lists all 22 governing terms). An extractor that takes the first temperature after an anchor term returns a programme beginning at the injector. We reject any candidate governed by a decoy term within a bounded window, and verify the rule on cases where a decoy occurs in the same sentence as the target.

Figure 1 shows the difficulty concretely. In that study the transfer line sits at 320 °C, above the oven's final temperature, and the ion source at 240 °C, inside its range, in the clause immediately following the programme. Nothing about magnitude, position or units separates the wanted values from the unwanted ones; only the governing noun does.

Fabrication from adjacency. A study whose final temperature is followed by the clause *"the total run time was 50 min"* yielded a fifty-minute hold that never happened. We exclude durations governed by run-time and ramp-duration language, a rule that caught a second instance immediately on introduction.

Substring collapse. The third class arose in the stationary-phase normaliser of Section 3 rather than in the extractor: a pattern containing `5ms` and bounded only on the right matches inside `DB-35ms`, a chemically different phase, and yields a plausible value rather than an obvious error. Every pattern that reads a chemical identifier is now bounded at both ends.

The reference-series extractor buys precision with recall. It parses a carbon range or list only when a mention of an alkane, a fatty acid methyl ester, a retention index or Kováts appears in a window running 120 characters before the match and 60 after it (Appendix B). That construction is why its precision is high, and it is also its limit: a series described without carbon numbers is not recovered at all. We report this as a design choice rather than a defect, because for a value that will license a comparison we prefer a narrow extractor with a visible miss rate to a broad one with an unknown error rate. The same reasoning governs abstention on two studies whose protocols name two distinct methods: rather than choose between them, the extractor emits nothing and the studies are withheld.

6 Evaluation

Yield. The programme extractor recovers 206 ordered programmes from the 355 studies, 58.0% of the corpus, comprising 529 temperature steps of which 370 carry a hold time, giving 135 distinct programmes. Sixty-six of the 206 are multi-ramp, which matters because every failure of the extractor's first version (Table 4) fell in that class. The series extractor recovers a normalised carbon series for 66 studies, 18.6%. Neither target has a structured field in this corpus, so both figures are gains over nothing that a machine could previously join on.

How often the prose states each target is a separate question, and we have answered it only heuristically. The programme's anchor pattern (oven, column temperature, temperature programme and their variants; Appendix A) matches the protocol text of 279 studies, 78.6%, and a series term (retention index, alkane, Kováts, FAME, fatty acid

Thermal programme	
Studies in corpus (gas chromatography)	355
Anchor phrase in protocol text (regex)	279
Programmes recovered	206
of which multi-ramp	66
Temperature steps recovered	529
of which carrying a hold time	370
Distinct programmes	135
Median initial, final temperature (°C)	60, 315
Median ramp rate (°C/min)	10
Retention index reference series	
Series term in protocol text (regex)	181
Studies with a series recovered	66
By kind: n-alkane / FAME	51 / 21
By form: range / list	63 / 9
Commonest span	C10–C40

Table 3: What the two extractors recovered from 355 MetaboLights gas chromatography studies, none of which has a structured field for either target. Kind and form counts exceed the study count because some protocols state more than one series.

Audit	**Verdict**	**Rate**	**95% CI**
Programmes, MetaboLights, 40 draws			
fully correct	26 of 40	65%	50–78
structural agreement	38 of 40	95%	83–99
free of fabrication	40 of 40	100%	91–100
Programmes, Workbench transfer, 25 draws			
fully correct	22 of 25	88%	70–96
Reference series, every recovery			
MetaboLights	57 of 66	86%	76–93
Workbench	5 of 5	100%	57–100
Development checks, single pass			
programme, final	10 of 14	71%	45–88
programme, tightened	11 of 12	92%	65–99
programme, first	7 of 12	58%	32–81
reference series	12 of 12	100%	76–100
internal std., withdrawn	4 of 12	33%	14–61

Table 4: Every precision audit in this work, with Wilson intervals in per cent (Wilson, 1927). The upper blocks are the adjudicated verdicts. Only the "fully correct" and series rows are precisions; structural agreement and freedom from fabrication are separate properties, defined in Section 6. The series blocks are censuses of every recovery, not samples.

methyl ester) occurs in 181. These are regular-expression counts, not annotations.[1] To measure recall instead, both raters read the full protocols of 40 studies drawn at random from all 355, and agree on all 160 judgements. Of these, 35 state an oven programme; the extractor returns a programme that agrees with the text in every temperature, rate and step order for 18 of them, a recall of 51% (95% interval 36–67; 49% if an omitted hold counts as a miss), and returns a programme for no study that states none. Of the 40, 7 state a reference series and the series extractor finds 5. On the same sample the anchor pattern matches 30 of the 35 studies that state a programme and 1 that does not, so the regex count of 279 undercounts the studies that state a programme. The extractor returns nothing for 17 stated programmes, 5 of them because the anchor never matches, and none of the 18 it returns invents a value.

Table 3 summarises what was recovered. The distributional properties are a weak but useful check that the extractor is reading oven programmes and not something else, since the median recovered programme runs from 60 °C to 315 °C at 10 °C per minute, which is an ordinary temperature-programmed separation.

Audit protocol. Precision cannot be read off the output distribution, which looks reasonable whether or not the values are right. We draw a random sample (seed 20260923) and show the raters described next each extracted value beside the sentence it was taken from, with up to 2,800 characters of the surrounding protocol for MetaboLights items. An extractor that changes is re-audited from a fresh draw, because a verdict on the previous version measures the previous version, and an extraction whose audited precision is poor is withdrawn rather than qualified.

Who adjudicated. Every audited item was judged by two blind LLM raters, Claude Opus 5.5 and Claude Fable 5.1, which agree on 133 of 136 verdicts; quoted protocol text settles the rest. The raters were tested rather than trusted: on 40 items, 24 of them deliberately corrupted, each catches 24 of the 24 with no false alarm, and a mechanical check with no model in it agrees with both on 129 items. The author's own reading of 48 cards agrees on 40 cards with the verdicts as they stood before it, 7 of them planted corruptions it caught, and one of its differences exposed an error both raters shared, now corrected. Appendix E gives the protocol in full, and the development checks at the foot of Table 4 carry no claim.

Three verdicts, kept apart. The programme audits use three separate verdicts. *Correct* means every step, its order, and the rates and holds at-

[1] Rovai (2026a) reports 284 studies for the series. That count also matched *internal standard*, which names a different object; excluding it leaves the 181 here.

MTBLS748, a dropped step

column temperature was set to 70 °C. A temperature ramp of 1 °C/min for 5 min followed by a temperature ramp of 5.63 °C/min to a final temperature of 330 °C with a hold time of 10 min

Output: 70 °C, then 5.63 °C/min to 330 °C, hold 10 min. The first ramp is lost; what survives is correct.

MTBLS12353, a wrong initial temperature

column temperature was maintained at 40 °C, and the injection volume was 2 µL. [. . .] *The oven temperature ramp was held at 90 °C for 1 min, raised to 100 °C at a rate of 25 °C/min*

Output opens at 40 °C, taken from a sentence that is not the programme, and the programme's own opening step at 90 °C for 1 min is dropped.

Figure 2: Two structurally wrong outputs from the 40-draw audit, quoted from their stored evidence. Neither invents a number: one loses a step and one binds the wrong sentence.

tached to it agree with the protocol as it is written, and this is the only programme figure we call precision. *Structural agreement* is weaker and reported separately: every temperature, every ramp rate and the order of every step agree, while a hold time stated in the text may be absent from the output. *Fabrication* means a value appears in the output that is in no reading of the text, and it is reported as its own rate because it is the failure the consumer of these values cannot detect.

On the 40-draw sample the extractor's precision is 26 of 40, 65%, with an interval of 50–78. Structural agreement is 38 of 40, and 0 of 40 outputs contain a fabricated value. Of the outputs that are not correct, 12 differ from the protocol only by an omitted hold time and 2 are structurally wrong. Because a reader cannot check a failure mode that is only described, the structural failures are reproduced in Figure 2. The second is decoy capture in a form the decoy list cannot catch, since the offending phrase is the anchor itself occurring in a sentence about something other than the oven.

Reference series. Every recovery was adjudicated rather than a sample. Of the 66 MetaboLights studies, 57 are correct and 8 get the span wrong; for the remaining 1, the protocol states no reference series. A list interrupted by a concentration or continued with *and*, as in *C22, C28 (each 0.1 mg/ml final concentration; Sigma Aldrich) and C32*, is cut at the interruption; and a mixture the text gives as two ranges, such as *C8-C16: 1 mg/mL, C18-C24: 0.5 mg/mL*, is output as two series. The five Workbench recoveries are taken up in Section 7.

A withdrawn target. A third extractor aimed at internal standard identities reported candidates in 138 studies. A random twelve returned four compounds and eight fragments of English, including *"and vortexed"* and *"see below"*. A precision near one third is not a measurement, so no yield is reported for it and nothing it produced appears in the release. A compound name needs a chemical entity recogniser and a registry lookup (Swain and Cole, 2016), not a pattern over tokens: the targets that shipped are numeric and positional, and the one that failed is nominal.

7 The same corpus, before and after

The question a consumer actually asks is how many more comparisons recovery licenses. We answer it on the census corpus itself.

Control. Before changing anything we recomputed the census of Section 3 from the same cache with the same field derivation. It reproduces exactly: 579 analyses, 167,331 pairs, 98,479 at L_2, 120 at L_3 and 2 at L_4. Had it not, nothing downstream would be reportable.

Transfer is not free, and the reasons are nameable. Applied to Workbench unchanged, the extractor of Section 5 structures 22 of 579 analyses, 3.8%, against 58.0% on MetaboLights. Two causes account for nearly all of the gap and neither is about the science. The extractor first locates the programme in a long protocol using an anchor phrase, and that phrase occurs in only 32 of the 239 populated programme fields, 13.4%, because a repository that gives the programme its own field has already done the locating. And the degree sign is written differently: Workbench records use `U+2103`, and some are mojibake, a UTF-8 degree sign decoded as Latin-1 arriving as `ˆA°`. Bypassing the anchor when the field is itself the window, and normalising the degree sign (Appendix C), raises the yield to 62 of 579 analyses, which is 25.9% of the 239 populated fields. An audit of 25 of those extractions finds 22 correct, 88% with an interval of 70–96 (Table 4); every failure omits a hold written as *minute* or *minutes* in words, which the hold pattern does not match.

Most of the remainder has no programme to extract. Table 5 classifies the 177 populated fields that yield nothing. A temperature range with no

What the populated field holds	n	%
A temperature range, no rate or hold	73	41.2
A single temperature	56	31.6
Other, including inlet-only values	34	19.2
A rate is present, extractor declined	6	3.4
Ramps given as durations, not rates	5	2.8
Nullish text in a populated slot	2	1.1
Defers to a cited publication	1	0.6
Total declined	**177**	

Table 5: The 177 populated Workbench programme fields from which no programme is recovered. Only the fourth and fifth rows, 11 fields, are extractor misses; the first, second and last rows, 130 fields, hold no programme at all.

rate and no hold accounts for 73, a single temperature for 56, and one field defers to a cited paper, reading *"According to temperature gradient. DOI: 10.1385/1-59745-003-0:439"*. Only 11 of the 177 are extractor misses with a recoverable programme present. A completeness audit scores all 239 populated fields as filled, and 130 of them hold no programme at all.

The five recoveries, one by one. The series extractor finds a reference series in the prose of five analyses whose series key is empty, raising the count from 19 to 24. Because the downstream result rests on these records, each is listed in Table 6 with the phrase it came from and its verdict; the full evidence windows are in Appendix D. Every L_3 pair in the census lies within one phase family, 5% phenyl, where 16 analyses state a series. Three of the five recoveries are on that family and add 18 pairs each, three of which are pairs among themselves, for 51 new pairs. The other two move no pair: one is on a family where no other analysis states a series, and the phase of the other is not resolved by the normaliser.

What recovery licenses, under three readings. Under the census's own predicate the pairs at L_3 rise from 120 to 171, an increase of 42.5%. Because L_3 asks only that a series be present on both sides (Section 3), we recomputed it twice more (Table 7). The 16 analyses on 5% phenyl that state a series are 13 stating FAME C8–C30, 2 stating FAME without a span and 1 stating n-alkanes C10–C40; the three recoveries add 1 FAME C8–C30 and 2 n-alkane C8–C40. Requiring the same kind of series separates the alkane analyses from the FAME ones, and requiring the same series leaves the FAME C8–C30 block and the single pair of C8–C40 alkanes. Under the strictest reading the gain is 14 pairs, 17.9%. The direction holds under every reading; its size depends on the predicate as much as on the extraction.

The scalar rung does not genuinely move. Both census L_4 pairs carry character-identical programme text; the transferred extractor recovers both programmes of one pair and neither of the other, whose fields write the degree sign as a superscript zero (`U+2070`), which the normalisation of Appendix C does not map. The recomputed count of 1 is therefore an extractor miss, not a stricter comparison, and on the evidence of this corpus two pairs meet the necessary conditions for a direct comparison of retention values.

8 Why the field is empty

Fixing the schema is the obvious remedy, but across four disciplines with machine-readable formats (Rovai, 2026a) the field that makes a record interpretable is present on 100% of seismic channels, 53.8% of metrology records, 16.6% of meteorological deployments and none of the 355 MetaboLights gas chromatography studies, although in none of the four is the field required (International Federation of Digital Seismograph Networks, 2019; World Meteorological Organization, 2019) and none is validated. We offer a hypothesis these domains suggest rather than establish: a field is complete when its primary consumer cannot complete the primary task without it. A seismometer's counts mean nothing without the transfer function; a chromatogram without its programme still yields peaks, and the programme is needed only to *compare*, which is somebody else's task, later. Recovery from prose is then a realistic remedy, because a methods section is written for a reader who must judge the work.

9 Release

The recovered values are released in version 0.2.0 of the Zenodo deposit (Rovai, 2026b) that accompanies Rovai (2026a), as two graphs in a vocabulary of OWL 2 modules with SHACL (World Wide Web Consortium, 2017) validation layers: 4,256 triples for the 206 thermal programmes and their 529 ordered steps, and 581 triples for the 72 reference series of 66 studies. Every verdict from every rater and check on every audited item, the five Workbench recoveries and the scripts that regenerate every figure in this paper are deposited with them.

Analysis	Recovered series	Evidence (verbatim)	Phase family	New L_3	Verdict
ST001369	FAME C8–C30	*a standard mix of fatty acid methyl esters (FAME C8-C30)*	5% phenyl	18	correct
ST000496	n-alkane C8–C40	*A standard alkane mixture (C8-C40) was injected at the beginning of the analysis*	5% phenyl	18	correct
ST001134	n-alkane C8–C40	*Retention indices (RIs) were calculated relative to those of a C8-C40 alkane series*	5% phenyl	18	correct
ST000980	FAME C8–C30	*a standard mix of fatty acid methyl esters (FAME C8-C30)*	not resolved	0	correct
ST001432	n-alkane C7–C30	*a standard alkane series mixture (C7-C30) was injected before every batch*	100% dimethyl	0	correct

Table 6: Every reference series recovered from prose on the census corpus. New L_3 counts pairs involving the analysis; the three 5% phenyl recoveries share three pairs, so the distinct total is 51.

L_3 requires	Before	After	Gain
a series on both sides	120	171	42.5%
the same kind of series	105	123	17.1%
the same series	78	92	17.9%

Table 7: Pairs at L_3 on the census corpus before and after recovery, under the census predicate and two stricter readings. L_4 is 2 before and 1 after under all three; the lost pair is an extractor miss (see text).

Every recovered value carries the sentence it was extracted from, so any extraction can be adjudicated by reading rather than rerunning, and the adjudicated value and sentence pairs seed a gold standard for the task.

Every recovered value is attributed, in PROV-O (Lebo et al., 2013), to the software agent that extracted it, and carries the time and the extraction run that generated it, rather than being written into a metadata field as though the depositor had stated it. An extraction is a claim about what a text says, and recording it as the depositor's own value launders a pipeline's inference into a scientist's assertion. Attribution keeps the two separable and lets a later extractor supersede an earlier one without rewriting the record it read. A SHACL shape rejects any recovered value that lacks the attribution, the time or the run. Version 0.1.0 of the deposit named the pipeline only in a file comment, and the shape reports 834 violations against it; against the version cited here it reports none, and no recovered value differs between the two.

The extraction graphs are validated on their own. Against all five SHACL layers the programme graph conforms, and two deliberate corruptions, a step with no temperature and a temperature written as text, are each rejected, so the step shapes can fail. The series graph is checked by the provenance shape alone, since no other shape targets its class. The deposit's wider bridge graph, which contains none of the extracted values, has no violations and 531 open warnings, all recording that a percept stored without a stimulus concentration cannot be compared with another panel's; the rule is left firing so that the count can serve as a progress metric, in the manner of crystallography's structured validation alerts (International Union of Crystallography, 2026).

10 Conclusion

A repository's quality report should say what its records can be used for, not how many of its fields are filled. Asked of a complete corpus of gas chromatography analyses, the answer is that two pairs in 167,331 meet the conditions for a direct comparison of values, and that one field, missing from the record and sometimes recoverable from the prose, removes almost every pair that shares a stationary phase. On the same corpus, three recovered series raise the pairs that can compare indices by 42.5%, or 17.9% when the series must be identical, while 130 populated programme fields that a completeness audit counts as filled hold no programme at all.

Limitations

The gain at the calibrated rung rests on three recovered records out of five. Because the rung is a property of a pair, three records move 51 pairs, which is the mechanism working as intended, but the figure is sensitive to those records and to the predicate: the 42.5% under the census predicate is 17.9% when the two series must be identical.

The comparability ladder is an operational approximation. Each predicate is a necessary condition read off the record, and none is sufficient: L_2 and L_4 do not test column dimensions or car-

rier flow, and the census form of L_3 does not test whether the two series are on the same scale. The logic of graded, separately tested levels transfers from psychometrics; the factor-analytic apparatus that establishes those levels there is neither carried across nor needed here.

Recall is measured on a sample of 40 studies, so its interval is wide, and it rests on the same two raters as precision. The 58.0% and 25.9% yields are recovery rates against their denominators, not recall.

The verdicts rest on two LLM raters and a mechanical check, with quoted text settling disagreements. One person's reading checked them and exposed an error both raters shared; on 4 cards the reading and the text disagree. Neither is sufficient alone, and no item was annotated by two people. Planted errors measure detection only for the kinds of error we anticipated.

Four failure modes are diagnosed and deliberately not fixed: holds written as *minute* or *minutes* in words, a degree sign written as a superscript zero, a series list interrupted by a concentration or continued with *and*, and a mixture written as two ranges. Each repair is a small change to a pattern. We leave them because changing the extractor would invalidate the audit just reported, and the paper's own rule is that a changed extractor is re-audited from a fresh draw.

In MetaboLights the stored programme window is 300 characters, which truncates the longest programmes; the raters saw the surrounding protocol text in those cases, but a reader of the release who has only the window may not be able to settle them.

An LLM used as the extractor, scored by the same raters, checks and planted errors, is the obvious comparison and our next experiment.

The extractor is a bounded token walk with lexical exclusion rules, developed on English protocol prose and carrying a decoy list drawn from one instrument configuration. It has been measured on two repositories, and the transfer cost was two concrete assumptions rather than a redesign, but a third repository, another language or another analytical technique would call for its own audit. The one target we explored that is nominal rather than numeric, internal standard identities, needs a chemical entity recogniser with a registry lookup, which marks where this family of methods ends and another begins.

Ethics statement

The work uses only material already published in public repositories, and no human subjects data was collected. One data source used elsewhere in the wider deposit declined further automated requests partway through a crawl its robots policy permitted; we stopped at that point, and we withhold the graph derived from it because the source declares no licence. Recovered values are published as attributed claims of an extracting pipeline and not as statements of the original depositors, so no scientist is recorded as having asserted something they did not write. LLM assistants were used to write and run the analysis code and to draft text, and made the adjudication passes described in Section 6: Claude Opus 5 the development pass, Claude Opus 5.5 and Claude Fable 5.1 the two blind passes. The author read the audit cards described there and is responsible for the content.

References

Akhil Pandey Akella, Hamed Alhoori, and David Koop. 2022. Reproducibility signals in science: A preliminary analysis. In *Proceedings of the First Workshop on Information Extraction from Scientific Publications*, pages 140–144, Online. Association for Computational Linguistics.

Necva Bölücü, Maciej Rybinski, and Stephen Wan. 2023. Investigating the impact of syntax-enriched transformers on quantity extraction in scientific texts. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 1–13, Bali, Indonesia. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Isabel Fortier, Parminder Raina, Edwin R. Van den Heuvel, Lauren E. Griffith, et al. 2017. Maelstrom research guidelines for rigorous retrospective data harmonization. *International Journal of Epidemiology*, 46:103–105.

Felix Grezes, Jennifer Lynn Bartlett, Kelly Lockhart, Alberto Accomazzi, Ethan Seefried, Anjali Pandiri, 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*, pages 5–15, Mumbai, India and virtual. 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.

Kenneth Haug, Reza M. Salek, Pablo Conesa, Janna Hastings, et al. 2013. MetaboLights: an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research*, 41:D781–D786.

Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, et al. 2010. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45:703–714.

International Federation of Digital Seismograph Networks. 2019. FDSN StationXML schema, version 1.2. https://www.fdsn.org/xml/station/.

International Union of Crystallography. 2026. checkCIF and the validation reply form. https://checkcif.iucr.org/.

Bikash Joshi, Anthi Symeonidou, Syed Mazin Danish, and Floris Hermsen. 2023. An end-to-end pipeline for bibliography extraction from scientific articles. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 101–106, Bali, Indonesia. Association for Computational Linguistics.

Tobias Kind, Gert Wohlgemuth, Do Yup Lee, Yun Lu, Mine Palazoglu, Sevini Shahbaz, and Oliver Fiehn. 2009. FiehnLib: Mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Analytical Chemistry*, 81(24):10038–10048.

Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. 2019. Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data*, 6:203.

Ervin Kováts. 1958. Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. *Helvetica Chimica Acta*, 41:1915–1932.

Timothy Lebo, Satya Sahoo, and Deborah McGuinness. 2013. PROV-O: The PROV ontology. W3C Recommendation, 30 April 2013, https://www.w3.org/TR/prov-o/.

Barbara McGillivray, Kaveh Aryan, Viola Harperath, Marton Ribary, and Mandy Wigdorowitz. 2025. Finding the paper behind the data: Automatic identification of research articles related to data publications. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 34–43, Mumbai, India and virtual. Association for Computational Linguistics.

W. O. McReynolds. 1970. Characterization of some liquid phases. *Journal of Chromatographic Science*, 8:685–691.

William Meredith. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58:525–543.

Faith Mutinda, Kongmeng Liew, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2022. PICO corpus: A publicly available corpus to support automatic data extraction from biomedical literature. In *Proceedings of the First Workshop on Information Extraction from Scientific Publications*, pages 26–31, Online. Association for Computational Linguistics.

Van-Thuy Phi, Dinh-Truong Do, Hoang-An Trieu, and Yuji Matsumoto. 2025. A hybrid LLM and supervised model pipeline for polymer property extraction from tables in scientific literature. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 94–102, Mumbai, India and virtual. Association for Computational Linguistics.

L. Rohrschneider. 1966. Eine Methode zur Charakterisierung von gaschromatographischen Trennflüssigkeiten. *Journal of Chromatography A*, 22:6–22.

Fabio Rovai. 2026a. Comparability in the public olfactory record: volatile measurement, human perception, and the join between them. *Preprint*, arXiv:2609.29702.

Fabio Rovai. 2026b. Olfactory measurement ontology: a vocabulary, conformance suite and graded audit of the public olfactory record. Zenodo deposit, version 0.2.0, doi:10.5281/zenodo.23270951.

Timo Pierre Schrader, Matteo Finco, Stefan Grünewald, Felix Hildebrand, and Annemarie Friedrich. 2023. MuLMS: A multi-layer annotated text corpus for information extraction in the materials science domain. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 84–100, Bali, Indonesia. Association for Computational Linguistics.

Rachel A. Spicer, Reza Salek, and Christoph Steinbeck. 2017. Compliance with minimum information guidelines in public metabolomics repositories. *Scientific Data*, 4:170137.

Manish Sud, Eoin Fahy, Dawn Cotter, Kenan Azam, et al. 2016. Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research*, 44:D463–D470.

Lloyd W. Sumner, Alexander Amberg, Dave Barrett, Michael H. Beale, Richard Beger, Clare A. Daykin, et al. 2007. Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3:211–221.

Nicholas Susemiehl and Joseph Mazzarella. 2025. AI for data ingestion into IPAC archives. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 87–93, Mumbai, India and virtual. Association for Computational Linguistics.

Matthew C. Swain and Jacqueline M. Cole. 2016. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56:1894–1904.

Yuan-Sen Ting, Alberto Accomazzi, Tirthankar Ghosal, Tuan Dung Nguyen, Rui Pan, Zechang Sun, and Tijmen de Haan. 2025. AstroMLab 5: Structured summaries and concept extraction for 400,000 astrophysics papers. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 170–185, Mumbai, India and virtual. Association for Computational Linguistics.

Masaya Tsunokake and Shigeki Matsubara. 2022. Classification of URL citations in scholarly papers for promoting utilization of research artifacts. In *Proceedings of the First Workshop on Information Extraction from Scientific Publications*, pages 8–19, Online. Association for Computational Linguistics.

H. van den Dool and P. Dec. Kratz. 1963. A generalization of the retention index system including linear temperature programmed gas-liquid partition chromatography. *Journal of Chromatography A*, 11:463–471.

Alain C. Vaucher, Federico Zipoli, Joppe Geluykens, Vishnu H. Nair, Philippe Schwaller, and Teodoro Laino. 2020. Automated extraction of chemical synthesis actions from experimental procedures. *Nature Communications*, 11:3601.

Yongdong Wang and Bruce R. Kowalski. 1992. Calibration transfer and measurement stability of near-infrared spectrometers. *Applied Spectroscopy*, 46:764–771.

Yu Watanabe, Koichiro Ito, and Shigeki Matsubara. 2025. Metadata generation for research data from URL citation contexts in scholarly papers: Task definition and dataset construction. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 72–79, Mumbai, India and virtual. Association for Computational Linguistics.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

Edwin B. Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209–212.

World Meteorological Organization. 2019. WIGOS metadata representation, version 1.0. `https://schemas.wmo.int/wmdr/`.

World Wide Web Consortium. 2017. Shapes constraint language (SHACL). W3C Recommendation, `https://www.w3.org/TR/shacl/`.

Sean Yang, Chris Tensmeyer, and Curtis Wigington. 2022. TELIN: Table entity LINker for extracting leaderboards from machine learning publications. In *Proceedings of the First Workshop on Information Extraction from Scientific Publications*, pages 20–25, Online. Association for Computational Linguistics.

A Extractor pseudocode

The ordered token walk of Section 5, stated so that the extractor can be reimplemented from this paper alone. DECOY and RUNTIME are closed lexical sets. DECOY holds *inlet, injector, injection port, transfer line, interface, ion source, source temp, detector, quadrupole, manifold, septum, MS source, auxiliary, aux, trap temp, desorption, incubat(ion), agitat(ion), storage, freezer, water bath* and *oven of the mass*. RUNTIME holds *run time, analysis time, acquisition time, cycle time, chromatography time, total time, solvent delay, equilibrat(ion), post-run, run lasted* and *during*. The anchor that locates the programme matches *oven* (optionally followed by *temperature* and *program* or *programme*), *column temperature*, *temperature program(me)*, *chromatograph… was programmed* and *thermal program(me)*.

```
input : protocol text T
output: ordered programme P, or abstain

# 1. scan once, left to right
S <- []
for each match m in scan(T):
    if m is a temperature   : S += TEMP(m)
    if m is a rate per min  : S += RATE(m)
    if m is a duration      : S += DUR(m)

# 2. reject governed candidates
for each token t in S:
    if governed(t, DECOY, window)  : drop t
    if t is DUR and
       governed(t, RUNTIME, window): drop t

# 3. bind by position, not by pattern
P <- []
for each TEMP(v) in S, in stream order:
    step <- Step(target = v)
    r <- nearest preceding unbound RATE
    if r exists and no TEMP lies between:
        step.ramp <- r ; mark r bound
    d <- nearest following unbound DUR
    if d exists and no TEMP lies between:
        step.hold <- d ; mark d bound
    P += step

# 4. abstain rather than guess
if T names two distinct methods: abstain
if P is empty                  : abstain
return P
```

B Reference-series pseudocode

```
input : chromatography text T
output: normalised carbon series, or abstain
```

```
out <- []
for each match m of RANGE in T:        # C7-C40
    win <- T[m.start-120 : m.end+60]
   if no ANCHOR in win: continue    # alkane | FAME
                                # | retention index
                                        # | Kovats
    lo, hi <- carbon numbers of m
    if lo >= hi or hi > 60 or lo < 4: continue
    kind <- FAME if FAME in win else n-alkane
    out += series(kind, lo, hi, range, win)

for each match m of LISTED in T:     # C8, C10, C12
    win <- T[m.start-120 : m.end+60]
    if no ANCHOR in win: continue
    nums <- carbon numbers of m in 4..60
    if |nums| < 3: continue
    kind <- FAME if FAME in win else n-alkane
    out += series(kind, min, max, list, win)

return dedup(out) on (kind, lo, hi, form)
```

The sanity bounds are what make abstention the default. A carbon range that runs backwards, exceeds C60 or begins below C4 is not a reference series, and two carbon numbers that merely happen to sit near each other are rejected by the anchor requirement rather than reported with low confidence.

C Transfer pseudocode

The adaptation of Section 7 in full. Only the two marked changes differ from Appendix A: decoy rejection, the run-time exclusion, binding by position and abstention are reused unmodified.

```
input : F, value of a dedicated programme field
output: ordered programme P, or abstain

# changed 1: normalise degree-sign conventions
F <- replace(F, U+2103, "degC")
       # the single-character degree-Celsius
F <- replace(F, "A-circumflex degree", "degree")
       # mojibake: UTF-8 read as Latin-1
F <- replace(F, U+00BA, "degree")
       # masculine ordinal used as a degree sign

# changed 2: the field is the window
win <- F[:460]
       # was: locate with ANCHOR, then window

return walk(tokenise(win), win)   # unchanged
```

D Evidence for the five Workbench recoveries

The evidence window stored with each recovery in Table 6, verbatim and including its truncation at both ends. ST001369 and ST000980 carry the same protocol text.

ST001369 FAME C8–C30
ss range of m/z 50-600 at a rate of 2.7 scan/s. Several IS injections, a standard mix of fatty acid methyl esters (FAME C8-C30), extraction blanks and 3 QCs samples were injected at the

ST000496 n-alkane C8–C40
by a 70 kV electron impact (EI), and 20 scans/sec were recorded over the mass range 85-500. A standard alkane mixture (C8-C40) was injected at the beginning of the analysis for subseque

ST001134 n-alkane C8–C40
aTOF® software package (Leco Corp., St Joseph, MI, USA). Retention indices (RIs) were calculated relative to those of a C8-C40 alkane series. Deconvolution, feature extraction and peak i

ST000980 FAME C8–C30
ss range of m/z 50-600 at a rate of 2.7 scan/s. Several IS injections, a standard mix of fatty acid methyl esters (FAME C8-C30), extraction blanks and 3 QCs samples were injected at the

ST001432 n-alkane C7–C30
detect retention time shifts and facilitate Kovats retention index (RI) calculation, a standard alkane series mixture (C7-C30) was injected before every batch during sample analysis. Th

E How the verdicts were checked

Each of the 136 audited items, the 40- and 25-draw programme samples and every reference-series recovery in both repositories, was judged by two LLM raters, Claude Opus 5.5 and Claude Fable 5.1, each blind to every other verdict. They agree on 65 of 65 programme verdicts ($\kappa = 1.00$) and 68 of 71 series verdicts ($\kappa = 0.71$; Cohen, 1960). Two checks test the raters rather than trust them. A mechanical reconciliation with no model in it lists every temperature, rate and duration in the window the extractor read and compares them with the output; it agrees with both raters on 129 items, and each of the other 7 is settled by words of its protocol (Table 8). A planted-error set of 40 items, 24 of them corrupted in known ways (a dropped hold or step, swapped targets, a fabricated rate or hold, a shifted span, a flipped kind) and shown under opaque identifiers, measures each checker: Opus catches 24 of 24 and Fable 24, both flag all 6 planted fabrications with no false alarm, and the mechanical check catches 22, missing only dropped holds. Where the raters agree, their verdict stands, except the 2 items corrected after the author's check (Appendix F); the 3 items on which they disagree are settled by the quoted words of the protocol (Table 8). A development pass made with an earlier model (Claude Opus 5) agrees with the final verdicts on 65 of 65 programme items. The author then read 48 cards, each asking whether the output agrees with the text on one named point: every non-correct verdict, 16 items found correct

and 8 outputs we had deliberately corrupted. The reading agrees with the verdicts as they stood before it on 40 cards, 7 of them planted corruptions it caught. It also found an error both raters shared, a mixture the text gives as two ranges and the output gives as two series, and the 2 items of that form are counted wrong. Of the other differences, the reading and the quoted text disagree on 4, and 2 turn on whether a marker added at another concentration belongs to the series (Appendix F).

F Every verdict

The author's reading led to two corrections, both items that both raters had judged correct: MTBLS3627 (*"C8-C16: 1 mg/mL, C18-C24: 0.5 mg/mL"*); MTBLS13384 (*"(FAME, C8-C16, C18-C30) for internal retention marker index"*). Where the reading still differs from the final verdicts: MTBLS866 (arguable: *"and C32 (0.4 mg/ml final concentration)"*); MTBLS748 (the reading and the text disagree: *"A temperature ramp of 1 °C/min for 5 min"*); MTBLS12353 (the reading and the text disagree: *"ACQUITY UPLC BEH Amide"*); MTBLS687 (arguable: *"and C32 (0.4 mg/ml final concentration)"*); MTBLS8006 (a planted corruption the reading missed: *"maintained for 4 min"*); MTBLS1228 (the reading and the text disagree: *"standard was added to the solution as internal standard"*); MTBLS7299 (the reading and the text disagree: *"to 305 °C at a rate of 20 °C/min and finally held at 305 °C for 5 min"*).

Every reference-series item that is not correct, or that either rater did not judge correct: MTBLS171 (final wrong kind or span; O wrong kind or span, F wrong kind or span); MTBLS208 (final wrong kind or span; O wrong kind or span, F wrong kind or span); MTBLS687 (final wrong kind or span; O wrong kind or span, F correct); MTBLS748 (final wrong kind or span; O wrong kind or span, F wrong kind or span); MTBLS866 (final wrong kind or span; O wrong kind or span, F correct); MTBLS1228 (final not a series; O correct, F not a series); MTBLS3627 (final wrong kind or span; O correct, F correct); MTBLS9884 (final wrong kind or span; O wrong kind or span, F wrong kind or span); MTBLS13384 (final wrong kind or span; O correct, F correct). All other series items are correct under all three.

Item	Mechanical	Opus 5.5	Fable 5.1	Settled	Words of the protocol
MTBLS13035	hold omitted	correct	correct	correct	*"between 5-32 min"*
MTBLS2323	correct	hold omitted	hold omitted	hold omitted	*"at which it was initially held for 4 min"*
MTBLS2324	correct	hold omitted	hold omitted	hold omitted	*"at which it was initially held for 4 min"*
MTBLS1228	correct	correct	not a series	not a series	*"standard was added to the solution as internal standard"*
MTBLS13081	wrong span	correct	correct	correct	*"calculated from an in-house C8–C40 alkane mix"*
MTBLS687	correct	wrong span	correct	wrong span	*"and C32 (0.4 mg/ml"*
MTBLS866	correct	wrong span	correct	wrong span	*"and C32 (0.4 mg/ml"*

Table 8: The items on which the mechanical check and the two raters are not unanimous, each settled by the quoted words of its protocol.

Set	Item	Final	O	F	D	Set	Item	Final	O	F	D
A	MTBLS1778	C	C	C	C	A	MTBLS2021	C	C	C	C
A	MTBLS849	H	H	H	H	A	MTBLS7299	C	C	C	C
A	MTBLS1206	C	C	C	C	A	MTBLS1610	H	H	H	H
A	MTBLS1980	C	C	C	C	A	MTBLS8105	C	C	C	C
A	MTBLS14861	C	C	C	C	A	MTBLS2323	H	H	H	H
A	MTBLS848	H	H	H	H	A	MTBLS10841	C	C	C	C
A	MTBLS2324	H	H	H	H	A	MTBLS532	C	C	C	C
A	MTBLS4894	C	C	C	C	B	ST002871	C	C	C	C
A	MTBLS779	C	C	C	C	B	ST003544	C	C	C	C
A	MTBLS13	H	H	H	H	B	ST003415	C	C	C	C
A	MTBLS12365	C	C	C	C	B	ST002540	C	C	C	C
A	MTBLS15	H	H	H	H	B	ST004201	C	C	C	C
A	MTBLS7006	C	C	C	C	B	ST003863	H	H	H	H
A	MTBLS780	C	C	C	C	B	ST001310	C	C	C	C
A	MTBLS13991	C	C	C	C	B	ST000586	C	C	C	C
A	MTBLS8958	C	C	C	C	B	ST003561	C	C	C	C
A	MTBLS95	C	C	C	C	B	ST004845	C	C	C	C
A	MTBLS4332	H	H	H	H	B	ST004136	C	C	C	C
A	MTBLS12437	C	C	C	C	B	ST004739	C	C	C	C
A	MTBLS13035	C	C	C	C	B	ST004845#2	C	C	C	C
A	MTBLS226	C	C	C	C	B	ST000092	C	C	C	C
A	MTBLS96	C	C	C	C	B	ST004089	C	C	C	C
A	MTBLS1579	H	H	H	H	B	ST004376	H	H	H	H
A	MTBLS808	C	C	C	C	B	ST000054	C	C	C	C
A	MTBLS14	H	H	H	H	B	ST004599	C	C	C	C
A	MTBLS8056	H	H	H	H	B	ST004056	C	C	C	C
A	MTBLS14913	C	C	C	C	B	ST004738	C	C	C	C
A	MTBLS1737	C	C	C	C	B	ST003998	C	C	C	C
A	MTBLS5	H	H	H	H	B	ST004643	C	C	C	C
A	MTBLS748	S	S	S	S	B	ST002993	H	H	H	H
A	MTBLS8006	C	C	C	C	B	ST003544#2	C	C	C	C
A	MTBLS12291	C	C	C	C	B	ST004123	C	C	C	C
A	MTBLS12353	S	S	S	S						

Table 9: Every programme verdict. Set A is the 40-draw MetaboLights audit and set B the 25-draw Workbench transfer audit. Final is the reported verdict; the raters agree on every programme item, so none needed settling. O and F are the blind passes by `claude-opus-5-5` and `claude-fable-5-1`, and D is the development pass. C is correct, H hold omitted, S structural error.

Modality-Specific and Relation-Aware Modeling for Scientific Figure Linking

Bodhisatta Maiti
Independent Researcher
Durham, North Carolina, USA

Debshree Chowdhury
Independent Researcher
Durham, North Carolina, USA

Abstract

We present a modality-specific system for the AstroCLIMB shared task, which predicts relations among astronomical figures and captions. Caption–caption, image–image, and image–caption pairs use different combinations of frozen pretrained representations and supervised classifiers. Image–caption prediction first applies a Qwen3-VL 4B/8B exact-correspondence gate before classifying the remaining paper relations. Subsequent improvements add astronomy-domain sentence embeddings to the caption branch and a learned visual-projection specialist to the image branch. Across eight Kaggle submissions, macro-F1 increases from 0.3641 to 0.6096. Adding astronomy-specific caption features improves five-fold pooled out-of-fold (OOF) macro-F1 from 0.6071 to 0.6388; the image specialist offers a smaller incremental test-set gain. The direct citation relation, `related_papers`, remains the most difficult. Results support selecting representations and decision mechanisms according to modality, while highlighting the limitations of recovering citation links from isolated content.

1 Introduction

Figures are central to scientific communication. They condense experimental results, comparisons, trends, and domain-specific evidence into a form that is often easier for a reader to interpret than the surrounding prose. At the same time, much of this information remains difficult to exploit computationally in scholarly search and literature analysis. The AstroCLIMB shared task addresses this gap through an unusual multimodal formulation: given a pair of scientific objects—two figures, two captions, or a figure and a caption—a system must determine whether they correspond to the same figure, originate from the same paper, come from citation-linked papers, or are unrelated (Grezes et al., 2026). The task therefore asks models to recover part of the structure of the scholarly citation graph using only the content of scientific figures and captions.

Motivated by these differences, we use separate modeling strategies for each modality and relation type rather than a single end-to-end classifier. For caption-caption pairs, we combine scientific text representations with vision-language embeddings and lightweight lexical features. For image-image pairs, we use a higher-capacity, higher-resolution SigLIP2 visual encoder, motivated by the improved semantic and visual representation capabilities of the SigLIP2 family (Tschannen et al., 2025). Image-caption pairs are handled differently. We first use Qwen3-VL (Bai et al., 2025) as a dedicated exact-correspondence detector for the `same_figure` relation. Pairs not identified as exact matches are then passed to a separate classifier for the remaining scholarly relations. In the final system, the exact-match decision combines scores from Qwen3-VL 4B and 8B models, while the residual classifier combines complementary scientific-text and multimodal signals. We deliberately avoid DOI, citation-graph, or other source-metadata lookup at inference time so that predictions are based on the figure-caption content provided by the benchmark.

This specialization substantially outperformed a uniform representation. Our initial frozen SigLIP2 baseline obtained 0.3641 Kaggle macro-F1; adding scientific-text features increased this to 0.4472, and introducing the dedicated image–caption correspondence gate raised it to 0.5660. Stronger visual representations and score-level combinations produced a final score of 0.6096 after astronomy-specific caption features and a complementary visual specialist. Across experiments, `related_papers` remained the most difficult relation, highlighting the gap between semantic relatedness and direct citation linkage.

Our work makes three main contributions. First, we develop a modality-specialized architecture

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 262–271
November 10, 2026 ©2026 Association for Computational Linguistics

for AstroCLIMB that assigns different representations and decision mechanisms to caption-caption, image-image, and image-caption pairs instead of forcing all three modalities through a shared classifier. Second, we show that vision-language models are particularly effective when their role is narrowed to exact figure-caption correspondence; direct use of the same models for broader paper-provenance classification is considerably less reliable. Third, through controlled ablations, cross-fitted predictions, and a separately evaluated Qwen gate, we identify direct citation prediction as the main remaining bottleneck and show that stronger generic visual representations improve same-paper and unrelated-paper discrimination much more consistently than citation-linked-paper recognition. Together, these results support a view of scientific multimodal relation prediction as a collection of distinct inference problems rather than a single similarity task.

2 Related Work

Scientific-document representations have used citation relationships during pretraining. SPECTER learns document embeddings from citation-informed supervision (Cohan et al., 2020); SciRepEval studies scientific-text representations across multiple tasks and formats (Singh et al., 2023). In contrast, our inference setting supplies only a figure or caption pair, not full papers or their citation neighborhoods. We apply pretrained representations directly to captions, without access to the underlying citation graph.

For visual and cross-modal evidence, SigLIP2 provides pretrained image–text embeddings (Tschannen et al., 2025), and Qwen3-VL supports multimodal question answering (Bai et al., 2025). We use these models for distinct purposes: SigLIP2 supplies pairwise features, while Qwen3-VL provides continuous evidence of exact figure–caption correspondence. Our experiments also compare an astronomy-specialized sentence encoder (NeuML, 2026) with more general scientific-text embeddings. The AstroCLIMB benchmark defines provenance by paper identity or direct citation (Grezes et al., 2026); such relationships cannot in general be identified from visual similarity alone.

3 Task and Dataset

3.1 Task Definition

AstroCLIMB is a multiclass, single-label classification task designed to recover relationships between scientific objects drawn from astronomy papers (Grezes et al., 2026). Each example contains a pair of objects, where an object is either a scientific figure or its textual caption. This results in three input configurations: caption–caption (CC), image–image (II), and image–caption (IC).

Each pair is assigned exactly one of four labels:

- `same_figure`: the image and caption correspond to the same scientific figure. This relation is defined only for IC pairs.
- `same_paper`: the two objects originate from the same source paper.
- `related_papers`: the objects originate from different papers for which one paper directly cites the other.
- `unrelated_papers`: none of the preceding relationships holds.

The relations are symmetric with respect to the order of the two input objects (Grezes et al., 2026). The distinction between `related_papers` and the other classes is particularly important for our study. Unlike `same_figure` or `same_paper`, a direct citation edge is not necessarily recoverable from semantic similarity alone: two papers may investigate the same object or use similar methods without directly citing one another.

3.2 Released Data

The Kaggle release[1] contains 10,000 labeled training pairs and 10,000 test pairs (Grezes et al., 2026). Captions are provided as English text, whereas figures are represented as encoded image strings that can be converted back to image objects. We use the Kaggle pair dataset as the sole source of task-specific supervision.

Inspection of the released training split shows a deliberately structured distribution across modalities. The CC and II subsets each contain 3,000 examples, with 1,000 examples from each of `same_paper`, `related_papers`, and `unrelated_papers`. The IC subset contains 4,000

[1] https://www.kaggle.com/competitions/astroclimb

Pair	Same fig.	Same paper	Related	Unrel.
CC	0	1000	1000	1000
II	0	1000	1000	1000
IC	1000	1000	1000	1000
Total	1000	3000	3000	3000

Table 1: Class distribution in the 10,000-pair AstroCLIMB training split. CC, II, and IC denote caption–caption, image–image, and image–caption pairs, respectively.

examples, with 1,000 examples from each of the four classes. The test split follows the same modality counts—3,000 CC, 3,000 II, and 4,000 IC pairs—although its class labels are hidden.

The broader AstroCLIMB resource[2] also provides figure-caption records and bibliographic metadata from the underlying astronomy corpus (Grezes et al., 2026). We intentionally do not use source DOIs, citation adjacency information, paper titles, author metadata, or other corpus-level identifiers as task supervision or as lookup information at inference time. In particular, we do not recover test-pair relationships by matching test objects back to the underlying citation graph. This restriction keeps the system focused on the information available in the benchmark pair itself and avoids introducing paper-level graph information that could trivially reveal some target relations.

3.3 Evaluation Protocol

The official metric is macro-F1 over the four target classes (Grezes et al., 2026). We use stratified five-fold out-of-fold (OOF) evaluation for lightweight classifiers and separate development and validation subsets for the IC exact-correspondence component. Submitted labels are determined from model scores without imposing fixed class quotas. CV splits are stratified at the *pair* level, not guaranteed to be disjoint at figure, caption, or source-paper level. Thus OOF estimates support within-benchmark ablation comparisons but do not establish unseen-paper generalization. We report pooled OOF macro-F1 separately from the mean and sample standard deviation of the five fold scores; hidden-test Kaggle values are reported separately.

4 System Architecture

A single classifier is unlikely to receive equally useful evidence from all three AstroCLIMB input configurations. A caption–caption pair is primarily a scientific-language problem, an image–image pair depends on visual and stylistic evidence, and an image–caption pair additionally permits direct cross-modal correspondence. We therefore use a modality-specialized architecture in which CC, II, and IC pairs follow separate prediction paths. All pretrained models are kept frozen; only lightweight downstream classifiers and decision rules are fit using the AstroCLIMB training pairs.

4.1 Pair Representation

For two normalized embeddings $\mathbf{e}_1$ and $\mathbf{e}_2$ from a shared representation space, we construct a symmetric pair representation

$$\phi(\mathbf{e}_1, \mathbf{e}_2) = [|\mathbf{e}_1 - \mathbf{e}_2|; \mathbf{e}_1 \odot \mathbf{e}_2; \cos(\mathbf{e}_1, \mathbf{e}_2)], \quad (1)$$

where $\odot$ denotes element-wise multiplication. The absolute difference and element-wise product preserve dimension-wise agreement between the two objects, while cosine similarity provides a compact global similarity signal. Because the benchmark relations are symmetric, the representation does not depend on the order of the two inputs.

The resulting features are classified using XGBoost (Chen and Guestrin, 2016). In the CC configurations we evaluated, the XGBoost head improved over the simpler classifier when combined pretrained representations were available, while remaining inexpensive to train.

4.2 Caption–Caption Specialist

For CC pairs, both inputs contain scientific language, so we combine a general vision-language text representation with a scientific-document representation. Each caption is first embedded using the text encoder of SigLIP2 (Tschannen et al., 2025). In parallel, we encode captions using SPECTER2, the scientific-document representation family introduced with the SciRepEval benchmark (Singh et al., 2023).

We construct separate pair features using Eq. 1 for the SigLIP2 and SPECTER2 representations and concatenate them with lightweight lexical overlap features. The latter capture surface-level evidence such as shared tokens and relative caption length. The combined representation is passed to a three-class XGBoost classifier that predicts `same_paper`, `related_papers`, or `unrelated_papers`. The `same_figure` relation is not defined for CC pairs.

[2] https://huggingface.co/datasets/adsabs/AstroCLIMB

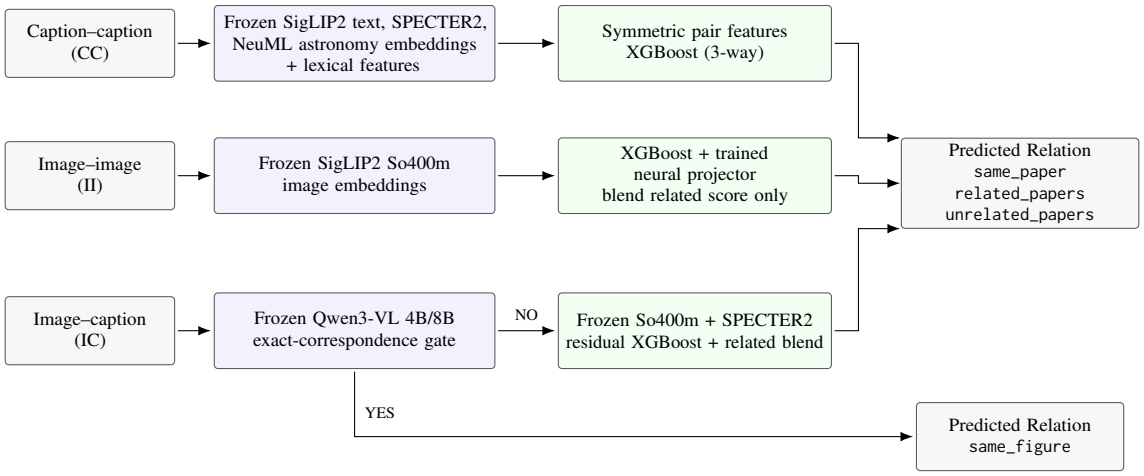


Figure 1: Final V8 modality-specific system. Pretrained encoders are frozen. The CC and II paper-relation classifiers and the II projection specialist learn from Kaggle training pairs. The IC gate routes exact matches directly to `same_figure`; remaining IC pairs use the unchanged V6 residual classifier to predict `same_paper`, `related_papers`, or `unrelated_papers`.

In V7, we additionally encode each caption with the frozen astronomy-specialized `NeuML/astrobert-small-embeddings` sentence model (NeuML, 2026). Its native sentence representation supplies a third symmetric pair-feature block alongside SigLIP2 and SPECTER2, while the original lexical features are retained. We compare replacing and augmenting SPECTER2; adding the astronomy-specific block performs best on OOF evaluation.

This branch is intended to separate topical similarity from stronger paper-level correspondence. In particular, the scientific-text representation provides information that is not well captured by the shorter text context of the vision-language encoder alone.

4.3 Image–Image Specialist

For II pairs, we use the image encoder from SigLIP2 So400m with 384×384 input resolution (Tschannen et al., 2025). Each image is encoded independently, and the two visual embeddings are converted into the symmetric representation in Eq. 1. An XGBoost classifier then predicts one of `same_paper`, `related_papers`, or `unrelated_papers`.

The evaluated So400m configuration uses both a larger encoder and higher image resolution; our experiments do not isolate the contribution of either change. Scientific figures contain small labels, panel structure, plotting conventions, and other visual cues that may be relevant to provenance. We use the So400m representation directly rather than concatenating it with the smaller SigLIP2 representation; our ablations show that naive feature-level fusion does not improve II performance.

For V8, we add a neural specialist on cached, frozen So400m embeddings. A shared two-layer projection maps each image to a normalized task-adapted space; the symmetric representation in Eq. 1 feeds a three-class multilayer perceptron. We average five trained models' scores at test time and blend only the `related_papers` score with the five-fold XGBoost ensemble (neural weight 0.90); the other two class scores remain those of XGBoost. The weight was chosen by searching OOF predictions, so the selected OOF blend score is not an independent estimate of generalization.

4.4 Image–Caption Routing

IC pairs require a different treatment because they are the only modality for which `same_figure` is possible. A flat four-class classifier forces one model to solve two qualitatively different problems: exact visual-text correspondence and paper-level provenance. We instead separate these decisions.

Exact correspondence gate. We first determine whether the caption describes the exact input figure. Qwen3-VL (Bai et al., 2025) is prompted with the figure and caption and asked a binary question: whether the caption corresponds to that exact figure. Rather than relying on a generated class label, we run deterministic generation (`max_new_tokens=1`, no sampling) immediately after the assistant-generation prefix. Both tokenizers encode uppercase `YES` and `NO` as single tokens

(IDs 14004 and 8996). From their first-step logits, ℓ_Y and ℓ_N, we compute the two-token normalized score $p = \exp(\ell_Y)/(\exp(\ell_Y) + \exp(\ell_N))$. This is a decision score, not a separately calibrated probability.

Our final system combines scores from the 4B and 8B Qwen3-VL variants. Let p_4 and p_8 denote their respective two-token normalized scores. Before the logit transformation, scores are clipped to $[10^{-8}, 1 - 10^{-8}]$ to avoid infinities. We combine them in log-odds space,

$$s_{\mathrm{IC}} = \sigma\left((1-\alpha)\mathrm{logit}(p_4) + \alpha\mathrm{logit}(p_8)\right), \quad (2)$$

where σ is the logistic function and α is selected on development data. If s_{IC} meets or exceeds a tuned threshold, the pair is assigned `same_figure`. Using the continuous model scores proved substantially more reliable than treating the generated binary answer as the final decision.

Residual paper-relation classifier. Pairs that do not pass the exact-correspondence gate are classified into `same_paper`, `related_papers`, or `unrelated_papers`. The image and caption are embedded in the shared SigLIP2 So400m space, and their cross-modal pair representation is constructed using Eq. 1. We additionally append the SPECTER2 representation of the caption, providing a longer-context scientific-language signal alongside the visual-text similarity features. An XGBoost classifier produces the three paper-relation scores.

We retain a small amount of complementary information from the earlier, lower-capacity IC classifier for the `related_papers` score. This probability-level combination is performed only after classification rather than by concatenating the underlying embedding spaces. The resulting three-way scores are used only when the Qwen correspondence gate rejects `same_figure`.

5 Experimental Setup

5.1 Pretrained Encoders

We use pretrained encoders as frozen feature extractors throughout the system. The initial SigLIP2 representation is obtained from `google/siglip2-base-patch16-224`, which produces 768-dimensional embeddings from 224×224 inputs. For the stronger visual and cross-modal branches, we use `google/siglip2-so400m-patch14-384`, with 1152-dimensional embeddings and 384×384 image inputs (Tschannen et al., 2025).

Scientific-text representations include `allenai/specter2_base` and, from V7 onward, `NeuML/astrobert-small-embeddings`. We use the base SPECTER2 encoder directly, without task-specific adapters, and represent each caption using the hidden state of the first token. SPECTER2 was introduced as part of the SciRepEval work on transferable scientific-document representations (Singh et al., 2023).

For IC exact-correspondence detection, we use `Qwen/Qwen3-VL-4B-Instruct` and `Qwen/Qwen3-VL-8B-Instruct` (Bai et al., 2025). The 8B model is loaded with 4-bit NF4 quantization and bfloat16 computation to fit within a single-GPU Colab environment. These models are used only for inference; no Qwen3-VL parameters are fine-tuned on AstroCLIMB.

All pretrained encoders remain frozen and embeddings are cached; the V8 projection and downstream classifiers are trained on Kaggle pairs in a single-GPU Google Colab environment.

5.2 Supervised Classifiers

The CC and residual IC branches use XGBoost classifiers; the II branch combines XGBoost and a neural specialist (Chen and Guestrin, 2016). Training examples are represented using the symmetric pair features described in Section 4, together with the additional modality-specific representations used by each branch.

For CC and II, the classifier predicts the three paper-level relations: `same_paper`, `related_papers`, and `unrelated_papers`. For IC Stage 2, training is restricted to the 3,000 non-`same_figure` IC examples, since the first-stage VLM gate is responsible for detecting exact figure–caption correspondence.

We use five-fold stratified cross-validation to generate out-of-fold predictions for the supervised classifiers. These predictions are used for model comparison and for constructing the probability-level combination in the IC residual classifier. The II related-score blend weight was also selected from OOF predictions; it is not included as an independently cross-validated result. Test predictions are generated by the corresponding trained branch and are not constrained to match the class proportions observed in the training set.

5.3 Image–Caption Development Protocol

The IC branch requires additional calibration because the output of the exact-correspondence detector is a continuous score rather than a direct four-class prediction. We therefore separate IC development from the out-of-fold evaluation used for the lightweight classifiers.

We construct a balanced 200-example IC development set containing 50 examples from each of the four target classes. This subset is used for selecting the exact-match threshold and, for the V6 and later systems, the relative contribution of the 4B and 8B Qwen3-VL models. We additionally maintain a separate balanced set of 400 IC examples, with 100 examples per class, for evaluating the exact-correspondence gate. This subset is not used when selecting the Qwen3-VL ensemble weight or correspondence threshold. For the hybrid evaluations on these rows, non-`same_figure` Stage-2 outputs are retrieved by row ID from the five-fold OOF predictions: the fitted classifier producing each such score excluded that row. Nevertheless, the residual-score blend coefficient was selected from pooled OOF predictions that included these rows; therefore, the hybrid results are cross-fitted diagnostics, not a fully independent validation of all model-selection choices.

For a figure–caption pair, each Qwen3-VL model receives the figure together with the caption and is asked whether the caption describes that exact figure. We obtain the two-token normalized correspondence decision score from the next-token `YES/NO` logits as defined in Section 4. For the final ensemble, the two model scores are combined in log-odds space as defined in Eq. 2. Development-set search selected an 8B ensemble weight of

$$\alpha = 0.65 \tag{3}$$

and a correspondence threshold of

$$\tau = 0.01778. \tag{4}$$

Pairs with $s_{\mathrm{IC}} \geq \tau$ are labeled `same_figure`; the remaining pairs are passed to the three-way IC Stage-2 classifier.

This protocol is important because several apparently stronger configurations on the 200-example development subset did not transfer to the held-out 400-example set. We therefore use the latter only as a validation check and do not retune the final system on it.

Version	Main change	Macro-F1
V1	Frozen SigLIP2 baseline	0.3641
V2	Scientific-text features	0.4472
V3	Qwen3-VL IC routing	0.5660
V4	So400m visual features	0.5843
V5	IC blend and gate calibration	0.5919
V6	Qwen 4B+8B ensemble	0.5982
V7	Astronomy-specific CC features	0.6081
V8	II neural specialist + V7 CC	**0.6096**

Table 2: Kaggle hidden-test leaderboard progression. Scores are rounded to four decimal places from submission records; the timing-only sample submission is omitted.

5.4 Residual IC Score Combination

The final IC Stage-2 classifier primarily uses the SigLIP2 So400m representation together with the SPECTER2 caption representation. An earlier IC classifier based on the lower-capacity representation retained complementary signal for `related_papers`. We therefore combine only the `related_papers` probability from the two classifiers:

$$p_{\mathrm{rel}} = \lambda p_{\mathrm{rel}}^{\mathrm{old}} + (1 - \lambda) p_{\mathrm{rel}}^{\mathrm{new}}, \tag{5}$$

with

$$\lambda = 0.35. \tag{6}$$

The `same_paper` and `unrelated_papers` probabilities are taken directly from the stronger So400m-based classifier. The final Stage-2 label is the class with the highest resulting score.

6 Results and Ablations

We report two types of results. Kaggle leaderboard scores measure end-to-end performance on the unlabeled test pairs, whereas local results are obtained either through five-fold out-of-fold (OOF) predictions or through the IC validation subsets described in Section 5. Since these evaluation sets differ, their absolute scores should not be compared directly.

6.1 End-to-End System Progression

Table 2 reports the eight completed Kaggle submissions (excluding a timing-only sample submission). The largest early gain comes from introducing the exact figure–caption gate (V3). V7 adds astronomy-specific sentence features only to CC, raising the hidden-test score from 0.5982 to 0.6081. V8 adds an II neural relatedness specialist while keeping the IC branch unchanged; its further gain is small (0.6096). These leaderboard differences are not paired statistical tests, since test labels are not released.

Branch	Configuration	$\mu \pm \sigma$	OOF
CC	SigLIP2 + linear	$.4059 \pm .0127$	.4058
CC	+ SPECTER2 + lexical	$.5966 \pm .0164$	.5970
CC	+ XGBoost (V6)	$.6070 \pm .0105$	.6071
CC	+ NeuML add (V7)	$\mathbf{.6386 \pm .0079}$	**.6388**
II	SigLIP2 base	$.4377 \pm .0141$	.4377
II	So400m XGBoost	$.4992 \pm .0172$	.4994
II	Base + So400m	$.4913 \pm .0179$	.4914
II	Neural projection alone	$.4887 \pm .0180$	.4917
IC	Flat classifier	$.3527 \pm .0101$	.3526
IC	Residual Stage 2 (V6)	$.4264 \pm .0172$	.4266

Table 3: Local five-fold stratified CV ablations. Mean $\pm$ sample SD uses fold macro-F1 values, while OOF reports pooled predictions. II's weight-selected V8 blend is excluded from this independent-fold table.

From V1 to V8, the absolute macro-F1 gain is 0.2455 (about 67.4% relative to V1). The central architectural change was the early separation of exact correspondence from provenance; later revisions add targeted representation and score-level improvements.

6.2 Modality-Specific Ablations

Table 3 reports pooled five-fold OOF macro-F1 and the mean $\pm$ sample SD of individual fold macro-F1. These are component-level experiments, not estimates of each complete V1–V8 pipeline. In particular, the IC hybrid experiments below use a separate 400-example Qwen evaluation with cross-fitted Stage-2 predictions, and appear in Table 5.

The clearest Phase-2 component gain is in CC: supplementing SigLIP2, SPECTER2, and lexical features with NeuML embeddings raises pooled OOF macro-F1 from 0.6071 to 0.6388. The CC improvement occurs on each of the five folds, although choosing the best observed configuration among multiple OOF ablations can introduce model-selection optimism. Alternative replacements and an `adsabs/astroBERT` variant (Grezes et al., 2021) did not exceed the NeuML-add configuration (Appendix A).

The original II So400m encoder raises OOF macro-F1 over the base image model, but naive feature concatenation is worse. The standalone neural projection reaches 0.4917, below the XGBoost 0.4994 baseline, yet improves the `related_papers` class (Table 6). The *OOF-selected* related-score-only blend reaches 0.5254, with weight chosen using those same predictions. It is reported as model-selection evidence only; the final Kaggle submission gains 0.0015 beyond V7.

Gate	Precision	Recall	F1
Qwen3-VL 4B	0.8824	0.9000	0.8911
Qwen3-VL 8B	0.8246	0.9400	0.8785
4B + 8B	**0.9029**	**0.9300**	**0.9163**

Table 4: Exact `same_figure` detection on the held-out 400-example IC validation set. The 4B row uses the calibrated gate employed before the final ensemble.

6.3 Exact Figure–Caption Matching

The behavior of Qwen3-VL further supports the decision to use the VLM as a specialist rather than as a general relation classifier. In an early balanced 60-example experiment, direct four-way Qwen3-VL classification obtained only 0.3416 macro-F1. However, the `same_figure` class alone achieved an F1 of 0.8750. We therefore reformulated the VLM component as binary exact-correspondence detection.

On the 200-example development set, Qwen3-VL-4B achieved an ROC-AUC of 0.9885 and a PR-AUC of 0.9643 for this binary task. The 8B model achieved still higher development values of 0.9943 ROC-AUC and 0.9863 PR-AUC. The larger model did not, however, translate into better held-out F1 when used alone.

As shown in Table 4, Qwen3-VL-8B alone favors recall at the cost of substantially lower precision. Combining the 4B and 8B scores produces a better balance and raises held-out exact-match F1 to 0.9163. This ensemble becomes the correspondence gate from V6 onward. The gate comparison is evaluated separately from the cross-fitted hybrid diagnostics in Table 5. A class-stratified bootstrap of the ensemble's retained 400-example confusion counts (93 true positives, 7 false negatives, 10 false positives) gives an approximate 95% interval of 0.877–0.952 for its F1. This interval characterizes the ensemble alone, not the paired improvement over 4B: aligned per-example predictions for the compared systems were not retained.

Importantly, the same models were much less reliable when asked to infer paper provenance directly. For example, an image–image same-paper detector based on Qwen3-VL achieved only 0.7494 ROC-AUC, and a small few-shot three-way II experiment obtained 0.2913 macro-F1. These results suggest that strong visual-language correspondence does not automatically translate into reliable citation or provenance reasoning.

IC hybrid variant	Diagnostic macro-F1
Qwen gate + old Stage 2	0.4863
Qwen gate + So400m Stage 2	0.5414
+ Stage-2 blend / calibration	0.5491

Table 5: IC hybrid diagnostics on 400 examples using the earlier Qwen3-VL 4B correspondence gate. Gate scores were evaluated on a separate validation set, while residual predictions for non-`same_figure` examples were generated out of fold. Because the related-score blending weight was selected using pooled OOF predictions that included these examples, the reported scores are not fully independent held-out estimates.

Configuration	Same	Rel.	Unrel.	F1
CC V6	.6813	.4498	.6903	.6071
CC V7 NeuML	.6803	.4936	.7424	.6388
II So400m	.5753	.3454	.5774	.4994
II neural alone	.5265	.4405	.5079	.4917
II blend*	.5667	.4378	.5717	.5254
IC residual Stage 2	.4540	.3425	.4833	.4266

Table 6: Per-class OOF F1. Labels abbreviate `same_paper`, `related_papers`, and `unrelated_papers`. *OOF-selected score-blend weight: descriptive model-selection result, not independent validation.

6.4 Relation-Level Behavior

Table 6 exposes which provenance relations change under V7/V8 ablations. The NeuML CC features particularly improve `related_papers` and `unrelated_papers`; the II projection raises related-paper F1 but reduces the other classes. The tuned II blend partly restores that balance. All entries are three-class OOF diagnostics, *not* end-to-end IC four-class results.

In the So400m II out-of-fold results, 415 of 1,000 true `related_papers` pairs were classified as `unrelated_papers` and 275 as `same_paper` (Appendix Table 8). The residual IC Stage-2 classifier exhibited a similar pattern, with 360 and 317 examples, respectively (Appendix Table 10). These component-level counts reveal confusion with both neighboring relations but do not establish the causes of individual errors.

6.5 Negative Ablations

Several alternatives were tested and discarded because their gains did not generalize or because they degraded local validation. Naively concatenating the original and So400m II features reduced macro-F1 from 0.4994 to 0.4914. Few-shot Qwen3-VL classification of II relations collapsed largely toward `unrelated_papers`, obtaining 0.2913 macro-F1. We also tested class-bias calibration for the residual IC classifier. Although the bias parameters were selected on separate OOF examples, held-out Stage-2 macro-F1 decreased from 0.4544 to 0.4500. We therefore retained the uncalibrated decision rule.

In Phase 2, substituting NeuML in the residual IC classifier also reduced pooled OOF macro-F1 from 0.4266 to 0.4226 (replacement) and 0.4131 (addition). Single-fold LoRA and neural-adapter pilots were not promoted to the final system and are not presented as five-fold results.

These negative results reinforce two design choices in the final system: representation improvements are most useful when matched to the appropriate modality, and the VLM is most effective when assigned the narrowly defined exact-correspondence decision rather than the full citation-relation problem.

7 Discussion

The main remaining difficulty is `related_papers`. Unlike exact correspondence or same-paper prediction, a direct citation edge is only indirectly observable from an isolated figure or caption. Two papers may contain highly similar scientific content without citing one another, while a citation may concern a method, catalog, instrument, or background result not visible in the sampled figure. This helps explain why `related_papers` remains the weakest paper-level class across CC, II, and IC.

The Qwen3-VL experiments further show that model capability depends strongly on how the model is used. Exact figure–caption correspondence provides concrete cross-modal evidence such as axes, annotations, plotted quantities, object names, and panel structure. Qwen3-VL performs well on this narrow decision but is considerably less reliable for broader provenance prediction. Similarly, the 8B model achieves stronger development AUC but lower standalone validation F1 than the 4B model; combining their partially complementary scores gives the best correspondence gate.

Our ablations also show that additional capacity or features do not automatically improve performance. Combining the base and So400m II representations reduced OOF macro-F1, and aggressive post-hoc calibration did not improve the residual IC classifier. The strongest gains instead came from matching representations and decision mechanisms to the evidence available in each modality.

Finally, our system intentionally excludes

paper identifiers, citation graphs, and bibliographic metadata at inference time. This makes `related_papers` inherently difficult because the target itself is graph-based. Paper-level text or citation context could improve future systems, but would constitute a richer inference setting than the content-only formulation studied here.

8 Conclusion

We developed a modality- and relation-aware AstroCLIMB system that distinguishes exact figure–caption correspondence from paper-level provenance. The dedicated Qwen3-VL IC gate produced the largest early improvement; astronomy-specific CC embeddings subsequently improved OOF performance, and a learned II relatedness specialist contributed a smaller test-set gain. The final V8 Kaggle macro-F1 is 0.6096, compared with 0.3641 initially. Across branches, `related_papers` remains difficult because citation linkage is not equivalent to visible semantic similarity. The results favor specialized, content-based components, while the pair-level validation protocol and score-selection procedures limit generalization claims.

Limitations

Pair-level cross-validation does not guarantee disjoint training and validation sets at the image, caption, or source-paper level. We did not quantify such overlap; therefore, the local scores do not establish generalization to previously unseen papers. The citation-link relation may not be recoverable from isolated figure or caption content. Selection among multiple OOF CC variants and tuning the II blend on pooled OOF predictions may overstate prospective gains. The IC hybrid table uses cross-fitted Stage-2 predictions, but its residual blend weight was selected on OOF results including the 400 diagnostic examples, so it is not fully independent of model selection. Repeated Kaggle submissions provided performance feedback and the final score should be interpreted within that shared-task setting; the small V7–V8 gain is not a significance test. The retained gate-summary counts do not support a paired confidence interval for 4B versus 8B+4B.

Acknowledgments

During the preparation of this manuscript, the authors used ChatGPT to assist with language editing, including grammar correction, spelling checks, restructuring of sentences, and rephrasing to improve clarity, flow, and readability. The authors reviewed and edited all text after using the tool and take full responsibility for the content of the publication.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

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.

Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J. Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S. Grant, Donna M. Thompson, Roman Chyla, Stephen McDonald, Timothy W. Hostetler, Matthew R. Templeton, Kelly E. Lockhart, Nemanja Martinovic, Shinyi Chen, Chris Tanner, and Pavlos Protopapas. 2021. Building astrobert, a language model for astronomy and astrophysics. *arXiv preprint arXiv:2112.00590*.

Felix Grezes, Kelly Lockhart, Alberto Accomazzi, Atilla Alkan, Ethan Seefried, and Tirthankar Ghosal. 2026. 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.

NeuML. 2026. AstroBERT small embeddings. Hugging Face model card. Model identifier: NeuML/astrobert-small-embeddings.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566, Singapore. Association for Computational Linguistics.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin,

Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

A Additional Ablations and Diagnostics

CC variant	$\mu \pm \sigma$	OOF
NeuML replace	.6327 ± .0094	.6328
NeuML added (V7)	.6386 ± .0079	.6388
AstroBERT replace	.5974 ± .0188	.5977
AstroBERT added	.6313 ± .0115	.6317

Table 7: Astronomy-specific CC representation ablations on identical stratified folds.

True class	Same	Rel.	Unrel.
Same paper	550	234	216
Related papers	275	310	415
Unrelated papers	87	251	662

Table 8: OOF confusion matrix for the frozen So400m II XGBoost model.

IC Stage 2 variant	Same	Rel.	Unrel.	Macro
So400m + SPECTER2	.4540	.3425	.4833	.4266
NeuML replaces SPECTER2	.4533	.3265	.4881	.4226
NeuML added	.4461	.3340	.4592	.4131

Table 9: Per-class OOF F1 of IC residual three-class models. These results exclude the `same_figure` gate.

True IC Stage 2	Same	Rel.	Unrel.
Same paper	451	296	253
Related papers	317	323	360
Unrelated papers	219	267	514

Table 10: Residual IC Stage-2 XGBoost OOF confusion matrix, before the Qwen exact-match gate.

Author Index

Accomazzi, Alberto, 1, 10, 61, 142
Alkan, Atilla Kaan, 1, 10, 17
Armeni, Kristijan, 122
Athanasiadis, Ioannis N., 38

Bang, My-Linh, 52
Bari, Anasse, 91
Bartlett, Jennifer Lynn, 142
Basseri, Benjamin, 61
Batalis, Steph, 77
Bisero, Sofia, 17
Boros, István, 23
Bovy, Jo, 174
Bölücü, Necva, 224

Cellier, M., 17
Centonze, Laura, 133
Chartab, Nima, 166
Chopra, Ankush, 210
Chowdhury, Debshree, 262
Cinosi, J., 17
Ciric, A., 17
Cornejo, Bernardo, 17

Dao-Sy, Duy-Minh, 52
Dawadi, Sundeep, 192
Duenser, Andreas, 224

Farzana, Most. Atkia, 112
Fedosov, Egor, 99

Ghimire, Prajwal, 192
Ghosal, Tirthankar, 1, 10, 142
Goyal, Rushank, 199
Grezes, Felix, 1, 10, 142

Hoque, Syed Mohaiminul, 104
Huynh, Trung-Kiet, 52

Irons, Jessica, 224

Jarmak, Stephanie, 61
Jaroschewski, Ilja, 17
Jin, Brian, 224

Kalia, Ananya, 216
Kelbert, Anna, 61
Kiendrebeogo, R. Weizmann, 17
Koblischke, Nolan, 174
Kovács, Ádám, 23
Kumar, Ashmani, 210
Kumar, Gaurav, 91

Lee, Changhyun, 224
Lockhart, Kelly, 1, 10, 142
Lu, Zehao, 38
Lynn, Meca, 166

Mahajan, Rutal, 216
Mahato, Aashish, 192
Maiti, Bodhisatta, 262
Mane, Raj Mahesh, 216

Nagel, David J., 91
Naiman, J. P., 180
Nguyen-Lam, Phu-Quy, 52
Nilsson, Ricky, 166

Pandiri, Anjali, 142
Pareek, Preksha, 216
Pham, Phu-Hoa, 52
Pollak, Senja, 122
Pranjić, Marko, 122
Protopapas, Pavlos, 61

Quinn, Katherine N., 77

Ramalakshmi, K S, 210
Ramírez, Irving Ernesto Quezada, 154
Recski, Gábor, 23
Regmi, Sunil, 192
Rovai, Fabio, 247

Saint-Paul, A., 17
Saud, Ramesh, 192
Saud, Shiv Ram, 192
Schüssler, Fabian, 17
Seefried, Ethan, 10, 142
Shakya, Manoj, 192
Sharma, Shubham, 210
Shetty, Aashna, 216
Sourav, Tonmoy Sarker, 112
Susemiehl, Nicholas, 166

Tan, Junyuan, 34
Timmer, Roelien C., 224

Toney-Wails, Autumn, 77
Tran, Chi-Nguyen, 52
Tóth, Szilveszter, 23

van der Plas, Thijs L., 38
van der Werf, Wopke, 38
Verdha, Nadia, 23

Wan, Stephen, 224
Xiong, Xingguo, 38

Yang, Huichen, 224

Zahid, M. Tahsan, 112
Zhang, Kemeng, 61
Zhu, Kewen, 34