WASP 2025

International Joint Conference on Natural Language Processing and Asia-Pacific Chapter of the Association for Computational Linguistics, 2025

Proceedings of the Workshop

December 23, 2025

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 979-8-89176-310-4

Introduction

We are excited to welcome you to WASP at IJCNLP-AACL 2025, the Third Workshop for Artificial intelligence for Scientific Publications. This year the conference is being held both online and in Mumbai, India on December 23, 2025.

Building on the success of the First Workshop on Information Extraction from Scientific Publications (WIESP) at AACL-IJCNLP 2022 and the Second WIESP at IJCNLP-AACL 2023, the Third Workshop on Artificial intelligence for Scientific Publications (WASP) at IJCNLP-AACL 2025 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 third 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 2025 includes one shared task where we invite teams (individuals and groups) to come up with a system to tackle bibliographic creation for space telescopes for TRACS: the dataset of Telescope Reference and Astronomy Categorization Shared task.

WASP 2025 received 31 submissions of which 21 were accepted (15 papers and 6 shared task system papers).

We are thankful to our program committee members for helping us curate a strong WASP 2025 program.

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

Organizing Committee

Program Chairs

Alberto Accomazzi, Center for Astrophysics | Harvard and Smithsonian

Program Chairs

Tirthankar Ghosal, Oak Ridge National Laboratory

Program Chairs

Felix Grezes, Center for Astrophysics | Harvard and Smithsonian

Program Chairs

Kelly Lockhart, Center for Astrophysics | Harvard and Smithsonian

Program Committee

Program Committee

Akhil Pandey Akella, AllSci Corp
Akiko Aizawa, National Institute of Informatics
Alberto Accomazzi, Center for Astrophysics | Harvard and Smithsonian
Antoine Gauquier, Ecole Normale Supérieure de Paris
Anurag Acharya, Pacific Northwest National Laboratory
Arne Binder, German Research Center for AI
Atilla Kaan Alkan, Center for Astrophysics | Harvard and Smithsonian
Bhargab Choudhury, Gauhati University
Biswadip Mandal, Amazon
Buse Sibel Korkmaz, Imperial College London
Daniel Acuna, University of Colorado at Boulder
Ekaterina Artemova, Toloka AI
Felix Grezes, Center for Astrophysics | Harvard and Smithsonian
Hamed Alhoori, Northern Illinois University
Iana Atanassova, Université de Franche Comté
Ibrahim Al Azher, Northern Illinois University
Jeyadev Needhidevan, New York University
Kelly Lockhart, Center for Astrophysics | Harvard and Smithsonian
Markus Zhang, Stanford University
Mengyang Liu, Independent Researcher
Michael Cochez, VU Amsterdam
Mina Basirat, University of Central Florida
Neil R. Smalheiser, University of Illinois at Chicago
Pawin Taechoyotin, University of Colorado at Boulder
Pierre Zweigenbaum, LISN, CNRS, Université Paris-Saclay
Rajesh Piryani, Akkodis
Sergey Feldman, Allen Institute for Artificial Intelligence
Sergi Blanco-Cuaresma, UniDistance Suisse
Shufan Ming, University of Illinois Urbana-Champaign
Shuntaro Yada, Tsukuba University, Tokyo Institute of Technology
Soham Chitnis, New York University
Sridevi Wagle, Pacific Northwest National Laboratory
Sujit Pal, Elsevier
Ted Pedersen, Nara Institute of Science and Technology
Thomas S Allen, Center for Astrophysics | Harvard and Smithsonian
Tim Schopf, Technische Universität Dresden
Tosho Hirasawa, Omron Sinic X
Wojtek Sylwestrzak, University of Warsaw
Yagmur Ozturk, Université Grenoble Alpes
Yuan Chang, University of the Chinese Academy of Sciences
Ziyue Li, University of the Chinese Academy of Sciences

Keynote Talk
Invited 1

Kartheik Iyer
Columbia University
2025-12-23 08:15:00 –

Abstract: Wandering through the Cosmic Library: Harnessing the embedding spaces of large language models for astronomical research and discovery"
Astronomical literature is expanding at an unprecedented rate, with thousands of papers added every month to preprint servers like arXiv.org and indexed by the NASA Astrophysics Data System (ADS). For academics and students, staying current with relevant work while keeping track of shifting trends therefore represents a critical challenge. This talk presents lessons learned from working with the UniverseTBD collaboration to develop Pathfinder, a complement to systems like ADS that uses large language models combined with retrieval-augmented generation (RAG) to enable semantic search and question-answering across the astronomy literature. I will discuss some of the unique challenges of applying NLP and LLMs to scientific publications in astronomy, including: (1) handling domain-specific terminology and mathematical notation, (2) grounding LLM responses in archival data to minimize hallucinations, and (3) leveraging embeddings to create interpretable semantic spaces for literature exploration. Drawing from Pathfinder's deployment (pfdr.app) and user feedback from the astronomy community, I will highlight how interpretable intermediate representations such as semantic embeddings and citation graphs can lend interpretability and rigor to otherwise black-box models, and help their adoption in research pipelines. Beyond astronomy, the development of these methods have broader implications for AI-assisted scientific discovery across disciplines. I will conclude by discussing open challenges in adapting large models in scientific contexts, the importance of retrieval mechanisms that preserve provenance, and the potential for LLM-powered tools to not just assist with literature review, but to help generate testable hypotheses and identify research gaps. As scientific publishing continues to accelerate across all fields, developing trustworthy and grounded systems for navigating the literature becomes increasingly essential.

Keynote Talk
Invited 2

Karin Verspoor
Royal Melbourne Institute of Technology
2025-12-23 12:00:00 –

Abstract: Impacts of AI on the Scientific Ecosystem"
Artificial Intelligence, in both predictive and generative forms, is increasingly being adopted to support — and in some cases, entirely perform — scientific research. In this talk, I will discuss both the significant opportunities that AI brings to science and the questions that AI raises for science. The talk will be grounded in some of my own work in use cases including bio-curation and literature-based discovery, as well as ongoing work exploring the limitations of LLMs, that may have particular impacts in the scientific arena.

Table of Contents

Overview of the Third Workshop for Artificial Intelligence for Scientific Publications

Kelly Lockhart‡, Alberto Accomazzi‡, Felix Grezes‡, Tirthankar Ghosal§
‡ Astrophysics Data System, Harvard-Smithsonian Center for Astrophysics, US
§Oak Ridge National Laboratory, US
kelly.lockhart@cfa.harvard.edu, accomazzi@cfa.harvard.edu
felix.grezes@cfa.harvard.edu, ghosalt@ornl.gov

Abstract

The Workshop for Artificial Intelligence for Scientific Publications (WASP), formerly Workshop on Information Extraction from Scientific Publications (WIESP), started in 2022 to provide a platform for researchers to discuss research on information extraction, mining, generation, and knowledge discovery from scientific publications using Natural Language Processing and Machine Learning techniques. The third WASP workshop was held at the 14th International Joint Conference on Natural Language Processing & 4th Asia-Pacific Chapter of the Association for Computational Linguistics in Mumbai, India on December 23rd, 2025, as a hybrid event. The WASP workshop saw great interest, with 29 submissions, of which 16 were accepted. The program consisted of the contributed research talks, 2 keynote talks, a panel discussion, and one shared task, Telescope Reference and Astronomy Categorization Shared task (TRACS).

1 Workshop description

The rise in scientific paper publications has greatly contributed to scientific advancement but has also complicated the ability of researchers to stay up-to-date in their fields. To navigate this vast amount of data and facilitate discovery, incorporating the metadata, full text, and citations into search engines is crucial. A popular and open example is the Astrophysics Data System (ADS; Kurtz et al., 2000), which offers many ways to discover research articles of interest within a curated collection of over 26 million records. However, navigating through this vast amount of data presents considerable challenges. To overcome them, extracting structured and semantically meaningful information from scientific publications becomes imperative.

The Workshop for Artificial Intelligence for Scientific Publications (WASP) was started to provide a platform for researchers to discuss research on information extraction, mining, generation, and knowledge discovery from scientific publications using Natural Language Processing and Machine Learning techniques.

The first WASP workshop was held under the name Workshop on Information Extraction from Scientific Publications (WIESP; Ghosal et al., 2022) in conjunction AACL-IJCNLP 2022. The second edition of WIESP was held along with IJCNLP-AACL 2023 (Ghosal et al., 2023). Much technological change has occurred since the first Workshop, especially around Generative Artificial Intelligence research. The Workshop's scope has expanded, along with the technology, and this year the inclusion of AI was cemented along with the new workshop name.

2 Program

The WASP 2025 workshop consisted of two keynote talks, contributed talks, a shared task, and a panel discussion. The main workshop received 29 submissions for contributed talks, of which 16 were accepted (55% acceptance rate). Since the workshop will be hybrid, there will be both in-person and virtual presentations at the conference venue and online. The papers accepted to the workshop cover a diverse array of research topics primarily centered on automating scholarly workflows, enhancing information extraction from scientific literature, ensuring the reliability of large language models (LLMs) in research, and advancing data management for open science initiatives.

Compared to the previous workshops in this series, the collection of research activities described in these works demonstrates a movement toward AI-assisted critical curation, where LLMs are employed not just to process and generate information, but are architecturally constrained and verified using external knowledge and validation signals derived from the scholarly ecosystem itself. Figure

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 1–4
December 23, 2025 ©2025 Association for Computational Linguistics

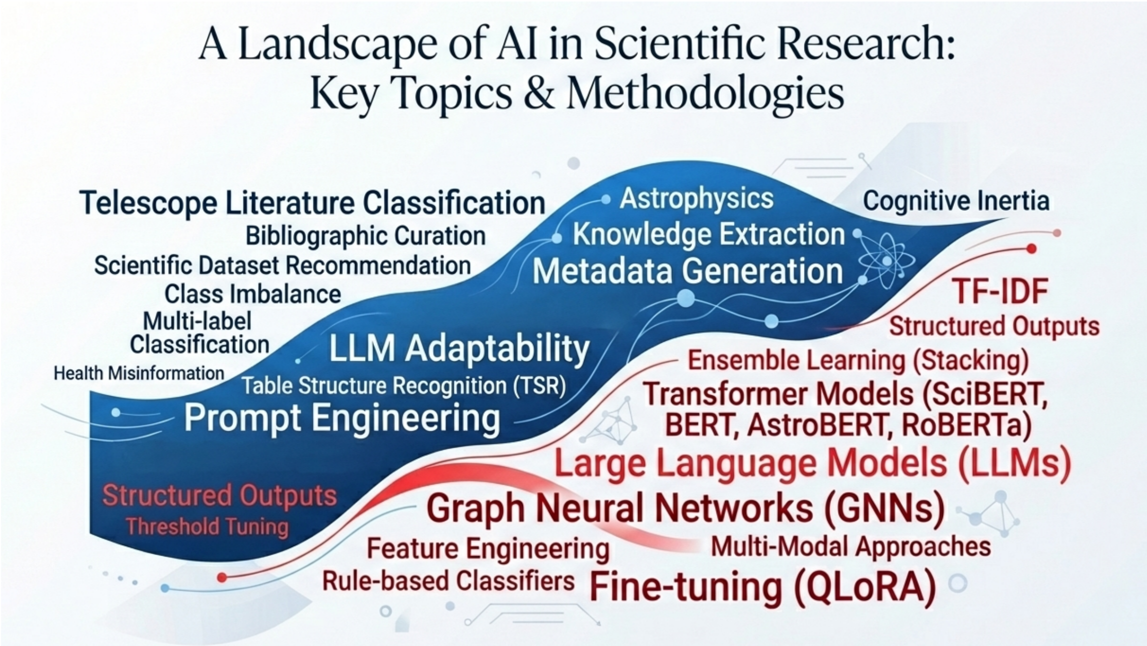


Figure 1: A representation of the scientific topics (blue) and methodologies (red) in the WASP papers.

1 provides a graphical representation of key topics and methodologies.

WASP 2025 also includes a panel discussion, *LLMs for "Trustworthy and Grounded" Scientific Discovery*. The panelists include our two keynote speakers, Karin Verspoor and Kartheik Iyer, along with Prasanna Balaprakash (Director of AI Programs, ORNL), and ChatGPT.

The full program, with links to papers, is available at https://ui.adsabs.harvard.edu/WIESP/2025/schedule.

3 Keynotes

This year we had two keynote lectures from researchers working in AI for scientific publications:

- Karin Verspoor, Dean, School of Computing Technologies, Royal Melbourne Institute of Technology, Australia
- Kartheik Iyer, NASA Hubble Fellow, Columbia University, USA

Speaker Karin Verspoor

Title "Impacts of AI on the Scientific Ecosystem"

Abstract Artificial Intelligence, in both predictive and generative forms, is increasingly being adopted to support — and in some cases, entirely perform — scientific research. In this talk, I will discuss both the significant opportunities that AI brings to science and the questions that AI raises for science. The talk will be grounded in some of my own work in use cases including bio-curation and literature-based discovery, as well as ongoing work exploring the limitations of LLMs, that may have particular impacts in the scientific arena.

Speaker Kartheik Iyer

Title "Wandering through the Cosmic Library: Harnessing the embedding spaces of large language models for astronomical research and discovery"

Abstract Astronomical literature is expanding at an unprecedented rate, with thousands of papers added every month to preprint servers like arXiv.org and indexed by the NASA Astrophysics Data System (ADS). For academics and students, staying current with relevant work while keeping track of shifting trends therefore represents a critical challenge. This talk presents lessons learned from working with the UniverseTBD collaboration to develop Pathfinder, a complement to systems like ADS that uses large language models combined with retrieval-augmented generation (RAG) to enable semantic search and question-answering across the astronomy literature. I will discuss some of the unique challenges of applying NLP and LLMs to scientific publications in astronomy, including: (1) handling domain-specific terminology and mathematical notation, (2) grounding LLM responses in archival data to minimize hallucinations, and (3) leveraging embeddings to create interpretable semantic spaces for literature exploration. Drawing from Pathfinder's deployment (pfdr.app) and user feedback from the astronomy community, I will highlight how interpretable intermediate representations such as semantic embeddings and citation graphs can lend interpretability and rigor to otherwise black-box models, and help their adoption in research pipelines. Beyond astronomy, the development of these methods have broader implications for AI-assisted scientific discovery across disciplines. I will conclude by discussing open challenges in adapting large models in scientific contexts, the importance of retrieval mechanisms that preserve provenance, and the potential for LLM-powered tools to not just assist with literature review, but to help generate testable hypotheses and identify research gaps. As scientific publishing continues to accelerate across all fields, developing trustworthy and grounded systems for navigating the literature becomes increasingly essential.

4 Telescope Reference and Astronomy Categorization Shared task (TRACS)

WASP 2025 hosted a shared task, Telescope Reference and Astronomy Categorization Shared task (TRACS; Grezes et al., 2025). The organizing committee was: Alberto Accomazzi, Tirthankar Ghosal, Kelly Lockhart, and Felix Grezes. The detailed overview paper is referred to and included in the proceedings. TRACS is available publicly on HuggingFace [1]. The scoring evaluation was run on the Kaggle platform[2].

TRACS Description: Astronomers typically gauge the scientific influence of observational facilities by examining publications that use the facilities' data. This depends on bibliographies that explicitly annotate and link data products to the relevant literature, enabling bibliometric analysis of data impact. Compiling such bibliographies is time-intensive and requires experts to comb the literature for names, acronyms, and identifiers, and then assess whether and how observations were used. Beyond impact assessment, these data-literature links are vital for researchers, as they form an important route to discovering and accessing data. By capitalizing on the expertise of librarians and archivists, telescope bibliographies can therefore directly support the scientific research workflow. In this context, we present the Telescope Reference and Astronomy Categorization Shared task (TRACS) and dataset, comprising more than 89,000 publicly available English texts drawn from space telescope bibliographies. These texts are labeled with a new, streamlined taxonomy developed in collaboration with experienced bibliographers. TRACS is intended as training material for contemporary Machine Learning and Artificial Intelligence methods that can assist data curators in building bibliographies. As an initial benchmark, we assess how existing Large Language Models perform on automatic bibliography curation. Both baseline and participant results underscore the difficulty of the problem and highlight the need for specialized tools. The TRACS shared task attracted 9 participating teams on Kaggle, of which 6 submitted system papers to WASP 2025.

[1] huggingface.co/datasets/adsabs/TRACS
[2] kaggle.com/competitions/tracs-wasp-2025

5 Conclusion

The rapid growth of scientific publishing presents both opportunity and difficulty for researchers who rely on accurate, interpretable, and well-structured access to scholarly literature. While significant progress has been made in information extraction, document understanding, and the responsible use of artificial intelligence, many challenges remain, particularly in reliably grounding automated systems in the scholarly ecosystem and in ensuring that AI-generated outputs can be trusted in research contexts. The contributions of WASP 2025 illustrate both the promise of current approaches and the continued need for rigorously validated, domain-aware methods.

By assembling researchers from NLP, information retrieval, and neighboring disciplines, WASP aims to advance these efforts and highlight emerging directions for AI-assisted scholarship. We hope that this workshop and its shared task will spur new collaborations, sharpen our understanding of open problems, and inspire the creation of robust tools that meaningfully support scientific discovery and the broader research community.

Program Committee

1. Akhil Pandey Akella, AllSci Corp
2. Akiko Aizawa, National Institute of Informatics
3. Alberto Accomazzi, Center for Astrophysics | Harvard & Smithsonian
4. Antoine Gauquier, Ecole Normale Supérieure

de Paris
5. Anurag Acharya, Pacific Northwest National Laboratory
6. Arne Binder, German Research Center for AI
7. Atilla Kaan Alkan, Center for Astrophysics | Harvard & Smithsonian
8. Bhargab Choudhury, Gauhati University
9. Biswadip Mandal, Amazon
10. Buse Sibel Korkmaz, Imperial College London
11. Daniel Acuna, University of Colorado at Boulder
12. Ekaterina Artemova, Toloka AI
13. Felix Grezes, Center for Astrophysics | Harvard & Smithsonian
14. Hamed Alhoori, Northern Illinois University
15. Iana Atanassova, Université de Franche Comté
16. Ibrahim Al Azher, Northern Illinois University
17. Jeyadev Needhidevan, New York University
18. Kelly Lockhart, Center for Astrophysics | Harvard & Smithsonian
19. Markus Zhang, Stanford University
20. Mengyang Liu, Independent Researcher
21. Michael Cochez, VU Amsterdam
22. Mina Basirat, University of Central Florida
23. Neil R. Smalheiser, University of Illinois at Chicago
24. Pawin Taechoyotin, University of Colorado at Boulder
25. Pierre Zweigenbaum, LISN, CNRS, Université Paris-Saclay
26. Rajesh Piryani, Akkodis
27. Sergey Feldman, Allen Institute for Artificial Intelligence
28. Sergi Blanco-Cuaresma, UniDistance Suisse
29. Shufan Ming, University of Illinois Urbana-Champaign
30. Shuntaro Yada, Tsukuba University, Tokyo Institute of Technology
31. Soham Chitnis, New York University
32. Sridevi Wagle, Pacific Northwest National Laboratory
33. Sujit Pal, Elsevier
34. Ted Pedersen, Nara Institute of Science and Technology
35. Thomas S Allen, Center for Astrophysics | Harvard & Smithsonian
36. Tim Schopf, Technische Universität Dresden
37. Tosho Hirasawa, Omron Sinic X
38. Wojtek Sylwestrzak, University of Warsaw
39. Yagmur Ozturk, Université Grenoble Alpes
40. Yuan Chang, University of the Chinese Academy of Sciences
41. Ziyue Li, University of the Chinese Academy of Sciences

References

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, Felix Grezes, Thomas Allen, Kelly Lockhart, Alberto Accomazzi, and Sergi Blanco-Cuaresma, editors. 2023. *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*. Association for Computational Linguistics, Bali, Indonesia.

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.

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.

Overview of TRACS: the Telescope Reference and Astronomy Categorization Dataset & Shared Task

Felix Grezes[1], Jennifer Lynn Bartlett[1], Kelly Lockhart[1], Alberto Accomazzi[1], Ethan Seefried[2], Anjali Pandiri[3], and Tirthankar Ghosal[2]
[1]Center for Astrophysics, Harvard & Smithsonian, USA
[2]Oak Ridge National Laboratory, USA, [3]Florida State University, USA
[1]{felix.grezes,jennifer.bartlett,kelly.lockhart,aaccomazzi}@cfa.harvard.edu
[2]{seefriedej,ghosalt}@ornl.gov, [3]ap23bl@fsu.edu

Abstract

To evaluate the scientific influence of observational facilities, astronomers examine the body of publications that have utilized data from those facilities. This depends on curated bibliographies that annotate and connect data products to the corresponding literature, enabling bibliometric analyses to quantify data impact. Compiling such bibliographies is a demanding process that requires expert curators to scan the literature for relevant names, acronyms, and identifiers, and then to determine whether and how specific observations contributed to each publication. These bibliographies have value beyond impact assessment: for research scientists, explicit links between data and literature form an essential pathway for discovering and accessing data. Accordingly, by building on the work of librarians and archivists, telescope bibliographies can be repurposed to directly support scientific inquiry. In this context, we present the Telescope Reference and Astronomy Categorization Shared task (TRACS) and its accompanying dataset, which comprises more than 89,000 publicly available English-language texts drawn from space telescope bibliographies. These texts are labeled according to a new, compact taxonomy developed in consultation with experienced bibliographers.

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 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 bibliographic 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

[1]scixplorer.org

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 5–15

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 the Telescope Reference and Astronomy Categorization Dataset & 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 Dataset Description

2.1 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 2.2 below.

Bibliographic data for the Chandra X-ray Observatory (`CHANDRA`) was provided by Erin Scott of 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 Jenny Novacescu of 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.

[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 introdu see it., ce 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 §2.2); 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.

2.2 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.

2.3 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.

3 Baseline Evaluation Task

An automated assistant able to emulate the supervised curation activities listed in the 5 above would provide a valuable contribution to the human effort involved. LLMs have shown flexibility in interpreting and classifying scientific articles which are the basis for this curation activity. They have also been successfully used for information extraction tasks, 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.

3.1 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.

3.1.1 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.

3.1.2 Usage Classification

The second subtask evaluates how each paper utilizes telescope data. As defined in Section 2.2, each entry includes four boolean labels: `science`, `instrumentation`, `mention`, and `not_telescope`. Each system must output a structured 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.

3.2 Evaluation Metrics

Each submission is automatically evaluated by matching predictions to reference labels via the `Id` field. Participants are ranked by macro-averaged F1-scores across both subtasks, adapting code from the standard Scikit Learn library (2011).

3.2.1 Telescope Classification

Performance on this subtask is measured using the **macro-F1** score across the four telescope categories (`CHANDRA`, `HST`, `JWST`, `None`), ensuring equal weighting for rare and frequent classes alike.

3.2.2 Usage Classification

For the second subtask, performance is ranked by the **macro-F1** averaged across the four usage categories, rewarding balanced sensitivity across the different forms of telescope data use.
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
```

3.3 Baseline Experiments

To establish initial performance benchmarks for the TRACS baseline task, we evaluated five state-of-the-art open large language models (LLMs): *GPT-OSS-20B* (OpenAI et al., 2025), *Mistral-7B-Instruct* (Jiang et al., 2023), *LLaMA-3.1-8B-Instruct* (Weerawardhena et al., 2025), *Zephyr-7B-Beta* (Tunstall et al., 2023), and *Solar-Pro-Preview* (Kim 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 architectural characteristics and motivations for each baseline model.

3.3.1 Telescope Classification

For the telescope prediction task, each baseline model was prompted with the relevant textual fields and asked to output one of the four valid labels (`CHANDRA`, `HST`, `JWST`, or `None`). Predictions were evaluated against the gold labels using **macro-F1** to ensure balanced treatment of all telescope categories, along with overall accuracy for reference.

As shown in Table 4, GPT-OSS-20B and LLaMA-3.1-8B achieved the strongest overall performance, demonstrating that general-purpose open LLMs can capture some telescope-specific cues without additional training. Meanwhile, smaller instruction-tuned models such as Mistral-7B and Zephyr-7B exhibited lower recall across minority classes, suggesting limited domain generalization in zero-shot settings.

3.3.2 Usage Classification

For the usage classification task, models were evaluated on their ability to assign one of four binary labels (`science`, `instrumentation`, `mention`, `not_telescope`) to each paper, indicating the role of telescope data. Each model was prompted with the same input fields and evaluated using **macro-F1** per usage category within each telescope split.

Table 5 reports the per-class F1-scores across telescope subsets. Performance varied widely across usage types, with higher recall observed for `science` and `mention` labels, while `instrumentation` and `not_telescope` were more challenging. These results highlight the difficulty of capturing fine-grained scientific intent from text without explicit domain supervision.

3.4 Analysis of Benchmarks

3.4.1 Qualitative Analysis

The results reveal several key patterns in how baseline models approach telescope and usage classification tasks.

Telescope Classification Challenges The modest macro-F1 scores (ranging from 7.00% to 11.50%) across all models indicate that distinguishing between telescope types from textual descriptions alone remains a substantial challenge in zero-shot settings. Notably, LLaMA-3.1-8B achieved the highest telescope accuracy (38.40%) but a lower macro-F1 (11.12%), suggesting a bias toward predicting the dominant `NONE` class. This pattern is consistent across models: the `NONE` F1 scores (ranging from 23.88% to 25.00%) substantially out-

Model	Architecture Summary	Motivation	Params
GPT-OSS-20B	Decoder-only transformer trained on diverse web and technical text, representing a general-purpose open-source GPT design.	Serves as the closest open analog to proprietary GPT-series models, providing a strong general baseline.	20B
Mistral-7B-Instruct	Decoder-only dense transformer using grouped-query attention (GQA) and sliding-window context mechanisms.	Known for efficient context handling and strong instruction tuning despite small parameter size.	7B
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
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
Solar-Pro-Preview	Hybrid attention decoder combining dense and mixture-of-experts routing layers.	Tests whether hybridized attention mechanisms improve performance on specialized scientific reasoning tasks.	22B

Table 3: Baseline models evaluated on the TRACS dataset. Each model was run "out-of-the-box" with identical prompts and token limits. The architecture and motivation columns highlight differences in model design and intended use.

Model	CHANDRA F1	HST F1	JWST F1	NONE F1	Macro F1	Tel. Accuracy
LLaMA-3.1-8B	**17.14**	14.05	**9.41**	23.88	11.12	**38.40**
Zephyr-7B-Beta	6.79	1.79	2.20	**25.00**	7.00	31.00
Solar-Pro-Preview	5.35	1.05	0.37	24.45	7.80	27.60
GPT-OSS-20B	16.73	**14.76**	8.84	24.91	**11.50**	19.80
Mistral-7B	9.11	6.06	6.08	24.91	8.20	19.80

Table 4: Telescope classification performance across four telescope categories. Reported are per-class macro F1-scores, overall macro F1, and overall accuracy (all in %). Models are ordered by descending telescope accuracy.

Model	Science F1 (%)			Instr. F1 (%)			Mention F1 (%)			Not-Tel. F1 (%)		
	CH.	HST	JW.	CH.	HST	JW.	CH.	HST	JW.	CH.	HST	JW.
LLaMA-3.1-8B	**73.41**	72.80	**34.60**	4.19	0.00	10.26	**36.53**	**24.71**	40.42	19.32	11.50	12.42
Zephyr-7B-Beta	54.12	48.66	16.06	22.75	17.65	4.44	9.88	3.33	7.03	**28.69**	10.74	6.97
Solar-Pro-Preview	28.66	7.52	2.80	40.58	**38.10**	10.00	3.35	0.40	0.27	0.00	0.61	0.00
GPT-OSS-20B	72.01	**75.34**	27.03	**44.34**	17.89	**20.00**	20.12	17.34	**45.36**	0.00	0.00	0.00
Mistral-7B	65.91	66.16	27.44	19.35	10.26	18.18	13.49	19.11	15.61	26.22	**26.93**	**23.34**

Table 5: Per-split usage classification F1-scores across telescope subsets (CH = CHANDRA, HST = Hubble Space Telescope, JW = JWST). Values are per usage class and split, averaged over all test papers in each subset. Models are ordered by descending telescope accuracy.

perform telescope-specific classes, with CHANDRA F1 scores reaching only 6.79–17.14%, HST scores of 1.05–14.76%, and JWST scores of 0.37–9.41%. The poor performance on minority telescope classes suggests that subtle linguistic markers distinguishing telescope types are not readily captured by general-purpose language models without domain-specific fine-tuning or few-shot examples.

Usage Classification Patterns Usage classification performance exhibits significant variance across both models and telescope types. The `science` category consistently achieves the highest F1 scores, with models reaching 28.66–73.41% for CHANDRA, 7.52–75.34% for HST, and 2.80–34.60% for JWST papers. This suggests that scientific usage contains more explicit textual indicators that align with pre-training distributions. In contrast, `instrumentation` detection proves highly inconsistent, with most models struggling (0.00–22.75% for CHANDRA) except for specialized cases where Solar-Pro-Preview and GPT-OSS-20B achieve 40.58% and 44.34% respectively on CHANDRA papers. The `mention` category shows moderate but unstable performance (0.27–45.36%), while `not_telescope` classification remains particularly challenging, with several models achieving 0.00% F1 and the best results reaching only 26.22–28.69%.

Model-Specific Behaviors GPT-OSS-20B demonstrates the most balanced performance profile, excelling at `instrumentation` detection (44.34% for CHANDRA) and achieving competitive scores on `science` classification, though it completely fails to identify `not_telescope` cases. LLaMA-3.1-8B shows strong performance on `science` classification and maintains reasonable scores across `mention` categories, but struggles with `instrumentation` (particularly 0.00% for HST). Smaller models like Zephyr-7B-Beta and Mistral-7B exhibit more conservative prediction patterns, achieving modest but non-zero scores across categories, suggesting less confident predictions that may result in better calibration for certain classes. Solar-Pro-Preview displays an unusual specialization pattern, performing well on `instrumentation` but nearly failing on `mention` and `not_telescope` categories.

These qualitative patterns underscore the need for domain-specific training data and suggest that telescope and usage classification require understanding of specialized astronomical terminology and research methodology that is not adequately represented in general pre-training corpora.

3.4.2 Error Analysis

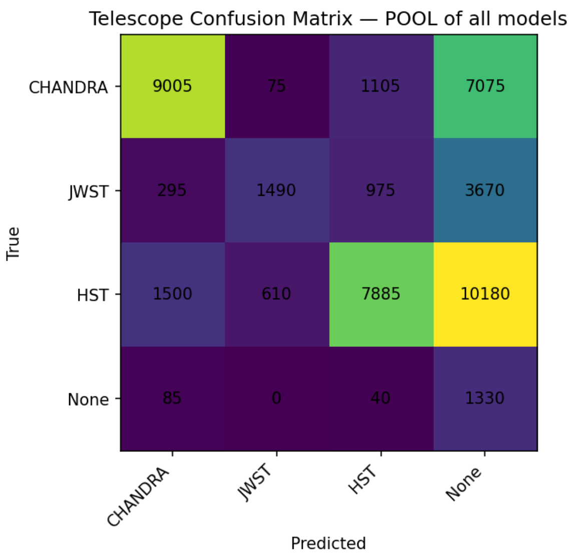


Figure 1: **Confusion matrix across all models: telescope accuracy.**

To better understand model failure modes, we conducted an error analysis using the pooled confusion matrix across all baseline models (Figure 1).

Systematic Over-Prediction of `NONE` The confusion matrix reveals a strong bias toward predicting the `NONE` class across all models. Of the 18,260 true CHANDRA papers, 7,075 (38.7%) were incorrectly classified as `NONE`, representing the single largest error category. Similarly, 3,670 JWST papers (56.1% of true JWST instances) and 10,180 HST papers (50.8%) were misclassified as `NONE`. This systematic over-prediction reflects the class imbalance in the dataset and suggests that models default to the majority class when telescope-specific textual cues are absent or ambiguous. The severe impact on minority classes (particularly JWST, with only 1,490 correct predictions out of 6,430 instances) indicates that zero-shot models struggle to identify distinctive markers for less-represented telescopes.

Cross-Telescope Confusion Patterns Beyond the `NONE` bias, substantial confusion exists between telescope classes themselves. CHANDRA papers show moderate confusion with HST (1,105 errors, 6.0% of true CHANDRA), while HST papers exhibit bidirectional confusion with CHANDRA (1,500 errors, 7.5%) and JWST (610 errors, 3.0%). Notably, the confusion matrix is asymmetric: while HST is frequently mispredicted as CHANDRA (1,500 instances), the reverse error occurs less frequently (1,105 instances). This asymmetry likely reflects differences in corpus frequency during pretraining, with CHANDRA-related terminology potentially more prominent in general astronomical corpora due to its longer operational history. The relatively low inter-telescope confusion for JWST (295 JWST papers predicted as CHANDRA, 975 as HST) suggests that when JWST is not classified as `NONE`, its textual markers are somewhat distinctive—though the high `NONE` misclassification rate remains the dominant error mode.

Usage Classification Challenges Analysis of per-class F1 scores across usage types reveals stark performance disparities. The `science` category achieves the highest scores across all models and telescopes (ranging from 63.1% to 77.9% support across splits), indicating that research papers describing scientific findings contain relatively explicit linguistic indicators. In contrast, `instrumentation` classification proves highly inconsistent, with precision, recall, and F1 varying dramatically by model—support is only 2.5% of papers, yet several models achieve 0.00 F1 while others reach above 40% on specific splits. This suggests that instrumental development papers employ technical jargon that some model architectures capture while others miss entirely.

The `mention` category (35.7% support) shows moderate performance, likely because papers merely citing telescope data use formulaic language patterns (e.g., "archival observations from..."). However, the `not_telescope` class remains challenging despite representing 18.6% of papers, with most models achieving near-zero F1 scores. Manual inspection of errors in this category revealed that papers discussing related instruments (e.g., ground-based telescopes, space missions without the target telescopes) use similar astronomical terminology, making discrimination difficult without explicit negative evidence.

Model-Specific Error Patterns Examining per-model usage classification reveals distinct behavioral profiles. Models achieving higher macro-F1 on telescope classification (GPT-OSS-20B and LLaMA-3.1-8B) do not consistently outperform on usage classification, suggesting these are partially independent capabilities. GPT-OSS-20B demonstrates strong `instrumentation` detection (44.3% F1 on CHANDRA) but completely fails on `not_telescope` (0.00% across all splits), indicating overly aggressive telescope assignment. Conversely, Mistral-7B shows more conservative predictions with non-zero performance across all categories, though at lower overall accuracy. This trade-off between precision and recall across usage categories highlights the difficulty of calibrating decision boundaries in zero-shot settings without task-specific examples.

Implications for Future Work These error patterns motivate several directions for improvement. The severe class imbalance necessitates sampling strategies or loss functions that explicitly counteract majority-class bias. The high rate of cross-telescope confusion suggests that models would benefit from few-shot examples highlighting distinctive features of each telescope's observational methodology. Finally, the near-complete failure on `not_telescope` classification indicates that negative training examples—papers that superficially resemble telescope studies but do not use the target instruments—are essential for learning proper decision boundaries. Future dataset iterations should include balanced sampling and explicit annotation of telescope mention spans to support more fine-grained extractive approaches.

4 Participant Systems

The TRACS shared task attracted 9 participating teams on Kaggle, of which 6 submitted system papers to WASP 2025.

- Varshney et al. (2025) propose a multi-model ensemble architecture integrating transformer models DeBERTa, RoBERTa, and TF-IDF logistic regression. They demonstrate the effectiveness of combining transformer-based contextual embeddings with traditional TF-IDF lexical features in a multi-label classification framework or telescope-paper linkage. The ensemble approach significantly improves performance, especially on challenging and imbalanced label categories such as instrumentation.

- Khatib et al. (2025) combined symbolic and neural approaches, utilizing a tuned Random Forest classifier stacked with domain-adapted semantic modeling (astroBERT) and four independent boosting meta-learners.

- Rawat et al. (2025) leveraged the domain-adapted SciBERT, stochastically sampled segments from the training data and used majority voting over the test segments at inference time. significantly outperforming the open-weight GPT baseline.

- Wu et al. (2025) built `amc` on top of existing LLMs, combining keywords, re-ranking, and reasoning to achieve the 3rd highest score on the leaderboard. They also explore how to interrogate historical datasets and surface potential label errors.

- Nguyen et al. (2025) compare traditional machine learning methods such as multinomial Naive Bayes with TF–IDF and CountVectorizer representations, to various modern transformer BERT-based models. Their experiments demonstrate that domain-adapted BERT variants significantly outperform traditional statistical machine learning methods.

- Naidu (2025) show that SciBERT, despite its context-length constraints, can be efficiently finetuned to TRACS. They achieve the highest score on the leaderboard, while discussing the effect of truncation and arguing that a lightweight model can outperform larger LLMs, achieving the top leaderboard score.

5 Results, Analysis, and Findings of TRACS

We report the results of the participating teams in table 6. Overall, SciBERT(Beltagy et al., 2019), astroBERT(Grezes et al., 2021), and other BERT based systems performed well, highlighting the utility of smaller open-source networks when finetuned networks when compared to closed, large general purpose LLMs. The top performer further described how these smaller models are also more efficient, with high-potential for real world applications.

Team	Test F1
Naidu (2025)	0.89
Nguyen et al. (2025)	0.85
Wu et al. (2025)	0.84
Khatib et al. (2025)	0.82
Varshney et al. (2025)	0.73
Rawat et al. (2025)	0.73
Random Baseline	0.24
GPT-OSS-20B	0.12

Table 6: Main TRACS@WASP 2025 shared task results. All scores computed using micro-averaging.

6 Conclusion and Future Directions

In this paper, we present TRACS, a novel dataset and associated shared for task automated bibliographic curation for astronomy, and briefly describe the 6 system papers submitted to TRACS@WASP 2025. For the dataset introduce a bibliographic taxonomy developed in collaboration with established bibliographers, grounded in real-world curatorial practices, and 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. From the participating systems, we find that finetuned BERT-based models have both the best performance and efficiency. The best model obtains 89% F1-score.

In the future, we plan to keep expanding TRACS with as many human-curated bibliographies as possible, including ground telescopes. We have already starting coordinating with curators at the European Southern Observatory to add the Very Large Telescope to the dataset. In addition to more data, we would also 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 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

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.

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. In *EMNLP*. Association for Computational Linguistics.

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.

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.

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. *arXiv e-prints*, arXiv:2112.00590.

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.

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.

Arshad Khatib, Aayush Prasad, Rudra Trivedi, and Shrikant Malviya. 2025. A hybrid stacking ensemble for astrophysical document classification. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling. *arXiv e-prints*, arXiv:2312.15166.

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.

Madhusudhana Naidu. 2025. Efficient context-limited telescope bibliography classification for the wasp-2025 shared task using scibert. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Lam Phu Quy Nguyen, Chi Nguyen Tran, Sy Duy Minh Dao, Phu Hoa Pham, and Trung Kiet and Huynh. 2025. Systematic evaluation of machine learning and transformer-based methods for scientific telescope literature classification. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

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.

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.

Shivam Rawat, Lucie Flek, and Akbar Karimi. 2025. Encoder fine-tuning with stochastic sampling outperforms open-weight gpt in astronomy knowledge extraction. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

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.

Ojaswa Varshney, Prashasti Vyas, Priyanka Goyal, Tarpita Singh, Ritesh Kumar, and Mayank Singh. 2025. Automated telescope-paper linkage via multi-model ensemble learning. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Sajana Weerawardhena, Paul Kassianik, Blaine Nelson, Baturay Saglam, Anu Vellore, Aman Priyanshu, Supriti Vijay, Massimo Aufiero, Arthur Goldblatt, Fraser Burch, Ed Li, Jianliang He, Dhruv Kedia, Kojin Oshiba, Zhouran Yang, Yaron Singer, and Amin Karbasi. 2025. Llama-3.1-FoundationAI-SecurityLLM-8B-Instruct Technical Report. *arXiv e-prints*, arXiv:2508.01059.

John F. Wu, Joshua E.G. Peek, Sophie J. Miller, Jenny Novacescu, Achu J. Usha, and Christopher A. Wilkinson. 2025. amc: The automated mission classifier for telescope bibliographies. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Exploring Health Misinformation Detection with Multi-Agent Debate

Chih-Han Chen[*]
National Taiwan University
chen.eric0208@gmail.com

Chen-Han Tsai
HTC DeepQ
maxwell_tsai@htc.com

Yu-Shao Peng
HTC DeepQ
ys_peng@htc.com

Abstract

Fact-checking health-related claims has become increasingly critical as misinformation proliferates online. Effective verification requires both the retrieval of high-quality evidence and rigorous reasoning processes. In this paper, we propose a two-stage framework for health misinformation detection: Agreement Score Prediction followed by Multi-Agent Debate. In the first stage, we employ large language models (LLMs) to independently evaluate retrieved articles and compute an aggregated agreement score that reflects the overall evidence stance. When this score indicates insufficient consensus—falling below a predefined threshold—the system proceeds to a second stage. Multiple agents engage in structured debate to synthesize conflicting evidence and generate well-reasoned verdicts with explicit justifications. Experimental results demonstrate that our two-stage approach achieves superior performance compared to baseline methods, highlighting the value of combining automated scoring with collaborative reasoning for complex verification tasks.

1 Introduction & Related Work

The proliferation of health-related content on digital platforms poses significant challenges to ensuring accurate medical information reaches the public. Verifying health claims is critical for safeguarding public well-being, as false or misleading information can cause substantial harm to individual and population health. Despite the vast volume of health content available online, only a small fraction is supported by robust scientific evidence, underscoring the urgent need for automated verification systems.

In open-domain fact-checking, traditional methods predominantly rely on BERT-based architectures (Devlin et al., 2019). Pipeline-based systems employ BERT models to retrieve relevant evidence sentences, followed by a classification module to predict claim veracity. Joint systems perform evidence retrieval and veracity prediction simultaneously within a unified model. While conceptually straightforward, these approaches require predefined knowledge databases and necessitate training encoder-based models from scratch (Vladika et al., 2024), limiting their flexibility and scalability.

The emergence of large language models (LLMs) has introduced new paradigms. Tian et al. (2024) deploy web retrieval agents to gather evidence dynamically, enabling LLMs to assess sufficiency and render verdicts. Singal et al. (2024) integrate retrieval-augmented generation (RAG) with in-context learning (ICL) for veracity prediction. Vladika et al. (2025) propose multi-turn LLM interactions that iteratively generate questions, retrieve evidence, and reason about claim validity. However, these approaches typically lack explicit evidence filtering mechanisms, relying directly on outputs from web search tools or dense retrieval models.

Recent work has explored multi-agent frameworks for fact-checking. Hong et al. (2025) leverage multiple agents to evaluate evidence quality and determine veracity, with provisions for regathering evidence when necessary. Hu et al. (2025), Liang et al. (2024), and (Liu et al., 2025) adopt Multi-Agent Debate (MAD) frameworks to enhance reasoning robustness and mitigate degenerate reasoning patterns.

Building upon these advances, we propose a two-stage multi-agent debate framework for health misinformation detection. Our approach first employs LLMs to retrieve and evaluate high-quality articles, computing an aggregated agreement score. When evidence exhibits significant disagreement—indicated by a score below a predefined threshold—the system initiates a structured multi-agent debate. Through iterative argumenta-

[*]Work done while interning at HTC DeepQ.

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 16–21
December 23, 2025 ©2025 Association for Computational Linguistics

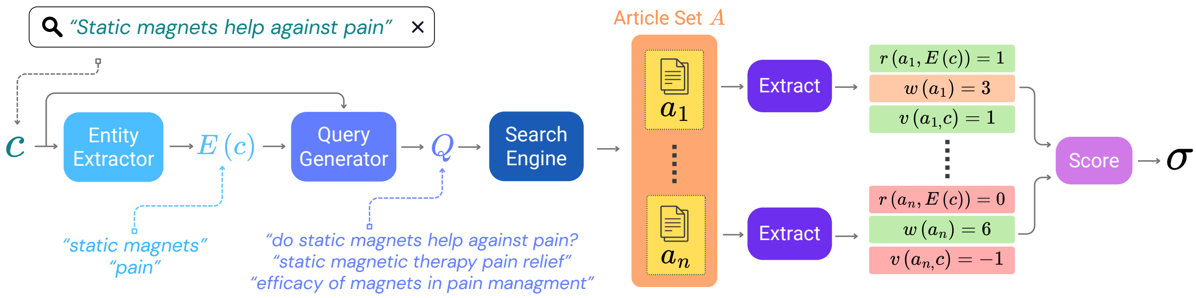


Figure 1: **Agreement Score Prediction (Stage 1).** For a given claim c, entities $E(c)$ are extracted and passed alongside c to a query generator to generate search queries Q. Articles relating to Q are collected into an article set A. We extract topic relevance r, article weights w, and article verdict v for each article $a \in A$. The results are aggregated, resulting in the final agreement score σ.

tion, agents collaboratively analyze conflicting evidence to produce well-justified verdicts grounded in explicit reasoning.

2 Methodology

In this section, we detail the implementation of our proposed two-stage health misinformation detection algorithm. The first stage takes a claim as input and retrieves a set of articles relating to the claim. Each article is classified as to whether it *Supports* or *Refutes* the claim, and the predictions are aggregated. When the agreement among predictions is high, the veracity of the claim is determined by majority vote. In the case of low agreement, we initiate the second stage multi-turn debate. Two opposing agents are provided with supporting evidence collected during the first stage, and a judge agent supervises the debate process until the claim's veracity can be determined. The details of each stage are presented in the following.

2.1 Agreement Score Prediction

Figure 1 illustrates the first stage framework of our approach. For a given claim c, we first extract a set of entities $E(c)$ from c using an LLM. The entities are keywords or phrases from c that the claim is focused on. The claim c and entities $E(c)$ are then provided to an LLM to generate a set of queries Q. Each query $q \in Q$ is sent to a search engine for article retrieval. The article sets retrieved from each query are de-duplicated and merged to form the article set A.

Given the obtained queries Q, entities $E(c)$, and article set A, we prompt an LLM to extract the following information from each article $a \in A$. Specifically, we look for:

1. **Topic Relevance:** Check whether the article a contains content relevant for all entities in $E(c)$. We define this relevance as $r(a, E(c)) \in \{0, 1\}$, where $r(a, E(c)) = 1$ if the article contains content relevant for all entities in $E(c)$ and $r(a, E(c)) = 0$ otherwise.

2. **Attribute Assessment:** Evaluate whether article a contains the following attributes: *Problem Statement*, *Experimental Setup*, *Findings*, *Statistical Significance*, *Limitations*, and *Results*. These 6 attributes reflect the structure of modern scientific publications. Specifically, an article that coverts the 6 attributes are often more thorough in its claims. We define the article weight as:

$$w(a) = \sum_{\alpha \in \text{Attributes}} \mathbf{1}[\alpha \in a] \in \{0, 1, \ldots, 6\}$$

where $\mathbf{1}[\cdot]$ is the indicator function for whether attribute α is in article a.

3. **Article Verdict:** Determine whether the contents of the article a *support* or *refute* the claim c. We denote $v(a, c) \in \{-1, 1\}$ where $v(a, c) = 1$ indicates *support* and $v(a, c) = -1$ indicates *refute*.

We then compute the *agreement score* $\sigma(c, A) \in [-1, 1]$ for claim c and article set A as:

$$\sigma(c, A) = \frac{1}{Z} \sum_{a \in A} r(a, E(c)) \cdot w(a) \cdot v(a, c),$$

where

$$Z = \sum_{a \in A} r(a, E(c)) \cdot w(a)$$

is the normalizing constant. We consider the case where $Z \neq 0$ by assuming quality relevant

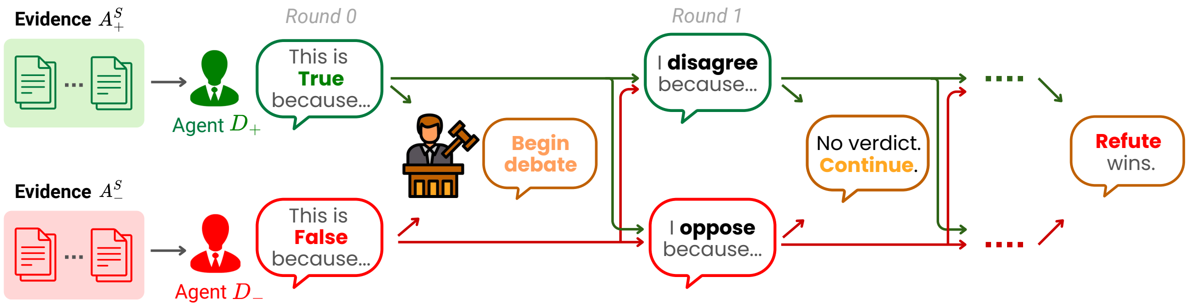


Figure 2: **Multi-Agent Debate (Stage 2).** Articles from the first Agreement Score Prediction stage are organized into supporting and refuting evidence sets A^S_+ and A^S_-, which are provided to agents D_+ and D_-, respectively. Each agent begins with an opening statement based on their evidence set, after which the judge initiates the debate. In each round, agents review their opponent's argument before providing a counterargument. After each round, the judge determines whether sufficient information exists to reach a verdict. If not, the debate continues for another round. The process concludes when the judge reaches a final verdict.

articles to be available within the search engine results.

We introduce a threshold $\tau > 0$ to quantify the *level of agreement* among the retrieved articles. If $|\sigma| \geq \tau$, this indicates that most articles consistently *support* or *refute* the claim. When such high level of agreement exists, the first stage directly outputs *support* for $\sigma \geq \tau$ and *refute* for $\sigma \leq -\tau$.

Conversely, an agreement score $|\sigma| < \tau$ indicates a significant level of disagreement among the articles. In this case, we pass the results to the second stage for debate.

2.2 Multi-Agent Debate

Figure 2 illustrates the second stage framework of our approach. We employ a multi-agent debate framework based on the work by Liang et al. (2024). The debate framework involves three agents: the **Support Agent** D_+, **Refute Agent** D_-, and **Judge Agent** J. Evidence is first prepared using the results from the first stage before initiating the debate.

Evidence Preparation: Given the article set A, we select two disjoint subsets A_+ and A_- from A such that:

$$A_+ = \{a \in A \mid v(a,c) = +1,\ r(a, E(c)) = 1\},$$
$$A_- = \{a \in A \mid v(a,c) = -1,\ r(a, E(c)) = 1\}.$$

Articles in A_+ and A_- are ranked in descending order using $w(a)$, and we limit each set to contain an equal number of articles. For each article in the remaining sets, we prompt an LLM using the claim c to extract passages from the original text that *supports* or *refutes* claim c along with its reason. We concatenate the LLM responses from all articles in sets A_+ and A_- into A^s_+ and A^s_-. We denote A^s_+ and A^s_- as the *supporting* and *refuting evidence* throughout the debate process.

Opening Statement: The support agent D_+ and refute agent D_- begins with an opening statement by presenting the evidence in A^S_+ and A^S_-. We denote the outputs of the support and refute agents as

$$S^{(0)}_+ = D_+(A^s_+), \quad S^{(0)}_- = D_-(A^s_-).$$

Each agent also maintains a conversation history H. Following the opening statement, we initialize each agent's history as

$$H^{(0)}_+ = \{S^{(0)}_+\}, \quad H^{(0)}_- = \{S^{(0)}_-\}.$$

The judge agent's history is initialized using the opening statements given by the two debate agents

$$H^{(0)}_J = \{S^{(0)}_+, S^{(0)}_-\}.$$

Next, the judge initiates the debate process, and we proceed to the first round of debate.

Debate Process: In every debate round, each agent responds to the opposing agent's statement $S^{(i-1)}$ using its past conversation history $H^{(i-1)}$. The outputs of the support and refute agent from the i-th round are given as

$$S^{(i)}_+ = D_+\big(S^{(i-1)}_-, H^{(i-1)}_+\big),$$
$$S^{(i)}_- = D_-\big(S^{(i-1)}_+, H^{(i-1)}_-\big).$$

The debate agent's histories are updated by concatenating the opposing agent's response along with the current response

$$H^{(i)}_+ = H^{(i-1)}_+ \oplus S^{(i-1)}_- \oplus S^{(i)}_+,$$
$$H^{(i)}_- = H^{(i-1)}_- \oplus S^{(i-1)}_+ \oplus S^{(i)}_-.$$

The judge agent J takes the response from both agents along with its own history $H_J^{(i-1)}$, and decides whether sufficient information exists to reach a verdict. Specifically,

$$\theta^{(i)} = J\Big(S_+^{(i)},\ S_-^{(i)},\ H_J^{(i-1)}\Big)$$

where $\theta^{(i)} \in \{\texttt{support}, \texttt{refute}, \texttt{continue}\}$. If the judge agent believes an argument is compelling enough, the verdict $\theta^{(i)} \in \{\texttt{support}, \texttt{refute}\}$ is returned. If neither argument is sufficiently convincing, the judge agent outputs $\theta^{(i)} = \texttt{continue}$, and the debate continues for another round.

The judge's history is also updated by appending the debate agent responses

$$H_J^{(i)} = H_J^{(i-1)} \oplus S_+^{(i)} \oplus S_-^{(i)}.$$

To prevent indefinitely long debates, we limit the process to a maximum of M rounds, after which the judge must reach a verdict $\theta^{(M)} \in \{\texttt{support}, \texttt{refute}\}$ based on the debate history.

3 Experiments and Setup

3.1 Datasets

We consider the following health-related datasets for our experiments.

SciFact (Wadden et al., 2020) contains expert-written biomedical claims derived from medical paper abstracts. We use the development subset, consisting of 188 claims: 124 supported and 64 refuted.

TREC-Health (Pugachev et al., 2023) is constructed from the TREC 2019 Decision Track (Abualsaud et al., 2020) and the TREC 2021 Health Misinformation Track (Clarke et al., 2021), both of which target challenges in search engine results related to health misinformation. The dataset includes 113 consumer health questions, of which 61 are supported and 52 are refuted.

HealthFC (Vladika et al., 2024) consists of everyday health-related claims spanning diverse topics. We use a subset of 327 claims: 202 supported and 125 refuted.

3.2 Metrics

We report macro-precision, macro-recall, and macro-F1 as evaluation metrics. These are standard in fact-checking tasks, as they provide a balanced analysis of prediction performance across labels.

3.3 Baseline Algorithms

We consider WebAgent (Tian et al., 2024) and StepByStep (Vladika et al., 2025) as benchmark algorithms. Among them, StepByStep represents the current state-of-the-art in health-related fact-checking. For fairness, all methods, including ours, use the Brave search engine (Brave Software, Inc.) and GPT-4o (OpenAI, 2024) as the underlying LLM. Each algorithm is executed three times, and we report the best performance.

For our framework, we set the parameters as follows: entity set size $|E(c)| = 2$, query set size $|Q| = 5$, article set size $|A| = 10$, agreement threshold $\tau = 0.7$, and debate round limit $M = 5$.

3.4 Comparison Results & Analysis

The experimental results are shown in Table 1. Our first-stage-only method achieves better performance comparable to WebAgent, although StepByStep remains challenging to surpass.

When the second-stage debate mechanism is incorporated, our approach yields substantial improvements over the first-stage-only variant: F1 scores increase by $+3.1$ on TREC-Health and $+8.1$ on HealthFC. This demonstrates that, in cases of uncertain agreement among retrieved articles, the debate mechanism enables more effective reasoning and leads to stronger overall performance.

Compared to StepByStep, our two-stage pipeline achieves higher F1 performance by $+0.8$ on TREC-Health and $+1.4$ on HealthFC. Notably, our method maintains a balance between precision and recall, whereas StepByStep tends to favor high recall at the expense of precision.

Table 2 reports results on the high-agreement subset. High coverage and strong performance in this setting show that the first stage can reliably resolve many claims. However, when evidence is sparse or contradictory, the second-stage debate provides the additional reasoning needed, underscoring its critical role in the framework.

4 Conclusion

We proposed a two-stage framework for health misinformation detection that combines agreement score prediction with multi-agent debate. The first stage leverages weighted agreement scoring to resolve claims directly, while the second stage provides explainable reasoning through debate.

Experiments on three health datasets demonstrate consistent improvements over strong base-

Method	SciFact			TREC-Health			HealthFC		
	P	R	F1	P	R	F1	P	R	F1
WEBAGENT (Tian et al., 2024)	80.1	83.2	80.6	76.2	75.6	75.7	78.0	78.3	78.1
STEPBYSTEP (Vladika et al., 2025)	**86.1**	**89.5**	**87.8**	69.9	**95.1**	80.6	72.6	**91.6**	81.0
OURS (1ST STAGE ONLY)	84.9	86.1	85.5	**83.8**	78.2	78.3	76.9	73.4	74.3
OURS (1ST STAGE + 2ND STAGE)	82.4	85.3	83.1	81.3	81.5	**81.4**	**82.1**	82.7	**82.4**

Table 1: Performance comparison across three datasets (SciFact, TREC-Health, and HealthFC) using macro precision (P), recall (R), and F1 score. Best results are in **bold**.

	SciFact	TREC-Health	HealthFC
Coverage	64.9%	50.1%	58.1%
F1 Score	92.0	88.6	84.0

Table 2: Results on the high-agreement subset. *Coverage (%)* denotes the proportion of claims settled without debate in the first stage, while *F1 Score* reports the score for those claims.

lines, including gains of +0.8 F1 on TREC-Health and +1.4 F1 on HealthFC, with a better balance between precision and recall. These results underscore the value of integrating evidence consistency with structured debate, advancing reliable and explainable health misinformation detection.

Limitations

While our two-stage framework achieves strong performance, it also entails certain limitations. First, as the approach relies on LLMs, the debate judge may still be affected by model biases or occasional hallucinations. Second, the multi-agent design requires multiple API calls, introducing extra computational cost; however, this cost is modest compared to the performance gains. Finally, our current evaluation is limited to binary-labeled datasets. Extending the framework to more nuanced settings, such as incorporating a *Not Enough Information* class, represents a promising direction for future work.

References

Mustafa Abualsaud, Christina Lioma, Maria Maistro, Mark D. Smucker, and Guido Zuccon. 2020. Overview of the trec 2019 decision track.

Brave Software, Inc. Brave search api documentation. `https://api.search.brave.com/res/v1/web/search`. Accessed: 2025-08-21.

Charles L. A. Clarke, Saira Rizvi, Mark D. Smucker, Maria Maistro, and G. Zuccon. 2021. Overview of the trec 2021 health misinformation track. In *Text Retrieval Conference*.

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.

Spencer Hong, Meng Luo, and Xinyi Wan. 2025. Emulate: A multi-agent framework for determining the veracity of atomic claims by emulating human actions. *Preprint*, arXiv:2505.16576.

Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Li Qing. 2025. Removal of hallucination on hallucination: Debate-augmented RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15839–15853, Vienna, Austria. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

OpenAI. 2024. Gpt-4o. `https://openai.com/index/hello-gpt-4o`. Accessed: 2025-08-21.

Alexander Pugachev, Ekaterina Artemova, Alexander Bondarenko, and Pavel Braslavski. 2023. Consumer health question answering using off-the-shelf components. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II*, page 571–579, Berlin, Heidelberg. Springer-Verlag.

Ronit Singal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed fact checking using RAG and few-shot in-context learning with

LLMs. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 91–98, Miami, Florida, USA. Association for Computational Linguistics.

Jacob-Junqi Tian, Hao Yu, Yury Orlovskiy, Tyler Vergho, Mauricio Rivera, Mayank Goel, Zachary Yang, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Web retrieval agents for evidence-based misinformation detection. In *First Conference on Language Modeling*.

Juraj Vladika, Ivana Hacajova, and Florian Matthes. 2025. Step-by-step fact verification system for medical claims with explainable reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 805–816, Albuquerque, New Mexico. Association for Computational Linguistics.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Zero-Shot Cross-Sentential Scientific Relation Extraction via Entity-Guided Summarization

Vani Kanjirangat[1], **Fabio Rinaldi**[1]
[1]SUPSI, IDSIA, Switzerland
{vani.kanjirangat, fabio.rinaldi}@supsi.ch

Abstract

Structured information extraction (IE) from scientific abstracts is increasingly leveraging large language models (LLMs). A crucial step in IE is relation extraction (RE), which becomes challenging when entity relations span sentences. Traditional path-based methods, such as shortest dependency paths, are often unable to handle cross-sentential relations effectively. Although LLMs have been utilized as zero-shot learners for IE tasks, they continue to struggle with capturing long-range dependencies and multi-hop reasoning. In this work, we propose using GPT as a zero-shot entity-guided summarizer to encapsulate cross-sentential context into a single-sentence summary for relation extraction. We perform intrinsic evaluations, comparing our approach against direct zero-shot prompting on biomedical scientific abstracts. On the Chemical-Disease Relation (CDR) dataset, our method achieves a 7-point improvement in overall F-score and 6 points for cross-sentential relations. On the Gene-Disease Association (GDA) dataset, we observe an 8-point gain for inter-sentential relations. These results demonstrate that entity-guided summarization with GPT can enhance zero-shot biomedical RE, supporting more effective structured information extraction from scientific texts.[1].

1 Introduction

In structured information extraction from scientific literature, identifying and extracting entity relations is a key intermediate step, for example, in building knowledge graphs. A typical IE pipeline includes named entity recognition (NER), entity linking/normalization, relation extraction (RE), optional event/fact extraction, and knowledge base population (KBP) (Dagdelen et al., 2024; Jaradeh et al., 2023). With advances in generative language models, zero-shot (ZSL) and few-shot learning (FSL) have become increasingly popular for IE and other NLP tasks (Dagdelen et al., 2024; Hou et al., 2024; Savelka, 2023; Shu et al., 2022; Wu et al., 2025). Parallel research has explored the limitations of zero-shot learning (ZSL) across various domains (Manikandan et al., 2023; Lauscher et al., 2020; Kanjirangat et al., 2024; Al Nazi et al., 2025). GPT-based models (Radford et al., 2019; Liu et al., 2023; Achiam et al., 2023) and open-source models such as Falcon (Almazrouei et al., 2023), Bloom (Le Scao et al., 2023), LLaMA (Touvron et al., 2023), and Mistral (Jahan et al., 2023) have demonstrated strong capabilities in knowledge-intensive tasks, including question answering and summarization. However, their performance in classification tasks can be limited by factors such as domain specificity. For example, they excel in sentiment analysis or intent classification (Wei et al., 2021) but often struggle with clinical or biomedical classification. These limitations are especially pronounced in complex tasks like relation identification and causality detection(Armengol-Estapé et al., 2021; Khondaker et al., 2023; Lai et al., 2023; Yang et al., 2023; Bi et al., 2025; Chen et al., 2025). Considering the above points, this work focuses on the relation extraction task under two key constraints: (i) addressing complex cross-sentential relations and (ii) focusing the task within the biomedical domain. Concerning relation extractions, efforts have been made to leverage LLMs, specifically focusing on improving prompting approaches (Li et al., 2023; Wadhwa et al., 2023; Laskar et al., 2025), which have demonstrated performance upgrades. The potentials and limitations of GPT models in biomedical information extraction have been reported in multiple studies. It has been shown that even though GPT-4 had achieved near state-of-the-art results in few-shot knowledge transfer in open-domain NLP tasks, it underperformed the domain-specific models such as BioBERT (Lee et al., 2020) or SciBERT (Beltagy et al., 2019), which are or-

[1]Experimental codes will be made available

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 22–33
December 23, 2025

ders of magnitude smaller than them (Chen et al., 2024; Moradi et al., 2021; Ateia and Kruschwitz, 2023; Nori et al., 2023; Waisberg et al., 2023). The limitations and capacity of zero-shot LLMs are less explored (Jahan et al., 2023; Shang et al., 2025) in addressing complex cross-sentential relations, even though such relations are plentiful in scientific literature.

Limited work explores generative RE, for instance, El Khettari et al. (2025) used this concept with instruction-tuned LLMs in the microbiome domain. In contrast, Zhang et al. (2025) utilizes entity-pair relation summarizations for triplet fact judgments, whereas the proposed approach focuses primarily on extracting inter-sentential relations and integrating cross-sentential spans of information in an entity-guided summary.

Our core idea is to strategically leverage these generative abilities to enhance zero-shot RE performance, as it remains a valuable strategy for querying LLMs, particularly for non-expert users. In this paper, we formulate two main **research questions:** (i) What are the zero-shot relation extraction capabilities of LLMs (GPT) for cross-sentential RE in the biomedical domain? (ii) How can we simplify and tackle the problem of cross-sentential RE with LLMs' generative capability?

For the current experiment, we used open-sourced GPT-4-0-mini primarily due to its computational efficiency and accessibility, which allowed for extensive experimentation under limited resource constraints, while having core instruction-tuning and generative reasoning capabilities. **RQ1** explores the limitations and potentials of GPT with simple zero-shot prompting in the context of biomedical RE. In **RQ2**, we use LLMs in RE, but not directly as a relation classifier; instead, we explore the *generation capability of LLMs, serving as a summarizer*. In this way, we propose to use GPT's zero-shot capacity to generate an *entity-guided summary that converts cross-sentential relations to intra-sentential relations*. This can also help alleviate the problem of capturing long-range dependencies and complex multi-hop navigation. The current focus is not on maximizing absolute task performance, but instead on better understanding the relative behavior, strengths, and limitations of the approaches under controlled settings.

2 Dataset

We used the BioCreative V *Chemical Disease Relation (CDR)*(Li et al., 2016)[2] and *Gene-Disease Association (GDA)* (Wu et al., 2019) datasets for our experiments. They include abstracts from the scientific biomedical literature. In CDR, we need to identify the binary relations between chemical-induced diseases (CID). The dataset can be considered a good representative of cross-sentential relations, attributed to its complexity and diversity of entity spans, which makes the task challenging. Among the test samples, we extracted 1,800 (negative) and 266 (positive) cross-sentential samples and 748 (positive) and 1,716 (negative) intra-sentential samples. To assess generalizability, we applied the approach to a subset of the GDA dataset (Wu et al., 2019). Since our approach primarily evaluates cross-sentential RE, we specifically selected 1,491 cross-sentential samples (i.e., entity pairs with cross-sentential relations in the given abstract). As cross-sentence and intra-sentence entity pairs can sometimes overlap, following existing works (Christopoulou et al., 2019; Verga et al., 2018; Zhao et al., 2020), we consider cross-sentence subsets to be approximate, rather than strictly disjoint from intra-sentence ones. The details are given in Appendix A.

3 Methods

In this section, we describe the proposed and the baseline approaches used in our controlled experimentation setup.

3.1 Direct Zero-shot Learning

As a baseline, we employ a vanilla zero-shot prompting approach to evaluate GPT's capabilities in biomedical RE. We use a simple prompt template that asks GPT to predict whether the entity pair has a relation, given the input text as the context. In this case, the inputs are the abstracts and the corresponding entity pairs, whose relation needs to be classified. For instance, in the CDR dataset, it asks: *"Does the Given chemical entity induce the given disease or not"*.

3.2 Proposed Approach

In the proposed work, we aim to use GPT's zero-shot generative power as an intermediate step to enhance the relation classification pipeline. The

[2] https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/

Long term hormone therapy for perimenopausal and postmenopausal women.
BACKGROUND: Hormone therapy (HT) is widely used for controlling menopausal symptoms. It has also been used for the management and prevention of cardiovascular disease, osteoporosis and dementia in older women but the evidence supporting its use for these indications is largely observational.
OBJECTIVES: To assess the effect of long-term HT on mortality, heart disease, venous thromboembolism, stroke, transient ischaemic attacks, breast cancer, colorectal cancer, ovarian cancer, endometrial cancer, gallbladder disease, cognitive function, dementia, fractures and quality of life.
SEARCH STRATEGY: We searched the following databases up to November 2004: the Cochrane Menstrual Disorders and Subfertility Group Trials Register, Cochrane Central Register of Controlled Trials (CENTRAL), MEDLINE, EMBASE, Biological Abstracts. Relevant non-indexed journals and conference abstracts were also searched.
SELECTION CRITERIA: Randomised double-blind trials of HT (oestrogens with or without progestogens) versus placebo, taken for at least one year by perimenopausal or postmenopausal women.
DATA COLLECTION AND ANALYSIS: Fifteen RCTs were included. Trials were assessed for quality and two review authors extracted data independently. They calculated risk ratios for dichotomous outcomes and weighted mean differences for continuous outcomes. Clinical heterogeneity precluded meta-analysis for most outcomes. MAIN RESULTS: All the statistically significant results were derived from the two biggest trials. In relatively healthy women, combined continuous HT significantly increased the risk of venous thromboembolism or coronary event (after one year's use), stroke (after 3 years), breast cancer (after 5 years) and gallbladder disease. Long-term oestrogen-only HT also significantly increased the risk of stroke and gallbladder disease. Overall, the only statistically significant benefits of HT were a decreased incidence of fractures and colon cancer with long-term use. Among relatively healthy women over 65 years taking continuous combined HT, there was a statistically significant increase in the incidence of dementia. Among women with cardiovascular disease, long-term use of combined continuous HT significantly increased the risk of venous thromboembolism. No trials focussed specifically on younger women. However, one trial analysed subgroups of 2839 relatively healthy 50 to 59 year-old women taking combined continuous HT and 1637 taking oestrogen-only HT, versus similar-sized placebo groups. The only significantly increased risk reported was for venous thromboembolism in women taking combined continuous HT; their absolute risk remained very low.
AUTHORS' CONCLUSIONS: HT is not indicated for the routine management of chronic disease. We need more evidence on the safety of HT for menopausal symptom control, though short-term use appears to be relatively safe for healthy younger women.

('progestogens', 'stroke')	The use of progestogens was associated with a significant increase in the risk of stroke in women taking hormone therapy.
('progestogens', 'dementia')	Among relatively healthy women over 65 years, the long-term use of combined continuous hormone therapy with progestogens significantly increased the risk of dementia.
(('oestrogen or oestrogens', 'breast cancer')	long-term use of oestrogen or oestrogens was linked to a significant increase in the risk of breast cancer in women
('oestrogen or oestrogens', 'colon cancer')	long-term use of oestrogen or oestrogens was associated with a decreased incidence of colon cancer in women

Figure 1: A CDR abstract with the chemical entities highlighted in yellow and disease entities in blue. The right side shows the (chemical, disease) entity pairs and the corresponding summaries produced by the zero-shot entity-guided summarizer (GPT).

existing models struggle to capture cross-sentential relations for various reasons: The relations that define the entities are not contained within a single sentence. In this case, multi-hop reasoning approaches are needed, which the model may not inherently possess. Secondly, the semantic encodings may not capture sufficient context for identifying such relations due to the presence of long-range dependencies. Thirdly, some sentences or contexts can even act as noise to the model due to the span of entities in multiple sentences. Further, the general path-based approaches used in relation extractions, such as shortest dependency path (SDP) methods, only directly apply to intra-sentential relations.

In the proposed approach, we deviate from the general approach of path-based or multi-hop reasoning (combined with or without encoder/decoder variants) by enabling LLMs' generative capabilities to adapt cross-sentential sentences to intra-sentential ways. Specifically, we want to *convert cross-sentential sentences to a single-sentence entity-guided summary*. Given the impressive results of GPT in generation tasks[3], we used *GPT as a zero-shot entity-guided single-sentence summarizer*. For instance, consider the abstract from the CDR dataset in Figure 1 (enlarged figures are in Appendix B) with the entity pairs under consideration marked. Here *estrogens* and *progestogens* are the chemical entities, and *{dementia, breast cancer, colon cancer, stroke}* are the disease entities [4]. It can be observed that the relations are cross-sentential, and entities can span across multiple sentences. The entity pairs and the corresponding zero-shot summary generated by GPT-4 are shown in Figure 1.

Dataset	Direct ZSL	Proposed ZSL
CDR (Cross)	0.35	**0.41** (+0.07) ↑
GDA (Cross)	0.49	**0.57** (+0.08) ↑

Table 1: Performance comparison of Direct ZSL and Proposed ZSL on cross-sentential biomedical RE (F-scores).

Considering the entity pair *(progestogens, stroke)*, the relation is not apparent, and proper reasoning is required to classify the relation. Firstly, the model should consider the sentence - *"double-blind trials of HT (oestrogens with or without progestogens)"*, which is the only mention of progestogens in the abstract, and should deduce (entity normalization) that *HT* refers to *"Hormone Therapy"*. Further, it should be related to the sentence - *"In relatively healthy women, combined continuous HT significantly increased the risk of venous thromboembolism or coronary event (after one year's use), stroke (after 3 years), breast cancer (after 5 years) and gallbladder disease."* for capturing the actual relation.

The proposed approach initially uses a prompt to generate a zero-shot entity-guided summary for each cross-sentential entity pair (Figure 3). For instance, for the previous example, we generated a summary that directly conveys the cross-sentential relationship (*"The use of progestogens was associated with a significant increase in the risk of stroke in women taking hormone therapy."*). Similarly, a negative relation is indicated for the entity pair *('estrogen or estrogens,' 'colon cancer')*. These generated summaries were used as inputs for the second step, where the actual relation classifica-

[3]https://github.com/openai

[4]Only the entities required for illustration are highlighted. There are more entity relations in this abstract.

ZSL	Type	F-score	Recall	Precision
Direct ZSL	Overall	0.49	0.83	0.34
	Intra	0.55	0.91	0.40
	Inter/cross	0.35	0.66	0.24
Proposed ZSL	Overall	**0.56**	0.80	0.43
	Intra	**0.65**	0.81	0.54
	Inter/cross	**0.41**	0.75	0.28

Table 2: Comparing proposed ZSL with direct ZSL in CDR dataset

Model	Type	F-score	Recall	Precision
BioBERT_Proposed	Overall	**0.57**	0.56	0.58
	Intra	**0.64**	0.62	0.65
	Inter/cross	**0.41**	0.40	0.41
BioBERT_Baseline	Overall	0.25	0.17	0.44
	Intra	0.36	0.29	0.48
	Inter/cross	0.24	0.21	0.28

Table 3: Fine-tuned Encoder-only Model Performance on CDR dataset

tion is performed (Figure 4). Note that the intra-sentences were directly extracted from the abstract by considering sentences that mention both entities.

4 Results & Comparisons

In Table 1, we report the zero-shot results on the cross-sentential RE in the CDR and GDA datasets obtained with the baseline GPT model (Direct ZSL) and compare them with those of the proposed approach (Proposed ZSL). In the baseline approach, the input is the abstract directly, while, for the proposed approach, it is the entity-guided summary generated by GPT for cross-sentential relations. For intra-sentences, we use the sentences where both entity mentions are present. A 7-point F-score improvement can be observed in the CDR dataset, while in GDA, an 8-point increase is reported.

To analyze the overall improvements, we conducted similar experiments using intra-sentential samples from the CDR dataset. From Table 2, it can be observed that the proposed ZSL approach presents a 7-point improvement compared to baseline or direct ZSL, in terms of overall F-scores. In terms of recall and precision, it can be observed that GPT generally prioritizes recall, which is understandable given its general-purpose nature. In terms of intra-sentential RE, a 10-point improvement is noted. These improvements indicate the scope of utilizing the inherent generative capacity of these LLMs for the downstream tasks, specifically for zero-shot.

Furthermore, we also compare the performance of the proposed entity-guided summaries when used as inputs to fine-tuned encoder-only models. In this case, we fine-tune a BioBERT model using the generated summaries (BioBERT_Proposed) and compare it with the one fine-tuned directly on the abstracts (BioBERT_Baseline). This is the same as the input to the Direct ZSL and Proposed ZSL approaches, which are reported in Table 2. We fine-tune BioBERT with sentence pair classification - where the <*text,entity_pair*> is the input. As discussed, for the baseline, this text is the abstract, and for the proposed, it will be the entity-guided summary. From Table 3, it can be observed that the model appears to capture more accurate information when using the proposed summaries as input. With the cross-sentential RE, BioBERT_Proposed presents significantly better results, with an improvement of almost 17 points over the baseline counterpart. Based on the experimental results reported in Table 3, it is evident that the proposed entity-guided summaries already improve the performance of the simple BioBERT models. More details of experimental settings are in Appendix D. These intrinsic evaluations under controlled settings show that the proposed approach helps the model capture relations more accurately, guiding the LLM to make better predictions. Our analysis suggests that summarizing cross-sentential information into a single sentence enables simpler, more effective representations, which in turn support more accurate scientific information extraction.

5 Conclusions

In the proposed work, we aim to evaluate the zero-shot capabilities of GPT in biomedical relation extraction, with a focus on cross-sentential relations. We utilized the chemical-induced disease and gene-disease association datasets, which comprise complex inter-sentential spans of entity relations, as a representative dataset. We observed that GPT, in its zero-shot capacity, has considerable scope for improvement in capturing these relations. A novel approach is proposed to utilize the generative capabilities of GPT as an intermediate step in the relation extraction pipeline by using it as a zero-shot entity-guided summarizer. This is used to encapsulate information on cross-sentential relations and convert these relations into intra-sentential ones. We observed a good performance improvement compared to baseline zero-shot performances. We believe that the proposed direction has considerable potential for exploration, where, instead of using GPT directly as a downstream classifier, it would be more reasonable to exploit its inherent generative ability by mapping it to intermediate steps in a logical manner.

6 Related Works

In the field of structured IE from scientific literature, recently LLMs are used widely (Dagdelen et al., 2024; Li et al., 2024; Garcia et al., 2024). The approaches range from simple feature-based extractions to transformer-based to current Large Language Model (LLM) based approaches. A unified schema representation was proposed in Li et al. (2023) to encourage LLMs to follow schemas, learn easily, and extract structured knowledge accurately.

In the existing literature, a wide range of approaches and studies consider the problem of biomedical information extractions (Sciannameo et al., 2024; Fornasiere et al., 2024; Reichenpfader et al., 2024). In the task of relation extractions, Zhang et al. (2018) proposed a hybrid model that uses Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) with Shortest Dependency Features (SDP). An SDP-based feature extraction for candidate cross-sentential sample extractions, coupled with BioBERT models, was presented by Kanjirangat and Rinaldi (2021). The use of biomedical ontologies to enhance neural network knowledge is another direction (Sousa et al., 2020; Sänger and Leser, 2025; Liu et al., 2025). Another promising direction was to explore graph-based models, graph LSTMs (Peng et al., 2017), graph kernels (Panyam et al., 2018), graph CNN with multi-head attentions (Zhao et al., 2021), and multi-view GNNs (Al-Sabri et al., 2022). BERT and its variants have been widely used for biomedical RE tasks (Thillaisundaram and Togia, 2019; Bhasuran, 2022; Su and Vijay-Shanker, 2020, 2022). However, the complex task of cross-sentential RE necessitated more sophisticated approaches. For instance, Wei and Li (2022) proposed a sequence-aware graph model with adaptive margin loss, while Zhu et al. (2024) leveraged dependency and constituency information using Tree-LSTM, GNN, and BERT models.

Generative models are now being explored in biomedical RE, where their performance has been reported to vary based on the complexity of the dataset and task at hand Zhang et al. (2024); Asada and Fukuda (2024). Some of the findings reported good performances, but were limited to intra-sentential relations. A few studies (e.g., (El Khettari et al., 2025)) have explored generative approaches to relation extraction (RE) using instruction-tuned large language models (LLMs). In contrast, (Zhang et al., 2025) focuses on leveraging entity-pair relation summarization for triplet fact evaluation. In our proposed approach, we primarily address inter-sentential relation extraction, emphasizing the integration of cross-sentential contextual spans within an entity-guided summarization framework.

In the proposed work, we focus on exploring the zero-shot capability of GPT in cross-sentential RE. Moving a step further, we propose an approach to possibly utilize the generative capability of GPT in the RE pipeline, which is the inherent potential of generative models. This deviates from the general trend of using these generative models directly for classifications, a use case that does not fully align with their intrinsic generative nature.

7 Limitations

The proposed approach could propagate errors from the summarization module, as we introduce it as an intermediate path in the relation extraction pipeline. An explicit evaluation of the zero-shot summarization component is challenging, which limits the understanding of the summarizer's performance. Currently, the experiments are done only on the CDR and GDA biomedical datasets. These could be considered as representative datasets for complex cross-sentential relations; however, a proper generalization of the proposed approach has to be verified by extending the experiments to other datasets with cross-sentential relations, Chemical Reaction (CHR) dataset (Peng et al., 2017), or general-purpose datasets, such as DocRed (Yao et al., 2019), Codred (Yao et al., 2021), CrossRE (Bassignana and Plank, 2022), etc. Furthermore, GPT responses can be limited by multiple factors, including sensitivity to prompts, context, post-processing, controversies, ambiguities, efficiency, and costs (Kocoń et al., 2023). In general, the low performance of GPT models can be attributed to several factors, including the lack of domain-specific training, entity disambiguation issues in biomedical data, and the need for multi-hop reasoning to address inter-sentential relations. While refining prompts can mitigate some issues, prompt sensitivity remains a challenge. Soft prompting techniques offer a potential solution to improve robustness, though naive zero-shot prompting still holds value for user-centric applications across various domains. We also have the scope of experimenting with different LLMs (open-sourced). Finally, considering the state-of-the-art approaches, we still have considerable scope for improvement, even though our approach focuses on zero-shot capability.

Acknowledgments

The work of Vani Kanjirangat has been supported by HASLER STIFTUNG through the project "RING-P- Relation Extractions at Inter-sentential N-ary levels using Graph Prompting", project No. 24002.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Zabir Al Nazi, Md Rajib Hossain, and Faisal Al Mamun. 2025. Evaluation of open and closed-source llms for low-resource language with zero-shot, few-shot, and chain-of-thought prompting. *Natural Language Processing Journal*, 10:100124.

Raeed Al-Sabri, Jianliang Gao, Jiamin Chen, Babatounde Moctard Oloulade, and Tengfei Lyu. 2022. Multi-view graph neural architecture search for biomedical entity and relation extraction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1221–1233.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, and 1 others. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2021. On the multilingual capabilities of very large-scale english language models. *arXiv preprint arXiv:2108.13349*.

Masaki Asada and Ken Fukuda. 2024. Enhancing relation extraction from biomedical texts by large language models. In *International Conference on Human-Computer Interaction*, pages 3–14. Springer.

Samy Ateia and Udo Kruschwitz. 2023. Is chatgpt a biomedical expert?–exploring the zero-shot performance of current gpt models in biomedical tasks. *arXiv preprint arXiv:2306.16108*.

Elisa Bassignana and Barbara Plank. 2022. Crossre: A cross-domain dataset for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3592–3604.

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.

Balu Bhasuran. 2022. Biobert and similar approaches for relation extraction. In *Biomedical Text Mining*, pages 221–235. Springer.

Jing Bi, Ziqi Wang, Haitao Yuan, Xiankun Shi, Ziyue Wang, Jia Zhang, MengChu Zhou, and Rajkumar Buyya. 2025. Large ai models and their applications: Classification, limitations, and potential solutions. *Software: Practice and Experience*, 55(6):1003–1017.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, and 1 others. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature communications*, 16(1):3280.

Shan Chen, Yingya Li, Sheng Lu, Hoang Van, Hugo JWL Aerts, Guergana K Savova, and Danielle S Bitterman. 2024. Evaluating the chatgpt family of models for biomedical reasoning and classification. *Journal of the American Medical Informatics Association*, 31(4):940–948.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. 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 4925–4936.

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.

Oumaima El Khettari, Solen Quiniou, and Samuel Chaffron. 2025. Summarization for generative relation extraction in the microbiome domain. In *Actes de l'atelier Traitement du langage médical à l'époque des LLMs 2025 (MLP-LLM)*, pages 68–82.

Raffaello Fornasiere, Nicolò Brunello, Vincenzo Scotti, Mark James Carman, and 1 others. 2024. Medical information extraction with large language models. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 1–10. Association for Computational Linguistics.

Gabriel Lino Garcia, Joao Renato Ribeiro Manesco, Pedro Henrique Paiola, Lucas Miranda, Maria Paola de Salvo, and Joao Paulo Papa. 2024. A review on scientific knowledge extraction using large language models in biomedical sciences. *arXiv preprint arXiv:2412.03531*.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pages 364–381. Springer.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. Evaluation of chatgpt on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers. *arXiv preprint arXiv:2306.04504*.

Mohamad Yaser Jaradeh, Kuldeep Singh, Markus Stocker, Andreas Both, and Sören Auer. 2023. Information extraction pipelines for knowledge graphs. *Knowledge and Information Systems*, 65(5):1989–2016.

Vani Kanjirangat, Alessandro Antonucci, and Marco Zaalon. 2024. On the limitations of zero-shot classification of causal relations by llms (work in progress). *Proceedings http://ceur-ws. org ISSN*, 1613:0073.

Vani Kanjirangat and Fabio Rinaldi. 2021. Enhancing biomedical relation extraction with transformer models using shortest dependency path features and triplet information. *Journal of biomedical informatics*, 122:103893.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: a comprehensive evaluation of chatgpt on arabic nlp. *arXiv preprint arXiv:2305.14976*.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, and 1 others. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Md Tahmid Rahman Laskar, Israt Jahan, Elham Dolatabadi, Chun Peng, Enamul Hoque, and Jimmy Huang. 2025. Improving automatic evaluation of large language models (llms) in biomedical relation extraction via llms-as-the-judge. *arXiv preprint arXiv:2506.00777*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and 1 others. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023. Revisiting large language models as zero-shot relation extractors. *arXiv preprint arXiv:2310.05028*.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Zixuan Li, Yutao Zeng, Yuxin Zuo, Weicheng Ren, Wenxuan Liu, Miao Su, Yucan Guo, Yantao Liu, Xiang Li, Zhilei Hu, and 1 others. 2024. Knowcoder: Coding structured knowledge into llms for universal information extraction. *arXiv preprint arXiv:2403.07969*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. Gpt understands, too. *AI Open*.

Xiaoyong Liu, Xin Qin, Chunlin Xu, and Huihui Li. 2025. A knowledge-enhanced model with syntactic-aware attentive graph convolutional network for biomedical entity and relation extraction. *International Journal of Machine Learning and Cybernetics*, 16(1):583–598.

Hariharan Manikandan, Yiding Jiang, and J Zico Kolter. 2023. Language models are weak learners. *Advances in Neural Information Processing Systems*, 36:50907–50931.

Milad Moradi, Kathrin Blagec, Florian Haberl, and Matthias Samwald. 2021. Gpt-3 models are poor few-shot learners in the biomedical domain. *arXiv preprint arXiv:2109.02555*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Nagesh C Panyam, Karin Verspoor, Trevor Cohn, and Kotagiri Ramamohanarao. 2018. Exploiting graph kernels for high performance biomedical relation extraction. *Journal of biomedical semantics*, 9(1):7.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Daniel Reichenpfader, Henning Müller, and Kerstin Denecke. 2024. A scoping review of large language model based approaches for information extraction from radiology reports. *npj Digital Medicine*, 7(1):222.

Mario Sänger and Ulf Leser. 2025. Knowledge-augmented pre-trained language models for biomedical relation extraction. *arXiv preprint arXiv:2505.00814*.

Jaromir Savelka. 2023. Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 447–451.

Veronica Sciannameo, Daniele Jahier Pagliari, Sara Urru, Piercesare Grimaldi, Honoria Ocagli, Sara Ahsani-Nasab, Rosanna Irene Comoretto, Dario Gregori, and Paola Berchialla. 2024. Information extraction from medical case reports using openai instructgpt. *Computer methods and programs in biomedicine*, 255:108326.

Yufei Shang, Yanrong Guo, Shijie Hao, and Richang Hong. 2025. Biomedical relation extraction via adaptive document-relation cross-mapping and concept unique identifier. *arXiv preprint arXiv:2501.05155*.

Lei Shu, Hu Xu, Bing Liu, and Jiahua Chen. 2022. Zero-shot aspect-based sentiment analysis. *arXiv preprint arXiv:2202.01924*.

Diana Sousa, Andre Lamurias, and Francisco M Couto. 2020. Using neural networks for relation extraction from biomedical literature. In *Artificial Neural Networks*, pages 289–305. Springer.

Peng Su and K Vijay-Shanker. 2020. Investigation of bert model on biomedical relation extraction based on revised fine-tuning mechanism. In *2020 ieee international conference on bioinformatics and biomedicine (bibm)*, pages 2522–2529. IEEE.

Peng Su and K Vijay-Shanker. 2022. Investigation of improving the pre-training and fine-tuning of bert model for biomedical relation extraction. *BMC bioinformatics*, 23(1):120.

Ashok Thillaisundaram and Theodosia Togia. 2019. Biomedical relation extraction with pre-trained language representations and minimal task-specific architecture. *arXiv preprint arXiv:1909.12411*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. 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 872–884, New Orleans, Louisiana. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589.

Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Sharif Amit Kamran, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. 2023. Gpt-4: a new era of artificial intelligence in medicine. *Irish Journal of Medical Science (1971-)*, 192(6):3197–3200.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Ying Wei and Qi Li. 2022. Sagdre: Sequence-aware graph-based document-level relation extraction with adaptive margin loss. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2000–2008.

Ling-I Wu, Yuxin Su, and Guoqiang Li. 2025. Zero-shot construction of chinese medical knowledge graph with gpt-3.5-turbo and gpt-4. *ACM Transactions on Management Information Systems*, 16(2):1–17.

Ye Wu, Ruibang Luo, Henry CM Leung, Hing-Fung Ting, and Tak-Wah Lam. 2019. Renet: A deep learning approach for extracting gene-disease associations from literature. In *International Conference on Research in Computational Molecular Biology*, pages 272–284. Springer.

Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. 2023. Chatgpt is not enough: Enhancing large language models with knowledge graphs for fact-aware language modeling. *arXiv preprint arXiv:2306.11489*.

Yuan Yao, Jiaju Du, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou, and Maosong Sun. 2021. Codred: A cross-document relation extraction dataset for acquiring knowledge in the wild. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4452–4472.

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.

Fu Zhang, Hongsen Yu, Jingwei Cheng, and Huangming Xu. 2025. Entity pair-guided relation summarization and retrieval in LLMs for document-level relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4022–4037, Albuquerque, New Mexico. Association for Computational Linguistics.

Jeffrey Zhang, Maxwell Wibert, Huixue Zhou, Xueqing Peng, Qingyu Chen, Vipina K Keloth, Yan Hu, Rui Zhang, Hua Xu, and Kalpana Raja. 2024. A study of biomedical relation extraction using gpt models. *AMIA Summits on Translational Science Proceedings*, 2024:391.

Yijia Zhang, Hongfei Lin, Zhihao Yang, Jian Wang, Shaowu Zhang, Yuanyuan Sun, and Liang Yang. 2018. A hybrid model based on neural networks for biomedical relation extraction. *Journal of biomedical informatics*, 81:83–92.

D. Zhao, J. Wang, Y. Zhang, X. Wang, H. Lin, and Z. Yang. 2020. Incorporating representation learning and multihead attention to improve biomedical cross-sentence n-ary relation extraction. *BMC Bioinformatics*, 21(1):312.

Di Zhao, Jian Wang, Hongfei Lin, Xin Wang, Zhihao Yang, and Yijia Zhang. 2021. Biomedical cross-sentence relation extraction via multihead attention and graph convolutional networks. *Applied Soft Computing*, 104:107230.

Xudong Zhu, Zhao Kang, and Bei Hui. 2024. Fcds: Fusing constituency and dependency syntax into document-level relation extraction. *arXiv preprint arXiv:2403.01886*.

A Dataset Details

The CDR dataset annotation identifies entities that hold a relation (class 1/positive), and all remaining entity pairs fall into the negative category/class 0. The CID relations can be either intra-sentential or cross-sentential. There are no mention-level annotations in the CDR dataset. Hence, we can use the entire abstract as the context or deduce methodologies to extract the context that can convey possible relations (based on the presence of entities).

The Gene–Disease Associations (GDA) dataset is a large-scale biomedical corpus constructed from MEDLINE abstracts using distant supervision. In line with Christopoulou et al. (2019), we partition the data into 23,353 documents for training and 5,839 documents for development. The task is formulated as a binary classification problem, where the goal is to determine whether a given gene and disease entity pair is associated or not. A notable characteristic of the dataset is that many associations span across multiple sentences, which makes it particularly suitable for assessing methods that aim to capture long-range dependencies and inter-sentential relations.

B Methods

The enlarged examples for CDR abstracts and the entity guided summaries are shown in Figures 2a and 2b.

C Prompt Templates

The prompt templates for vanilla and the proposed approaches are given in Figures 3 and 4.

D Experiments

We used GPT4-o-mini [5] for our experiments (Approximately 150 USD was spent). The experiments were conducted on an HPC cluster with 1 GPU (NVIDIA A100 80GB PCI). For BERT-based experiments, we used BioBERT v1.1 (+ PubMed 1M), which refers to the BioBERT model trained on PubMed for 1M steps as the pre-trained model. The experiments were done using PyTorch Hugging-Face implementations [6] by fine-tuning the model on the respective datasets. The model is fine-tuned for 10 epochs, using the Adam optimizer and a learning rate of 2e-5 on the training data.

[5] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

[6] https://github.com/huggingface/transformers

Long term hormone therapy for perimenopausal and postmenopausal women.
BACKGROUND: Hormone therapy (HT) is widely used for controlling menopausal symptoms. It has also been used for the management and prevention of cardiovascular disease, osteoporosis and dementia in older women but the evidence supporting its use for these indications is largely observational.
OBJECTIVES: To assess the effect of long-term HT on mortality, heart disease, venous thromboembolism, stroke, transient ischaemic attacks, breast cancer, colorectal cancer, ovarian cancer, endometrial cancer, gallbladder disease, cognitive function, dementia, fractures and quality of life.
SEARCH STRATEGY: We searched the following databases up to November 2004: the Cochrane Menstrual Disorders and Subfertility Group Trials Register, Cochrane Central Register of Controlled Trials (CENTRAL), MEDLINE, EMBASE, Biological Abstracts. Relevant non-indexed journals and conference abstracts were also searched.
SELECTION CRITERIA: Randomised double-blind trials of HT (oestrogens with or without progestogens) versus placebo, taken for at least one year by perimenopausal or postmenopausal women.
DATA COLLECTION AND ANALYSIS: Fifteen RCTs were included. Trials were assessed for quality and two review authors extracted data independently. They calculated risk ratios for dichotomous outcomes and weighted mean differences for continuous outcomes. Clinical heterogeneity precluded meta-analysis for most outcomes. MAIN RESULTS: All the statistically significant results were derived from the two biggest trials. In relatively healthy women, combined continuous HT significantly increased the risk of venous thromboembolism or coronary event (after one year's use), stroke (after 3 years), breast cancer (after 5 years) and gallbladder disease. Long-term oestrogen-only HT also significantly increased the risk of stroke and gallbladder disease. Overall, the only statistically significant benefits of HT were a decreased incidence of fractures and colon cancer with long-term use. Among relatively healthy women over 65 years taking continuous combined HT, there was a statistically significant increase in the incidence of dementia. Among women with cardiovascular disease, long-term use of combined continuous HT significantly increased the risk of venous thromboembolism. No trials focussed specifically on younger women. However, one trial analysed subgroups of 2839 relatively healthy 50 to 59 year-old women taking combined continuous HT and 1637 taking oestrogen-only HT, versus similar-sized placebo groups. The only significantly increased risk reported was for venous thromboembolism in women taking combined continuous HT; their absolute risk remained very low.
AUTHORS' CONCLUSIONS: HT is not indicated for the routine management of chronic disease. We need more evidence on the safety of HT for menopausal symptom control, though short-term use appears to be relatively safe for healthy younger women.

(a) A CDR abstract with chemical entities highlighted in yellow and disease entities in blue.

('progestogens', 'stroke’)	The use of progestogens was associated with a significant increase in the risk of stroke in women taking hormone therapy.
('progestogens', 'dementia’)	Among relatively healthy women over 65 years, the long-term use of combined continuous hormone therapy with progestogens significantly increased the risk of dementia.
(('oestrogen or oestrogens', 'breast cancer’)	long-term use of oestrogen or oestrogens was linked to a significant increase in the risk of breast cancer in women
('oestrogen or oestrogens', 'colon cancer')	long-term use of oestrogen or oestrogens was associated with a decreased incidence of colon cancer in women

(b) (chemical, disease) entity pairs and the corresponding summaries produced by the proposed zero-shot entity-guided summarizer (GPT).

Figure 2: Illustration of the proposed approach: (a) a CDR abstract with highlighted entities; (b) entity pairs and generated summaries.

```
system_msg="You are a helpful summarizer who performs an entity-guided summarization
based on given entity pairs."

"Based on the given text and entity pair, perform an entity-guided single-sentence
summarization of the text.
Give focus on the terms or keywords that can distinguish whether
the given entities can have a relation or not?.
The output should be a single sentence with the entity mentions in it."""

instructions_msg="You are a helpful summarization assistant. You will be provided
with the text and a (chemical,disease) entity pair.

 Text:<Text>{text}</Text>
 Entity_pair:<Text>{ent_pair}</Text>

Provide the final summary within the tags <summary> </summary>."

 '''
```

Figure 3: A zero-shot prompt-template for an Entity-Guided Summarizer(The prompts will vary slightly based on the experimental datasets. This prompt is tailored for the CDR dataset).

```
system_msg = "You are a helpful medical assistant who tells whether a given chemical

induce a given disease or not."

instructions_msg= You will be provided with the text and a list of chemical and
disease entities.

            Text: <Text>{text}</Text>
            Chemical_list:{chem}
            Disease_list:{dis}

For each pair of (chemical, disease), predict whether the chemical induce the disease

or not?.
You should predict 1 if the chemical induce the disease and 0 if not.
Your response should be only based on the given text.

Provide all your final answers within the tags <Answer> </Answer> with entity pairs
expressed as a tuple with its corresponding prediction."
```

Figure 4: A zero-shot prompt-template for a Relation Classification (The prompts will vary slightly based on the experimental datasets. This prompt is tailored for the CDR dataset).

Finding the Paper Behind the Data: Automatic Identification of Research Articles related to Data Publications

Barbara McGillivray[1]
King's College London
barbara.mcgillivray@kcl.ac.uk

Kaveh Aryan[1]
King's College London
kaveh.aryan@kcl.ac.uk

Viola Harperath
King's College London
viola.harperath@kcl.ac.uk

Marton Ribary
Royal Holloway, University of London
marton.ribary@rhul.ac.uk

Mandy Wigdorowitz
University of Alabama
University of Cambridge
University of Johannesburg
mwigdorowitz@ua.edu

[1]*Joint first authors*

Abstract

Data papers are scholarly publications that describe datasets in detail, including their structure, collection methods, and potential for reuse, typically without presenting new analyses. As data sharing becomes increasingly central to research workflows, linking data papers to relevant research papers is essential for improving transparency, reproducibility, and scholarly credit. However, these links are rarely made explicit in metadata and are often difficult to identify manually at scale. In this study, we present a comprehensive approach to automating the linking process using natural language processing (NLP) techniques.

We evaluate both set-based and vector-based methods, including Jaccard similarity, TF-IDF, SBERT, and reranking with large language models. Our experiments on a curated benchmark dataset reveal that no single method consistently outperforms others across all metrics, in line with the multifaceted nature of the task. Set-based methods using frequent words (N=50) achieve the highest top-10% accuracy, closely followed by TF-IDF, which also leads in MRR and top-1% and top-5% accuracy. SBERT-based reranking with LLMs yields the best results in top-N accuracy. This dispersion suggests that different approaches capture complementary aspects of similarity (lexical, semantic, and contextual), showing the value of hybrid strategies for robust matching between data papers and research articles. For several methods, we find no statistically significant difference between using abstracts and full texts, suggesting that abstracts may be sufficient for effective matching. Our findings demonstrate the feasibility of scalable, automated linking between data papers and research articles, enabling more accurate bibliometric analyses, improved tracking of data reuse, and fairer credit assignment for data sharing. This contributes to a more transparent, interconnected, and accessible research ecosystem.

1 Introduction

Data sharing and reuse have become increasingly central to research practices, motivating the development of mechanisms to manage, disseminate, and cite datasets effectively. One response has been the emergence of *data papers*, scholarly publications dedicated to describing datasets in detail, including their structure, provenance, and potential applications (Jiao et al., 2023). Unlike traditional research papers, data papers typically do not present new analyses, but instead contain the context of the creation of the dataset being described, the method for its creation, a description of the dataset itself, a measure of its quality and an explanation of its reuse potential (Reymonet, 2017; Kembellec, Gérald and Le Deuff, Olivier, 2022; Li and Jiao, 2022; Liu, 2022). Previous work has found a relatively high amount of variation in the content and structure of data papers (Li et al., 2020; Jihyun, 2020; Li and Chen, 2018). Data papers have been shown to make datasets more discoverable, citable, and reusable across disciplines (Kosmopoulos and Schöpfel, 2024). Although data papers are sometimes perceived as a recent innovation, their development has been gradual: specialized data journals such as the *Journal of Chemical Engineering Data* were established as early as 1956, while more concerted growth in data journals occurred in the past couple of decades (Candela et al., 2015; Walters, 2020). These journals incentivize open sharing according to best practices, offering authors recognition, citation opportunities, and enhanced reuse potential for their resources.

Following a pyramid model of data-driven research projects (McGillivray et al., 2022), the base comprises project repositories containing scripts, notes, and raw files; the next layer is the structured dataset deposited in a public repository; the third layer is the data paper itself, which documents, contextualizes, and links to the dataset; and the

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 34–43

apex is the research paper, which interprets and analyzes the data. This pyramid makes explicit that the research paper represents only one possible interpretation of the underlying data, while the structured datasets and data papers facilitate transparency, reproducibility, and alternative analyses.

While this model highlights the continuum from raw data to interpretation, the links between its layers are often implicit or missing in metadata. In this study, we focus on automatically reconstructing one of the most critical links: the pairing of a data paper with a related research article. By "related", we mean that the research article is substantively connected to the dataset described in the data paper, for example, because the article is authored by the same team, builds on the same project, or cites the dataset. One such example is a research paper performing semantic profiling of legal language and cluster analysis on Justinian's *Digest*, a historical sourcebook of Roman law compiled under the order of Emperor Justinian in 533 CE (Ribary and McGillivray, 2020). This analysis is based on a relational database of the (mostly) Latin text of the *Digest* created by the same author and reported in a data paper in the same year (Ribary, 2020). By explicitly identifying these pairs of research and data papers, we provide the foundation for more complete reconstructions of dataset–paper–article triangles which is essential for critically assessing the impact of data sharing and enabling reuse.

2 Previous work

Despite their importance, links between data papers and the research papers that use the associated datasets are often implicit or missing from bibliographic metadata. Manual curation of these links is labour-intensive and difficult to scale given diverse data-sharing practices and the growing volume of publications. Previous work has addressed aspects of this problem from both bibliometric and computational perspectives. McGillivray et al. (2022) proposed simple heuristic rules for identifying meaningful links between data and research outputs in a manual fashion which were also followed to create the gold standard ground-truth dataset for the present study as reported below in Section 3.1. Ekman et al. (2025) conducted a qualitative analysis of the narrative practices in data papers. Li and Jiao (2021) analysed the rhetorical moves within abstracts of data papers published in the journals *Data in Brief* and *Scientific Data*, including the research article to which the dataset is connected. They found that the related research articles are only mentioned in *Data in Brief* abstracts, but this use has decreased over time, while the description of the data, as well as the introduction and method are among the most frequently used rhetorical moves. Kai et al. (2025) calculate TF-IDF to extract keywords that are distinctive of data papers in relation to their citing research papers using a sample of 10 papers, finding that many of the keywords that are characteristic of data papers did not appear in the abstracts, pointing to the importance of analysing the full texts to gain a better picture of these relations.

No previous study has proposed a method for automatically identifying research articles connected to data papers specifically. Instead, there has been growing research on linking datasets to research papers using a combination of Named Entity detection and disambiguation (Heddes et al., 2021), matching through textual embeddings (Färber and Leisinger, 2021) and large language models with retrieval (Datta et al., 2025).

In this work, we present the results of a series of experiments on fully automatic approaches to link English-language research papers and data papers, and thereby reconstruct the pairs connecting data papers and research articles. We systematically evaluate a spectrum of methods, ranging from simple keyword-based text mining to large language model (LLM)-based approaches. Evaluation on a curated gold standard of research–data paper pairs using metrics such as Mean Reciprocal Rank (MRR) and accuracy shows that our approach robustly identifies links, supporting reproducibility, data reuse, and more comprehensive measurement of scholarly impact.

Our task is related to citation recommendation, dataset discovery, and scientific document linking. While existing research, including models like SciLinkBERT (Yu et al., 2025) and other citation-based approaches (Bouziani et al., 2024), effectively utilizes cross-document relationships to enhance tasks like relation extraction and summarization, their focus remains on general citation networks. However, these approaches typically model broad citation structures rather than the specific, functional connection between a data paper and a related research article. This distinction is important because relatedness here is not captured simply by citation counts or co-occurrence, but by a substantive link that situates the dataset within ongoing research workflows. We contribute to this area by targeting a highly specific and often overlooked functional relationship: the link between a research article and its related data paper. This focus is critical for accurately tracking data reuse and measuring scholarly impact, and requires a dedicated, hybrid approach that goes beyond standard citation analysis. Our work has a number of potential practical applications, including enriching repository and publisher metadata with explicit dataset–article links, enriching dataset discovery tools that help researchers find analyses associated with published data, and enabling open science policy compliance by making data use and credit more transparent.

3 Methods

Figure 1 provides an overview of our data collection and processing pipeline, explained in detail in this section. The code is available on `https://github.com/BarbaraMcG/golden-triangle/tree/main/NLP%20Paper`.

3.1 Curated dataset

To identify a comprehensive list of data journals, we started with those compiled by Candela et al. (2015). From this list, we added data journals that were not included in the original list and selected only data journals that publish primarily data papers. The final list consisted of 11 data journals actively publishing in 2022 (see Table 1 for an overview). These data journals could also be found in OpenAlex by their journal name.

McGillivray et al. (2022) present a manually curated dataset containing 107 pairs of data papers and datasets. The dataset included a subset of 38 triangles where the pair of data paper and dataset could be linked up with an associated research paper. These links were curated from two sources: the *Journal of Open Humanities Data* (JOHD) and the *Research Data Journal for the Humanities and Social Sciences* (RDJ). Each of the pairs were manually validated to ensure that the research paper substantially builds upon the dataset described in the corresponding data paper.

The manual curation process was developed from the heuristic rules established in McGillivray et al. (2022). We linked a research paper to a data paper if:

1. at least one of the following four conditions was satisfied:
    (a) the research paper appeared in the reference list of the data paper;
    (b) the research paper was cited in the dataset repository;
    (c) the research paper cited the data paper;
    (d) the research paper cited the dataset.
2. and the following two conditions were also satisfied:
    (a) at least one person was an author of both the data and the research paper;
    (b) the research paper was a substantial, analytical interpretation of the dataset associated with the data paper.

We recognise that rule 2a and 2b need some justification. Rule 2a expresses the requirement that the data paper and research paper are products of the same research effort as opposed to the reuse of data by others for a new research question. This heuristic rule, therefore, creates a link between a data paper and a research paper where data is interpreted by the person who created that data in the first place. Rule 2b that requires "substantial, analytical interpretation of the dataset" is a matter of subjective judgement, and one which resists to be easily translated to a computer script, but such is the nature of the data and the association of data and research papers we work with. Taking the example of the research and data pair mentioned in section 1, substantial analytical interpretation is understood to be a deep manipulation of the data which generates new insights and goes beyond a simple reference. That is, the research paper goes into significant detail about how the data was used, reorganised and processed to answer a research question that the authors set for themselves. As we continue to expand the ground truth in our future work, we aim to create heuristic guidelines to improve consistency among annotators who are currently enjoying a large degree of discretion appropriate to this early stage of the project. For the purposes of the current study, manual curation involved skim-reading research articles to capture such substantial analytical treatment of data. This resulting curated dataset served as our benchmark to assess how different methods of automatic pairing perform.

We sample 159 data papers from three of the largest data journals (see Table 1). Our curated dataset consists of pairs of data papers published in 2022 and related research papers, 91 from

Figure 1: Data Extraction and Processing Steps

Table 1: Data journals used as sources of data papers. The third column contains the number of articles published in the journal in 2022 and the last column contains the number of curated pairs per journal.

Journal	Publisher	2022	Pairs
Biodiversity Data Journal	Pensoft	242	0
Data	MDPI	190	47
Data in Brief	Elsevier	1059	21
Data Science Journal	Ubiquity Press	20	0
Database	Oxford Academic	109	0
Geoscience Data Journal	Wiley	37	0
JOAD	Ubiquity Press	7	0
JOHD	Ubiquity Press	29	0
JOPD	Ubiquity Press	20	0
RDJ	Brill	8	0
Scientific Data	Springer Nature	765	91
Total		**2486**	**159**

Scientific Data, 47 from MDPI's *Data*, and 21 from *Data in Brief*. For our subsequent analysis we filtered out 28 cases where one data paper was matched to multiple research papers, to reduce the complexity of the task for the matching algorithms. The final number of pairs considered was 131. The dataset has been deposited to Figshare (Ribary and Wigdorowitz, 2025).

3.2 Data extraction

To support the experiments, we devised a data extraction pipeline using an openly available repository of research articles. Options for data sources we have considered include OpenAlex, PubMed, Zenodo and Dimensions. After analysis of each source looking at papers they cover, content volume, and metadata we could extract, we chose OpenAlex (Priem et al., 2022) as the source for our dataset. This was due to its comprehensive coverage of multiple disciplines and its extensive volume of more than 200 million works. Initially we queried the OpenAlex API. However, this approach posed challenges due to limitations on query volume, making it difficult to retrieve data at scale, and leading us to download its data dump.

The pipeline consisted of the following steps:

1. Extraction: We used the OpenAlex data dump to retrieve metadata from the papers.

2. Filtering: After extracting all papers, we retained only those that were published open access, were research articles, proceedings papers, or book chapters (excluding review articles and other types).

3. Paper categorization: We automatically categorized the collected papers into data papers (defined as all articles published in our list of data journals described in section 3.1) and research papers (all other articles extracted in the previous step).[1] We focussed on data papers published in 2022, as this year had the greatest coverage in the set of papers used for the ground truth. A five-year range was chosen before 2022 and to the date of analysis (2025), since data papers are published before and after their related research papers (McGillivray et al., 2022).

4. Author overlap: For each research paper, we checked whether any of its authors also appeared among the authors of the data papers. We only retained research papers that shared at least one author with a data document, helping to reduce redundancy and increase the relevance of the comparison set.

5. Length filtering: To ensure a fair comparison and remove anomalies, papers with abstracts shorter than 300 characters or longer than 2,000 characters were excluded.

6. Full-text extraction: We extracted the full text of the papers selected in the above step. This ensured that the storage and computational capacities were used effectively. The full texts were obtained from OpenAlex, which provided open access locations along with landing page URLs that could be used for download.

Table 2 presents the statistics for our dataset construction pipeline, showing the progression from initial OpenAlex metadata extraction through filtering and full-text extraction.

[1] We acknowledge that this heuristic may miss data papers

Table 2: Number of research papers, data papers and total number of papers at the different steps of our data extraction pipeline.

Data extraction Step	Total	Research Papers	Data Papers
Filtering	49,224,956	—	—
Categoration + Author Overlap	457,599	455,105	2,494
Length Filtering	305,308	303,063	2,245
Full-text Extraction	244,571	243,350	1,221

For each paper, we retrieved the following metadata fields: Title, Abstract, DOI, OpenAlex ID, Author IDs, Publication Year and Journal/Source of publication. In what follows, we will use the terms "data papers" and "research papers" to refer to the final set resulting from applying all 6 pipeline steps described above. For any given data paper d, we will refer to its set of candidate matching research papers r_i as the research papers selected by the 6 pipeline steps and which shared at least one author in common with d and were published up to 5 years before and after d.

3.3 Matching algorithms

We experimented with different methods to calculate the similarity between data papers (d) and research paper (r) and therefore identify research papers that are related to a given data paper. To evaluate both surface-level and semantic similarity in linking data papers to research articles, starting from the set of data papers in the dataset, we implemented and evaluated two complementary approaches: set-based matching and vector-based matching, as summarised in Table 3. Set-based approaches and TF-IDF offer interpretable and computationally efficient ways to measure lexical overlap between documents by identifying shared terminology. SBERT approaches capture deeper semantic relationships by representing documents in high-dimensional spaces.

3.4 Set-Based Matching

For the set-based matching methods we preprocessed the texts with tokenization, lowercasing, and stopword removal to ensure that similarity measures focus on semantically meaningful content rather than common function words. We applied two Jaccard-based metrics to three sets: the top N (N=10, 20, and 50) most frequent words in texts of research papers and data papers, the named entities (NEs) entracted from the texts using the `spaCy` [2] library in Python, and on all tokens in the texts. These methods were applied to both abstracts and full texts.

published in generalist journals or misclassify non-data papers in data journals.

3.4.1 Jaccard

For each data paper d and research paper r among its candidate matches, let Tokens(d) and Tokens(r) denote the sets of unique tokens extracted from d and r, respectively. We computed the Jaccard similarity based on the sets of unique tokens extracted from each document:

$$S_{\text{J}}(d, r) = \frac{|\text{Tokens}(d) \cap \text{Tokens}(r)|}{|\text{Tokens}(d) \cup \text{Tokens}(r)|} \quad (1)$$

3.4.2 Multi-set Jaccard

To account for term frequency, we also implemented a multiset version of Jaccard similarity, which compares token counts rather than just presence or absence (da Fontoura Costa, 2021). Let T be the total number of distinct tokens in the union of the data paper d and research paper r. The token frequencies in each document are represented as vectors: $[d_1, d_2, \ldots, d_T]$ and $[r_1, r_2, \ldots, r_T]$, where d_i and r_i denote the frequency of token i in d and r, respectively.

The multiset Jaccard similarity is defined as follows and rewards documents that not only share vocabulary, but also use it with similar frequency:

$$S_{\text{MJ}}(d, r) = \frac{\sum_{i=1}^{T} \min(d_i, r_i)}{\sum_{i=1}^{T} \max(d_i, r_i)} \quad (2)$$

3.5 Vector-based Matching

Vector-based matching methods represent documents as numerical vectors and compute similarity using distance metrics such as cosine similarity. While TF-IDF captures lexical overlap and term salience, SBERT captures deeper semantic relationships through contextualized embeddings.

[2] https://spacy.io/ (last accessed 05/10/2025).

Table 3: Overview of algorithms for matching data papers with research papers. We implemented two groups of methods: set-based and vector-based. Each method was applied to different scopes of textual content. We analysed three values for N:10, 20, and 50.

Method	**Top N Frequent Words**	**Named Entities (NEs)**	**All Tokens**
Set-based Methods			
Jaccard	✓	✓	✓
Multi-set Jaccard	✓	✓	✓
Vector-based Methods			
TF-IDF	–	–	✓
SBERT	–	–	✓
SBERT re-ranked	–	–	✓
SBERT re-ranked with LLM	–	–	✓

3.5.1 TF-IDF

We first applied Term Frequency–Inverse Document Frequency (TF-IDF) (Manning et al., 2008). Each document is represented as a sparse vector where each dimension corresponds to a term in the corpus vocabulary. The TF-IDF score reflects how important a term is to a document relative to its frequency across the corpus.

Let $\text{Tokens}(d)$ and $\text{Tokens}(r)$ denote the sets of unique tokens extracted from data paper d and research paper r, respectively. Let $R(d)$ be the set of candidate research papers for d. The corpus $C(d)$ consists of d and all papers in $R(d)$. For each token t in document $x \in C(d)$, the TF-IDF weight is:

$$\text{TF-IDF}_d(t, x) = \text{tf}(t, x) \times \text{idf}_d(t) \tag{3}$$

where $\text{tf}(t, x)$ is the term frequency of token t in document x, and $\text{idf}_d(t)$ is the inverse document frequency defined as:

$$\text{idf}_d(t) = \log \frac{1 + |C(d)|}{1 + |\{y \in C(d) : t \in \text{Tokens}(y)\}|} + 1 \tag{4}$$

computed over the corpus of the data paper and its candidate research papers, $C(d)$.

For each data paper d and research paper $r \in R(d)$, we compute cosine similarity between their TF-IDF vectors:

$$\begin{aligned} &S_{\text{TF-IDF}}(d, r) \\ &\quad = \cos(\text{TF-IDF}(d), \text{TF-IDF}(r)) \\ &\quad = \frac{\text{TF-IDF}(d) \cdot \text{TF-IDF}(r)}{\|\text{TF-IDF}(d)\| \cdot \|\text{TF-IDF}(r)\|} \end{aligned} \tag{5}$$

This method captures term salience and is sensitive to shared terminology, but does not account for synonymy or contextual meaning.

3.5.2 SBERT

To capture deeper semantic relationships, we used Sentence-BERT (SBERT), a transformer-based model that produces dense, contextualized embeddings for sentences and documents (Reimers and Gurevych, 2019). We computed cosine similarity between SBERT embeddings of the titles and abstracts of each data–research paper pair:

$$S_{\text{SBERT}}(d, r) = \cos(\text{emb}(d), \text{emb}(r)) \tag{6}$$

SBERT has been shown to outperform other embedding methods on semantic similarity and transfer learning tasks (Reimers and Gurevych, 2019).

3.5.3 Reranking

To further refine the initial similarity rankings, we implemented two reranking approaches that leverage more sophisticated architectures to capture nuanced relationships between data papers and research papers that may be missed by the initial vector-based methods.

SBERT Cross Encoder Reranking: We employed a cross encoder architecture (Reimers and Gurevych, 2019) that jointly processes pairs of data paper and research paper abstracts. Unlike bi-encoders that generate independent embeddings for each document, cross encoders allow attention mechanisms to operate across both documents simultaneously, capturing fine-grained interactions between their content. The cross encoder takes concatenated representations of the document pair as input and outputs a relevance score. This should lead in principle to superior performance in semantic matching tasks at the cost of increased computational complexity due to longer input sequences.

Listwise LLM-based Reranking: We implemented a listwise approach using Large Language

Models to re-order the top-k research paper candidates from SBERT rankings (Ma et al., 2023). We used GPT-4o-mini (OpenAI, 2024) with a temperature of 0.1 and maximum token limit of 1000. The LLM receives the data paper abstract and a numbered list of candidate research paper abstracts, then outputs a re-ranked ordering based on which papers most likely substantially use or analyze the described dataset. Reranking was performed in a zero-shot setting using prompts that emphasized analytical relevance and dataset usage patterns (see Appendix A for the complete prompt template).

4 Evaluation

In this paper we refer to the curated dataset of manually validated data–research paper pairs as our evaluation set. Since our approach does not involve training a supervised model, we do not distinguish between training, validation, and test subsets. We evaluate all methods using Mean Reciprocal Rank (MRR), a standard metric in information retrieval (Voorhees, 1999). Given a data paper d and a set of candidate research papers $R(d)$, with the correct match $r^* \in R(d)$, the reciprocal rank is defined as:

$$\text{RR}(d) = \frac{1}{\text{rank}(r^*)}$$

The MRR is the average of reciprocal ranks over all data papers in the evaluation set. In case of ties, we use the expected reciprocal rank under uniform random tie-breaking, i.e., the average of the reciprocals of the tied positions. Higher MRR values indicate better performance.

We also evaluated based on the retrieval accuracy on top-N and top-$N\%$ selection. Ties in ranking scores were handled in a tie-aware fashion. If the correct data paper was part of a tie block fully above the cutoff (e.g., top-10), the prediction was counted as correct. If the tie block straddled the cutoff, we assigned fractional credit proportional to the number of tied items within the cutoff (e.g., if three papers tied for ranks 9–11 and two were within the top-10, the correct paper received a score of $2/3$). If the entire tie block fell outside the cutoff, the prediction was considered incorrect. This approach prevents arbitrary tie-breaking and ensures consistency across metrics.

4.1 Comparing abstracts and full texts

One of the key practical considerations in designing systems to match data papers with research papers is the availability and granularity of textual content. While full texts may offer richer information, abstracts are more readily accessible and computationally efficient to process. To determine whether this trade-off affects matching performance, we conducted statistical significance tests on a subset of our set-based and vector-based method, comparing results obtained from abstracts and full texts.

We applied the paired sign test (Gibbons, 1993) to compare the performance of methods when using abstracts versus full texts. This non-parametric test was chosen because it makes minimal assumptions about the data distribution under the null hypothesis. It assesses whether there is a statistically significant difference in performance (i.e., ranking) between the two conditions across all data papers in our curated dataset.

Table 4: Significance test results comparing abstracts vs. full texts for a subset of methods. Each method was applied to different scopes of textual content: top N frequent words (Freq) for $N = 10$, and all tokens (All). $\alpha = 0.05$.

Method	**Scope**	***p*-value**	**Stat**
Jaccard	Freq	0.59	54.55%
	All	0.058	63.64%
Multi-Jaccard	Freq	0.59	54.55%
	All	0.237	58.62%
TF-IDF	All	0.77	47.06%

Table 4 reports the results of statistical significance tests (with significance threshold $\alpha = 0.05$) comparing the performance of set-based methods applied to abstracts vs. full texts. The *Stat* column shows the percentage of times the abstract method achieved a better ranking than the full-text method. Across all scopes and methods, the p-values indicate that the differences are not statistically significant, suggesting that using abstracts yields comparable performance to using full texts.

5 Results

Table 5 presents the performance of all matching methods when applied to abstracts. It shows that the best-performing methods vary depending on the evaluation metric and reflects the diverse strengths of each method. This dispersion reflects the complementary strengths of different approaches: lexical overlap, semantic similarity, and contextual reasoning each contribute uniquely to match qual-

Table 5: Matching performance using abstracts. Metrics include Mean Reciprocal Rank (MRR), top-N accuracy, and top percentile accuracy. Each method was applied to different scopes of content: top N frequent words (Freq), named entities (NE), and all tokens (All). For SBERT re-ranked methods, only the top 50 candidates were re-ranked.

Method	Scope	MRR	Top-5	Top-10	Top-50	Top-1%	Top-5%	Top-10%
Jaccard	Freq=10	0.38	44.40%	52.61%	77.14%	39.68%	60.76%	72.39%
	Freq=20	0.31	45.99%	55.73%	81.92%	35.37%	64.33%	77.49%
	Freq=50	0.44	54.01%	63.54%	87.37%	47.66%	72.71%	**82.90%**
	NE	0.10	14.06%	14.84%	22.53%	36.19%	54.76%	60.49%
	All	0.40	53.91%	62.50%	89.06%	46.88%	67.19%	79.69%
Multi-Jaccard	Freq=10	0.38	44.40%	52.61%	77.14%	39.68%	60.76%	72.39%
	Freq=20	0.31	45.99%	55.73%	81.92%	35.37%	64.33%	77.49%
	Freq=50	0.44	54.01%	63.54%	87.37%	47.66%	72.71%	**82.90%**
	NE	0.07	8.60%	12.06%	21.94%	25.35%	31.20%	34.82%
	All	0.41	50.78%	64.06%	89.06%	45.31%	71.88%	79.69%
TF-IDF	All	**0.45**	41.18%	70.31%	84.38%	**51.56%**	**75%**	82.81%
SBERT	All	0.40	53.91%	68.75%	92.19%	35.16%	53.91%	73.44%
SBERT re-ranked	All	0.39	53.91%	65.62%	92.19%	31.25%	56.25%	70.31%
SBERT re-ranked + LLM	All	0.44	**62.50%**	**71.88%**	**92.19%**	34.38%	62.50%	75.00%

ity. TF-IDF achieved the highest MRR (0.45) and top-1% and top-5% accuracy (51.56% and 75%, respectively), outperforming other vector-based methods in those metrics. SBERT re-ranked with LLMs showed the best overall performance in top-N accuracy, with top-10 and top-50 scores reaching 71.88% and 92.19%, respectively.

As expected, among set-based methods performance generally improves with larger token scopes (e.g., Freq=50 and All), which suggests that richer lexical context enhances matching accuracy. Named Entity-based matching underperforms across all metrics, showing that entity-level overlap alone is insufficient to capture the nuanced relationships between data and research papers. Vector-based methods show strong performance, with SBERT re-ranked using LLMs achieving the highest scores in top-N accuracy. This highlights the value of semantic understanding and contextual reasoning in identifying meaningful links. Interestingly, the performance gap between SBERT and SBERT re-ranked is modest, which suggests that initial semantic similarity captures much of the relevant signal, and reranking offers incremental gains. For the subset of methods tested (set-based approaches and TF-IDF), the lack of statistically significant differences between abstracts and full texts supports the feasibility of using abstracts for these approaches, especially when full texts are unavailable or costly to process. Further evaluation is needed for semantic embedding methods.

6 Limitations and Conclusion

Our findings demonstrate that it is feasible to automatically identify research papers related to data publications using NLP-based methods and that the best-performing method varies depending on the metric used. They also underscore the importance of evaluating methods across multiple metrics to avoid over-reliance on a single performance indicator. Hence, hybrid systems combining multiple matching strategies may offer the most robust performance. Set-based approaches, particularly Multi-set Jaccard and Jaccard with frequent words or all tokens, offer interpretable and computationally efficient solutions. TF-IDF achieves the highest MRR, indicating strong precision in ranking the correct match highly. Vector-based methods, especially SBERT with LLM-based reranking, provide superior performance in terms of top-N accuracy, though at higher computational cost. For set-based and TF-IDF methods, abstracts appear sufficient for effective matching, which has practical implications for scalability, since abstracts are more readily available and less resource-intensive to process than full texts.

As to this study's limitations, we rely on metadata availability, which may not generalize to all disciplines or publication venues. Future research could explore avenues to generalize our approach. First, expanding the curated dataset to include more data journals and multilingual content would help assess cross-domain applicability. Second, incorporating citation networks, dataset repository metadata, and author affiliations could enrich the matching process and reduce reliance on textual similarity alone. Third, exploring the analysis of matches between data papers and research papers where more than one research paper corresponds to the same data paper. Finally, developing hybrid models that combine lexical, semantic, and structural features may improve performance, especially in cases where abstracts are sparse or ambiguous.

7 Authors' contributions

BMcG designed the study, supervised the project, wrote sections 1, 2, 4, 5, and 6, reviewed and edited the manuscript. KA designed the taxonomy of methods and statistical tests; conducted experimentation on article abstracts; processed the OpenAlex data dump; wrote sections 3.5.2, 3.5.3, and Appendix A; and parts of sections 3.3 – 3.5. VH downloaded article full texts and conducted experiments for sections 3.4 and 3.5.1; wrote Figure 1, Table 2 and part of sections 3.2 – 3.5. MR conceptualized the study, the manual linking process, compiled the curated pairs, wrote parts of section 1 and 3, reviewed and edited the manuscript. MW conceptualized the study, compiled the curated pairs, wrote Table 1, and parts of section 3, reviewed and edited the manuscript.

References

Nacime Bouziani, Shubhi Tyagi, Joseph Fisher, Jens Lehmann, and Andrea Pierleoni. 2024. REXEL: An end-to-end model for document-level relation extraction and entity linking. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 119–130, Mexico City, Mexico. Association for Computational Linguistics.

Leonardo Candela, Donatella Castelli, Paolo Manghi, and Alice Tani. 2015. Data journals: A survey. *Journal of the Association for Information Science & Technology*, 66(9):1747–1762.

Luciano da Fontoura Costa. 2021. Further generalizations of the jaccard index. *ArXiv*, abs/2110.09619.

Priyangshu Datta, Suchana Datta, and Dwaipayan Roy. 2025. Raging against the literature: Llm-powered dataset mention extraction. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, JCDL '24, New York, NY, USA. Association for Computing Machinery.

Stefan Ekman, Olle Sköld, and Isto Huvila. 2025. Functions of paradata in data papers. *Journal of Documentation*, 81(7):253–272.

Michael Färber and Ann-Kathrin Leisinger. 2021. Recommending datasets for scientific problem descriptions. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 3014–3018, New York, NY, USA. Association for Computing Machinery.

Jean Gibbons. 1993. Location tests for single and paired samples (sign test and wilcoxon signed rank test). In *Nonparametric Statistics*, pages 5–25. SAGE Publications, Inc., Thousand Oaks, California. Accessed 2025-10-02.

Jenny Heddes, Pim Meerdink, Miguel Pieters, and Maarten Marx. 2021. The automatic detection of dataset names in scientific articles. *Data*, 6(8).

C. Jiao, K. Li, and Z. Fang. 2023. How are exclusively data journals indexed in major scholarly databases? an examination of four databases. *Scientific Data*, (737).

Kim Jihyun. 2020. An analysis of data paper templates and guidelines: types of contextual information described by data journals. *Science Editing*, (7):16–23.

Naoto Kai, Hayato Tomisu, Toshiki Shimbaru, and Tomoki Yoshihisa. 2025. Study on extracting keywords that reveal the value of research data through comparisons between academic and data papers. In *Advances in Internet, Data and Web Technologies*, pages 1–8, Cham. Springer Nature Switzerland.

Kembellec, Gérald and Le Deuff, Olivier. 2022. Poétique et ingénierie des data papers.

Christine Kosmopoulos and Joachim Schöpfel, editors. 2024. *Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux*. Collection Humanités numériques et science ouverte. Septentrion Presses Universitaires.

Kai Li and Pei-Ying Chen. 2018. The narrative structure as a citation context in data papers: A preliminary analysis of scientific data. *Proceedings of the Association for Information Science and Technology*, 55(1):856–858.

Kai Li, Jane Greenberg, and Jillian Dunic. 2020. Data objects and documenting scientific processes: An analysis of data events in biodiversity data papers. *Journal of the Association for Information Science and Technology*, 71(2):172–182.

Kai Li and Chenyue Jiao. 2021. How are data paper abstracts constructed? Preliminary analysis of rhetorical moves in data paper abstracts from Scientific

Data and Data. In *Proceedings of 18th International Conference on Scientometrics & Informetrics*.

Kai Li and Chenyue Jiao. 2022. The data paper as a sociolinguistic epistemic object: A content analysis on the rhetorical moves used in data paper abstracts. *Journal of the Association for Information Science and Technology*, 73(6):834–846.

Xiaozheng Liu. 2022. Discussion on the structural model and constituent elements of data papers. *Resources Data Journal*, 1:2–9.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *Preprint*, arXiv:2305.02156.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Scoring, term weighting, and the vector space model. In *Introduction to Information Retrieval*, page 100–123. Cambridge University Press.

Barbara McGillivray, Paola Marongiu, Nilo Pedrazzini, Marton Ribary, Mandy Wigdorowitz, and Eleonora Zordan. 2022. Deep impact: A study on the impact of data papers and datasets in the humanities and social sciences. *Publications*, 10(4):39.

OpenAI. 2024. Gpt-4o technical report. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-10-02.

J. Priem, H. Piwowar, and R. Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. ArXiv.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

Nathalie Reymonet. 2017. Améliorer l'exposition des données de la recherche : la publication de data papers. Ce texte présente la structure et le contenu d'un “ data paper ” ainsi que des exemples de revues qui publient de tels articles.

Marton Ribary. 2020. A relational database of roman law based on justinian's digest. *Journal of Open Humanities Data*.

Marton Ribary and Barbara McGillivray. 2020. A corpus approach to roman law based on justinian's digest. *Informatics*, 7(4).

Marton Ribary and Mandy Wigdorowitz. 2025. Manually curated links between data papers and research papers. Figshare. DOI: `https://doi.org/10.6084/m9.figshare.3058945`.

E. M. Voorhees. 1999. TREC-8 Question Answering Track Report. In *Proceedings of the 8th Text Retrieval Conference*, pages 77–82.

William H. Walters. 2020. Data journals: incentivizing data access and documentation within the scholarly communication system. *Insights: the UKSG journal*, 33(18).

Ju-Yeon Yu, Donghun Yang, and Kyong-Ha Lee. 2025. Scilinkbert: A bert-style language model for understanding scientific texts with citations. *Journal of Supercomputing*, 81(1356).

A LLM Reranking Prompt Template

The following prompt template was used for GPT-4o-mini reranking with temperature=0.1 and max_tokens=1000:

LLM Reranking Prompt Template

```
You are helping to find research papers
that are related to a given data paper.

DATA PAPER:
Title: {query_title}
Abstract: {query_text}

Below are {len(candidate_list)} research
papers ranked by semantic similarity.
Your task is to rerank them based on how
likely each research paper is to be
related to the data paper above (i.e.,
the research paper likely uses or
references the dataset described in the
data paper).

CANDIDATES TO RERANK:
{candidates_text}

Please provide your reranking as a
comma-separated list of numbers, with
the most relevant paper first.
For example: 3,1,7,2,5,4,6

Your reranking:
```

Where `{query_title}`, `{query_text}`, `{len(candidate_list)}`, and `{candidates_text}` are dynamically filled with the data paper title, abstract, number of candidates, and formatted list of candidate papers respectively.

A benchmark for end-to-end zero-shot biomedical relation extraction with LLMs: experiments with OpenAI models

Aviv Brokman, **Xuguang Ai**, **Yuhang Jiang**, **Shashank Gupta**, **Ramakanth Kavuluru**

Division of Biomedical Informatics, Department of Internal Medicine
University of Kentucky, USA.
Correspondence: `ramakanth.kavuluru@uky.edu`

Abstract

Extracting relations from scientific literature is a fundamental task in biomedical NLP because entities and relations among them drive hypothesis generation and knowledge discovery. As literature grows rapidly, relation extraction (RE) is indispensable to curate knowledge graphs to be used as computable structured and symbolic representations. With the rise of LLMs, it is pertinent to examine if it is better to skip tailoring supervised RE methods, save annotation burden, and just use zero shot RE (ZSRE) via LLM API calls. In this paper, we propose a benchmark with seven biomedical RE datasets with interesting characteristics and evaluate three Open AI models (GPT-4, o1, and GPT-OSS-120B) for end-to-end ZSRE. We show that LLM-based ZSRE is inching closer to supervised methods in performances on some datasets but still struggles on complex inputs expressing multiple relations with different predicates. Our error analysis reveals scope for improvements.

1 Introduction

Do we need training data to perform relation extraction (RE)? Since ChatGPT was introduced in December 2022, this has been a prominent question on minds of many NLP researchers, especially those that focus on structured information extraction from biomedical literature. With the recent success of zero-shot methods in other areas of NLP, RE is ripe for investigation, and biomedicine is a particularly compelling domain as relations are central to knowledge discovery.

RE is the high-value NLP task of identifying semantic relationships between entities within text. Consider the following sentence taken from the drug combination extraction (DCE) (Tiktinsky et al., 2022) dataset: *"Furthermore, in non-metastatic castration-resistant prostate cancer (M0 CRPC), two second-generation anti-androgens, apalutamide, and enzalutamide, when used in combination with ADT, have demonstrated a significant benefit in metastasis-free survival."* Two beneficial drug combinations are described here: (1) `apalutamide` with `ADT`, and (2) `enzalutamide` with `ADT`. In RE, we want to extract these relations into a structured form; in this case a tuple of drugs administered in combination along with a signifier of the normative effect of the drug:

- `{drugs: (apalutamide, ADT), effect: positive}`
- `{drugs: (enzalutamide, ADT), effect: positive}`

Thus, RE can be viewed as the conversion of unstructured data into structured data representing relationships between entities. In biomedicine, entities of interest are mainly genes, mutations, proteins, chemicals, drugs, diseases, and symptoms. The relationships that can hold between them are myriad, but some obvious relationships of importance are drug interactions, protein interactions, disease-causing mutations and chemicals, drug side effects, and disease-treating drugs. Many RE efforts assume that the entities and their types are already provided as part of the input. They do RE by giving the input text and a pair of entity spans in it and ask for the relation type linking them. The technical name for this would be relation classification (RC). However, for RE to be fully automated and evaluated fairly, the input must be just free text and the burden of (a) spotting the entities and (b) linking them with predicates, both fall on the method. This is called end-to-end RE (E2E RE) and is much harder than RE when entities are pre-annotated. For the rest of this paper, whenever we refer to RE we mean the end-to-end variety.

Biomedical publications are being generated at breakneck speed — PubMed indexes nearly 40 million articles and over four thousand more are

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 44–55
December 23, 2025 ©2025 Association for Computational Linguistics

indexed daily. Biomedical relations are so valuable that there are teams of workers employed to read biomedical text and populate databases with them. With so much text available, automating RE would allow us to mine biomedical relations at scale, rapidly enlarging databases.

1.1 Zero-shot RE

Traditionally, RE has been conducted by fine-tuning with hundreds to thousands of examples. In this paradigm, every narrow RE task requires a dataset to be curated for it. Dataset curation for biomedical RE is a laborious process — annotators need highly specialized knowledge, it takes time to develop clear and consistent annotation guidelines, and annotators need to be trained on the guidelines (Luo et al., 2022a; Li et al., 2016). The process is so laborious for annotators that, typically, named entity recognition and RE tools are used to make suggestions to annotators to speed up the process (Luo et al., 2022a; Li et al., 2016). For all these reasons, few-shot (FS) and zero-shot (ZS) RE are desirable. They remove the need for a training dataset, which is usually the largest in a training/validation/test split. If modeling choices do not need to be made and if performance need not be measured, validation and test sets could be omitted as well. The promise of high quality low-resource RE is a proliferation of databases and an increase in their richness and reliability.

Generative large language models (LLMs) have grown to dominate NLP research activity in recent years, with new multi-billion parameter models being released regularly. Owing to large amounts of diverse training data, vast quantities of parameters, and techniques that align models with human goals like instruction finetuning (Ouyang et al., 2022a; Wei et al., 2022; Sanh et al.) and RLHF (Ouyang et al., 2022b), these models can now perform a wide array of tasks, in a ZS manner, with impressive results. However, LLMs are not adept at producing long output in a consistent format, a key challenge when converting generated text into structured relations, without specific guiding mechanisms. To address this, previous studies have employed two main strategies: (1) prompting the language model to generate text in a predetermined format, followed by the use of predefined regular expressions to extract relation components (Luo et al., 2022b; Gupta et al., 2025), or (2) directly specifying a structured output within the prompt itself (Wadhwa et al., 2023). Though these approaches have shown promise, Wadhwa et al. (2023) found that even for few-shot sentence-level RE, GPT-3.5 often generates plausible relations that, while recognizable to humans as correct, do not precisely match the gold standard relations. This discrepancy should be expected to be more pronounced in document-level ZS settings. Without fine-tuning, fulfilling synergistically demanding requirements of long, exact, consistent extraction from text becomes more challenging.

Zero-shot relation extraction (ZSRE) poses a unique challenge to LLMs because it requires the generation of long, exact text in a consistent format. This makes it much more challenging to generate an exact output than most other exact-output tasks NLP researchers have been tackling, such as question answering (QA), where the answers are a single token or a phrase (Touvron et al., 2023; Wang et al., 2023a; Achiam et al., 2023). Also, ZS generation has been successful at tackling problems where long text must be generated, like summarizing (Touvron et al., 2023; Wang et al., 2023a; Achiam et al., 2023). In such tasks, there is no single correct generation, so the fact that LLMs produce diverse output is not a problem; these are often evaluated by humans or judgment by other more LLMs (though this practice is controversial). But in the real-world scenario of RE from a document, there may be many relations present (the BioRED (Luo et al., 2022a) dataset contains abstracts with > 50 relations), and the LM must generate long text with stringent requirements on what the text must consist of.

Due to the aforementioned challenges, LLMs generate relations that are correct to a human but do not match annotations in the test dataset exactly, which artificially deflates calculated performance. For example, if an annotated relation contains the entity `hypertensive` as the disease in a relation and an LM extracts `hypertension`, this would be considered incorrect in the usual performance evaluation. Wadhwa et al. (2023) deal with this problem using manual evaluation of their relation extraction systems, but this is not desirable because (1) it is expensive and (2) because we want a system that can be used at scale to populate databases automatically, without human intervention.

1.2 Related work

Most RE methods focus on constructing embeddings for candidate relations, followed by a classification step. A parallel line of research has developed around the use of copy-mechanisms in a sequence-to-sequence (seq2seq) framework. Seq2seq tasks involve generating an output/target sequence from a given input/source sequence. This method is predominantly favored in areas like machine translation, where the format aligns naturally with the task. However, researchers have adapted RE to fit into the seq2seq paradigm by transforming structured relations into predefined sequences of tokens (Zeng et al., 2018; Zhang et al., 2020; Nayak and Ng, 2020; Zeng et al., 2020; Giorgi et al., 2022). For instance, Giorgi et al. (2022) transform the relation `{gene: ESR1, disease: schizophrenia, predicate:association}` into the sequence `ESR1 @gene schizophrenia @disease @association`*. Subsequently, models are trained to generate such sequences, and decoding them becomes straightforward. The key to the success of these models across various architectures lies in the incorporation of copy mechanisms. In the context of copy mechanism-based RE, the fundamental component is an LSTM that, at each time step, opts to select either a token from the source sequence or a limited additional vocabulary, such as punctuation or special tokens like `@gene`.

In more recent developments, the use of LLMs has emerged as a novel seq2seq approach for RE (Luo et al., 2022b; Gupta et al., 2025). The authors of BioGPT, for example, have fine-tuned their model using soft prompts to generate relations within natural language sentences, such as `The relation between <head entity> and <tail entity> is <relation type>`. These constructs with place holders for entities and relation types (also called predicates) are often called output *templates*. The filled-in output template is then processed using regular expressions to extract the relations from the LM's generations. This method presents a significant advantage over traditional relation representation and copy-mechanism approaches primarily because it does not require mention annotations during training. Such a feature reduces the workload for annotators on additional datasets, as they can shift their focus solely to relation annotation rather than annotating every entity mention. Building on this, Wadhwa et al. (2023) modified this approach by designing target sequences as Python-interpretable tuples of relations, rather than in the form of natural sentences, for sentence-level RE tasks.

The remarkable performance of large, human-aligned language models in FS and ZS tasks has sparked interest in exploring their potential for low-resource RE. This emerging area of research particularly focuses on the capabilities of OpenAI's GPT models. Wadhwa et al. (2023) investigate the use of the instruction-finetuned GPT-3.5 for sentence-level biomedical RE. Their FS in-context learning experiments yield results that are competitive with state-of-the-art approaches. In a similar vein, Wang et al. (2023b) applied GPT-3.5 for sentence-level RC. Further advancing this line of inquiry, Jahan et al. (2023) conduct RE experiments using both GPT-3.5 and GPT-4, testing them on two RE dataset test sets, though in one they filter out all examples with no relations.

1.3 Our contributions

Given the effectiveness of ZS generation in other NLP tasks, in this paper, we investigate its utility in the high-value task of biomedical RE. We comprehensively test the effectiveness of OpenAI GPT-4 (Achiam et al., 2023), OpenAI o1 (Jaech et al., 2024), and the open-weights GPT-OSS (OpenAI, 2025) on seven RE datasets that vary in domain, length of text, diversity of entity and relation types, whether relations are entity-level (EL) or mention-level (ML), and whether relations are described across multiple sentences. We analyze model errors to determine the strengths and weaknesses of this approach. The code, datasets, and the LLM prompts for all our experiments are available here: `https://github.com/bionlproc/ZeroShotRE`.

Our research differs from prior studies in several ways. Previous research predominantly explored general biomedical tasks — a valuable effort — but restricting the study to one or two datasets is insufficient to explore the intricacies of RE. Our experiments encompass a broader spectrum of biomedical RE tasks, across seven datasets. We employ datasets that include relations confined within single sentences as well as those with relations spanning across multiple sentences. Moreover, some datasets we study feature relations between entities, while others contain relations between *mentions* of entities. This variety introduces a range of complexities, including varying levels of difficulty and differing quantities of relations within each text. Such diversity underlines that, although these tasks are all categorized under RE, they each pose unique challenges to RE methodologies. Another crucial

*This is a slight adaptation from the original paper, simplified for clarity aspect of our work is the comprehensive evaluation of performance across entire test sets, facilitating direct comparisons with other studies, whether they are fine-tuned, FS, or ZS.

2 Materials and methods

2.1 Task definition

Let (x, y) be an example in the test dataset, where x is text and y is the set of annotated relations expressed within x. Depending on the dataset, the text and relations may have different structures. x is most commonly a title-abstract pair but maybe a sentence along with a larger passage containing it or a single string of text. y depends on the dataset as well. In general, a relation consists of a set of typed entities along with a relation type connecting them. In most datasets, relations hold between two entities, but in the DCE (Tiktinsky et al., 2022) dataset, they hold between a variable number of entities. Entities and relations are typed, but the number of types varies by dataset.

Depending on the dataset, relations are either annotated at the mention level (ML) or entity level (EL). ML relations hold between textual mentions (exact spans) of entities. Textual mentions may consist of a span of text or multiple in the case of discontinuous entity mentions. EL relations hold between normalized entities, that is, the entities are provided in the form of an ID number corresponding to a biomedical concept from a controlled vocabulary. In most commonly used EL datasets (including all that we use), textual mentions of the biomedical concepts with their normalized ID number are also provided. Biomedical databases of relations are typically structured with EL relations.

For ZSRE, we guide the LLM $\mathcal{M}$ to predict y using template T. That is,

$$\widehat{y} = \mathcal{M}\left(T\left(x\right)\right), \tag{1}$$

where T is a user chosen natural language instruction along with an output template.

2.2 Extraction

Most prior work on generative RE has relied on traditional supervision. In such scenarios, the choice of template is of moderate importance, because $\mathcal{M}$ is finetuned to learn the nature of the problem and the structure of the output. Without a supervision signal, it is challenging to guide $\mathcal{M}$ to (1) understand the nature of the problem and (2) output relations in a consistently structured form. The latter is important for biomedicine if RE is to be automated, and important for research as it permits performance metrics to be calculated. To address these challenges, T adds a complete description of the RE task as well as instructions to produce output in the form of a JSON object, a description of the format the JSON object should take, and an example of what a filled-in JSON could look like.

Given that producing consistent structured output from a language model that has been principally trained to produce natural language is a problem faced in information extraction in general, a few attempts have been made to solve it (Newhouse, 2023; Sengottuvelu, 2023) or produce a structured output posthoc (Yurtsev, 2023). Given that we use GPT-4 (Achiam et al., 2023), o1 (Jaech et al., 2024) and GPT-OSS (OpenAI, 2025) as $\mathcal{M}$ and the fact that OpenAI added functionality in their API for obtaining JSON objects as output in two different modes, we use their tools. The first tool they developed requires the user to provide a schema (in the form of a JSON object) delineating the structure the output JSON should exhibit. The more recent tool infers the schema from the prompt. We refer to these modes as *explicit* and *inferred* modes and experiment with both on GPT-4. For OpenAI o1 and GPT-OSS (Jaech et al., 2024; OpenAI, 2025), we test only the mode that performed better on GPT-4, due to budgetary constraints. Unlike GPT-4, the other two models are designed with test-time chain-of-thought based reasoning ability; it learns to "recognize and correct its mistakes" and " break down tricky steps into simpler ones."[†]

2.3 Evaluation

As a task, RE is complex, and there are many reasonable ways to measure performance. This has led to a proliferation of measures, but also confusion and conflation of them — so much so that rigorous study of the issue has been made (Taillé et al., 2020). Unfortunately, the state of affairs has only worsened as (1) researchers have not heeded this work, (2) papers have faded descriptions of details necessary for reproducibility, (3) EL RE has been introduced, and (4) seq2seq methods, which have grown in popularity, lend themselves to new performance measures.

Among the three main methods that have been published for EL RE — JEREX (Eberts and Ulges, 2021), seq2rel (Giorgi et al., 2022), and BioGPT

[†]https://openai.com/index/learning-to-reason-with-llms/

(Luo et al., 2022b) — no two calculate F1 in the same way. JEREX measures performance very strictly: a predicted relation is considered correct if it matches a gold relation exactly, and entities within the relations are judged correct when mentioned boundaries are correct. Seq2rel's "strict" measure is similar, except that rather than entity mentions being judged on boundary correctness, they are judged on whether the predicted strings match gold entity mention strings, and duplicate gold mention strings are collapsed to a single mention. JEREX correctness therefore implies seq2rel "strict" correctness, but not vice versa. We note that since seq2rel uses a copy mechanism that points directly to tokens, nothing prevents them from making an exact comparison with JEREX. However, Giorgi et al. (2022) compared their performance with JEREX using slightly different "strict" metrics as indicated earlier. They additionally use a "relaxed" measure of correctness that only requires a majority of predicted entity mentions to match that of a gold entity.

The generative approach of BioGPT lends itself to the extraction of a single entity mention rather than all of them, and therefore Luo et al. (2022b) deem a predicted relation correct if the extracted mentions match the *longest* mention in the *dataset*, rather than the example text. Further distinguishing their performance measure from previous papers, Luo et al. (2022b) filter examples with no gold relations from the dataset. Despite these differences, they compare their performance with seq2rel. At this point, it is not clear, on any dataset, which of these methods has the highest performance; nor is it clear that they all *can* be compared with one another, even if done with utmost care. To make matters worse, Jahan et al. (2023) do not describe their evaluation methodology or provide code.

Our method most closely resembles BioGPT, but we believe that an extracted entity mention matching any gold one should be considered correct; so we develop yet another performance measure and strive for the utmost clarity in explaining it. In the EL RE context, we consider a predicted relation to match a gold relation if (1) each extracted entity mention participating in a relation matches any gold entity mention, (2) entity types are correct, and (3) relation type is correct (this is trivial when there is only one relation type.) In the ML RE context, gold entities consist of a single mention, so (1) becomes simpler: an extracted entity mention must match the gold entity mention. For EL RE datasets, we honor the annotation at the EL by mapping entities of predicted relations to their normalized ID numbers (based on gold annotations) and removing duplicate predictions before assessing matches to gold relations. True positives are predicted relations matching gold relations; false positives are predicted relations that do not match any gold relations; and False negatives are unmatched gold relations. We calculate precision, recall, and F1-score for each dataset.

2.4 Datasets

Table 1 shows the basic properties of the datasets we studied. Three of them contain EL relations; these datasets naturally contain relations with entity mentions across multiple sentences. The remaining four datasets contain intra-sentence ML relations, though relation types may be more easily extracted when the surrounding context is available.

The **ADE** dataset (Gurulingappa et al., 2012) consists of sentences extracted from MEDLINE case reports describing adverse effects resulting from drug use, extracted from medical case reports. It contains two entity types: drugs and adverse effects and one relation type, adverse drug event. There is no official split of the dataset.

DCE (Tiktinsky et al., 2022) documents the efficacy of drug combination therapies, presenting a unique RE challenge in that relations contain a variable number of entity types. Each instance consists of an abstract, within which a focal sentence is identified that contains multiple drug references. The drug references are classified as either being positive, for a beneficial drug combination, non-positive, for a combination with a neutral or negative effect, or non-combination, when the drugs are not given in combination. Following the practice of the original authors, DCE performance is evaluated using two metrics: `Positive Combination F1` score and `Any Combination F1` score. The `Positive Combination F1` treats the relation type `positive` as the positive class, while the `Any Combination F1` score lumps `positive` and `non-positive` relation types together, and treats them as the positive class.

The primary aim of **ChemProt** (Krallinger et al., 2017) is to extract intra-sentence relations between chemical compounds and proteins/genes from biomedical abstracts. Relation types holding between these entities can be described as `upregulator`, `downregulator`, `agonist`, `antagonist`, or `substrate of`. Over

Datasets	Type	Input Type	# Entity Types	# Predicates	Examples w/o relations
ADE (Gurulingappa et al., 2012)	Mention-Level	Sentence	2	1	No
DCE (Tiktinsky et al., 2022)		Abstract	1	3	Yes
ChemProt (Krallinger et al., 2017)		Abstract	2	5	Yes
DDI (Herrero-Zazo et al., 2013)		Abstract	4	4	Yes
CDR (Li et al., 2016)	Entity-Level	Abstract	2	1	No
GDA (Wu et al., 2019)		Abstract	2	1	No
BioRED (Luo et al., 2022a)		Abstract	4	8	Yes

Table 1: Basic properties of the biomedical datasets tested, including whether relations were annotated at the mention-level or entity-level, whether the input text is a sentence or an abstract, the number of entity types and predicates, and whether the dataset contains instances with no relations.

25% of abstracts contain no relations.

`DDI` (Herrero-Zazo et al., 2013) annotates intra-sentence interactions between four types of pharmacological substances: brand-name drugs, generic drugs, drug categories, and substances not approved for human use. Drug-drug interactions are either descriptions of pharmacokinetic mechanisms, descriptions of effect/pharmacodynamic mechanisms, recommendations about drug combinations, or documented interactions without additional details. Nearly two-thirds of instances contain no relations. The `CDR` (Li et al., 2016) and `GDA` (Wu et al., 2019) datasets respectively annotate diseases induced by chemicals/drugs or associated with genes in PubMed abstracts. Both datasets contain EL relations with a single relation type.

`BioRED` (Luo et al., 2022a) dataset annotates eight non-directional relation types holding between genes, gene variants, chemicals, and diseases. The relation types are `positive correlation`, `negative correlation`, `association`, `binding`, `co-treatment`, `drug interaction`, `comparison`, and `conversion`; certain relation types are only valid for a subset of all combinations of entity types. Instances in BioRED often contain many relations, sometimes in excess of 90. Presumably, for this reason, there are only 100 test instances.

We show ZS prompts for `ChemProt` and `CDR` in Table 4 in the Appendix. Prompts for all seven biomedical datasets are made available here: `https://github.com/bionlproc/ZeroShotRE/tree/main/prompts`.

3 Results

We present results for GPT-4, OpenAI o1, and GPT-OSS-120B in Table 2. Performance varies substantially across datasets, with consistently higher scores on ADE, DCE, CDR, and GDA compared to ChemProt, DDI, and BioRED. A clear pattern emerges: datasets with only 1 or 2 relation types yield much higher performance, while those with 4–8 relation types, often accompanied by a larger set of entity types, show lower performance. These observations may reflect spurious correlations; future work could test this hypothesis by slicing datasets by entity and relation types to examine whether performance improves.

When comparing models, GPT-4 generally underperforms relative to both o1 and GPT-OSS-120B. GPT-OSS consistently surpasses GPT-4 across all datasets, while o1 also outperforms GPT-4 except on ADE, CDR, and GDA—datasets that are simpler and contain only a single relation type. One explanation is that o1 and GPT-OSS-120B better accommodate tasks with greater relational and entity complexity through more robust reasoning. As a result, both surpass GPT-4 on ChemProt, BioRED, and DDI, each of which contains 4–8 distinct relation types. On average, o1 is over 3 F1 points better than GPT-4, with BioRED showing the largest relative gain (more than doubling GPT-4 performance). GPT-OSS-120B achieves the strongest overall results, attaining the best F1 scores on most datasets. It is encouraging to see a open-weights model that is better than large closed models (by an average four F1 points over the o1 model). Precision–recall trends are largely consistent across models, though o1 and GPT-OSS-120B tend to favor precision on ChemProt and DDI, while GPT-4 exhibits relatively higher recall.

As discussed in Section 2.3, a valid comparison of performance between current biomedical RE methods is generally not possible. However, all of the performance measures are obviously positively correlated, so we collate performance from

Datasets	GPT-4 (Inferred)			GPT-4 (Explicit)			OpenAI o1 (Inferred)			GPT-OSS-120B (Inferred)		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ADE (Gurulingappa et al., 2012)	75.3	60.4	67.0	76.7	62.8	69.1	73.5	62.8	67.7	76.7	68.8	**72.5**
DCE (Pos.) (Tiktinsky et al., 2022)	58.9	68.7	63.4	61.3	66.7	63.9	61.5	74.7	67.5	64.7	73.3	**68.8**
DCE (Any) (Tiktinsky et al., 2022)	55.6	76.1	64.2	49.2	71.8	58.4	69.6	74.6	72.1	70.6	77.0	**73.7**
ChemProt (Krallinger et al., 2017)	24.1	24.6	24.3	19.7	23.0	21.2	37.0	20.7	26.5	28.9	31.0	**30.0**
DDI (Herrero-Zazo et al., 2013)	27.7	33.6	30.4	27.6	33.2	30.1	46.1	29.3	35.8	36.2	53.3	**43.1**
CDR (Li et al., 2016)	48.9	42.3	45.3	49.3	42.4	45.6	46.6	41.2	43.7	52.1	48.5	**50.2**
GDA (Wu et al., 2019)	46.0	63.4	53.3	46.1	65.2	54.0	40.2	57.6	47.3	45.5	67.4	**54.4**
BioRED (Luo et al., 2022a)	12.6	7.1	9.1	15.1	7.3	9.8	30.8	18.6	**23.2**	27.0	17.9	21.5
Average	41.7	43.4	41.9	41.4	43.3	41.6	48.5	43.6	44.9	47.7	51.7	**49.0**

Table 2: Main results for end-to-end ZSRE experiments. As DCE is evaluated in two ways (see Section 2.4), their performance values are averaged before being included in the calculation for the "Average" row.

other publications with reasonably transparent evaluation methodology in Table 3. We find that *supervised* methods using far smaller LMs perform similarly or better than our method, an unsurprising result (Wadhwa et al., 2023). However, due to the aforementioned difficulties of comparison, the only obvious discrepancy occurs in ChemProt, where our method fared poorly. ChemProt encodes fine-grained and mechanistic relations between chemicals and proteins where the predicates are semantically close. This may require explicit learning of subtle lexical cues and biochemical context, which is difficult in the ZS setting, leading to the large performance gap relative to the supervised score.

We analyzed errors from all datasets to glean insights into the pitfalls of ZSRE. We first note which aspects of RE were highly successful. GPT-4 was nearly perfectly faithful to the structured schema we described in our templates and generated entity and relation types were nearly always selected from the set of types we described in the templates. Predicted entity mentions were usually assigned the correct entity type as well. Predicted mentions are rarely not present in the text. Last, it was uncommon for relation types to be incorrect when entities participating in relations were correct.

All models tend to under-predict relations when an instance contains more than a few gold relations. Figure 1 in the Appendix depicts this pattern for CDR on GPT-4, a representative example. It shows that the average number of relations predicted per test instance lags further behind the number of gold relations as the number of gold relations increases, and that this naturally results in decreased recall. We attribute this to the observation that generative models tend to perform worse with long sequences (Hochreiter et al., 2001; Li et al., 2023a).

A common error across datasets and models we encountered was that of partial matching entity mentions, in which a predicted relation nearly matches a gold relation, but predicted mentions either include extra words not found in gold mentions or exclude words found in them. For example, we extracted the incorrect chemical-disease relation (`chemical: methamphetamine, disease: methamphetamine-induced psychosis`), which would have been correct had we extracted the disease as `psychosis`. Future research should focus on extracting correct boundaries for entity mentions, as this was a major source of error.

Frequently, false positives appear to be correct relations missed by the annotators. This has been previously documented in RE datasets (Tran and Kavuluru, 2019; Tan et al., 2022) and has been shown to artificially deflate performance. Given the high frequency of missed relations, it may be prudent to re-annotate biomedical benchmark RE datasets in the mode of Tan et al. (2022).

For the most part, models predicted mentions that either were exact spans of source text or concatenated discontinuous spans. However, in some cases, they used domain knowledge, predicting a text string not found in the source text. In one extracted relation from GDA, we predicted the entity `interleukin-10`, which did not appear in the text in this form, whereas the gold version was `interleukin (IL)-10`.

Besides error patterns holding across datasets, errors arose particular to specific datasets. Our method largely failed on BioRED, with frequent, obviously incorrect, predicted entity mentions and types as well as relations claiming opposite rela-

Methods	ADE	DCE	ChemProt	DDI	CDR	GDA	BioRED
Yan et al. (Yan et al., 2021)	83.2	-	-	-	-	-	-
Seq2Rel (Giorgi et al., 2022)	-	$66.7^{P\dagger}/71.1^{A\dagger}$	-	-	$40.2^{\dagger}/52.4^{\ddagger}$	$55.2^{\dagger}/70.5^{\ddagger}$	-
BioGPT (Luo et al., 2022b)	-	-	-	40.8	46.2	-	-
PURE (Zhong and Chen, 2021)	-	-	69.0	-	-	-	-
GPT-4 (zero-shot)	69.1	$63.9^{P}/64.2^{A}$	24.3	30.4	45.6	54.0	9.8
OpenAI o1 (zero-shot)	67.7	$67.5^{P}/72.1^{A}$	26.5	35.8	43.7	47.3	23.2
GPT-OSS-120B (zero-shot)	72.5	$68.8^{P}/73.7^{A}$	30.0	43.1	50.2	54.4	21.5

Table 3: Comparison of performance (F1) of OpenAI ZS scores with previous finetuned methods. Note that F1 does not have identical meaning across methods (see Section 2.3). Superscripts P and A refer to the "Positive" and "Any Combination" evaluation settings for DCE (Jiang and Kavuluru, 2023). Superscripts † and ‡ refer to the "strict" and "relaxed" evaluations described in Section 2.3.

tionships between two given entities. In DCE, drug combinations were frequently missing drugs that participate in a relation. Also, relation types were often incorrectly assigned; we suspect that this is caused by the domain-specific knowledge, like the interpretation of lab result quantities, sometimes required to correctly assign relation type.

4 Discussion

In the LLM era, a fundamental question is whether and in what settings can we simply use ZS predictions from frontier LLMs without the tedious and expensive creation of training data and custom supervised models. The answer to this is heavily dependent on particular task on hand in terms of expectations on recall and precision and the consequences of false positives/negatives. For example, if the relations are being used for knowledge discovery, focusing more on precision can minimize creation of misleading hypotheses. However, to conduct systematic reviews on information encoded in the relations (e.g., drug–side-effects), the relations extracted should have high recall.

As a first step to assess the general end-to-end ZSRE competence of LLMs, we created a new benchmark and conducted experiments with frontier LLMs. Our high level takeaway is that for shorter instances with fewer relations (less information density), ZSRE is closer to fully supervised models; this is more so if the entities are shorter (fewer tokens) such as mostly single token drug entities in DCE and ADE datasets. Entity complexity is also high in ChemProt dataset, which may have contributed to the vast discrepancy (40 points in F1) between supervised and ZS performances. Relatively, high density longer inputs (e.g., BioRED) lead to almost unusable performance at this point. Further assessments that account for partial matches of entities may be warranted but unless that is done carefully, the results may not be meaningful. For instance, partial matches that do not involve the head word of an entity phrase are mostly incorrect and misleading at best.

All results using OpenAI GPT models for publicly available datasets since GPT-3 (Brown et al., 2020) come with a caveat: we do not know what data the models were trained on. These LLMs train on massive amounts of scraped web data, and most datasets we used are available on the web in some form. It is possible that they may have been trained on these datasets. However the very low scores obtained for BioRED and ChemProt indicate that this contamination is unlikely. Please note that here we are not focusing on the textual inputs (sans labels) being part of the LLM pre-training corpora; this was shown to not cause any contamination in general (Li et al., 2024). The focus here is on whether particular task-specific labels were part of the training process. Recently, methods to detect membership of specific texts in the pretraining corpora of LLMs have been introduced (Rastogi et al., 2025) but foolproof tools that assess whether ground truth labels of supervised tasks are leaked are not available. Recent evidence shows that label contamination is possible for question answering tasks where the answer is a single word or a short phrase, but if rich, structured outputs (as in end-to-end RE) are required, memorization is not as prevalent (Wang et al., 2025). While label contamination should be kept in mind, that potential should not be grounds for not evaluating ZS performances; other teams have been exploring the same with OpenAI models (Li et al., 2023b; Zhang et al., 2024), albeit not in an end-to-end manner. In light of recent calls for action (Jacovi et al., 2023), we have encrypted the datasets using in this project before hosting them in our GitHub space. Others can still reproduce our results by running our scripts, which would decrypt them on the fly in a programmatic manner.

5 Limitations

Our results and associated implications apply only to English datasets, but we hope others will follow up with benchmarks in other languages. With regards to LLMs studied, we only considered OpenAI models while there are more frontier options (e.g., from Anthropic, Google, and xAI), which we could not work with due to time and cost constraints. Also, there are datasets with more complex annotation schemes than BioRED such as the cancer genetics and pathway curation datasets of the BioNLP 2013 shared task (Nédellec et al., 2013). However, the download links for them are not active and additional efforts are needed to carefully recover and experiment with them. Finally, as we already acknowledged in Section 4, there is some potential risk of label contamination although we believe it is minimal for RE tasks where memorization is nontrivial.

Acknowledgment

This work is supported by the U.S. NIH National Library of Medicine through grant R01LM013240. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Markus Eberts and Adrian Ulges. 2021. An end-to-end model for entity-level relation extraction using multi-instance learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3650–3660.

John Giorgi, Gary Bader, and Bo Wang. 2022. A sequence-to-sequence approach for document-level relation extraction. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 10–25.

Shashank Gupta, Xuguang Ai, Yuhang Jiang, and Ramakanth Kavuluru. 2025. Comparison of pipelines, seq2seq models, and llms for rare disease information extraction. In *International Conference on Applications of Natural Language to Information Systems*, pages 49–63. Springer.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, and Corso Elvezia. 2001. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. *researchgate*.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084.

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*.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 326–336.

Yuhang Jiang and Ramakanth Kavuluru. 2023. End-to-end n-ary relation extraction for combination drug therapies. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 72–80. IEEE.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio López, Umesh Nandal, and 1 others. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

Bingzhi Li, Guillaume Wisniewski, and Benoît Crabbé. 2023a. Assessing the Capacity of Transformer to Abstract Syntactic Representations: A Contrastive Analysis Based on Long-distance Agreement. *Transactions of the Association for Computational Linguistics*, 11:18–33.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023b. Revisiting large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6877–6892.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. 2024. An open-source data contamination report for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022a. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022b. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Tapas Nayak and Hwee Tou Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8528–8535.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP shared task 2013 workshop*, pages 1–7.

Ben Newhouse. 2023. Structural alignment: Modifying transformers (like gpt) to follow a json schema. `https://github.com/newhouseb/clownfish`.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022a. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022b. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Saksham Rastogi, Pratyush Maini, and Danish Pruthi. 2025. STAMP your content: Proving dataset membership via watermarked rephrasings. In *Forty-second International Conference on Machine Learning*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, and 1 others. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Rahul Sengottuvelu. 2023. Jsonformer: A bulletproof way to generate structured json from language models.

Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. 2020. Let's stop incorrect comparisons in end-to-end relation extraction! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3689–3701.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting docred-addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487.

Aryeh Tiktinsky, Vijay Viswanathan, Danna Niezni, Dana Meron Azagury, Yosi Shamay, Hillel Taub-Tabib, Tom Hope, and Yoav Goldberg. 2022. A dataset for n-ary relation extraction of drug combinations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3190–3203.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv e-prints*, pages arXiv–2307.

Tung Tran and Ramakanth Kavuluru. 2019. Neural metric learning for fast end-to-end relation extraction. *arXiv preprint arXiv:1905.07458*.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589.

Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. 2025. Generalization vs memorization: Tracing language models' capabilities back to pretraining data. In *The Thirteenth International Conference on Learning Representations*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023a. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508.

Yuqing Wang, Yun Zhao, and Linda Petzold. 2023b. Are large language models ready for healthcare? a comparative study on clinical language understanding. In *Machine Learning for Healthcare Conference*, pages 804–823. PMLR.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Ye Wu, Ruibang Luo, Henry CM Leung, Hing-Fung Ting, and Tak-Wah Lam. 2019. Renet: A deep learning approach for extracting gene-disease associations from literature. In *Research in Computational Molecular Biology: 23rd Annual International Conference*, pages 272–284. Springer.

Zhiheng Yan, Chong Zhang, Jinlan Fu, Qi Zhang, and Zhongyu Wei. 2021. A partition filter network for joint entity and relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 185–197.

Eugene Yurtsev. 2023. Kor. `https://github.com/eyurtsev/kor`.

Daojian Zeng, Haoran Zhang, and Qianying Liu. 2020. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9507–9514.

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514.

Jeffrey Zhang, Maxwell Wibert, Huixue Zhou, Xueqing Peng, Qingyu Chen, Vipina K Keloth, Yan Hu, Rui Zhang, Hua Xu, and Kalpana Raja. 2024. A study of biomedical relation extraction using gpt models. *AMIA Summits on Translational Science Proceedings*, 2024:391.

Ranran Haoran Zhang, Qianying Liu, Aysa Xuemo Fan, Heng Ji, Daojian Zeng, Fei Cheng, Daisuke Kawahara, and Sadao Kurohashi. 2020. Minimize exposure bias of Seq2Seq models in joint entity and relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 236–246.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61.

A Appendix

In Table 4, we give examples of ZS prompts for **`ChemProt`** and **`CDR`** datasets each describing the task, predicate label definitions, along with JSON output templates with some dummy filled in examples. In Figure 1, we show how the number of predicted relations does not keep up as the number of gold relations increases (x-axis) and hence recall decreases as the number of gold relations increase in an instance.

Dataset	Prompt
ChemProt	Your task is to extract all relevant triples from an input biomedical text. Each triple has a chemical mention, a gene/protein mention, and a predicate linking the two mentions. The predicate belongs to one of the following 5 predicates: “CPR:3”, “CPR:4”, “CPR:5”, “CPR:6” and “CPR:9”. These 5 predicates are further specified as below: “CPR:3” includes UPREGULATOR, ACTIVATOR and INDIRECT UPREGULATOR “CPR:4” includes DOWNREGULATOR, INHIBITOR and INDIRECT DOWNREGULATOR “CPR:5” includes AGONIST, AGONIST ACTIVATOR and AGONIST INHIBITOR “CPR:6” includes ANTAGONIST “CPR:9” includes SUBSTRATE, PRODUCT OF and SUBSTRATE PRODUCT OF Note that chemical or gene/protein mentions should have appeared from the original input text. Make sure that each relation is based on mentions within the same sentence in an abstract. The output triples should be saved as per the following format: {“relations”: [{“chemical”: “chemical1”, “gene”: “gene1”, “relation”: “relation1”}, {“chemical”: “chemical2”, “gene”: “gene2”, “relation”: “relation2”}, ...] } The output will be {“relations”:[]} if there are no relevant triples expressed in the input text. With this format, a hypothetical example output for a biomedical text could be the following: {“relations”: [{“chemical”: “polyamines”, “gene”: “caspase”, “relation”: “CPR:3”}, {“chemical”: “DL-alpha-difluoromethylornithine”, “gene”: “ornithine decarboxylase”, “relation”: “CPR:4”}, {“chemical”: “putrescine”, “gene”: “ODC”, “relation”: “CPR:9”}] }
CDR	Your task is to extract all chemical-disease relations from a text in which the chemical/drug induces the disease. Note that the chemical or disease names should have appeared in the original input text. The output should be saved as per the following format: {“relations”: [{“chemical”: “chemical1”, “disease”: “disease1”}, {“chemical”: “chemical2”, “disease”: “disease2”}, ...] } The output will be {“relations”:[]} if there are no chemical-disease pairs in which the chemical induces the disease expressed in the input text. With this format, a hypothetical example output for a biomedical text could be the following: {“relations”: [{“chemical”: “Lidocaine”, “disease”: “cardiac asystole”}, {“chemical”: “daunorubicin”, “disease”: “neutropenia”}] }

Table 4: ZS prompts for **ChemProt** and **CDR**. All prompts for seven biomedical datasets are released in our GitHub website.

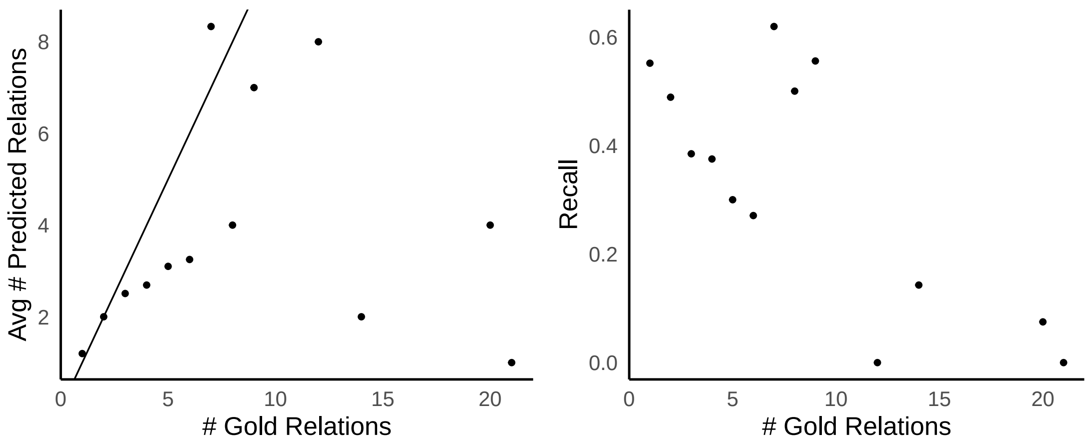


Figure 1: (Left) The average number of GPT-4 predicted relations per test instance is plotted against the number of gold relation in the instance for the CDR dataset. The line $y = x$ is overlayed for ease of interpretation. (Right) Recall is calculated for subsets of the data by the number of gold relations.

Bridging the Gap: Instruction-Tuned LLMs for Scientific Named Entity Recognition

Necva Bölücü[1], **Maciej Rybinski**[2], **Stephen Wan**[1]
[1]CSIRO Data61, Sydney, Australia
[2]ITIS, University of Málaga, Málaga, Spain

Abstract

Information extraction (IE) from scientific literature plays an important role in many information-seeking pipelines. Large Language Models (LLMs) have demonstrated strong zero-shot and few-shot performance on IE tasks. However, there are challenges in practical deployment, especially in scenarios that involve sensitive information, such as industrial research or limited budgets. A key question is whether there is a need for a fine-tuned model for optimal domain adaptation (i.e., whether in-domain labelled training data is needed, or zero-shot to few-shot effectiveness is enough). In this paper, we explore this question in the context of IE on scientific literature. We further consider methodological questions, such as alternatives to cloud-based proprietary LLMs (e.g., GPT and Claude) when these are unsuitable due to data privacy, data sensitivity, or cost reasons. This paper outlines empirical results to recommend which locally hosted open-source LLM approach to adopt and illustrates the trade-offs in domain adaptation.

We focus on several instruction-tuning frameworks leveraging IE benchmark datasets to capture task-specific knowledge whilst maintaining model generalisability. We refer to this class of LLM models as *Specialised LLMs* (s-LLMs). We show that instruction-tuned (IE task-adapted) s-LLMs can outperform open-source and proprietary LLMs for entity extraction from scientific documents. Furthermore, this improvement gain is substantial, highlighting the value of the in-domain (continual) fine-tuning.

1 Introduction

Information Extraction (IE) from the scientific literature (e.g., scientific documents, technical reports) is a critical component of scientific information-seeking pipelines (Luan et al., 2018; Nasar et al., 2018; Cai et al., 2025). IE supports tasks such as knowledge-base construction (e.g., BRENDA (Chang et al., 2021) and ChEMBL (Papadatos et al., 2015)), advancing knowledge discovery (Horawalavithana et al., 2022), and supporting predictive modelling (Li et al., 2022). In such pipelines, Named Entity Recognition (NER) is often the initial step used to extract structured output from unstructured text, enabling downstream tasks, such as relation extraction (RE) (Luan et al., 2018) or knowledge-graph construction (Zhang and Soh, 2024). As a result, improving NER accuracy is critical, as errors introduced at this stage can propagate and impact the reliability of the entire pipeline.

As IE pipelines evolve, they are increasingly designed as agentic systems, where multiple specialised models, or agents, collaborate to complete complex tasks (Belcak et al., 2025; Sharma and Mehta, 2025). Within such systems, smaller fine-tuned models play a key role: they can be assigned to specific subtasks, such as NER, RE, or validation, and interact with other agents to balance accuracy, efficiency, and scalability. In this context, NER is not only a technical bottleneck but also a foundational capability for multi-agent scientific systems, motivating the study of models that can be adapted to domain-specific tasks while remaining lightweight and composable.

Recent advances in large language models (LLMs) such as GPT-5[1] and Claude 3.7 Sonnet[2] have improved our ability to extract information from scientific documents. Commercial APIs built in proprietary LLMs offer a strong performance. Using these models becomes problematic, however, in scenarios that involve sensitive data (e.g., biomedical records, confidential industrial research), as privacy cannot be guaranteed. Consequently, many research and industrial settings rely on open-source models as a practical alternative.

Although open-source LLMs provide significant

[1]https://openai.com/gpt-5
[2]https://www.anthropic.com/news/claude-3-7-sonnet

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 56–71

flexibility, their zero-shot performance for IE tasks often remains insufficient for practical IE scenarios, as errors propagate to downstream tasks. In-context learning (ICL) enables task and domain adaptation through the inclusion of prototypical examples in the prompt (Li et al., 2023; Ghosh et al., 2024) *without actually performing supervised learning* (no parameter update) called few-shot learning. While ICL markedly improves over zero-shot performance, studies show that it still lags behind state-of-the-art results for IE tasks (Li et al., 2023; Ma et al., 2023; Xu et al., 2024; Wadhwa et al., 2023; Wan et al., 2023; Gao et al., 2023; Jiao et al., 2023; Huang et al., 2024; Wang et al., 2024; Gui et al., 2024b). For the domain of science literature, similar trends have been observed; ICL improves results but does not match supervised fine-tuning models (SLM and LLM) (Xiao et al., 2024; Zhou et al., 2024; Li et al., 2024; Zhang et al., 2025b), and simpler fine-tuned models (e.g., RoBERTa (Liu et al., 2019)) can outperform LLMs using ICL (Jimenez Gutierrez et al., 2022; Bölücü et al., 2023).

To bridge this gap, researchers increasingly turn to instruction-tuned LLMs for IE, which we refer to as *specialised LLMs* (s-LLMs). These models are trained using instruction-tuning on task-specific benchmark datasets (Zhou et al., 2024; Gui et al., 2024b; Wang et al., 2023; Zhang et al., 2025a), where each training instance pairs an instruction, an input text, and a structured output that reflects the benchmark's annotation scheme. Instruction-tuning provides task-level adaptation and enhances zero- and few-shot generalisation, while still enabling local deployment—an essential requirement for domains involving sensitive or proprietary data. Typically built on open-source LLMs such as Llama[3] and Qwen[4], s-LLMs provide cost-effective alternatives to proprietary systems like GPT-4 (Gui et al., 2024b,a; Yuan et al., 2025), making them suitable for applications such as industrial research.

The s-LLMs require a large set of benchmark datasets for instruction-tuning, which is not straightforward and requires substantial computational resources. Therefore, it is not practical to instruction-tune a new model for each conceivable domain for IE. For this reason, in this study, we evaluate the adaptability of *already instruction-tuned* IE-specialised models to scientific domains. Specifically, we focus on three examples of this class of approach: IEPile (Gui et al., 2024b), UniNER (Zhou et al., 2024), and YAYI-UIE (Xiao et al., 2024) (Section 3). These models have been instruction-tuned using a collection of datasets, including scientific datasets (see Table 6), and are designed to generalise across a wide range of IE tasks (e.g., NER, RE, and Event Extraction (EE)) and domains (e.g., social media, biomedical).

Hence, we investigate the following research questions.

- **RQ1:** How well do s-LLMs adapt to the scientific NER task, a subtask of IE, compared to the out-of-the-box (open-source and proprietary) LLMs?
- **RQ2:** What is the additional performance gain of continual (in-domain) tuning of s-LLMs on specific domains compared to their open-source (vanilla) counterparts?

To address the research questions, we evaluate the performance of these models (s-LLMs) and compare them to "out-of-the-box" open-source LLMs[5] (e.g., Llama (Touvron et al., 2023), Baichuan (Yang et al., 2023)), as well as proprietary LLMs (e.g., Claude (Anthropic, 2024), GPT (OpenAI, 2024)). We focus specifically on the case of scientific NER, using four datasets: MeasEval, SciERC, STEM-ECR, and WLPC, each representing a different scientific subdomain or text modality (Section A.4 for an overview of the datasets) in zero-shot, few-shot, and supervised settings. We compare the s-LLMs to baselines under different domain adaptation regimes (zero- and few-shot, continual tuning).

In summary, the **contributions** of this paper include:

- Comparative analysis of instruction-tuned LLMs against their open-source (vanilla) counterparts and proprietary LLMs under different 'learning' regimes (corresponding to different availability of training data).
- Exploratory experiments of models on the NER task to reveal the impact of task-specific instruction-tuning.
- Practical guidelines for researchers aiming to use LLMs for scientific IE.

[3]https://huggingface.co/meta-llama
[4]https://huggingface.co/Qwen

[5]That is, without any further specialisation beyond the foundation model training.

To the best of our knowledge, this is the first extensive evaluation of instruction-tuned LLMs for IE from scientific literature, providing a comprehensive analysis that compares foundation, open-source, and proprietary LLMs and their domain adaptation capabilities across diverse datasets under zero-shot, few-shot and supervised fine-tuning settings.

2 Related Work

LLMs (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023) have already been successfully used in IE (Gao et al., 2023). LLM-based IE methods are divided into In-Context Learning (ICL) and Supervised Fine-tuning (SFT) based approaches. ICL-based models (Jimenez Gutierrez et al., 2022; Li et al., 2023; Wang et al., 2025) rely on prompting with a few labelled examples in addition to instructions, while SFT-based models utilise annotated datasets for fine-tuning LLMs (Zhou et al., 2024; Xiao et al., 2024; Gui et al., 2024b; Li et al., 2024). Research indicates that ICL-based models tend to exhibit relatively inferior effectiveness on IE tasks compared to SFT-based models (Jimenez Gutierrez et al., 2022; Wang et al., 2022; Zhou et al., 2024).

To improve the task and domain adaptability of LLMs, instruction-tuning has become a common technique. This involved fine-tuning LLMs on instruction-based benchmark datasets (a set of datasets specific to a task or domain). Instruction-tuning has been explored across various domains, including Dialogue (Gupta et al., 2022), Intent Classification and Slot Filling (Rosenbaum et al., 2022), Sentiment Analysis (Varia et al., 2023), and Emotion Classification (Liu et al., 2024).

In the context of IE, several studies have advanced instruction-tuning approaches. Zhou et al. (2024) introduce UniNER, which reformulates IE as a Question-Answer (QA) task and instruction-tune Llama using knowledge-distilled datasets from ChatGPT within conversation-style setup, targeting the NER task across diverse domains. Gui et al. (2024b) propose a schema-based instruction-tuning framework for IE (NER, RE and EE) and present IEPile, a bilingual IE instruction benchmark for instruction-tuning. Additionally, Xiao et al. (2024) extend IEPile benchmark by adding more Chinese IE datasets and introduce chat-enhanced instruction tuning that helps gain a fundamental understanding of open-world understanding. Wang et al. (2023) curate the IE INSTRUCTIONS benchmark containing expert-written instructions for diverse IE tasks and apply instruction-tuning for IE tasks. Finally, Lu et al. (2023) focus on *open-world entity profiling*, which is a sub-domain of open-world IE, and construct the INSTRUCTOPEN-WIKI benchmark for the task. They instruction-tune BLOOM to obtain a task-specialised model named PIVOINE.

3 IE-specialised LLMs

Preliminaries Instruction tuning is a supervised fine-tuning (SFT) method in which LLMs are trained on datasets containing human-readable task instructions alongside input-output examples to guide the outputs of LLMs. Each training datapoint, $d = \langle instruction, input, output \rangle$, in the dataset D consists of: (i) an explicit instruction describing the task to be performed; (ii) the corresponding input data; and (iii) the desired output in a defined format.

Unlike standard SFT, which fine-tunes a model on input-output pairs for a specific task without explicit instructions, instruction-tuning conditions the model on natural language task descriptions. This enables better generalisation to unseen domains for the same task (Zhou et al., 2024; Gui et al., 2024b).

Several instruction-tuned LLMs have recently been developed to improve IE performance across diverse domains. We introduce these models (*specialised LLMs for IE*, henceforth 'IE s-LLMs' or simply 's-LLMs') with some discussion of how the approaches vary the basic instruction fine-tuning problem framing.

IEPile (Gui et al., 2024b)[6] proposed a schema-based instruction-tuning, where a schema defines the information to be modelled and extracted, such as entity types, relations, events, etc. This method involves *hard-negative schema construction* and *batched instruction generation*. The schemas are defined as positive (relevant types) and negative (non-relevant types), where negative types can be considered as a kind of "negative" case from a machine learning perspective; the model should not make predictions for this type. To control the complexity of each instruction, the method applies a batching strategy that limits the number of schemas included per instruction using a tunable hyperparameter. The IEPile model training specifically

[6]https://github.com/zjunlp/IEPile

chooses *hard negatives*, labels that are easily confused with positive (i.e., relevant) labels. At inference time, the union set of all schema types across dataset D is presented for prediction.

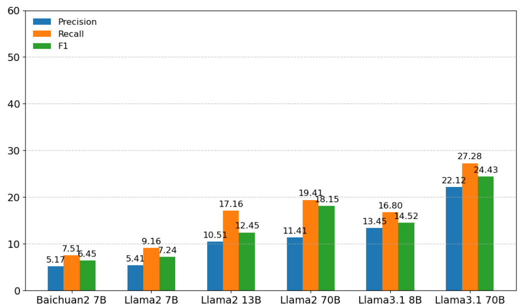
For instruction-tuning, the IEPile benchmark is constructed from a bilingual dataset D that comprises 26 English and 7 Chinese datasets. The dataset spans 3 different tasks: NER, RE, and EE, as exemplified by the datasets ConLL2004 (Carreras and Màrquez, 2004), FabNER (Kumar and Starly, 2022), and BC5CDR (Li et al., 2016), respectively. As a result, the instruction will differ for these datasets, with content specific to each of the task descriptions.

UniNER (Zhou et al., 2024)[7] is a framework that uses ChatGPT for knowledge distillation to generate instruction-tuning data for the NER task. It uses broad-coverage, unlabeled web text and distils this information into an instruction-tuned model built on an open-source LLM (LLaMA), resulting in the UniversalNER models.

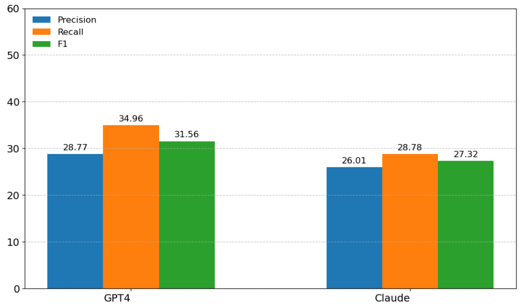
Unlike traditional NER, which frames the task as entity detection, UniNER reformulates it as a question-answering (QA) task. The model input is a question about what entity is present in the accompanying text (e.g., What describes t_1 in the text?), and the output is the corresponding entity span. These QA pairs are generated using GPT, which is prompted to answer such questions based on given texts. The responses are collected as "conversation" transcripts and subsequently segmented into QA tuples t, forming a training dataset for instruction tuning. In data construction, they apply *negative sampling* where non-relevant entity types are included in the dataset. This process creates a distilled dataset suitable for fine-tuning LLaMA-2 (Touvron et al., 2023), resulting in instruction-tuned models that generalise well across domains.

Additionally, the authors introduce a benchmark dataset D consisting of 43 datasets from a wide range of domains, including biomedicine, law, and finance, to evaluate models.

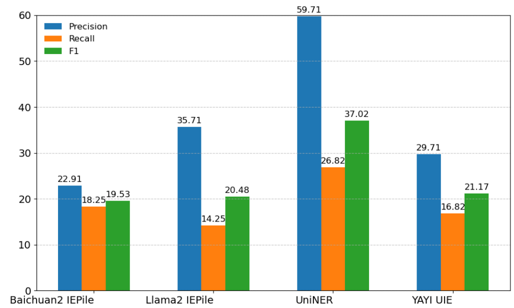
YAYI-UIE (Xiao et al., 2024)[8] is an instruction-tuning framework that consists of two steps: (i) instruction-tuning for chat, where an open-source dialogue data with instructions and a self-constructed corpus is used to train a chat-enhanced language model to gain a fundamental understanding of open-world language and enhance Chinese language capabilities. A key step in the chat-based training is to filter low-quality samples, such as meaningless, incomplete, sensitive, or duplicate samples; (ii) instruction-tuning for IE, where the chat-based model is used to tune for IE tasks with a benchmark dataset. The benchmark D includes a combined dataset of 16 Chinese IE datasets and the InstructUIE benchmark (Wang et al., 2023) for IE instruction-tuning, spanning data from diverse sources such as finance, politics, and security.



(a) Open-source LLMs



(b) Proprietary LLMs



(c) s-LLMs

Figure 1: Zero-shot performance comparison of open-source, proprietary, and IE s-LLMs on the SciERC dataset.

Statistical details of scientific datasets used in instruction-tuning of IE (s-LLMs) are given in Appendix A.4 in Table 6.

[7] github.com/universal-ner/universal-ner
[8] huggingface.co/wenge-research/yayi-uie

Method	SciERC		Stem-ECR		MeasEval		WLPC	
	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot
Proprietary LLMs								
GPT-4	31.56	41.12	21.39	**35.59**	15.67	24.47	<u>52.95</u>	**59.16**
Claude 3.5 Sonnet	27.32	34.19	21.70	34.06	14.70	22.47	36.21	41.05
Open-source LLMs								
Baichuan2	6.45	11.56	8.18	14.12	12.48	16.18	6.75	14.10
Llama2 7B	7.24	13.20	9.47	15.01	11.18	18.14	10.67	19.78
Llama2 13B	12.45	22.38	10.52	19.42	11.36	19.21	12.44	21.45
Llama2 70B	18.12	26.89	12.89	20.17	12.45	20.56	15.74	22.49
Llama3.1 8B	14.52	24.80	10.14	20.17	11.72	20.10	9.32	19.17
Llama3.1 70B	24.43	29.45	14.73	21.45	12.14	21.49	18.42	25.19
IE s-LLMs								
Baichuan2-IEPile	19.53	32.15	15.22	23.10	16.78	26.32	28.79	35.17
Llama2-IEPile	20.48	18.49	18.49	25.42	24.18	29.12	30.40	36.56
UniNER-7B	<u>37.01</u>	**46.43**	17.26	24.37	11.29	19.47	30.13	35.49
YAYI-UIE	22.55	35.19	<u>25.89</u>	32.17	<u>24.62</u>	**30.15**	33.13	37.10

Table 1: Zero-shot and few-shot (1-shot) performance on NER datasets (of note, SciERC is in-domain for the three s-LLMs (IEPile, UniNER and YAYI-UIE); the other datasets are out-of-domain). The best zero-shot results for each dataset are <u>underlined</u>, and the best few-shot results for each dataset are **boldfaced**.

4 Results and Analyses

We design our experiments to evaluate the performance of s-LLMs on the scientific NER task[9]. Our goal is threefold: (1) evaluation of zero-shot and few-shot (1-shot) capabilities of s-LLMs against their open-source (vanilla) counterparts and proprietary LLMs; (2) comparison of continual fine-tuning (in-domain) of s-LLMs' performance against their open-source counterparts' fine-tuning; and (3) exploration of the generalisability of in-domain adapted models to a specific dataset to other scientific datasets. Finally, we present experimental results on the general domain to compare them with findings from the scientific domain.

4.1 Experiments in Zero-shot and Few-shot Settings for Scientific Domain

Our first experiment focuses on examining whether the entity extraction capability learned by s-LLMs is transferable across scientific domain datasets under zero-shot and few-shot settings (**RQ1**). Table 1 reports performance across datasets. Of these, only **SciERC** was used during the instruction-tuning of the s-LLMs and is thus considered the *in-domain* (seen in training) dataset (see Table 6)[10]. The remaining datasets (STEM-ECR, MeasEval and WLPC) are *out-of-domain* (unseen in training), representing unseen entity type sets (covariate shift) and datasets.

Table 1 enables direct comparison across the prior work for the first time. Here, our results include results for open-source and proprietary LLMs that are state-of-the-art at the time of writing.

To begin with, as expected, we note that proprietary LLMs (GPT-4 and Claude 3.5 Sonnet) stand out as strong baselines across the board. Their performance is particularly impressive given the presumed absence of task-specific fine-tuning. Their effectiveness is the best/second-highest performance on most datasets. This demonstrates their ability to generalise across interpretable entity types (e.g., SciERC: *Material, Method, Metric,* $\cdots$; STEM-ECR: *Data, Material,* $\cdots$; WLPC: *Ph, Size, Action,* $\cdots$). However, our aim in this paper is to explore the best methods to obtain alternatives to these cloud-based models, which may be locally hosted by an organisation (particularly if they are responsible for sensitive data). We thus turn our focus to open-source LLMs.

In general, we find that zero-shot inference from IE s-LLMs is better than using open-source LLMs without any task specialisation. For SciERC, UniNER-7B (based on Llama2-7B) achieves a higher F_1 score than both open-source and proprietary LLMs. This demonstrates the benefit of task-specific instruction-tuning. Note that the SciERC dataset is used for the NER task in the UniNER model, whereas it is used as the RE dataset for the IEPile (based on Llama2-13B & Baichuan2-13B)

[9]Experimental settings are given in Appendix A.
[10]Results on it are not strictly zero-shot.

and YAYI-UIE (based on Baichuan2-13B) models (see Table 6). Indeed, the s-LLMs, UniNER-7B and YAYI-UIE, generally outperform the proprietary models for all the datasets except WLPC (which includes text from technical documentation instead of scientific publications), which is particularly interesting given the generally smaller parameter size of s-LLMs compared to GPT-4 and Claude 3.5 Sonnet. However, we note that the margin only has a maximum difference of approximately 9 F_1 points in the case of MeasEval (YAYI-UIE vs GPT-4). In a few-shot setting (1-shot), all models (open-source, proprietary and s-LLMs) benefit from ICL examples, leading to performance gains over zero-shot baselines. Excluding proprietary LLMs, the trend remains consistent. s-LLMs outperform their open-source(vanilla) counterparts.

To understand why s-LLMs exhibit performance gains, we analyse the precision and recall metrics for the models (open-source, proprietary and s-LLMs), presented in Figure 1. This figure presents a comparative analysis of zero-shot performance on the SciERC dataset. Notably, the s-LLMs lead to increased precision, at the expense of recall. In the case of the UniNER approach, the precision gains strongly outweigh any drop in recall. This indicates that targeted training on IE tasks enhances the models' ability to identify relevant entities with greater accuracy. Additionally, these models tend to be relatively conservative and precise in their positive predictions, though they may miss some relevant instances.

In conclusion, while s-LLMs benefit from fine-tuning, they still face generalisation challenges in scientific domains (i.e., the low recall). Moreover, although the s-LLMs are competitive against proprietary LLMs, the performance gap remains narrow in some cases, underscoring the need for further advancements in training and fine-tuning strategies to improve robustness. As a result, we turn our attention to the continued fine-tuning of the IE capability of both open-source LLMs and IE s-LLMs for supervised domain adaptation.

4.2 On the Benefits of Continual In-domain Fine-tuning for Scientific Writing

The results from the previous section show that IE s-LLMs remain competitive against proprietary LLMs under zero-shot and few-shot settings. However, despite their strengths, a performance gap remains compared to SFT models in scientific domain datasets, indicating that there is still room for further improvement.

In this section, we ask: does continual in-domain tuning on the *target* dataset lead to additional performance gains, or do IE s-LLMs already reach peak performance on scientific datasets through their general instruction-tuning? (**RQ2**) In the context of our motivation in Section 1, one might consider how further fine-tuning of a local model on a sensitive or private dataset might improve results.

Following prior work (Zhou et al., 2024; Gui et al., 2024b), we refer to this addition as continual in-domain fine-tuning, a next step after instruction-tuning that further adapts the model to a specific dataset and denote this in our results tables as SFT (for supervised fine-tuning).

To explore the impact of continual in-domain fine-tuning, we fine-tune both open-source (vanilla) LLMs and IE s-LLMs using the training sets of the scientific datasets (Appendix A.4). Table 2 presents the results of the SFT regime compared to zero-shot performance, alongside GPT-4 (zero-shot) and BERT-base (fine-tuned) as baselines; BERT-base represents the task-specific supervised models commonly used across studies in the literature (Xiao et al., 2024; Zhou et al., 2024; Gui et al., 2024b). The table shows that all SFT models improve significantly on all datasets compared to their untuned counterparts. Notably, they outperform GPT-4 in zero-shot settings by a considerable margin. For example, for the STEM-ECR dataset, the difference is over 55 F_1 points, demonstrating clearly that fine-tuning is still a preferred approach in the presence of annotated training data.

The results demonstrate that in-domain fine-tuning on a specific dataset helps, whether this is the original open-source LLMs or the s-LLMs. However, performance gains from in-domain fine-tuning are greater when starting with IE s-LLMs, indicating learning from the multiple datasets used in the s-LLMs training is transferable, demonstrating the benefits of instruction tuning and subsequent in-domain optimisation. Among the in-domain fine-tuned models, the YAYI-UIE model achieves the highest Micro F_1 score among the SFT models across all datasets, showing its strong performance in NER. This reflects its ability to handle diverse scientific NER tasks, possibly related to its larger benchmark datasets covering a wide range of IE tasks and domains in instruction tuning. YAYI-UIE differs from other methods (UniNER and IEPIle) in that dialogue data is used to perform general instruction tuning to train a chat-enhanced

	SciERC		STEM-ECR		MeasEval		WLPC	
Model	**Zero-shot**	**SFT**	**Zero-shot**	**SFT**	**Zero-shot**	**SFT**	**Zero-shot**	**SFT**
Open-source LLMs								
Baichuan2	6.45	52.18	8.18	51.08	12.48	48.47	6.75	35.23
Llama2 7B	7.24	53.14	9.47	50.98	11.18	52.78	10.67	39.56
Llama2 13B	12.45	55.45	10.52	57.14	11.36	54.10	12.44	42.21
Llama2 70B	15.12	56.48	12.89	59.24	12.45	53.18	15.74	45.40
Llama3.1 8B	14.52	56.20	10.14	56.74	11.72	54.10	9.32	43.18
Llama3.1 70B	24.43	55.26	14.73	58.31	12.14	52.85	18.42	46.12
LLMs optimised for IE tasks								
Baichuan2-IEPile	19.53	73.18	15.22	75.12	16.78	59.14	28.79	60.19
Llama2-IEPile	20.48	76.08	18.49	78.17	24.18	64.10	30.40	62.58
UniNER-7B	<u>37.01</u>	78.41	17.26	79.02	11.29	66.18	30.13	60.45
YAYI-UIE	21.17	**80.47**	<u>25.89</u>	**82.52**	<u>24.62</u>	**69.71**	33.13	**64.17**
BERT-base	-	62.81 ±0.85	-	68.17 ±0.76	-	55.43 ±1.15	-	39.52 ±0.52
GPT-4	31.56	-	21.39	-	15.67	-	<u>52.95</u>	-

Table 2: Strict Micro F_1 on NER datasets for zero-shot and SFT settings. The best zero-shot results for each dataset are <u>underlined</u>, and the best SFT results for each dataset are **boldfaced**.

model using a dialogue corpus in both English and Chinese instead of using an instruction model.

Of note, SFT appears somewhat ineffective for base open-source LLMs. Specifically, the BERT baseline yielded higher effectiveness on most datasets. While IE s-LLMs achieved the best performance among all SFT models, this comes at a cost. These models require extensive data resources for training (YAYI-UIE: *49 datasets*, IEPile: *33 datasets*, and UniNER: *43 datasets*) and significant computational resources for instruction-tuning and supervised fine-tuning compared to fine-tuning PLMs for the NER task. The model complexity alone can limit their accessibility and scalability for researchers or practitioners with resource constraints.

This highlights a fundamental trade-off between effectiveness and efficiency: IE s-LLMs deliver state-of-the-art performance but with higher training and inference cost, while smaller models like BERT offer a practical balance of accuracy and affordability.

In summary, continual fine-tuning remains critical for achieving optimal performance in scientific IE. When paired with general instruction tuning, this two-stage process supports both generalisability and domain specialisation (dataset adaptation), enabling robust and adaptable solutions for real-world applications.

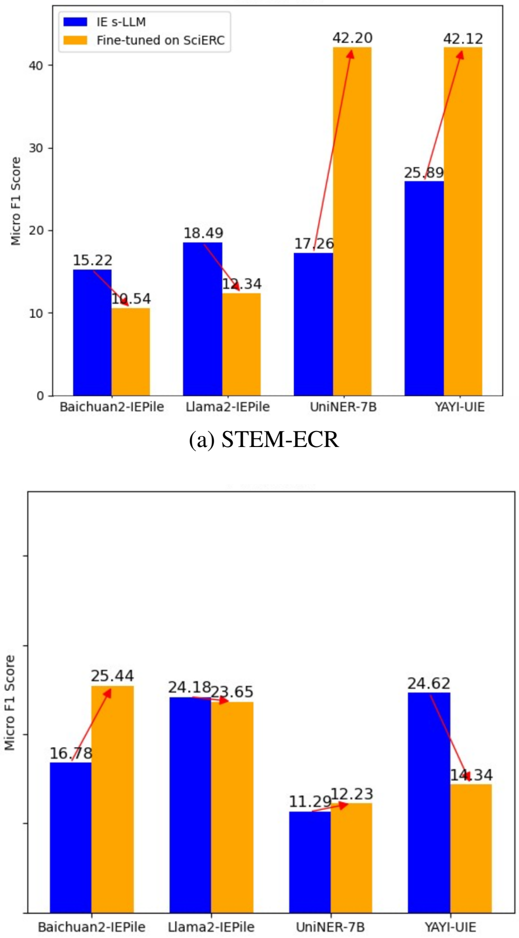


Figure 2: Strict Micro F_1 on NER datasets obtained from IE s-LLMs and fine-tuned on the SciERC dataset.

4.3 Generalisability of Fine-tuned s-LLMs

To assess whether the continual in-domain fine-tuning also leads to generalisable models (to other scientific datasets), we take the IE s-LLM models obtained by continual fine-tuning on the SciERC dataset (`X-SciERC`) and use these models for zero-shot inference on the MeasEval and STEM-

ECR datasets. We choose the SciERC dataset because there is an entity type overlap with the STEM-ECR dataset (‘Material’, ‘Method’), but not with the MeasEval dataset. The results are presented in Figure 2.

The findings indicate that IEPile models fine-tuned in-domain on the SciERC dataset (`X-IEPile-SciERC`) exhibit lower performance on the STEM-ECR dataset, while the UniNER and YAYI-UIE models demonstrate improved performance. The reason behind this might be the knowledge distillation used in the instruction-tuning of UniNER and the larger benchmark used in the tuning of YAYI-UIE and UniNER models. For the MeasEval dataset, the `UniNER-7B-SciERC` model provides a slight improvement, and `Baichuan2-IEPile-SciERC` outperforms the zero-shot `Baichuan2-IEPile`. In contrast, the continually trained YAYI-UIE model yields a performance drop.

From these results, we conclude that the general applicability of the model depends on how close the out-of-domain data is to the data used for continual training. As the SciERC and STEM-ECR entity types share some overlap (being about general concepts relating to the scientific method), we observe better cross-domain effectiveness in UniNER and YAYI-UIE models. In contrast, for the MeasEval dataset, given its particular focus on quantitative measurements, we see no meaningful improvements stemming from out-of-domain training, and, in one case (the YAYI-UIE model), we actually observe a marked performance drop.

4.4 General Domain Evaluation

To assess the generalisability of our findings to domains beyond scientific information extraction, we evaluated s-LLMs using the CrossNER (Liu et al., 2021) and CoNLL2003 (Tjong Kim Sang and De Meulder, 2003) datasets (statistical details are provided in Table 5). ConLL2003 is used for in-domain and CrossNER is used for out-of-domain, as outlined by Wang et al. (2023); Zhou et al. (2024).

As shown in Table 3, s-LLMs outperform their open-source (vanilla) counterparts on the in-domain dataset (CoNLL2003). However, GPT-4 (a proprietary LLM) outperforms these models on the CrossNER dataset. This performance gap may be related to two possible factors: (i) the model’s undisclosed pretraining data, which may include broader coverage of domains or overlap with similar data; and (ii) a similar trend observed in scientific domain evaluations (Section 4.1), where s-LLMs struggle with generalisation to unseen datasets.

These findings show a key limitation of s-LLMs: instruction-tuning improves performance within domains present in the instruction-tuning data; however, it does not guarantee robustness to domain shifts. In contrast, large-scale proprietary LLMs like GPT-4 benefit from diverse pretraining data or emergent generalisation capabilities (although these are difficult to verify given the lack of transparency around the training data and regime).

Model	CoNLL2003	AI	Literature	Music	Politics	Science
Proprietary LLMs						
GPT-4	68.68	**61.95**	52.32	**70.79**	63.99	62.66
Claude 3.5 Sonnet	55.10	32.78	30.18	43.52	45.37	47.12
Open-source LLMs						
Baichuan2	20.50	4.17	12.14	16.89	20.47	8.52
Llama2 7B	17.06	5.19	13.87	17.42	11.96	9.24
Llama2 13B	33.47	13.92	28.92	33.96	36.97	23.85
Llama2 70B	43.39	39.10	40.67	49.30	53.49	39.50
Llama3.1 8B	62.48	40.12	42.17	48.82	30.15	45.12
Llama3.1 70B	70.47	51.42	56.08	64.02	38.34	52.49
IE s-LLMs						
Baichuan2-IEPile	70.41	56.12	50.52	59.18	53.17	55.10
Llama2-IEPile	72.40	53.47	62.15	58.72	55.67	57.68
UniNER-7B	**81.14**	60.25	**62.98**	66.35	**65.30**	**69.23**
YAYI-UIE	78.18	51.60	43.38	61.46	47.43	48.45

Table 3: Zero-shot performance on general domain NER datasets (CoNLL2003 is in-domain; CrossNER is out-of-domain). The best results are **boldfaced**.

4.5 Practical Recommendations

Based on our evaluation of s-LLMs compared to proprietary and open-source LLMs for the scientific domain, we make the following recommendations for practitioners, especially those working in privacy-sensitive or resource-constrained environments in the domain of scientific literature information extraction:

1. **Domain adaptation as a solution for local (open-source) models.** For open-source LLMs, task adaptation (instruction-tuning) is required to enhance the task-specific zero- and few-shot generalisation capabilities of LLMs; i.e., open-source models without it perform poorly, perhaps too poorly for prototyping.
2. **s-LLMs as a starting point for dataset adaptation.** For in-domain adaptation in the scientific domain, starting with an s-LLM that has already adapted to the task yields stronger performance. The prior multi-task training often provides a useful foundation that can

be transferred across domains. On the flip-side, direct instruction tuning of a base open-source model provides limited value (or requires much more training data, see below).

3. **YAYI-UIE demonstrates the best overall performance and generalisation across s-LLMs.** Among s-LLMs, YAYI-UIE achieves the highest and most consistent results after continual in-domain fine-tuning, making it a strong choice for scientific IE applications.
4. **Task adaptation with a larger benchmark.** Gathering in-domain training data and using it for instruction-tuning is still the most effective way of task adaptation. For LLMs, instruction-tuning for task adaptation appears to require a prior step (instruction tuning), as direct in-domain fine-tuning vanilla open-source LLMs appears to yield subpar results.
5. **Smaller PLMs remain viable cost-effective alternatives.** Although s-LLMs offer improved performance, smaller PLMs like BERT can still provide competitive results, if in-domain training data can be sourced. Their lower computational demands make them practical options for projects with limited resources.

5 Conclusion

In this paper, we investigate instruction-tuned IE specialised LLMs (s-LLMs), specifically focusing on their performance in scientific entity extraction compared to open-source and proprietary LLMs. The experimental results show that s-LLMs perform better than their open-source (vanilla) counterparts, showing that instruction-tuning benefits in task-adaptation. However, s-LLMs still face a generalisation problem in the scientific domain. Continual in-domain fine-tuning of IE s-LLMs leads to the best results, particularly for specific scientific datasets of interest. In our experiments, these models outperformed proprietary ones by up to an order of magnitude, achieving over 55 F_1 points in zero-shot and 20 F_1 points in few-shot settings.

We also observe that models like YAYI-UIE perform well across a variety of datasets, highlighting their adaptability to unseen datasets in zero-shot and few-shot settings. However, the choice of s-LLM and its suitability for a given dataset remains a hyperparameter defined in the study. Despite the success of s-LLMs, PLMs (BERT) continue to offer competitive and cost-effective alternatives for NER, particularly when in-domain train data is available, often outperforming open-source LLMS in-domain tuned directly for specific tasks.

This work highlights the strengths and weaknesses of s-LLMs in scientific NER and provides a comparative analysis across zero-shot, few-shot and fine-tuned settings. However, our study is limited in scope: we focused exclusively on sentence-level NER within the scientific domain and relied on publicly available s-LLMs without modifications. As such, the performance and limitations of these models inherently constrain our findings. Additionally, due to resource limitations, we did not evaluate large proprietary LLMs such as GPT-4 or Claude under fine-tuned conditions. We also did not explore the problem of catastrophic forgetting in s-LLMs, which is important to understand how well these models retain knowledge and problem-solving skills learned from previous tasks.

Future work will extend this evaluation to other IE tasks such as relation and event extraction, and investigate how combining the strengths of different s-LLMs (e.g., UniNER's strong zero-shot performance vs. YAYI-UIE's fine-tuning responsiveness) can lead to more robust pipelines. Expanding the diversity and number of datasets may also help in identifying better general-purpose starting points for scientific information extraction.

Limitations

Our study is centred exclusively on the sentence-level Named Entity Recognition (NER) task. Specifically, we concentrate on the scientific domain, which may require further exploration to apply our findings to other domains. Additionally, due to resource constraints, we were unable to fine-tune large language models with more parameters (e.g., GPT-4, Claude). We use the IE s-LLMs provided by the papers. The limitations derived from these models are also limitations of our study.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Online. Accessed: 2024-08-13.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. 2025. Small Language Models are the Future of Agentic AI. *Preprint*, arXiv:2506.02153.

Necva Bölücü, Maciej Rybinski, and Stephen Wan. 2023. impact of sample selection on in-context learning for entity extraction from scientific writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5090–5107, Singapore. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. 33:1877–1901.

Ruichu Cai, Junhao Lu, Zhongjie Chen, Boyan Xu, and Zhifeng Hao. 2025. Handling Missing Entities in Zero-Shot Named Entity Recognition: Integrated Recall and Retrieval Augmentation. 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 10790–10802, Albuquerque, New Mexico. Association for Computational Linguistics.

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 89–97, Boston, Massachusetts, USA. Association for Computational Linguistics.

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.

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.

Jennifer D'Souza, Anett Hoppe, Arthur Brack, Mohmad Yaser Jaradeh, Sören Auer, and Ralph Ewerth. 2020. The STEM-ECR Dataset: Grounding Scientific Entity References in STEM Scholarly Content to Authoritative Encyclopedic and Lexicographic Sources. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2192–2203, Marseille, France. European Language Resources Association.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange. 2020. The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1255–1268, Online. Association for Computational Linguistics.

Jun Gao, Huan Zhao, Yice Zhang, Wei Wang, Changlong Yu, and Ruifeng Xu. 2023. Benchmarking large language models with augmented instructions for fine-grained information extraction. *arXiv preprint arXiv:2310.05092*.

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.

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.

Honghao Gui, Shuofei Qiao, Jintian Zhang, Hongbin Ye, Mengshu Sun, Lei Liang, Jeff Z Pan, Huajun Chen, and Ningyu Zhang. 2024a. Instructie: A bilingual instruction-based information extraction dataset. In *International Semantic Web Conference*, pages 59–79. Springer.

Honghao Gui, Lin Yuan, Hongbin Ye, Ningyu Zhang, Mengshu Sun, Lei Liang, and Huajun Chen. 2024b. IEPile: Unearthing Large Scale Schema-Conditioned Information Extraction Corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 127–146, Bangkok, Thailand. Association for Computational Linguistics.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey Bigham. 2022. InstructDial: Improving Zero and Few-shot Generalization in Dialogue through Instruction Tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 505–525, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Corey Harper, Jessica Cox, Curt Kohler, Antony Scerri, Ron Daniel Jr., and Paul Groth. 2021. SemEval-2021 Task 8: MeasEval – Extracting Counts and Measurements and their Related Contexts. In *Proceedings of* the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 306–316, Online. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

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.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2024. TextEE: Benchmark, Reevaluation, Reflections, and Future Challenges in Event Extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12804–12825, Bangkok, Thailand. Association for Computational Linguistics.

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.

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.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chaitanya Kulkarni, Wei Xu, Alan Ritter, and Raghu Machiraju. 2018. An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 97–106, New Orleans, Louisiana. Association for Computational Linguistics.

Aman Kumar and Binil Starly. 2022. "FabNER": information extraction from manufacturing process science domain literature using named entity recognition. *Journal of Intelligent Manufacturing*, 33(8):2393–2407.

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.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023. CodeIE: Large Code Generation Models are Better Few-Shot Information Extractors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Toronto, Canada. Association for Computational Linguistics.

Zixuan Li, Yutao Zeng, Yuxin Zuo, Weicheng Ren, Wenxuan Liu, Miao Su, Yucan Guo, Yantao Liu, Lixiang Lixiang, Zhilei Hu, Long Bai, Wei Li, Yidan Liu, Pan Yang, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. 2024. KnowCoder: Coding Structured Knowledge into LLMs for Universal Information Extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8758–8779, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A Joint Neural Model for Information Extraction with Global Features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. EmoLLMs: A Series of Emotional Large Language Models and Annotation Tools for Comprehensive Affective Analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 5487–5496, New York, NY, USA. Association for Computing Machinery.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. CrossNER: Evaluating Cross-Domain Named Entity Recognition. volume 35, pages 13452–13460.

Keming Lu, Xiaoman Pan, Kaiqiang Song, Hongming Zhang, Dong Yu, and Jianshu Chen. 2023. PIVOINE: Instruction Tuning for Open-world Entity Profiling. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15108–15127, Singapore. Association for Computational Linguistics.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified Structure Generation for Universal Information Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. 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.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10572–10601, Singapore. Association for Computational Linguistics.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2018. Information extraction from scientific articles: a survey. *Scientometrics*, 117(3):1931–1990.

OpenAI. 2024. Hello gpt-4o. Accessed: 2024-08-13.

Yixin Ou, Ningyu Zhang, Shengyu Mao, Runnan Fang, Yinuo Jiang, Ziwen Xu, Xiaolong Weng, Lei Li, Shuofei Qiao, and Huajun Chen. 2023. EasyInstruct: An Easy-to-use Framework to Instruct Large Language Models. `https://github.com/zjunlp/EasyInstruct`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. 35:27730–27744.

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.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022. LINGUIST: Language Model Instruction Tuning to Generate Annotated Utterances for Intent Classification and Slot Tagging. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 218–241, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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*, CIKM '21, page 4574–4583, New York, NY, USA. Association for Computing Machinery.

Raghav Sharma and Manan Mehta. 2025. Small Language Models for Agentic Systems: A Survey of Architectures, Capabilities, and Deployment Trade offs. *Preprint*, arXiv:2510.03847.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Siddharth Varia, Shuai Wang, Kishaloy Halder, Robert Vacareanu, Miguel Ballesteros, Yassine Benajiba, Neha Anna John, Rishita Anubhai, Smaranda Muresan, and Dan Roth. 2023. Instruction Tuning for Few-Shot Aspect-Based Sentiment Analysis. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 19–27, Toronto, Canada. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting Relation Extraction in the era of Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context Learning for Relation Extraction using Large Language Models. In *Proceedings of the*

2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore. Association for Computational Linguistics.

Jiaqi Wang, Yuying Chang, Zhong Li, Ning An, Qi Ma, Lei Hei, Haibo Luo, Yifei Lu, and Feiliang Ren. 2024. Techgpt-2.0: A large language model project to solve the task of knowledge graph construction. *arXiv preprint arXiv:2401.04507*.

Liwen Wang, Rumei Li, Yang Yan, Yuanmeng Yan, Sirui Wang, Wei Wu, and Weiran Xu. 2022. Instructionner: A multi-task instruction-based generative framework for few-shot ner. *arXiv preprint arXiv:2203.03903*.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named Entity Recognition via Large Language Models. pages 4257–4275.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023. InstructUIE: multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

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, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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.

Xinglin Xiao, Yijie Wang, Nan Xu, Yuqi Wang, Hanxuan Yang, Minzheng Wang, Yin Luo, Lei Wang, Wenji Mao, and Daniel Zeng. 2024. Yayi-uie: A chat-enhanced instruction tuning framework for universal information extraction.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Lin Yuan, Jun Xu, Honghao Gui, Mengshu Sun, Zhiqiang Zhang, Lei Liang, and Jun Zhou. 2025. Improving natural language understanding for llms via large-scale instruction synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25787–25795.

Bowen Zhang and Harold Soh. 2024. Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9820–9836, Miami, Florida, USA. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2025a. Instruction Tuning for Large Language Models: A Survey.

Zikang Zhang, Wangjie You, Tianci Wu, Xinrui Wang, Juntao Li, and Min Zhang. 2025b. A Survey of Generative Information Extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4840–4870, Abu Dhabi, UAE. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A Frustratingly Easy Approach for Entity and Relation Extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. In *The Twelfth International Conference on Learning Representations*.

A Experiments

A.1 Baseline Models

We compare the s-LLMs against two categories of foundation LLMs:

1. **Proprietary LLMs:** We use GPT4 (GPT-4o) (Achiam et al., 2023) and Claude (Claude 3-5 Sonnet) (Anthropic, 2024).

2. **Open-source base LLMs:** We include the open-source (vanilla) counterparts of s-LLMs in our evaluation, including Baichuan2 (Baichuan2-7B-Chat) (Yang et al., 2023), and Llama (Llama2-7B-Chat, Llama2-13B-Chat, Llama2-70B-Chat, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct) (Touvron et al., 2023).

In addition, we also compare the performance of LLMs against a fine-tuned **PLM**, i.e., BERT (Devlin et al., 2019) (BERT-base), which consists of an encoder and a span-based classifier on top of the encoder (Zhong and Chen, 2021).

A.2 Evaluation Metrics

We follow prior studies (Lu et al., 2022; Lin et al., 2020) and use strict entity-level micro-F_1 as our evaluation metric, where both the entity boundary and entity type must be correctly predicted.

A.3 Training Environment

We use NVIDIA H100 GPUs for inference and fine-tuning of open-source LLMs and s-LLMs. Our experiments are conducted on a node with two NVIDIA H100 GPUs.

A.4 Datasets

A.4.1 Scientific Domain Datasets

We use four sentence-level datasets, each with a slightly different focus for scientific IE:

1. **MeasEval**[11] (Harper et al., 2021) is a dataset collected from scientific documents from 10 different domains (e.g., agriculture, chemistry and materials science), annotated for four entity types: Quantity, Measured Property, Measured Entity, Qualifier.

2. **SciERC**[12] (Luan et al., 2018) is a dataset collected from the Artificial Intelligence (AI) domain, describing general AI, NLP, Speech Recognition (SR), Machine Learning (ML), and Computer Vision (CV). The entity types: Generic, Material, Method, Metric, OtherScientificTerm and Task.

3. **STEM-ECR**[13] (D'Souza et al., 2020) is a dataset containing scientific abstracts annotated at the sentence-level, covering ten domains (e.g., agriculture and astronomy). Entity types are Material, Data, Process and Method[14].

4. **WLPC**[15] (Kulkarni et al., 2018) is a dataset of technical writing (as opposed to peer-reviewed scientific publications) collected from wet lab protocols for biology and chemistry experiments, providing entity, relation, and event annotations.

The descriptive statistics of all four datasets are listed in Table 4.

[11] https://github.com/harperco/MeasEval

[12] http://nlp.cs.washington.edu/sciIE

[13] https://data.uni-hannover.de/dataset/stem-ecr-v1-0

[14] Although originally there are 7 entity types, we follow previous work (D'Souza et al., 2020) and leave Task, Object, and Results entity types out.

[15] https://github.com/chaitanya2334/WLP-Dataset

Data Split	MeasEval	SciERC	STEM-ECR	WLPC
# Train	542	1,861	942	8.581
# Dev	155	275	118	2,589
# Test	294	551	118	2,861
# Sentences	991	2,687	1,178	14,301
#Word Count	34,779	65,334	25,968	181,908
# Unique Entity Types	4	6	4	18

Table 4: Statistical details of datasets. "#" denotes the number of samples in the specific dataset.

Characteristics of Datasets We note that the first three of the datasets focus on text found in scientific publications, though the scope of the entity detection may be different. For example, the MeasEval dataset focuses on the general concept of quantitative measurements in empirical investigations (e.g., Measured Property). SciERC and STEM-ECR include a combination of specific concepts from the science disciplines as well as general concepts from the scientific method (e.g., Material, Method), although the publication set of SciERC is narrower than that of STEM-ECR. Finally, the WLPC dataset focuses on experimental reports with entity types that differ from the other datasets (given the physical experiment focus), including measure-based (e.g., Numerical, Generic-Measure, Size, Ph, Measure-Type) and science discipline-specific object entities (e.g., Action, Amount, Location).

Of these datasets, only the SciERC was used in the instruction fine-tuning steps for the three models, as the NER task for the UniNER model and the RE task for the IEPile and YAYI-UIE models (Table 6). That is, the data points for the entities and entity types of MeasEval, STEM-ECR, and WLPC datasets were **not seen** during the initial instruction fine-tuning of the s-LLMs.

A.4.2 General Domain Datasets

The descriptive statistics of general domain datasets are given in Table 5.

Dataset	Domain	Type	# Test
CrossNER Politics	Political	9	650
CrossNER Literature	Literary	12	416
CrossNE Music	Musical	13	465
CrossNER AI	AI	14	431
CrossNER Science	Scientific	17	543
CoNLL2003	News	4	3,453

Table 5: The statistical details of the CrossNER dataset. "#" denotes the number of samples in the specific dataset.

Model	Base LLM	Dataset	# Entity Type
IEPile	Llama2-13B & Baichuan2-13B	FabNER (NER) (Kumar and Starly, 2022)	12
		SciERC (RE) (Luan et al., 2018)	4
		SemEval (RE) (Hendrickx et al., 2010)	-
UniNER-7B	Llama2-7B	WLP (Kulkarni et al., 2018)	16
		SoMeSci (Schindler et al., 2021)	14
		SciREX (Jain et al., 2020)	4
		SciERC (Luan et al., 2018)	4
		SOFC (Friedrich et al., 2020)	3
		FabNER (Kumar and Starly, 2022)	12
		DEAL (Grezes et al., 2022)	30
YAYI-UIE	Baichuan2-13B	FabNER (NER) (Kumar and Starly, 2022)	12
		SciERC (RE)	4

Table 6: Statistical details of scientific datasets used in instruction-tuning of IE s-LLMs. "#" denotes the number of entity types in the entity type set. Details are from Zhou et al. (2024).

A.4.3 Benchmark Datasets

Statistical details of scientific datasets used in instruction-tuning of IE (s-LLMs) are given in Table 6. You can find the complete list of datasets in the respective original papers.

A.5 Models and Fine-tuning

For further supervised fine-tuning (SFT) experiments, we use IE s-LLMs (UniNER, IEPile, YAYI-UIE), which are open-source LLMs instruction-tuned for IE tasks and open-source LLMs. Specifically, we employ LoRA (Hu et al., 2022) for parameter-efficient fine-tuning. We follow the previous works for the hyperparameters of SFT (Gui et al., 2024b; Zhou et al., 2024). We set the LoRA rank and alpha parameters to 16 and 32, respectively. The dropout ratio is set to 0.05. The learning rate is set to 5e-5. We limit the input source length to 400 and the target length to 512. The training epoch size is 10, and the batch size is 2.

Baseline BERT-base PLM is fine-tuned utilising the Hugging Face[16] (Wolf et al., 2020) library. The hyperparameters used in the fine-tuning PLM are the batch size of 32, the max length of 128, the learning rate of 1e-5, and 15 epochs of training.

A.6 Zero-Shot and Few-Shot Settings

We conduct zero-shot and few-shot experiments on open-source and proprietary LLMs using the NER prompt of EasyInstruct[17] (Ou et al., 2023). We use random sampling for a few-shot setting, where we select 1 sample from the train set. We set the temperature to 0.0 for results with less variability and set the top probability to 0.95. We use the original prompt templates used in the training of the respective IE s-LLMs in the experiments with these models to align with the setup of the respective NER-specific training regimes.

[16] https://huggingface.co
[17] https://github.com/zjunlp/EasyInstruct

Prompts We follow EasyInstruct (Ou et al., 2023) in our experiments for open-source and proprietary LLMs. For each dataset, we use its defined entity types and samples (text) from the test set.

IEPile:

```
User: You are an expert in named entity
     recognition. Please extract
    entities that match the schema
    definition from the input. Return
    an empty list if the entity type
    does not exist. Please respond in
    the format of a JSON string.,
    schema: {entity_types}, input: {
    Text}
```

UniNER:

```
User: Text: {Text}
Assistant: I've read this text.
User: What describes {entity_type} in
    the text?
```

YAYI-UIE:

```
User: Text: {Text}
From the given text, extract all the
    entities and types. Please format
    the answer in JSON {{{', '.join(
    entity_types)}: [entities]}}
```

General:

```
User: You are a highly intelligent and
    accurate {domain} domain Named-
    entity recognition(NER) system. You
     take Passage as input and your
    task is to recognize and extract
    specific types of {domain} domain
    named entities in that given
    passage and classify into a set of
    following predefined entity types:
{entity_types}
our output format is only [{'E': type
    of entity from predefined entity
    types, 'W': entity in the input
    text},...] form, no other form.
Input: {Text}
```

Metadata Generation for Research Data from URL Citation Contexts in Scholarly Papers: Task Definition and Dataset Construction

Yu Watanabe[1], Koichiro Ito[1], Shigeki Matsubara[1,2],
[1]Graduate School of Informatics, Nagoya University,
[2]Information Technology Center, Nagoya University,
Correspondence: watanabe.yu.x3@s.mail.nagoya-u.ac.jp

Abstract

This paper proposes a new research task aimed at automatically generating metadata for research data, such as datasets and code, to accelerate open science. From the perspective of "Findable" in the FAIR data principles, research data is required to be assigned a global unique identifier and described with rich metadata. The proposed task is defined as extracting information about research data (specifically, *name*, *generic mention*, and *in-text citation*) from texts surrounding URLs that serve as identifiers for research data references in scholarly papers. To support this task, we constructed a dataset containing approximately 600 manually annotated citation contexts with URLs of research data from conference papers. To evaluate the task, we conducted a preliminary experiment using the constructed dataset, employing the In-Context Learning method with LLMs as a baseline. The results showed that the performance of LLMs matched that of humans in some cases, demonstrating the feasibility of the task.

1 Introduction

Open science is a movement to promote the utilization of research data by making them publicly available (G7 OSWG, 2023). To utilize research data, such as datasets and code, effectively, it is necessary to assign metadata. One solution to accelerate this process is to extract information on research data from texts referring to the data, such as scholarly papers.

The FAIR Guiding Principles (Wilkinson et al., 2016) outlines the criteria for achievement in open science. FAIR stands for "Findable," "Accessible," "Interoperable," and "Reusable." The most fundamental principle is "Findable," and the requirements for research data to be findable are that a unique identifier is assigned and that rich metadata are described. However, no previous study on extracting information about research data from scholarly papers has explicitly considered the above requirements.

This paper proposes a new research task of extracting information about research data from scholarly papers. We define the task based on DataCite (DataCite Metadata Working Group, 2024), a global standard metadata schema. Specifically, the task is defined as extracting information corresponding to *name*, *generic mention*, and *in-text citation* of research data. This information is extracted from the citation context, i.e., the paragraph containing URL citations.

To perform the proposed task, we manually annotated approximately 600 paragraphs of text (citation contexts) containing URLs citing research data from conference papers. We then conducted a preliminary experiment to evaluate our task. In the experiment, we adopted In-Context Learning (ICL) using LLMs as the baseline method and compared it with the performance of humans. The results demonstrated that the performance of LLMs for *generic mention* and *in-text citation* was comparable to that of humans.

2 Extraction of Information about Research Data

2.1 Metadata of Research Data

Research data are data collected or generated through research activities. In this study, data, such as datasets and code, were treated as research data. For research data to meet the most fundamental principle in the FAIR, i.e., "Findable," it is required to be assigned a unique identifier and described with rich metadata.

2.2 Utilization of Scholarly Papers for Metadata Generation

In scholarly papers, information about research data is provided when the data are created or used for a study. When mentioning the created research

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 72–79

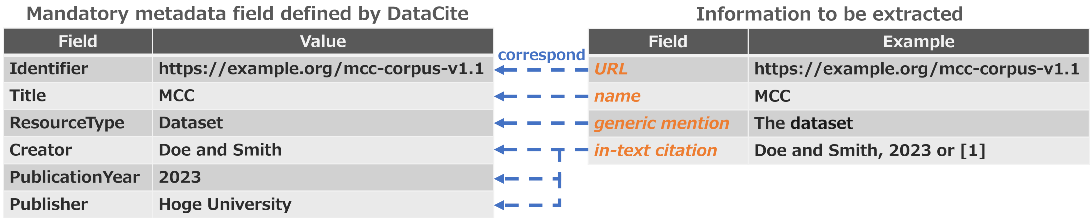


Figure 1: Correspondence between information fields to be extracted and DataCite mandatory metadata fields. Creator, PublicationYear, and Publisher can be obtained not only from the body text of the citing paper, but also from metadata of the cited paper, such as its authors and affiliations. This metadata is often available in the reference list or the header section of the cited paper. *In-text citation* provides access to the cited paper.

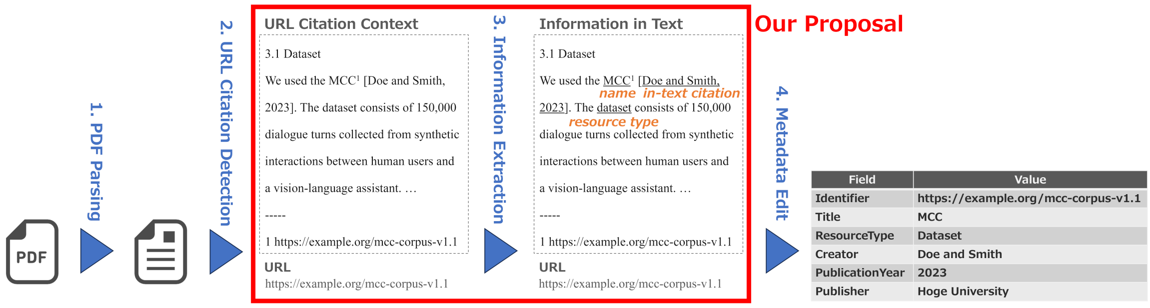


Figure 2: Flow of generating metadata on research data from scholarly papers.

data, the creator notes the name and description of the data and provides access to it, e.g., using a URL. When mentioning the used research data, the user provides the identifier, an overview of the research data, and its usage in the study. This information about research data is often included in the text surrounding mentions of research data in papers. Thus, the information extracted from the text can be used as a source to generate metadata.

2.3 Requirements for Information to be Extracted from Papers

In this study, we assumed metadata generation based on DataCite, which is a global standard and universal metadata schema. Furthermore, DataCite offers an additional advantage of interoperability, as its fields can be mapped to other metadata schemas such as the Dublin Core[1]. DataCite defines a metadata schema with six mandatory fields: "Identifier," "Creator," "Title," "Publisher," "PublicationYear," and "ResourceType." Thus, at minimum, it is necessary to extract the information corresponding to these metadata fields.

Based on the above, we define the following four conditions for the information to be extracted from scholarly papers.

1. Identifier for research data
2. Name of research data
3. Information about the type of research data
4. Information related to the creation of the research data

Figure 1 shows the correspondence between these conditions and mandatory fields in DataCite. The information in the above conditions 1 and 2 can be used for "Identifier" and "Title," respectively. The information in the condition 3 can be used to classify "ResourceType." From the information in the condition 4, it may be possible to generate the "Creator," "Publisher," and "PublicationYear."

2.4 Related Work

Previous studies have tackled the task of extracting information about research data from scholarly papers. Most of these studies extracted information by identifying the names and mentions of research data (Luan et al., 2018; Jain et al., 2020; Schindler et al., 2021; Hou et al., 2021; Pan et al., 2023; Otto et al., 2023; Stavropoulos et al., 2023; Pan et al., 2024; Watanabe et al., 2024). The name and mention detection realized comprehensive extraction of information from scholarly papers, satisfying the condition 2. However, this approach does not necessarily satisfy the condition 1 because it may not include identifiers such as URLs.

[1] http://purl.org/dc/elements/1.1/

Table 1: Fields of information to be extracted.

Field	Explanation
name	name given to research data
generic mention	generic reference to research data
in-text citation	in-text reference marker for research data

In contrast, other studies have obtained information on research data from URL citations in the text of scholarly papers. For example, Tsunokake and Matsubara classified whether URLs in scholarly papers cite research data or not (Tsunokake and Matsubara, 2021). Zhao et al., Tsunokake and Matsubara, and Wada et al. classified types of research data cited by URL using the text surrounding URL citations (Zhao et al., 2019; Tsunokake and Matsubara, 2022; Wada et al., 2024). These studies satisfy the condition 1 because URL is regarded as an identifier. They also satisfy the condition 3 by classifying the type of research data. However, the conditions 2 and 4 are not satisfied because they did not target information excluding URL and type.

3 Extraction from Citation Contexts

3.1 Prerequisites for the Task

The flow of generating metadata on research data from scholarly papers is shown in Figure 2. The procedure is summarized as follows.

1. Parse the paper in PDF format and convert it to semi-structured text.
2. Detect URLs that refer to research data among all URLs in the text and extract segments containing the URLs as body texts.
3. Extract the information about research data from the body text.
4. Edit the extracted strings and generate metadata on the research data.

In the above procedure, step 3 represents the task proposed in this study. A detailed definition of the proposed task is given in Section 3.2. For step 1, several tools have been developed to parse and convert scholarly papers in PDF format to text format (Lopez, 2009; The Apache Software Foundation, 2009; Abekawa and Aizawa, 2016; Mistral AI Team, 2025). Regarding step 2, some URL citations refer to related web pages or scholarly papers rather than the research data.

Table 2: Statistics of the dataset.

Annotation unit		Value
#(paragraph, URL)		601
# span	*name*	571
	generic mention	435
	in-text citation	202

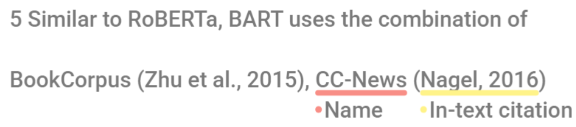


Figure 3: Annotation interface.

To address this issue, we will adapt a previously proposed URL citation classification method (Tsunokake and Matsubara, 2021). Editing in step 4 is left for future work because it requires advanced techniques, e.g., integrating information extracted from multiple papers.

3.2 Task Settings

We define step 3, the proposed task, as follows.

Input: a pair of a URL citing research data and a URL citation context. If the URL appears in a footnote or the bibliography, its text is concatenated to the body text as the input URL citation context.

Output: strings included in the input text corresponding to *name*, *generic mention*, and *in-text citation* of research data. Table 1 explains these three fields of information.

This task takes text with URL citation as input; thus, it satisfies the condition 1. In addition, the information *name*, *generic mention*, and *in-text citation* satisfy the conditions 2 to 4, respectively.

4 Dataset Construction

For the annotation, we used a dataset constructed in a previous study (Tsunokake and Matsubara, 2022) that targets URL citations of research data. This dataset contains URLs citing research data and their corresponding paragraph texts (i.e., citation contexts), extracted from papers published in notable natural language processing conferences[2]. If URLs appeared in footnotes or bibliographies, the corresponding paragraphs were extracted. In this study, we used a total of 601 URL-paragraph pairs, where the URLs refer to datasets or code[3].

[2] https://aclanthology.org/

[3] Whether URLs refers to research data was determined manually in the previous study.

We asked an expert in corpus annotation in NLP to assign information *name*, *generic mention*, and *in-text citation* to paragraphs (if any). Assigning information of URL-cited research data was done by annotating spans and labels. Table 2 shows the statistics of the dataset. We used the doccano (Nakayama et al., 2018) annotation tool, where the worker annotated the text, as shown in Figure 3.

5 Preliminary Experiment

To verify the feasibility of the proposed task, we conducted a preliminary experiment.

5.1 Experimental Data

The constructed dataset was split into training, development, and test data based on the papers' publication years. The test data included paragraphs from papers published in 2021, the latest year in the dataset. The development data were split such that the proportion of publication years was uniform (excluding the test data). The training data were obtained by excluding the test and development data. Finally, the ratio of the training, development, and test data was 397:107:97.

5.2 Extraction Methods

In this experiment, we compared the extraction performance of LLMs with that of humans. For the human extraction, we asked another worker who majored in NLP to extract information.

We adopted ICL (Brown et al., 2020) with LLMs as the baseline method. We set few-shot settings because the performance of LLMs is affected by the given demo samples. The demo samples were selected based on the similarity between the test input text and the candidate texts.

The prompts comprised an instruction, a demonstration, and a test input[4]. The instruction provided the task definition and label in Section 3 and Table 1, respectively. The demonstration included samples retrieved from the training data.

5.3 Implementation and Evaluation

We used Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen3-8B (Qwen Team, 2025) as open LLMs, and GPT-4.1 (OpenAI, 2025) as a closed LLM. In all LLMs, the decoding method was greedy, and the output format was the JSON schema. In the few-shot setting, e5-Mistral-7b-instruct (Wang et al., 2023) retrieved five samples for each input from the training data.

The evaluation was performed for each information field shown in Table 1. We used entity-based F1 as the evaluation metric. Note that we used the edit distance to determine the match between the output and the ground truth because the format of the model output is not span[5].

Table 3: Comparison of extraction performance between LLMs and humans.

	Llama	Qwen	GPT	Human
name	39.33	41.94	45.98	59.86
generic mention	18.63	29.11	42.60	40.40
in-text citation	38.78	39.58	61.33	63.77
Macro average	32.25	36.88	49.97	54.68

5.4 Results

Table 3 shows the performance of all models, alongside the human performance. Overall, the best-performing LLM, GPT-4.1, still underperformed humans by approximately 5 points. Focusing on each information field, both LLMs and the human achieved high performance for *name* and *in-text citation*. Notably, the extraction of *name* showed a gap of approximately 10 points, suggesting that this aspect remains challenging for LLMs. In contrast, for *in-text citation*, GPT-4.1 achieved an F1 of 61.33%, demonstrating extraction performance comparable to that of humans. For *generic mention*, while the open LLMs' performance was lower than the human performance, GPT-4.1 outperformed humans with an F1 of 42.60%.

6 Error Analysis

To reveal the challenges of information extraction, we conducted an error analysis on the outputs of the best-performing model, GPT-4.1, as well as the human. To analyze errors in detail, we introduced the four categories defined in the Message Understanding Conference (Chinchor and Sundheim, 1993).

Correct (COR): both span and label are perfectly matched.

Partial (PAR): span is partially matched, and labels are matched.

Missing (MIS): ground truth is missed by a system.

[4]The prompt is provided in Appendix A.

[5]The Levenshtein distance was used, and a similarity greater than 0.8 was considered a match.

URL	http://opus.nlpl.eu/
Citation Context	... This data is derived from two main sources: (1) open-source repository of parallel corpora, OPUS [Cite_Footnote_3] (Tiedemann, 2012) and (2) ParaCrawl (Esplà et al., 2019). From OPUS, we use the JW300 corpus (Agić and Vulić, 2019), OpenSubtitles (Lison and Tiedemann, 2016), XhosaNavy, Memat, and QED (Abdelali et al., 2014). Despite the existence of this paral-lel data, these text datasets were often collected from large, relatively unclean multilingual corpora, ... *footnote*: 3 http://opus.nlpl.eu/
ground truth	{*name*: "OPUS", *generic mention*: "this paral-lel data", *in-text citation*: "Tiedemann, 2012"}
GPT	{*name*: ["OPUS", "JW300", "OpenSubtitles", "XhosaNavy", "QED", "Memat"], *generic mention*: "paral-lel data", *in-text citation*: "N/A"}
Human	{*name*: "OPUS", *generic mention*: "N/A", *in-text citation*: "Tiedemann, 2012"}

Figure 4: Representative example of the observed error cases. "[Cite_Footnote_3]" denotes a footnote citation tag (which would normally be the number 3).

Table 4: Number of error categories for each information field.

	name		*generic mention*		*in-text citation*	
	GPT	Human	GPT	Human	GPT	Human
COR	37	43	22	17	21	22
PAR	3	1	14	3	2	0
SPU	67	36	68	14	16	11
MIS	27	23	29	45	13	14

Spurious (SPU): a system produces a response that doesn't exist in the ground truth.

As in the experiment in Section 5, we used the edit distance for span matching.

Figure 4 shows a representative example of the observed error cases. In this example, the correct data name is "OPUS" (COR case), but GPT additionally extracted unrelated names such as "JW300" and "OpenSubtitles" (SPU case). For *generic mention*, the human failed to extract "this parallel data" (MIS case), while GPT produced a partial extraction by outputting "parallel data" (PAR case). Regarding *in-text citation*, GPT failed to extract the citation in this example, again resulting in a MIS error.

Table 4 shows the number of error categories for each information field. For *name*, the human produced approximately six more COR cases than GPT, indicating more accurate extraction. In contrast, GPT produced a substantially larger number of SPU cases, suggesting that it is more likely to extracting incorrect information and that its extraction precision remains challenging. For *generic mention*, both GPT and the human yielded far more SPU and MIS cases than COR, demonstrating that extracting this field is generally challenging. Moreover, GPT tends to generate a huge number of incorrect mentions (higher SPU), whereas the human more frequently fail to extract valid mentions (higher MIS). For *in-text citation*, both GPT and the human produced a high number of COR cases, indicating that this information can be extracted reliably by both humans and models.

7 Conclusion

This paper proposed the task of extracting information about research data from URL citation contexts in scholarly papers, and constructed a dataset thorough text annotations according to the DataCite schema. The result of the preliminary experiment demonstrated that the performance of LLMs matched that of humans in some cases, indicating the feasibility of the proposed task.

8 Limitations

Task We defined the output of the task as the information to generate the mandatory metadata fields in the DataCite schema. However, the schema also includes recommended and optional fields, such as "subject" and "size," which could potentially be extracted from scholarly papers. To generate richer metadata, we should expand the scope of the task to cover a wider range of metadata fields.

Dataset The dataset constructed in this study is limited to conference papers in the field of natural language processing and their associated research data. However, research data, such as datasets or code, are also frequently mentioned in papers from diverse domains, specifically digital libraries and medical research. To improve the domain adaptability of the proposed task, we should extend the dataset to cover a broader range of domains.

Experiment The evaluation in this study was designed as a preliminary investigation, and consequently, the reported performance should be considered exploratory. To perform a more comprehensive evaluation of the proposed task, future experiments should be conducted, including evaluations of several supervised approaches.

9 Ethical Considerations

In this project, annotation workers were employed by a staffing agency in Japan. The workers annotated a total of 601 paragraph-URL pairs. Workers were paid approximately 800 yen ($5) per pair.

Acknowledgments

This work was partially supported by the Grant-in-Aid for Scientific Research (B) (No. 23K21844) of JSPS.

References

Takeshi Abekawa and Akiko Aizawa. 2016. SideNoter: Scholarly Paper Browsing System based on PDF Restructuring and Text Annotation. In *Proceedings of the 26th International Conference on Computational Linguistics: System Demonstrations (COLING 2016)*, pages 136–140.

Tom Brown, Benjamin Mann, Nick Ryder, and 1 others. 2020. Language Models are Few-Shot Learners. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 1877–1901.

Nancy Chinchor and Beth Sundheim. 1993. MUC-5 evaluation metrics. In *Proceedings of the 5th Message Understanding Conference (MUC-5)*.

DataCite Metadata Working Group. 2024. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.6.

G7 OSWG. 2023. Annex 1: G7 Open Science Working Group (OSWG).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXive:2407.21783*.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2021. TDMSci: A Specialized Corpus for Scientific Literature Entity Tagging of Tasks Datasets and Metrics. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)*, pages 707–714.

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 (ACL 2020)*, pages 7506–7516.

Patrice Lopez. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *Proceedings of the 13th European Conference on Digital Libraries (ECDL 2009)*, pages 473–474.

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 (EMNLP 2018)*, pages 3219–3232.

Mistral AI Team. 2025. Mistral OCR.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. Doccano: Text Annotation Tool for Human.

OpenAI. 2025. Introducing GPT-4.1 in the API.

Wolfgang Otto, Matthäus Zloch, Lu Gan, Saurav Karmakar, and Stefan Dietze. 2023. GSAP-NER: A Novel Task, Corpus, and Baseline for Scholarly Entity Extraction Focused on Machine Learning Models and Datasets. In *Proceedings of the Findings of the Association for Computational Linguistics (EMNLP 2023)*, pages 8166–8176.

Huitong Pan, Qi Zhang, Cornelia Caragea, Eduard Dragut, and Longin Jan Latecki. 2024. SciDMT: A Large-Scale Corpus for Detecting Scientific Mentions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14407–14417.

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.

Qwen Team. 2025. Qwen3 Technical Report. *arXiv preprint arXive:2505.09388*.

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 (CIKM 2021)*, page 4574–4583.

Petros Stavropoulos, Ioannis Lyris, Natalia Manola, Ioanna Grypari, and Haris Papageorgiou. 2023. Empowering Knowledge Discovery from Scientific Literature: A novel approach to Research Artifact

Analysis. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 37–53.

The Apache Software Foundation. 2009. Apache PDFBox.

Masaya Tsunokake and Shigeki Matsubara. 2021. Classification of URLs Citing Research Artifacts in Scholarly Documents based on Distributed Representations. In *Proceedings of 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE 2021)*, pages 20–25.

Masaya Tsunokake and Shigeki Matsubara. 2022. Classification of URL Citations in Scholarly Papers for Promoting Utilization of Research Artifacts. In *Proceedings of the 1st Workshop on Information Extraction from Scientific Publications (WIESP 2022)*, pages 8–19.

Kazuhiro Wada, Masaya Tsunokake, and Shigeki Matsubara. 2024. On an Intermediate Task for Classifying URL Citations on Scholarly Papers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics and Language Resources and Evaluation (LREC-COLING 2024)*, pages 12359–12369.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving Text Embeddings with Large Language Models. *arXiv preprint arXiv:2401.00368*.

Yu Watanabe, Koichiro Ito, and Shigeki Matsubara. 2024. Capabilities and Challenges of LLMs in Metadata Extraction from Scholarly Papers. In *Proceedings of the 26th International Conference on Asia-Pacific Digital Libraries (ICADL 2024)*, volume 1, page 280–287.

Mark Wilkinson, Michel Dumontier, IJsbrand Aalbersberg, and 1 others. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3(160018).

He Zhao, Zhunchen Luo, Chong Feng, Anqing Zheng, and Xiaopeng Liu. 2019. A Context-based Framework for Modeling the Role and Function of On-line Resource Citations in Scientific Literature. 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 2019)*, pages 5206–5215.

A Prompt

In the experiment, we employed the chat templates defined by each LLM's tokenizer. Figure 5 shows the prompt constructed using the Llama chat template. For Llama and Qwen, we embedded the target information fields into the prompt using function calling. For GPT-4.1, we incorporated them using the response format.

```
#system
Your task is to extract information about research data cited by the given URL
    from section title, body text and footnote/reference.

You have access to the following functions. To call a function, please respond
    with JSON for a function call.
Respond in the format
{
  "name": "information_extraction",
  "description": "Your task is to extract information about research data cited
      by the given URL from section title, body text and footnote/reference.",
  "parameters": {
    "title": "InfoSchema","type": "object",
    "properties": {
      "name": {"title": "Name",
        "type": "array","items": { "type": "string" },
        "description": "A name or title by which the research data is known. May
             be the title of a dataset or the name of a piece of software or an
            instrument. If no names are given, return N/A"
      },
      "genericmention": {"title": "Genericmention",
        "type": "array","items": { "type": "string" },
        "description": "Generic mention refers to a common noun phrase that
            references the research data. If no generic mentions are given,
            return N/A"
      },
      "citationtag": {"title": "Citationtag",
        "type": "array","items": { "type": "string" },
        "description": "Citation tag is a tag that indicates the citation of a
            scholarly paper related to research data. If no citation tags are
            given, return N/A."
      }
    },
    "required": ["name","genericmention","citationtag"]
  }
}

#demonstration
{pairs of demo input and demo output}

#user
Given URL: https://example.org/mcc-corpus-v1.1
Section Title: 3.1 Dataset
Body Text: We used the MCC[Cite_Footnote_1] (Doe and Smith, 2023). The dataset
    consists of 150,000 dialogue turns collected from synthetic interactions
    between human users and a vision-language assistant. ···
Footnote or Reference Text: 1 https://example.org/mcc-corpus-v1.1
```

Figure 5: A simplified version of the used prompt.

Dynamic Reference Extraction and Linking across Multiple Scholarly Knowledge Graphs

Nicolau Duran-Silva[1,2], Pablo Accuosto[1],
[1]SIRIS Lab, Research Division of SIRIS Academic, Barcelona, Spain,
[2]LaSTUS Lab, TALN Group, Universitat Pompeu Fabra, Barcelona, Spain,

Abstract

References are an important feature of scientific literature; however, they are unstructured, heterogeneous, noisy, and often multilingual. We present a modular pipeline that leverages fine-tuned transformer models for reference location, classification, parsing, retrieval, and reranking across multiple scholarly knowledge graphs, with a focus on multilingual and non-traditional sources such as patents and policy documents. Our main contributions are: a unified pipeline for reference extraction and linking across diverse document types, openly released annotated datasets, fine-tuned models for each subtask, and evaluations across multiple scholarly knowledge graphs, enabling richer, more inclusive infrastructures for open research information.

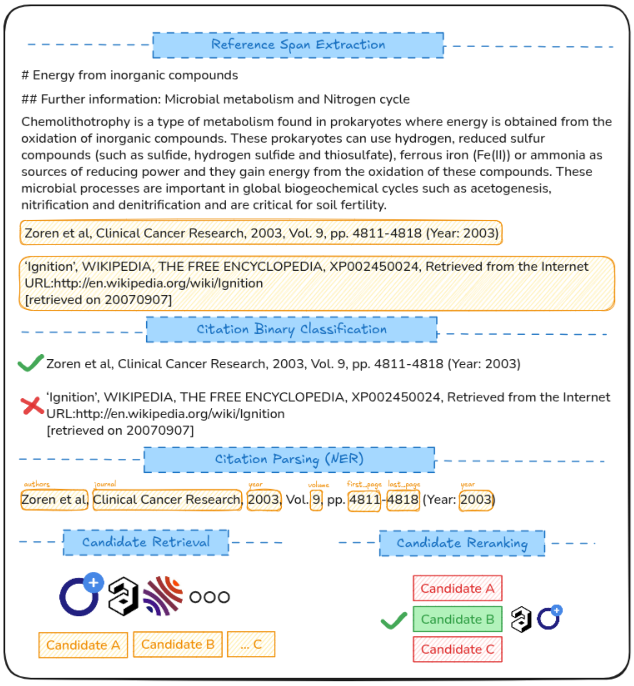


Figure 1: Overview of the pipeline and subtasks.

1 Introduction

Citations and references have been described as one of the most important features of scientific literature (Backes et al., 2024). They ground claims and reference previous work, connect research across disciplines, form the basis for the construction of scholarly knowledge graphs (SKGs), and enable bibliometrics and research impact evaluation and assessment (Leydesdorff et al., 2013; Cioffi and Peroni, 2022; Tkaczyk et al., 2018). Beyond scholarly articles, the number of documents that contain references to scientific work is increasing rapidly, ranging from project proposals, narrative CVs, patents, policy documents and public uses, and even social media and news (Lin et al., 2023; Cong et al.). In the context of open research information and open science, finding and linking references in multi-source documents is crucial for creating richer datasets and infrastructures.

However, extracting references from such diverse sources remains a challenge. Raw references appear in different citation styles (Tkaczyk et al., 2018), are often noisy or incomplete (missing DOI, title, or authors), and occur in multiple languages. Moreover, no single SKG offers complete coverage, making robust research object normalisation non-trivial.

Extraction and linking of scholarly references is an information extraction problem, and a key task of scholarly document processing (Backes et al., 2024). In an era of fake news and LLM hallucinations, research and new tools for grounding references and finding background support are fundamental. Existing tools focus mainly on parsing PDF articles. Although effective in controlled settings, they remain limited and are not very flexible in more diverse settings of references and document types. Recent experiments with LLMs (Backes et al., 2024) have shown mixed results, and previous research has underscored that deep-learning citation-parsing tools suffer from a lack of training data (Grennan and Beel, 2020).

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 80–86
December 23, 2025 ©2025 Association for Computational Linguistics

In this work, we explore encoder-based language models for reference extraction and linking across multiple SKGs. We present a unified pipeline that combines reference location, reference parsing, retrieval, and re-ranking, and introduce ensemble-based linking to improve robustness across OpenAlex, OpenAIRE, CrossRef, and PubMed. To support this, we release new annotated datasets and fine-tuned models for each subtask, together with benchmark results demonstrating their effectiveness in multilingual and noisy-document settings. These resources enable reference extraction not only from scholarly articles but also from non-traditional sources, broadening the scope of SKG construction and downstream applications.

We have released our code, datasets and models fine-tuned in the context of this paper [1].

2 Related Work

A wide range of tools have been proposed for locating and parsing bibliographic references from PDF versions of scholarly articles (Cioffi and Peroni, 2022). Methods have relied on rule-based methods or shallow machine-learning approaches such as CRFs or SVMs (Zou et al., 2010; Tkaczyk et al., 2018), with widely used tools like ParsCit, AnyStyle, GROBID, CERMINE, Scholarcy, and Science Parse. Cioffi et al. (Cioffi and Peroni, 2022) differentiate between tools that can parse a single reference, those for parsing a list of references, and frameworks for parsing references from PDFs. Recent surveys (Backes et al., 2024; Cioffi and Peroni, 2022) report that GROBID and AnyStyle remain strong baselines, but also highlight that most tools focus on parsing rather than full extraction and linking, are restricted to a single database, and offer limited multilingual support. In addition, deep-learning approaches have been hindered by the lack of large annotated datasets (Grennan and Beel, 2020), and LLM-based attempts show mixed results (Backes et al., 2024). Biblio-Glutton (bib, 2018–2024) offers an open framework for reference resolution against authoritative records such as CrossRef, PubMed, HAL, and Unpaywall. While highly effective for processing scholarly articles, it remains tied to specific sources. In contrast, we explore encoder-based models designed to handle more diverse document types and reference settings.

[1] https://github.com/sirisacademic/references-tractor/

3 Materials and Methods

The modular pipeline comprises five steps (subtasks) to extract and link references, which are described below:

1. **Reference Location**: detect citation-bearing spans in raw documents (policy reports, patents, scholarly works, blogs), marking both the broader *citation-span* and the inline *citation-ref*, *author(s)*, *year*, and *citation-ID* (e.g., "(Smith et al., 2019)" or "[12]").

2. **Reference Classification**: the task of classifying citation-like text segments as academic references (e.g., journal articles, scholarly books, conference papers) or non-academic references (e.g., web pages, patents, generic abstracts). It is a binary classification that filters citations to scholarly works from other raw reference data, relevant for heterogeneous sources that cite a diverse set of documents.

3. **Reference Parsing (NER)**: a Named Entity Recognition (NER) model extracts key fields from the citation, parsing it into structured fields using a fine-tuned NER model. The extracted fields can include `TITLE`, `AUTHORS`, `VOLUME`, `ISSUE`, `YEAR`, `DOI`, `ISSN`, `ISBN`, `FIRST_PAGE`, `LAST_PAGE`, `JOURNAL`, and `EDITOR`.

4. **Reference Retrieval**: parsed fields are used to dynamically build queries to scholarly APIs.

5. **Reference Pairwise Reranking**: re-ranks pairs of the input reference and retrieved candidates from scholarly knowledge graphs.

3.1 Datasets

To support each component of the pipeline, we created five supervised datasets that cover the key subtasks: reference location, reference classification, reference parsing, pairwise reranking, and end-to-end multi-SKG linking. Table 1 provides an overview.

Dataset	Labels	Samples
Reference Location	5	1,922
Reference Classification	2	3,999
Reference Parsing (NER)	12	2,688
Reference Reranking	2	3,276
MultiSKG Linking	–	200

Table 1: Datasets overview.

Reference Location Dataset represents 1,922 annotated text segments from policy documents, patents, websites, news, and scientific papers, in both plain text and markdown formats. Each segment was manually annotated with the full citation span and the inline citation expression, enabling extraction of the reference span and its in-text context, including citation ID, year, and author mentions.

Reference Classification Dataset addresses the filtering step that separates scholarly citations from other sequences. We sample ∼5k non-patent literature entries from the PATSTAT database, covering common `NPL_TYPE` categories (`a`: unspecified, `b`: book, `s`: serial/journal, `w`: web). Each string is labeled TRUE (academic: journal article, scholarly book, conference paper, etc.) or FALSE (non-academic: web pages, office actions, manuals, etc.). Annotation follows a semi-supervised procedure: GPT-3.5 produces initial pseudo-labels, which we compare with the raw categories; we then split the corpus into two folds for cross pseudo-labelling, and human annotators resolve disagreements in Argilla (Daniel and Francisco, 2023) (see Appendix A, *Binary Classification prompt*). The final dataset is multilingual (mainly `en` and `zh`) and approximately balanced (55% TRUE, 45% FALSE), with a train/test split of 90/10.

Reference Parsing (NER) Dataset consists of 2,688 raw citation strings annotated with entity labels: `TITLE`, `AUTHORS`, `VOLUME`, `ISSUE`, `YEAR`, `DOI`, `ISSN`, `ISBN`, `FIRST_PAGE`, `LAST_PAGE`, `JOURNAL`, and `EDITOR`. The samples were gathered from non-patent literature entries in the PATSTAT database to ensure coverage of different citation formats and degrees of metadata completeness. The dataset is multilingual and was annotated following a semi-supervised approach. Pseudo-labels were generated with GPT-3.5 and refined by human annotators with Argilla (see Appendix A, *Reference Parsing (NER) prompt*).

Reference Pairwise Reranking Dataset provides 3,276 reference pairs. Each example is a pair of strings—raw reference and candidate—where the candidate is an APA-normalised reference constructed from OpenAlex metadata (authors, year, title, venue, volume, pages, DOI). Labels are binary (`1`=*same*; `0`=*different*). The corpus was built in two steps: (i) manual annotation of 1,276 candidate pairs to collect positive and hard negative examples, and (ii) to improve generalisation, we synthesise hard negatives by crossing citations with non-matching candidates.

MultiSKG Linking Dataset serves as a gold standard for end-to-end linking to multiple Scholarly Knowledge Graphs, we considered: OpenAlex (Priem et al., 2022), OpenAIRE (Manghi et al., 2012), CrossRef, and PubMed.[2] The dataset consists of 200 manually annotated references to the four target knowledge graphs providing unique identifiers for each source. Two annotators cross-annotated all references. Samples in the dataset vary in complexity, from well-structured to minimal metadata, including ambiguous and hard-to-match references, to evaluate real-world diversity.

3.2 Models & Training

We fine-tune transformer encoder models (Vaswani et al., 2017; Devlin et al., 2019), with model choices guided by baseline models (BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DeBERTa-v3 (He et al., 2020), ModernBERT (Warner et al., 2025)), multilingual coverage (mBERT (Pires et al., 2019), XLM (Lample and Conneau, 2019)), and efficiency to support large-scale runs (multilingual DistilBERT (Sanh et al., 2019)). Our models are fine-tuned using the Hugging Face `Transformers` library, with early stopping and model selection based on validation performance. Hyperparameter configurations for each subtask (classification, NER, reranking) are reported in Appendix B.

3.3 Candidate Retrieval & Selection

The candidate retrieval component builds structured queries from the parsed citation fields and issues them to multiple scholarly knowledge graph APIs. Our approach includes:

- **Incremental metadata search:** Queries are constructed progressively, starting from high-confidence fields (e.g., DOI, title + year) and falling back to partial metadata combinations (e.g., authors + venue, title substrings) when primary identifiers are missing. We address this with multi-API retrieval, querying OpenAlex, OpenAIRE, Crossref, PubMed, and HAL, each offering different coverage, domain focus, and search capabilities.

- **Candidate reranking:** Retrieved candidates are scored with a fine-tuned pairwise model

[2] OpenAlex: `https://openalex.org/`, OpenAIRE: `https://explore.openaire.eu/`, Crossref: `https://www.crossref.org/`, and PubMed: `https://pubmed.ncbi.nlm.nih.gov/`.

(Section 3.1), which takes the raw reference and a candidate record as input and predicts whether they refer to the same publication. This learned approach combines lexical cues (title, authors, venue, year) with semantic similarity from transformer encoders, and the prediction score is used for reranking.

- **Ensemble linking:** After reranking, top-scoring candidates are cross-compared across APIs. When DOIs are present, we perform a majority-vote consensus to mitigate single-API inconsistencies and maximise coverage.

4 Evaluation

4.1 Experimental Setup

Our datasets, described in Section 3.1, were split 80/10/10 into train, development, and test. Models were fine-tuned as described in Section 3.2. We report results using macro-F1, computed on the held-out test split. For NER tasks, we compute token-level F1 scores on entity spans. For reference linking, we evaluate on the MultiSKG dataset by requiring exact DOI/ID matches as correct.

4.1.1 Task-level Evaluation

Model	Location	Classification	Parsing	Reranking
DistilBERTm	.755	.935	.949	.904
BERTm-base	.773	**.944**	.957	.902
RoBERTa-base	.788	.940	**.962**	**.915**
XLM-base	–	.914	.957	.901
DeBERTa-v3-base	**.792**	.932	.961	.903
ModernBERT	.732	.936	.955	**.915**

Table 2: Task-level results across models (macro-F1).

We first evaluate each subtask independently. Table 2 shows that RoBERTa and DeBERTa-v3 achieve consistently strong performance across NER and reranking, while BERTm provides the best overall performance on the classification task. DistilBERT offers competitive results with lower computational cost, and XLM demonstrates robust multilingual generalisation.

4.2 Linking Evaluation

We evaluate per-API accuracy with an error breakdown. As shown in the results in Table 3, the ensemble achieves the highest accuracy.

API	Accuracy	C_Match	I_Miss	I_Match
OpenAlex	.745	127	15	30
OpenAIRE	.675	105	19	34
PubMed	.590	48	12	5
CrossRef	.640	104	23	39
Ensemble	**.755**	122	24	19

Table 3: Linking evaluation results, reporting accuracy on strict DOI/ID match. Error breakdown as C_Match (correct matches), C_NoRes (correct empty), I_Miss (missed matches), and I_Match (incorrect matches).

5 Discussion

While overall performance across the subtasks is strong, the linking evaluation reveals several ambiguous cases that complicate strict accuracy metrics. Many of the errors occur during the reranking step and are actually ambiguous matches: although correct DOIs are often retrieved, metadata mismatches (e.g., page ranges, abbreviated venues, missing affiliations) can lead to false negatives. For example, "*Yamagishi et al., J. Phycol. 43: 519–527 (2007)*" illustrates how strict page-number matching can cause the reranker to fail, even when the DOI is correct. Additional errors arise from different versions or duplicate entries with different unique IDs, suggesting that recall-based evaluation might better reflect the system's performance. Some errors are due to partial parsing, while others are caused by missing records in certain SKGs. While the pipeline's true impact lies in its ability to handle cross-database complexities, improving the reranking step would result in better handling of ambiguous matches.

6 Conclusions

We propose a novel pipeline for multilingual reference extraction and linking, using fine-tuned transformer models to enhance scholarly knowledge graph coverage. The approach combines transformer models, incremental retrieval, and ensemble reranking for robust performance in noisy, multilingual settings. We aim to create open citation datasets from policy documents and patents, and expand linking to national and discipline-specific SKGs. Future work will focus on scaling for larger datasets and exploring span-based techniques and long-context models to improve citation extraction from lengthy documents, broadening its applicability to open research infrastructures.

7 Limitations and Future Work

While the proposed pipeline demonstrates strong performance across individual subtasks, several limitations guide ongoing development. The datasets are relatively small (1,922 spans for location, 2,688 for NER, 3,276 for reranking, and 200 gold references for multi-KG linking) and rely on semi-supervised annotation with GPT-based pseudolabels and human adjudication. Larger, more diverse datasets with reported inter-annotator agreement are needed to strengthen claims of generalization across domains and languages.

Our end-to-end evaluation uses strict DOI/ID matching on a limited multilingual sample. As discussed in Section 5, many errors arise from metadata inconsistencies across knowledge graphs rather than true matching failures. Future work should incorporate relaxed matching criteria, additional metrics (top-k recall, MRR), and systematic comparison with established reference extraction systems on standardized benchmarks.

The pipeline assumes text or markdown input and does not explicitly handle PDF layout or OCR errors, which limits applicability to certain document types. Integration with PDF extraction tools would broaden the scope. Additionally, the reranking component could be improved to better handle metadata ambiguity (abbreviated venues, page range variations) through fuzzy matching and multi-field attention. Finally, explicit mechanisms for detecting potentially fabricated or hallucinated references would strengthen the system's reliability.

All models, code, and datasets are openly available, and ongoing experiments will be progressively added to the project repository.

Acknowledgments

Supported by the Industrial Doctorates Plan of the Department of Research and Universities of the Generalitat de Catalunya, by Departament de Recerca i Universitats de la Generalitat de Catalunya (ajuts SGR-Cat 2021). This work was co-funded by the EU HORIZON SciLake (Grant Agreement 101058573). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the funders.

References

2018–2024. biblio-glutton. *Preprint*, swh:1:dir:a5a4585625424d7c7428654dbe863837aeda8fa7.

Tobias Backes, Anastasiia Iurshina, Muhammad Ahsan Shahid, and Philipp Mayr. 2024. Comparing free reference extraction pipelines. *International Journal on Digital Libraries*, 25(4):841–853.

Alessia Cioffi and Silvio Peroni. 2022. Structured references from pdf articles: assessing the tools for bibliographic reference extraction and parsing. In *International Conference on Theory and Practice of Digital Libraries*, pages 425–432. Springer.

Ting Cong, Er-Te Zheng, Zekun Han, Zhichao Fang, and Rodrigo Costas. Social media uptake of scientific journals: A comparison between x and wechat. *Journal of Information Science*, page 01655515251359759.

Vila-Suero Daniel and Aranda Francisco. 2023. Argilla - Open-source framework for data-centric NLP.

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.

Mark Grennan and Joeran Beel. 2020. Synthetic vs. real reference strings for citation parsing, and the importance of re-training and out-of-sample data for meaningful evaluations: Experiments with GROBID, GIANT and CORA. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pages 27–35, Wuhan, China. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Loet Leydesdorff, Ismael Rafols, and Chaomei Chen. 2013. Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal–journal citations. *Journal of the American society for Information science and Technology*, 64(12):2573–2586.

Zihang Lin, Yian Yin, Lu Liu, and Dashun Wang. 2023. Sciscinet: A large-scale open data lake for the science of science research. *Scientific Data*, 10(1):315.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Paolo Manghi, Nikos Houssos, Marko Mikulicic, and Brigitte Jörg. 2012. The data model of the openaire scientific communication e-infrastructure. In *Research Conference on Metadata and Semantic Research*, pages 168–180. Springer.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Dominika Tkaczyk, Andrew Collins, Paraic Sheridan, and Joeran Beel. 2018. Evaluation and comparison of open source bibliographic reference parsers: A business use case.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Jie Zou, Daniel Le, and George R Thoma. 2010. Locating and parsing bibliographic references in HTML medical articles. *International Journal on Document Analysis and Recognition (IJDAR)*, 13(2):107–119.

A Pseudo-Annotation Prompts

Binary Classification prompt

```
Given a piece of text, classify it into one of the
      following categories:

- TRUE: include publications, review, "academic" book
       chapters, conference papers.
- FALSE: references to other types of sources

Instructions: Analyze the content of the provided
      text and assign the appropriate category,
      returning TRUE or FALSE

Examples of TRUE:
- Matthews et al., Homeostasis Model Assessment:
      Insulin Resistance and Beta-cell Function from
      Fasting Plasma Glucose and Insulin
      Concentrations in Man, Diabetologia, 28, (1985),
       pp. 412-419.
- Liu, et al.; Design of carbonylative polymerization
       of heterocycles. Synthesis of polyesters and
      poly(amide-block-ester)s J. Am. Chem. Soc. 2004,
       vol. 126, pp. 14716-14717; 6 pages.
- JULIAN FIERREZ-AGUILAR ET AL: 'Incorporating Image
      Quality in Multi-algorithm Fingerprint
      Verification', 1 January 2005, ADVANCES IN
      BIOMETRICS LECTURE NOTES IN COMPUTER SCIENCE;;
      LNCS, SPRINGER, BERLIN, DE, PAGE(S) 213 - 220,
      ISBN: 978-3-540-31111-9, XP019026878
- KUANG-HUA CHANG: 'E-Design computer-Aided
      Engineering Design', 2015, ELSEVIER ACADEMIC
      PRESS
- HWANG J S ET AL: 'Heteroepitaxy of gallium nitride
      on (0001), (1012) and (1010) sapphire surfaces',
       JOURNAL OF CRYSTAL GROWTH, ELSEVIER, AMSTERDAM,
       NL LNKD- DOI:10.1016/0022-0248(94)90263-1, vol.
       142, no. 1-2, 1 September 1994 (1994-09-01),
      pages 5 - 14, XP024439721, ISSN: 0022-0248, [
      retrieved on 19940901]
- KUMA HIROYUKI ET AL: 'Liquid phase immunoassays
      utilizing magnetic markers and SQUID
      magnetometer', CLINICAL CHEMISTRY AND
      LABORATORY MEDICINE, vol. 48, no. 9, 1 January
      2010 (2010-01-01), DE, XP055783197, ISSN:
      1434-6621, Retrieved from the Internet <URL:
      http://dx.doi.org/10.1515/CCLM.2010.259> DOI:
      10.1515/CCLM.2010.259
- PAN ET AL.: 'Sustainable production of highly
      conductive multilayer graphene ink for wireless
       connectivity and loT applications', NATURE
      COMMS, vol. 9, 2018, pages 5197
- LELOIR, L.F., ARCH BIOCHEM, vol. 33, no. 2, 1951,
      pages 186 - 90

Examples of FALSE:
- Final Office Action, U.S. Appl. No. 13/316,351,
      dated Jul. 31, 2013, 20 pages.
- U.S. Appl. No. 13/006,270, filed Jan. 13, 2011 Non-
      Final Office Action dated Sep. 12, 2014, 41
      pages.
- Matrx Metalloproteinase, from Wikipedia,the free
      encyclopedia (8 pages), retrieved from the
      Internet on Dec. 17, 2009 at http://en.
      wikipedia.org/wiki/Matrix-metalloproteinase.
- 'Double Layer DVD+R Multi-Media Command Set
      Description, Version 1.00', 4 June 2004, ROYAL
      PHILIPS ELECTRONICS, EINDHOVEN, THE NETHERLANDS,
       XP002386267
- 'The Leukocyte Antigen Facts Book', 1997, HARCOURT
      BRACE & CO.
- DOUGLAS GRAHAM: 'Folding a bandana into fade mask',
       6 April 2020 (2020-04-06), XP055859991,
      Retrieved from the Internet <URL:https://www.
      youtube.com/watch?v=dI3343Gb9YA> [retrieved on
      20211110]
- Banknote Paper', WEBPAGES G&D, pages 9PP,
      XP055351061, Retrieved from the Internet <URL:
      https://www.gi-de.com/en/products_and_solutions/
      products/banknote_paper/banknote-paper.jsp>
- PHILIPS: 'Fallback mode for Rel-7 FDD MIMO scheme',
       3GPP TSG RAN WG1 MEETING #46 TDOC R1-061952

Predict the category for this text:
{INPUT_TEXT}
```

Reference Parsing (NER) prompt

```
Can you parse this citation string:
"{INPUT_TEXT}"

in the following attributes:
- authors
- title
- editor
- volume
- issue
- publication date
- publisher
- journal
- first_page
- last_page
- doi
- isbn
- issn
- link online

Only return attributes in bullet points with a not
      empty value
```

B Fine-tuning hyperparameters

B.1 Text Classification

We fine-tune transformer encoder models for the **Reference Classification** task by adding a classification head with two output labels, implemented with HuggingFace `Transformers`. Each model was trained on a single NVIDIA A100 GPU for up to 6 epochs with early stopping (patience 2) with main hyperparameters described in Table 4.

Hyper-parameter	**Value**
Learning Rate	2e-5
Learning Rate Decay	Linear
Weight Decay	0.01
Warmup Steps	0
Batch Size	32
Max. Training Epochs	6
Metric for best model	F1-macro

Table 4: Fine-tuning hyperparameters for the Reference Classification task.

B.2 NER

For the **Reference Location** and **Reference Parsing** tasks, we fine-tune transformer encoder models with a token classification head, using subword-level alignment. All models were trained on a single NVIDIA A100 GPU with early stopping (patience 2). Table 5 summarises the main hyperparameters.

Hyper-parameter	**Value**
Learning Rate	2e-5
Learning Rate Decay	Linear
Weight Decay	0.01
Warmup Steps	0
Batch Size	32
Max. Training Epochs	25
Max Sequence Length	512
Metric for best model	F1
Early Stopping Patience	2

Table 5: Fine-tuning hyperparameters for the Reference Location and Reference Parsing tasks.

B.3 Pairwise Reranking

For the **Pairwise Reranking** task, the goal is to classify pairs of references (`reference1`, `reference2`) as either referring to the same publication (`1`) or different publications (`0`). Each pair is encoded as a single sequence by concatenating the two reference strings with a special separator token (`[SEP]`). We fine-tune transformer encoder models with a sequence classification head (two output labels). Models were trained on a single NVIDIA A100 GPU with early stopping (patience 2). Table 4 reports the training hyperparameters.

AI for Data Ingestion into IPAC Archives

Nicholas Susemiehl and Joseph Mazzarella
IPAC, California Institute of Technology, 1200 E. California Blvd., Pasadena, CA 91125, USA

Abstract

The data archives at IPAC, including the NASA Extragalactic Database (NED) and NASA Exoplanet Archive (NEA), have served as repositories for data published in the astronomical literature for decades. Throughout this time, extracting data from journal articles has remained a challenging task and future large data releases will exasperate this problem. We seek to accelerate the rate at which data can be extracted from journal articles and reformatted into database load files by leveraging recent advances in natural language processing enabled by AI. We are developing a new suite of tools to semi-automate information retrieval from scientific journal articles. Manual methods to extract and prepare data, which can take hours for some articles, are being replaced with AI-powered tools that can compress the task to minutes. A combination of AI and non-AI methods, along with human supervision, can substantially accelerate archive data ingestion. Challenges remain for improving accuracy, capturing data in external files, and flagging issues such as mislabeled object names and missing metadata.

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 which have served the scientific community since 1990 and 2011 respectively. NED has collected over 1.1 million distinct objects, including galaxies, quasars, and gamma ray bursts. NEA seeks to provide a complete list of confirmed exoplanets, which now number over 6,000, and their stellar hosts. New data flow through similar pipelines for both NED and NEA as they are prepared for ingestion into the archives' internal databases. Newly-published articles are found via queries to the listing services of the Astrophyiscs Data System (ADS)[3]. These articles are then fed through a relevance classification model, which seeks to predict whether or not the data from an article should be ingested into the archive. A scientist then selects the relevant papers from a user interface displaying the relevance classifier results. Next, the appropriate data is extracted from the article and transformed into the particular load file formats for NED and NEA before being ingested into the databases. Throughout most of the history of these archives, the data extraction and load file creation process has been done manually, largely because astronomical journal articles vary widely in structure and semantics. While this manual process has been functional, both archives currently have backlogs of unprocessed published journal articles and keeping up with newly published literature can be difficult. To add to this, anticipated exoplanet candidate detection yields from future missions have the potential to substantially increase NEA's holdings. Data releases from missions such as Gaia (Perryman et al., 2014), Roman (Penny et al., 2019; Wilson et al., 2023), PLATO (Matuszewski et al., 2023), and Earth 2.0 (Ge et al., 2024) include estimated yields of thousands to hundreds of thousands of candidate exoplanets. 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 has yielded tools that are increasingly-capable of mining data from the text of scientific journal articles. This work initially investigated Word2Vec (Mikolov et al., 2013) and its extension Doc2Vec (Le and Mikolov, 2014). Word2Vec/Doc2Vec have largely been improved upon by transformer-architecture large language models (LLMs), which use attention mechanisms to create more dynamic

[1] https://ned.ipac.caltech.edu/
[2] https://exoplanetarchive.ipac.caltech.edu/
[3] https://ui.adsabs.harvard.edu/

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 87–93
December 23, 2025

contextual understanding of text (Devlin et al., 2018; Liu et al., 2019; Brown et al., 2020; Touvron et al., 2023a). LLMs can function as foundation models and be finetuned or directed via prompt engineering to complete downstream tasks.

Other works have shown that foundational LLMs pretrained on domain-specific data (in particular, astronomy - c.f. Grezes et al., 2022, 2024; Bhattacharjee et al., 2024; de Haan et al., 2025) outperform other LLMs not attuned to this domain on downstream tasks related to it. This project therefore uses models pretrained on the astronomical literature. We initially considered the encoder AstroBERT (Grezes et al., 2024, 2022) and decoder AstroLlama (Dung Nguyen et al., 2023), but decided to use the encoder INDUS (Bhattacharjee et al., 2024) and decoder AstroSage-Llama-3.1-8B (AstroSage; de Haan et al., 2025) instead. INDUS and AstroSage are based on more advanced models than the preceding AstroBERT and AstroLlama: astroBERT uses the architecture of BERT (Devlin et al., 2018) and INDUS is based on RoBERTa (Liu et al., 2019), while AstroLlama uses the Llama-2 (Touvron et al., 2023b) architecture and AstroSage is based on that of Llama-3.1 (Grattafiori et al., 2024).

The goal of this work is to produce a tool which accelerates the processes of data extraction and load file creation for NED and NEA. There is no expectation that the load files created using these tools will be perfect, so automated issue flagging, human supervision, and periodic re-training of models will be integral to this process.

2 Methods

A variety of methods, both AI-based and not, are being deployed at the different stages of the archive data ingestion pipeline.

2.1 Data Retrieval

Each module of this work uses the full text of a journal article. These are downloaded from ADS using their API service and converted from the PDF format to plain text using the PyMuPDF loader provided by LangChain[4]. We used the INDUS tokenizer to convert this text into the appropriate format when using INDUS.

[4]https://docs.langchain.com/oss/python/langchain/overview

2.2 Relevance Classification

Both NED and NEA already use machine learning classifiers to predict the probability that an article is relevant to their holdings. The relevance classifier used by NED (Chen et al., 2022a) is based on the Stanford Classifier (Finkel et al., 2005), while the NEA tool (Susemiehl & Christiansen, in prep.) inputs Doc2Vec embeddings to a logistic regression model. Both of these tools have successfully automated this task. However, their accuracy has been declining due to changes in content and structure of newer journal articles, and transformer-based LLMs finetuned to this task have the potential to more accurately predict paper relevance. Due to the active development cycle of this project, this tasks is being reserved for after after the completion of other modules (see Future Work S4.1).

2.3 Data Extraction

Once a relevant article is identified, the data it presents must be extracted into load files that can be ingested into the NED and NEA databases. The data in these articles, such as a planet's mass or a galaxy's redshift, can be contained in the main text or tables within an article, and also in external files linked to some articles. Transformer-based LLMs have as input one-dimensional strings of text, so the two-dimensional structure of tables is lost during training/inference. We therefore employ different methods for text and tabular data extraction.

2.3.1 Object Name Detection

The detection of the names of astronomical objects which are presented in a given article is a fundamental task in this work. To this end, we finetuned INDUS (Bhattacharjee et al., 2024) instances using the HuggingFace Python library (Wolf et al., 2019) on token classification tasks for both archives.

While both NED and NEA hold large databases of object names and their corresponding article identifiers, the locations of the names within these articles is not recorded. In order to frame this task as a supervised learning problem, it is necessary to label each token in an article as either an object or not an object. While human annotation is commonly employed to label training data in similar cases, this is an expensive endeavor. We sought to leverage the large set of object names and articles held by NED and NEA by automatically labeling the tokens within each article. We converted each object name in the NED and NEA lists to generic forms using regular expressions. The formulated regular expressions allow for variations in separators, abbreviations, numerical digits, and planet letter suffixes from the published object names to the canonical forms in the archives. A challenge of this technique is in eliminating both false positive and false negative labeling, as mislabeled tokens pollute the training data set and limit model performance during inference.

Following the regular expression-based token labeling, the training data sets were composed of 8268 articles (300.4 million tokens) for the NED model and 2230 articles (89.6 million tokens) for the NEA model. A hyperparameter search of 10 trials was performed over the learning rate, dropout, weight decay, and random seed. The INDUS models were then finetuned for 10 epochs using the HuggingFace framework. The NED model achieved an F1 score on an unseen test set of 0.95 while the NEA model scored an F1 of 0.94 on its test set. However, an investigation of the learning curve (Figure 1) reveals that both models failed to learn the validation data. This is corroborated by a high occurrence of incorrect labels while qualitatively assessing the models' performances. We found that this finetuned version of INDUS outperforms the base model on the name identification task, suggesting the finetuning process was still useful. External validation tools, which compare potential names to expected name formats, are used to reduce false positive predictions during usage of the tool. It seems likely that this poor model performance is caused by pollution of the training data set during the labeling phase, so future work will investigate means to improve this process.

2.3.2 Text Extraction

Necessary data are often included within the body text of an article. These can be numerical values (e.g. coordinates) or words/phrases (e.g. telescope). Examples of data types regularly found in the body of an article include type of an extragalactic object and the method used to detect an exoplanet. The usage of synonyms, abbreviations, and acronyms for these values is common in the literature. Given the unstructured nature of the body text and the difficulty in composing a token-level training dataset for heterogeneous labels, supervised finetuning approaches may be less applicable. Instead, generative AI is useful because of its ability to read large contexts and answer questions pertaining to data extraction from prompts. We prompt the decoder LLM AstroSage (de Haan et al., 2025) to return the

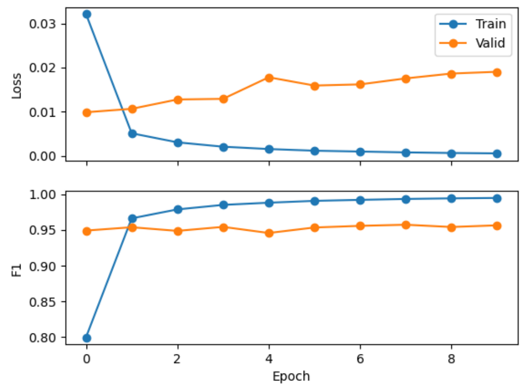


Figure 1: A typical example of a learning curve from finetuning the object name detection model on NEA data (NEA and NED learning curves are qualitatively similar. The flat validation loss curve indicates that the model failed to learn during finetuning.

various data types needed to populate a database load file. This is done in zero-shot context without finetuning. Any article longer than the effective context length of AstroSage (Llama-3.1 8B), 32,000 tokens (Hsieh et al., 2024), is broken into sections of 512 tokens each. A retrieval augmented generation system then selects the most relevant 512 token-sized chunks of text to fill the 32,000 token context (i.e. the top ∼62 chunks are combined to serve as the prompt's context). Otherwise, sufficiently short articles are included entirely within the prompt's context. Grammar-constrained decoding is leveraged to force the output of the LLM into a JSON format with keys corresponding to the needed data types. The LLM-outputted values are further required to be chosen from a particular set of possible options for some categorical data types (e.g. object class). These methods control the output to align with the data types expected to be included in the database load files.

2.3.3 Table Extraction

The structured nature of tables is advantageous for extracting data from them. A variety of detected and derived object parameters, such as redshift and mass, are commonly found within the tables of NED and NEA articles. LLMs, while useful for unstructured data (text), reduce the 2-D grid of a table down to a 1-D string which makes the correct alignment between labels and their respective values difficult. Including positional and descriptive tags within the table text included in prompts to the AstroSage was not found to improve the correct association between extracted values and their respective objects in this work, so methods not solely reliant on an LLM were investigated for this task.

Tables are identified within articles and extracted using the GMFT package[5], which converts the PDF tables to Pandas dataframes (Wes McKinney, 2010). Next, every cell in the dataframe is labeled as either an astronomical object using the object name detection model or a parameter label. We achieve parameter label assignments by matching each as-published potential label to a dictionary of previously-seen labels, converting both to embeddings, and calculating a cosine similarity score. The label type corresponding to the highest matching previously-seen label is then assigned to the as-published label in the table's dataframe (if there is a match score greater than 0.9). With both the objects and parameter labels of a table identified, the dataframe cell containing the respective value is assumed to be the lower-right intersection of the object and parameter label positions in the 2-D grid of the dataframe. This enables direct, automated extraction of data values while maintaining alignment between the object and parameter labels. However, the dictionary of previously-seen parameter labels needs to be expanded whenever substantially different label presentations are encountered.

2.4 Load File Creation

Once data are extracted from the text and tables of an article, they are cleaned and reformatted into database load files using programmatic methods in Python. There are additional components to these files which can be inferred without the use of the above methods.

2.4.1 Other data

NED and NEA load files contain sections of "metadata" regarding the objects to be ingested. This includes, for example, the addition of aliases for a given object. The aliases which need to be added for an object can be inferred by comparing existing entries in the NEA database to those in external databases (e.g. Simbad). Other metadata, like the internal updates to a system's orbital configuration, can be inferred by querying of the NEA databases. An example of inferrable data for NED is the coordinate system (sexagesimal or decimal degree), which can be ascertained via regular expression matching of the retrieved coordinate value.

[5]https://github.com/conjuncts/gmft

3 Results

Prototype versions of these tools have been developed to enable the creation of a database load file with minimal operator input, enabling the semi-automated extraction of data from articles into database load files. Preliminary testing of the tools shows promising performance in its ability to save time. The accuracy of an AI-generated load file is computed by comparing the presence and equality of extracted values to those in the respective human-created load file. These comparisons are made between dataframes containing the extracted data, so the score is robust to minor formatting differences within the files. Early testing has shown accuracies around 20%, but this low score is often not reflective of the often small effort needed to correct an AI-generated file. Future work will seek to expand the usage of this accuracy metric to robustly quantify the performance of these tools.

3.1 Computational Performance

This work was has used a Quadro RTX 6000 GPU for model finetuning and inference. The run duration of the data extraction tools increase with the number of objects and the length of the text. Inference using INDUS typically takes less than one minute per table, while prompts using AstroSage return responses in roughly 5-10 minutes per object. The slow completion speed of AstroSage prompts motivates the investigation of methods not based on decoder models to extract data from the text of an article (see S4.1) in less time, which will also aid in large-scale performance quantification.

4 Conclusions

This early work has shown that AI-powered tools, when combined with other programmatic methods and supervised by humans, can enhance the data ingestion pipelines at NED and NEA. While the results from early versions of these tools can suffer in accuracy, the time it takes to generate and correct a file can also be less than the time it would take to make the file by hand. Transformer-based AI is useful at several junctures of this work, but reliance on these methods alone were found to be insufficient for some subtasks of this project. Both automated and human verifications within the pipeline are needed due to the inaccuracy of AI-derived solutions (although there is potential for improvement). There are also practical limitations to the effectiveness and accuracy of automated data extraction from the literature due to issues with the way data are sometimes published. Examples include: ambiguous object names, which are typically truncated coordinate-based names that cannot be accurately cross-identified using the NED and NEA name resolvers; data with missing uncertainties; omission of the reference frame for some measurements; and critical data linked to URLs that are no longer working. Many of these issues can be solved if authors and referees are more careful about following best practices for publishing data in the astronomical literature (Chen et al., 2022b). This work is in active development and will continue to be improved upon in the coming months.

4.1 Future Work

Prototype versions of these tools are being tested in production contexts. Operators have been asked to provide feedback which will be addressed to make improvements.

We will also seek to improve the automated training data labeling process for the object name detection model. Name validation tools will be used to eliminate false positive token labels and a broader search (i.e. searching each article for every name type) will reduce false negative token labels. This finetuned INDUS model and accompanying training data will be shared on the HuggingFace platform once its performance is improved.

Supervised finetuning of INDUS and AstroSage for the extraction of other data types will also be investigated, as decoder finetuning has been shown to increase the accuracy of related tasks (Zhao et al., 2024). This can be done at the document level for most data types, as the location of extracted data within an article is not retained by NED or NEA.

Additionally, data from external sources provided in links within articles will be accounted for where possible, as well as the units of numerical values (including automatic conversion). The evolving nature of the language used in astronomical journal article as new methods are employed or missions launched necessitates the periodic retraining of literature models. This will begin by replacing the old Standford/Doc2Vec-based relevance classification models with the encoder LLM INDUS, as discussed in S2.2. Other extensions, such as the consideration of images, will be approached in the future. While models adapted to the domain of astronomy have been shown to achieve better performance on astronomy-related tasks than models trained on broader contexts (e.g. Grezes et al., 2022, 2024; Bhattacharjee et al., 2024; de Haan et al., 2025), this work would benefit from a comparison between models like INDUS and AstroSage to modern frontier models from groups such as OpenAI and DeepSeek.

5 Acknowledgements

This work is supported by the NASA/IPAC Extragalactic Database (NED) and NASA/IPAC Exoplanet Archive (NEA), which are funded by the National Aeronautics and Space Administration under Award Number 80NSSC21M0037 and operated by the California Institute of Technology. We thank Douglas McElroy and Denis Zaytsev for important early contributions to this work.

References

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. 2022a. Classification of Astrophysics Journal Articles with Machine Learning to Identify Data for NED. , 134(1031):014501.

Tracy X. Chen, Marion Schmitz, Joseph M. Mazzarella, Xiuqin Wu, Julian C. van Eyken, Alberto Accomazzi, Rachel L. Akeson, Mark Allen, Rachael Beaton, G. Bruce Berriman, Andrew W. Boyle, Marianne Brouty, Ben H. P. Chan, Jessie L. Christiansen, David R. Ciardi, David Cook, Raffaele D'Abrusco, Rick Ebert, Cren Frayer, and 26 others. 2022b. Best Practices for Data Publication in the Astronomical Literature. , 260(1):5.

Tijmen de Haan, Yuan-Sen Ting, Tirthankar Ghosal, Tuan Dung Nguyen, Alberto Accomazzi, Azton Wells, Nesar Ramachandra, Rui Pan, and Zechang Sun. 2025. Achieving GPT-4o level performance in astronomy with a specialized 8B-parameter large language model. *Scientific Reports*, 15(1):13751.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, arXiv:1810.04805.

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. *arXiv e-prints*, arXiv:2309.06126.

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.

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.

F. Grezes, S. Blanco-Cuaresma, A. Accomazzi, M. J. Kurtz, G. Shapurian, E. Henneken, C. S. Grant, D. M. Thompson, R. Chyla, S. McDonald, T. W. Hostetler, M. R. Templeton, K. E. Lockhart, N. Martinovic, S. Chen, C. Tanner, and P. Protopapas. 2024. Building astroBERT, a Language Model for Astronomy & Astrophysics. In *Astromical Data Analysis Software and Systems XXXI*, volume 535 of *Astronomical Society of the Pacific Conference Series*, page 119.

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. *arXiv e-prints*, arXiv:2212.00744.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models? *arXiv e-prints*, arXiv:2404.06654.

Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. *arXiv e-prints*, arXiv:1405.4053.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, arXiv:1907.11692.

F. Matuszewski, N. Nettelmann, J. Cabrera, A. Börner, and H. Rauer. 2023. Estimating the number of planets that PLATO can detect. , 677:A133.

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

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. *arXiv e-prints*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv e-prints*, arXiv:2307.09288.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv e-prints*, arXiv:1910.03771.

Hang Zhao, Qile P. Chen, Yijing Barry Zhang, and Gang Yang. 2024. Advancing Single and Multi-task Text Classification through Large Language Model Fine-tuning. *arXiv e-prints*, arXiv:2412.08587.

A Hybrid LLM and Supervised Model Pipeline for Polymer Property Extraction from Tables in Scientific Literature

Van-Thuy Phi[1], Dinh-Truong Do[1,2*], Hoang-An Trieu[1,2*], Yuji Matsumoto[1]

[1] RIKEN Center for Advanced Intelligence Project
[2] Japan Advanced Institute of Science and Technology
{thuy.phi, truong.do, an.trieu, yuji.matsumoto}@riken.jp

Abstract

Extracting structured information from tables in scientific literature is a critical yet challenging task for building domain-specific knowledge bases. This paper addresses extraction of 5-ary polymer property tuples: (POLYMER, PROP_NAME, PROP_VALUE, CONDITION, CHAR_METHOD). We introduce and systematically compare two distinct methodologies: (1) a novel two-stage Hybrid Pipeline that first utilizes Large Language Models (LLMs) for table-to-text conversion, which is then processed by specialized text-based extraction models; and (2) an end-to-end Direct LLM Extraction approach. To evaluate these methods, we employ a systematic, domain-aligned evaluation setup based on the expert-curated PoLyInfo database. Our results demonstrate the clear superiority of the hybrid pipeline. When using Claude Sonnet 4.5 for the linearization stage, the pipeline achieves a score of 67.92% F1@PoLyInfo, significantly outperforming the best direct LLM extraction approach (Claude Sonnet 4.5 at 56.66%). This work establishes the effectiveness of a hybrid architecture that combines the generative strengths of LLMs with the precision of specialized supervised models for structured data extraction.

1 Introduction

The field of materials science, particularly polymer science, generates vast amounts of data published in scientific articles. This data, often embedded in tables, is crucial for developing new materials, training predictive models, and enabling data-driven discovery. Automated Information Extraction (IE) systems are essential for curating this knowledge into structured, machine-readable databases like PoLyInfo (Otsuka et al., 2011).

Recent studies by Phi et al. (2024) and Do et al. (2025) introduced a new corpus and developed a practical system for extracting polymer-related concepts and properties from unstructured text, demonstrating the high performance of supervised models like W2NER (Li et al., 2022) for Named Entity Recognition (NER) and ATLOP (Zhou et al., 2021) for Relation Extraction (RE) on their PolyNERE corpus. However, these models are inherently designed for plain text and cannot be directly applied to the semi-structured format of tables. Conversely, Large Language Models (LLMs) are adept at parsing diverse data formats but often lack the accuracy of fine-tuned models for domain-specific tasks.

This paper bridges this gap by investigating a hybrid approach that synergizes the strengths of both paradigms for the complex task of table extraction. Our primary contributions are:

- We propose a two-stage method that first leverages an LLM's structural understanding to convert table rows into natural language paragraphs. This linearized text is then processed by the advanced text-based IE system components identified by Phi et al. (2024) and Do et al. (2025).
- We systematically compare five advanced LLMs in both our hybrid pipeline and a direct end-to-end extraction approach using carefully engineered, task-specific prompts.

* These authors contributed equally.

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 94–102
December 23, 2025

- We introduce a new PoLyInfo-based benchmark for evaluating property extraction from tables, providing near-comprehensive coverage (~66% of property names) of critical, standardized properties in the domain.
- Our results demonstrate that the hybrid pipeline significantly outperforms the direct LLM approach, establishing it as a more robust method for this task.

2 Related Work

Traditional neural approaches have achieved strong performance in domain-specific text extraction tasks. The W2NER architecture (Li et al., 2022) has shown particular effectiveness in capturing complex entity structures in scientific text—such as flat, overlapping, and discontinuous entities—commonly found in materials science literature, as demonstrated by Do et al. (2025). For relation extraction, ATLOP (Zhou et al., 2021) reformulates the task as entity-pair linking, delivering robust performance on specialized corpora like PolyNERE. Domain-adapted language models, such as MatSciBERT, have further improved results for materials science applications. However, these specialized models remain constrained to plain text input, limiting their direct applicability to tabular data.

Recent research has demonstrated the remarkable zero-shot and few-shot capabilities of LLMs for NER and RE. Most approaches attempt direct, end-to-end extraction, where the model is prompted to output structured data from a given input. However, this method forces a single model to handle multiple complex sub-tasks (parsing, entity recognition, etc.), which can lead to hallucinations or conversational outputs ill-suited for scientific data extraction (Kumar et al., 2025).

Converting tabular structures for LLM processing has emerged as a critical research area, with various serialization methods showing different effectiveness depending on table complexity. Recent work has shown that table linearization quality significantly impacts downstream extraction performance, though optimal strategies remain domain-dependent (Deng et al., 2024).

Our work bridges these areas by proposing a hybrid pipeline that leverages LLMs for table-to-text conversion while utilizing specialized supervised models for robust extraction, specifically addressing the gap in scientific table information extraction for polymer property data.

3 Methodology

The input for our system is a multi-modal prompt, combining a high-fidelity table image with its corresponding textual caption and footnotes. Table images are extracted directly from scientific documents using the MinerU parser (Wang et al., 2024). This image-based approach is motivated by Circi et al. (2024), who demonstrated that visual layout cues enable vision-enabled LLMs to more accurately extract complex relationships from scientific tables compared to text-only inputs.

We formalize the task as extracting a set of 5-ary property information tuples from a given scientific table. This formalization is grounded in the schema of the PoLyInfo database (Otsuka et al., 2011), the largest expert-curated database for polymers. The target is a set of tuples T = {t_1, t_2, ..., t_n}, where each tuple t_i consists of five key entity types:

- POLYMER: The name of the polymer material (e.g., “*polyethylene*”, “*poly(p-diethynylbenzene)*”).
- PROP_NAME: The name of the physical or chemical property being described (e.g., “*glass transition temperature*”, “*density*”).
- PROP_VALUE: The measured value of the property, typically including units (e.g., “*25 MPa*”, “*1.097 g/cm3*”).
- CONDITION: The experimental conditions under which the property was measured (e.g., “*at 25°C*”, “*under nitrogen atmosphere*”).
- CHAR_METHOD: The characterization technique or method used for the measurement (e.g., “*DSC*”, “*tensile testing*”).

These five types represent the core elements required to form a complete and usable entry in a materials science knowledge base. The primary challenge in the context of tables lies in correctly associating information that is structurally fragmented. The goal of our system is to accurately parse the combination of visual and textual information to compose a comprehensive set of valid 5-ary tuples. We compare two distinct approaches to solve this task.

3.1 Method 1: Hybrid LLM and Supervised Model Pipeline

This method decomposes the task into two sequential stages, leveraging the optimal model type for each sub-task.

Stage 1: LLM-based Table-to-Text Conversion: An LLM is given the multi-modal prompt (table image, caption, footnotes) and is instructed to act as a domain expert to linearize each data row into a descriptive paragraph. A novel aspect of our approach is the carefully engineered prompt (see Appendix A), which transforms the LLM into a specialized pre-processor for our supervised models. The prompt's key innovation is a conditional grouping strategy: it instructs the LLM to create separate, self-contained paragraphs for each material (POLYMER or its composite), and further subdivides these by CHAR_METHOD only if a method is explicitly stated. This hierarchical grouping is crucial as it prevents the ambiguous association of multiple properties with their respective measurement contexts—a common challenge for downstream relation extraction models.

Furthermore, by enforcing a strict, single-line output format and text normalization rules (e.g., "*T_g*" to "*Tg*"), the prompt ensures the generated text is a consistent and machine-readable intermediate representation, optimized for the models in the subsequent stage.

Stage 2: Supervised Text-based Tuple Extraction: The text generated from Stage 1 is then processed by a fixed, pre-trained text extraction system composed of supervised models trained on the PolyNERE corpus (Phi et al., 2024), selected based on their proven high performance.

We employ a W2NER model (Li et al., 2022), which is adept at handling the flat, overlapped, and discontinuous entity structures common in scientific text. This architecture is similar to that used in the PolyMinder system (Do et al., 2025). To further optimize for the materials science domain, we pair it with the MatBERT encoder (Walker et al., 2021).

We utilize the ATLOP model (Zhou et al., 2021), a choice validated by its strong performance in prior work (Phi et al., 2024; Do et al., 2025). To effectively capture the complex relationships present in the text, the model is paired with the powerful DeBERTa-v3-large encoder (He et al., 2020).

3.2 Method 2: Direct Tuple Extraction using LLMs

This approach follows a conventional end-to-end paradigm. The same multi-modal prompt is passed to a vision-enabled LLM. The prompt (see Appendix B) instructs the model to analyze the table's visual structure and associated text to directly output a list of all identifiable property tuples. To ensure a fair comparison, this prompt is also highly engineered with a similar set of detailed instructions and critical rules. This method relies entirely on the LLM's in-context reasoning to perform all sub-tasks simultaneously and serves as a direct baseline to evaluate the effectiveness of our hybrid pipeline.

4 Experiments

4.1 Datasets

The ground truth for our evaluation was constructed through a manual alignment process. We sourced curated polymer property data from the expert-driven PoLyInfo database (Otsuka et al., 2011) and mapped it to relevant content within a corpus of 37 tables from 29 scientific papers. Our final golden set comprises 293 property information tuples. Each tuple contains three essential entities (POLYMER, PROP_NAME, and PROP_VALUE), supplemented with optional CONDITION and CHAR_METHOD entities when available in the PoLyInfo entry. We confirmed that the 37 evaluation tables have no overlap with the PolyNERE training corpus, ensuring that supervised models in Stage 2 were tested on entirely unseen content.

Our analysis shows that the PoLyInfo-based golden annotations cover ~66% of all property names found across the evaluated tables. Specifically, we manually counted 132 property names appearing in the row and column headers of the 37 tables. The PoLyInfo database is an expert-curated resource where domain experts selectively extract and store only the most critical and standardized property information from scientific papers. Of the 132 property names in our tables, 87 (66%) have corresponding entries in PoLyInfo and were used to construct our golden set of 293 tuples. The remaining 45 property names (34%) may represent less critical properties that were not prioritized by expert curators for inclusion in PoLyInfo. Our evaluation is therefore near-comprehensive in its assessment of the most important, standardized properties deemed critical by domain experts for polymer characterization.

For the hybrid pipeline, predicted binary relations from the ATLOP model are merged into 5-ary tuples based on the relation schema defined in Phi et al. (2024). During manual evaluation of these composed tuples (for both methods), we observed a consistent one-to-one mapping between a golden tuple and a corresponding prediction for each (POLYMER, PROP_NAME) pair (see Appendix E). A prediction is marked as True (T) only if all five of its constituent entities exactly match the golden tuple; otherwise, it is marked as False (F).

4.2 Results

Task	Model	Encoder	P	R	F1
NER	W2NER	MatBERT	**78.79**	**79.81**	**79.30**
	Baseline	MatSciBERT	78.05	76.53	77.28
RE	ATLOP	DeBERTa-v3-large	**87.93**	**86.89**	**87.40**
	Baseline	MatSciBERT	83.99	82.49	83.23

Table 1: NER and RE performance on the PolyNERE test set. RE uses gold entities.

Based on the observed one-to-one mapping in our evaluation setup, the number of False Positives and False Negatives are equivalent for the set of evaluated golden tuples. Consequently, Precision and Recall converge to the same value. We therefore report this unified metric as F1@PoLyInfo, representing the percentage of correctly extracted tuples from the set of important, PoLyInfo-defined properties:

*F1@PoLyInfo (%) = # True / (# True + # False) * 100*

We trained the supervised W2NER and ATLOP models using established hyperparameters from prior work (30 epochs, batch size 8, Adam optimizer). All LLM inferences were performed with deterministic settings (temperature=0, top_p=1).

We first establish the performance of our pipeline's core supervised models by evaluating them on the PolyNERE test set against the PolyMinder baseline (Do et al., 2025). Table 1 shows our selected models significantly outperform the established baseline for text-based extraction in this domain. Our W2NER+MatBERT configuration improves the NER F1 score by +2.02 points, while our ATLOP+DeBERTa-v3-large model shows a more significant +4.17 F1 point gain for RE. These results confirm their role as a powerful foundation for processing the linearized table data.

Model	Hybrid Pipeline			LLM Extraction		
	True	False	F1	True	False	F1
Claude Sonnet 4.5	199	94	**67.92**	166	127	**56.66**
GPT-4.1	112	181	38.23	119	174	40.61
GPT-4o mini	142	151	48.46	141	152	48.12
Gemini 2.5 Flash	164	129	55.97	73	220	24.91
Qwen2.5-VL 32B	158	135	53.92	86	207	29.35

Table 2: Model performance results.

We then evaluated the two end-to-end methodologies on our table extraction task. The results are summarized in Table 2.

The hybrid pipeline proves to be the superior strategy for the majority of the tested models. The advanced LLMs, Claude Sonnet 4.5, Gemini 2.5 Flash, and Qwen2.5-VL 32B Instruct, all saw dramatic performance increases when used in the hybrid pipeline. Specifically, Gemini 2.5 Flash and Qwen2.5-VL 32B improved by an absolute +31.06% and +24.57%, respectively, indicating that decomposing the complex task is critical for these models.

The best performance in our study was achieved by the hybrid pipeline, with Claude Sonnet 4.5 in the linearization stage reaching 67.92% F1@PoLyInfo. This represents a substantial +11.26% absolute improvement over its already strong direct extraction performance. An important exception to the general trend is GPT-4.1, for which the direct extraction method performed slightly better (40.61%) than the hybrid pipeline (38.23%). Similarly, the performance of GPT-4o mini was nearly identical across both methods. This suggests that for certain models, error propagation in a two-stage process—where suboptimal text generation in Stage 1 negatively impacts the supervised models—can outweigh the benefits of task decomposition. A detailed case study in Appendix D analyzes the specific failure modes of the pipeline for GPT-4.1.

The direct extraction method proved significantly more challenging for the majority of LLMs, with steep performance drops for models like Gemini 2.5 Flash and Qwen2.5-VL 32B highlighting the immense difficulty of simultaneously parsing a 2D structure and composing complex relations in a single step.

The hybrid pipeline's success lies in assigning the right task to the right model. The LLM excels at the generative, context-aware task of converting a table into fluent text. The supervised W2NER and ATLOP models, which are pre-trained and fine-tuned for their specific tasks, then excel at precise, closed-set extraction from this clean, textual input. This hybrid architecture proves more robust and accurate for most models, though it is not a universally guaranteed improvement, as seen with GPT 4.1.

5 Conclusion

In this work, we compared a hybrid pipeline (LLM linearization and supervised NER/RE) against a direct LLM approach for property extraction from tables, finding the hybrid architecture to be the more robust strategy on our PoLyInfo-based benchmark. Our best pipeline configuration achieves 67.92% F1@PoLyInfo, demonstrating that task decomposition with specialized supervised models yields superior performance compared to end-to-end LLM approaches.

Limitations

First, the evaluation set, while carefully curated, is of moderate size (293 tuples from 37 tables) and focused exclusively on the polymer science domain, and performance may vary on other types of scientific tables. Second, the hybrid pipeline's performance is highly dependent on the quality of the LLM-generated text in Stage 1, and as shown with GPT-4.1, poor linearization can create a bottleneck. Third, the success of our hybrid pipeline relies on the availability of well-trained text analyzers for NER and RE. This approach presupposes that high-quality, domain-specific supervised models are available for the second stage. Finally, our prompts were carefully designed with domain-specific instructions, but we did not systematically evaluate sensitivity to prompt variations. Evaluation requires manual normalization of tuples before matching, making comprehensive prompt experiments labor-intensive. Future work could explore automated evaluation methods for systematic prompting strategy comparison.

References

Circi, D., Khalighinejad, G., Chen, A., Dhingra, B., & Brinson, L. (2024). Extracting Materials Science Data from Scientific Tables. In *ACL 2024 Workshop Language+Molecules*.

Do, T. D., Trieu, A. H., Phi, V. T., Le Nguyen, M., & Matsumoto, Y. (2025, January). PolyMinder: A Support System for Entity Annotation and Relation Extraction in Polymer Science Documents. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations* (pp. 1-8).

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint* arXiv:2006.03654.

Kumar, P., Kabra, S., & Cole, J. M. (2025). MechBERT: Language Models for Extracting Chemical and Property Relationships about Mechanical Stress and Strain. *Journal of Chemical Information and Modeling*.

Li, J., Fei, H., Liu, J., Wu, S., Zhang, M., Teng, C., Ji, D. and Li, F., 2022, June. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 36, No. 10, pp. 10965-10973).

Naihao Deng, Zhenjie Sun, Ruiqi He, Aman Sikka, Yulong Chen, Lin Ma, Yue Zhang, and Rada Mihalcea. 2024. Tables as Texts or Images: Evaluating the Table Reasoning Ability of LLMs and MLLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 407–426, Bangkok, Thailand. Association for Computational Linguistics.

Nicholas Walker, Amalie Trewartha, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin Persson, Gerbrand Ceder, and Anubhav Jain. 2021. The impact of domain-specific pre-training on named entity recognition tasks in materials science. Available at SSRN 3950755.

Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y., & Yamazaki, M. (2011, September). PoLyInfo: Polymer database for polymeric materials design. In *2011 International Conference on Emerging Intelligent Data and Web Technologies* (pp. 22-29). IEEE.

Phi, V. T., Teranishi, H., Matsumoto, Y., Oka, H., & Ishii, M. (2024, May). PolyNERE: A novel ontology and corpus for named entity recognition and relation extraction in polymer science domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 12856-12866).

Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F. and Zhang, B., 2024. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Zhou, W., Huang, K., Ma, T., & Huang, J. (2021, May). Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 16, pp. 14612-14620).

A Prompt for LLM-based Table-to-Text Conversion (Method 1)

You are analyzing a scientific table image. Convert it into structured natural language text that will be processed by Named Entity Recognition (NER) and Relation Extraction (RE) models.

TABLE CAPTION: [INSERT CAPTION TEXT HERE]
FOOTNOTES: [INSERT FOOTNOTES TEXT HERE]

TASK: Create separate paragraphs for each material to prevent entity confusion. If different properties are measured using different characterization methods (found in caption, footnotes, or column headers), create separate paragraphs for each material-method combination.

CRITICAL: Only separate by characterization method if methods are explicitly stated. If no methods are mentioned, write all properties for a material in one paragraph.

OUTPUT STRUCTURE:
1. First sentence: Introduce the table using the caption
2. Then, for EACH material:
*- **If characterization methods are specified**: Write separate paragraphs for each method*
*- **If NO methods are specified**: Write one paragraph with all properties*
3. Add blank line between paragraphs

REQUIREMENTS FOR EACH PARAGRAPH:
- Start with the material name EXACTLY as it appears in the table
*- **If characterization method is specified**: Include it after material name*
*- **If NO method is specified**: Omit method phrase entirely*
- List properties with their values and units
- Include any conditions from the caption, footnotes, or column headers
- Write each paragraph as a SINGLE continuous line
*- **Format with method**: "For [material name] measured by [CHAR_METHOD] [condition phrase]: [property name] is [value unit], [property name] is [value unit], ..."*
*- **Format without method**: "For [material name] [condition phrase]: [property name] is [value unit], [property name] is [value unit], ..."*

ENTITY TYPES TO INCLUDE:
1. POLYMER: Material/polymer name exactly as written in the table
2. PROP_NAME: Complete property name from column header
3. PROP_VALUE: Numerical value WITH unit (e.g., "7.29 MPa", "266.53%", "45.2 wt%")
4. CONDITION: Experimental conditions starting with a preposition (e.g., "at X°C", "with n=Y", "under annealing")
5. CHAR_METHOD: Measurement or characterization method as a noun phrase (e.g., "SEC", "DSC", "tensile testing")

CRITICAL RULES:
- Use material names EXACTLY as they appear in the table (no expansion or modification)
*- **DO NOT treat property names as characterization methods***
*- **Only use "measured by" when an actual measurement technique is specified (e.g., SEC, NMR, DSC, XRD, TEM, SEM, FTIR)***
*- **Column headers showing property names (e.g., "Tensile strength", "Density", "Modulus") are NOT characterization methods***
- Separate by characterization method only when methods are explicitly mentioned
- Copy exact numbers and units from the table
- Include units WITH values (e.g., "7.29 MPa" not just "7.29")
- Each paragraph must be a single continuous line - NO line breaks within a paragraph
- Add blank line between paragraphs only
- DO NOT use subscript notation with underscores (e.g., M_n, T_g, T_c). Instead use simplified notation (e.g., Mn, Tg, Tc)
- Condition phrases must start with a preposition (e.g., "at", "under", "with", "in", "by")
- CHAR_METHOD must be a noun phrase (e.g., "DSC", "tensile testing", "X-ray diffraction")

EXAMPLE FORMAT:

Case 1 - NO characterization methods specified:
This table presents [property category] of [material type] materials.

For [Material-A] [condition phrase if any]: [property-1] is [X.XX unit], [property-2] is [Y.YY unit], [property-3] is [Z.ZZ unit].

For [Material-B] [condition phrase if any]: [property-1] is [X.XX unit], [property-2] is [Y.YY unit], [property-3] is [Z.ZZ unit].

Case 2 - Characterization methods ARE specified:
This table presents [property category] of [material type] materials.

For [Material-A] measured by [CharMethod-1]: [property-1] is [X.XX unit], [property-2] is [Y.YY unit].

For [Material-A] measured by [CharMethod-2] [condition phrase if any]: [property-3] is [Z.ZZ unit], [property-4] is [W.WW unit].

For [Material-B] measured by [CharMethod-1]: [property-1] is [X.XX unit], [property-2] is [Y.YY unit].

For [Material-B] measured by [CharMethod-2] [condition phrase if any]: [property-3] is [Z.ZZ unit], [property-4] is [W.WW unit].

OUTPUT: Return ONLY the converted text. No explanations or additional commentary.

B Prompt for Direct Tuple Extraction (Method 2)

You are analyzing a scientific table image. Extract ALL property measurements from the table as structured tuples.

TABLE CAPTION: [INSERT CAPTION TEXT HERE]
FOOTNOTES: [INSERT FOOTNOTES TEXT HERE]

TASK: Extract ALL property measurements from the table as 5-element tuples.

TUPLE FORMAT:
(POLYMER, PROP_NAME, PROP_VALUE, CONDITION, CHAR_METHOD)

REQUIREMENTS FOR EACH TUPLE:
- Extract one tuple for EACH property measurement (one row × one column = one tuple)
- Include the complete material name in every tuple
- Copy exact values with units from table cells
- Extract any conditions or methods from the caption, footnotes, or column headers
- Process systematically: for each material (row), extract all properties (columns)

ENTITY TYPES TO INCLUDE:
1. POLYMER: Material/polymer name exactly as written in the table (e.g., "PE", "Sample A", "Composite-5")
2. PROP_NAME: Complete property name from column header (e.g., "tensile strength", "glass transition temperature")
3. PROP_VALUE: Numerical value WITH unit (e.g., "X.XX MPa", "YY.Y%", "Z.ZZ ± 0.XX unit")
4. CONDITION: Experimental conditions starting with a preposition (e.g., "at X°C", "with n=Y", "under annealing", "in air")
5. CHAR_METHOD: Measurement or characterization method as a noun phrase (e.g., "tensile testing", "thermal analysis", "SEC", "DSC")

CRITICAL RULES:
- Use material names EXACTLY as they appear in the table (no expansion or abbreviations)
- Repeat material names in every tuple for clarity
- Copy exact numbers and units from the table
- Include units WITH values (e.g., "7.29 MPa" not just "7.29")
- Extract conditions/methods from caption, footnotes, and headers
- CONDITION must start with a preposition (e.g., "at", "under", "with", "in", "by")
- CHAR_METHOD must be a noun phrase (e.g., "DSC", "tensile testing", "X-ray diffraction")
- If condition or method not specified, use empty string ""
- One measurement = one tuple
- DO NOT use subscript notation with underscores (e.g., M_n, T_g, T_c). Instead use simplified notation (e.g., Mn, Tg, Tc)

EXAMPLE FORMAT (using placeholder values):
("PE", "property 1", "value unit", "at condition", "method name")
("PE", "property 2", "value unit", "", "")
("Sample C", "property 1", "value ± error unit", "at condition 1, with condition 2", "characterization method")

OUTPUT: Return ONLY the tuple list. One tuple per line. No explanations or additional commentary.

C Examples of Evaluated Tables

Table 1. Molecular Characteristics of CE Copolymers

polymer	M_n[a] (kg/mol)	M_w/M_n[a]	cyclohexyl ethylene[b] (wt %)	ethyl branches per 100 backbone carbon atoms in ethylene units[b]	T_g[c] (°C)	ρ[d] (g/cm³)
PE	65	1.04		3.5	−31	0.907
CE50	49	1.05	51	3.5	−20	0.918
CE60	67	1.05	60	2.6	7	0.928
CE70	52	1.05	72	3.3	30	0.937
CE80	65	1.05	84	2.4	43	0.947
PCHE	70	1.05	100		144	0.960

[a]Measured with SEC using the parent SB copolymer with universal calibration. The Mark−Houwink parameters for PS and PB are K_{PS} = 8.63 × 10^{-3} mL/g, α_{PS} = 0.736 and K_{PB} = 25.2 × 10^{-3} mL/g, α_{PB} = 0.727.[33] The K and α of SB copolymers are estimated using the weight-averaged values of the homopolymer counterparts. [b]Calculated from the integration of characteristic peaks in ^{1}H NMR spectra. [c]Determined with DSC. [d]Measured with density gradient column at 23 °C.

Figure 1: Example table from the evaluation set, featuring complex headers and footnotes linking properties to characterization methods (SEC, DSC, NMR).

Table 1. Characteristics of Both PFS and PWN2010

	GPC			IEC [mmol g^{-1}]		TGA	DSC	water uptake[c]	
polymer	M_n	M_w	PD	calcd	found	T_{decomp} [°C]	T_g [°C]	[wt %]	$[\lambda]^d$
PFS	30 200	59 000[a]	1.92	0		285	105.5	0	0
PWN2010	9 000	67 000	7.5	7.8[b]	7.0	340	>330	18	2.5

[a] Value is received from the supplier (Monomer Polymer & Dajac Laboratories, USA). [b] Calculation is based on 100% substitution (1 $-PO_3H_2$/aromatic ring). [c] Value at RH = 50%, T = 30 °C. [d] $\lambda = [H_2O]/[-PO_3H_2]$.

Figure 3: Input table for the error analysis.

Table 1. Thermal properties of LPEEK/HPEEK blends with various HPEEK contents

LPEEK/HPEEK (w/w)	T_g (°C)[a]	T_g (°C)[b]	T_c (°C)	T_m (°C)	X_c (%)	$T_d{}^5$ (°C)
100/0	145	167	283	334	15.4	560
99/1	145	170	288	337	15.5	566
98/2	145	171	291	337	17.2	570
97/3	144	172	292	338	17.2	570
96/4	144	173	293	338	17.6	570
95/5	144	174	294	338	19.5	565

[a] Measured using DSC.
[b] Measured using DMA.

Figure 2: Example table illustrating a common ambiguity where two columns share a property name (Tg) but are distinguished by different characterization methods (DSC, DMA) in the footnotes.

D Error Analysis: Understanding the GPT-4.1 Performance Anomaly

As noted in Section 4.2, GPT-4.1 was a notable exception to the general trend, performing slightly better with direct extraction than within the hybrid pipeline. This appendix provides a case study to dissect the specific failure modes that explain this result, highlighting the critical role of the table-to-text linearization stage and the potential for error propagation. We use the table in Figure 3 and its corresponding PoLyInfo-based golden annotations in Figure 4 as a running example.

Analysis of the Hybrid Pipeline (Method 1): In the hybrid pipeline, the quality of the final extraction is entirely dependent on the quality of the text generated by the LLM in Stage 1. Figure 3 shows a table where characterization methods (TGA, DSC) are specified as column headers.

Generated Text from GPT-4.1: For the input table in Figure 3, GPT-4.1 produced the following linearized text for Stage 2:

> *"This table presents characteristics of both PFS and PWN2010.*
> *For PFS at RH = 50%, T = 30 °C: Mn is 30 200, Mw is 59 000, PD is 1.92, IEC calcd is 0 mmol g⁻¹, IEC found is 0 mmol g⁻¹, Tdecomp is 285 °C, Tg is 105.5 °C, water uptake [wt %] is 0, water uptake [λ] is 0.*
> *For PWN2010 at RH = 50%, T = 30 °C: Mn is 9 000, Mw is 67 000, PD is 7.5, IEC calcd is 7.8 mmol g⁻¹, IEC found is 7.0 mmol g⁻¹, Tdecomp is 340 °C, Tg is >330 °C, water uptake [wt %] is 18, water uptake [λ] is 2.5."*

This generated text contains two critical flaws that lead to downstream extraction failures:

- **Omission of Characterization Methods:** The LLM failed to identify "*TGA*" and "*DSC*" from the column headers as characterization methods and did not include them in the generated paragraphs. Because this information is completely absent from the text, it is impossible for the downstream supervised W2NER and ATLOP models to extract the CHAR_METHOD entity. This results in an immediate and unavoidable False evaluation for four of the six golden tuples shown in Figure 4.

- **Incorrect Value-Unit Representation**: The linearization format "*...water uptake [wt %] is 0...*" separates the property's unit from its value. The supervised NER model, which relies on surface text patterns, struggles with this structure. It is likely to identify PROP_VALUE as just "*0*" and incorrectly associate "*[wt %]*" with the PROP_NAME. This creates a mismatch with the golden annotation in Figure 4, which correctly defines PROP_NAME as "*water uptake*" and PROP_VALUE as "*0 wt%*".

These linearization errors propagate through the pipeline, preventing the supervised models in Stage 2 from performing correctly and resulting in a low overall score.

POLYMER	PFS
PROP_NAME	Tg
PROP_VALUE	105.5 ◦C
CONDITION	
CHAR_METHOD	DSC
POLYMER	PFS
PROP_NAME	water uptake
PROP_VALUE	0 wt%
CONDITION	at RH = 50%, T = 30 °C
CHAR_METHOD	
POLYMER	PFS
PROP_NAME	Tdecomp | Thermal decomposition temperature
PROP_VALUE	285 ◦C
CONDITION	
CHAR_METHOD	TGA
POLYMER	PWN2010
PROP_NAME	Tg
PROP_VALUE	>330 ◦C
CONDITION	
CHAR_METHOD	DSC
POLYMER	PWN2010
PROP_NAME	water uptake
PROP_VALUE	18 wt%
CONDITION	at RH = 50%, T = 30 °C
CHAR_METHOD	
POLYMER	PWN2010
PROP_NAME	Tdecomp | Thermal decomposition temperature
PROP_VALUE	340 ◦C
CONDITION	
CHAR_METHOD	TGA

Figure 4: The PoLyInfo-based golden annotations for the table in Figure 3. These tuples serve as the ground truth for the error analysis, highlighting failures in CHAR_METHOD and PROP_VALUE extraction.

Analysis of Direct LLM Extraction (Method 2): In the direct extraction method, the LLM is responsible for parsing the table and generating tuples in one step. Below are the corresponding outputs from GPT-4.1 for the golden tuples in Figure 4. Incorrectly predicted entities are shown in **bold**.

("PFS", "Tdecomp", "285 °C", "", "TGA")

("PFS", "Tg", "105.5 °C", "", "DSC")

*("PFS", "**water uptake [wt %]**", "**0**", "at RH = 50%, at T = 30 °C", "")*

("PWN2010", "Tdecomp", "340 °C", "", "TGA")

("PWN2010", "Tg", ">330 °C", "", "DSC")

*("PWN2010", "**water uptake [wt %]**", "**18**", "at RH = 50%, at T = 30 °C", "")*

From these outputs, we observe:

- **Correct CHAR_METHOD Extraction**: For properties with simple headers like "*Tdecomp*" and "*Tg*", the direct method performs perfectly, correctly identifying "*TGA*" and "*DSC*" as the CHAR_METHOD. This gives it an advantage over the flawed pipeline output, where this information was lost.
- **Incorrect PROP_NAME and PROP_VALUE Parsing**: Similar to the pipeline's issue, the direct method also struggles with the complex "*water uptake*" header. It incorrectly merges the unit "*[wt %]*" into the PROP_NAME and extracts only the numerical part ("*0*" or "*18*") as the PROP_VALUE, leading to a mismatch.

This case study explains the GPT-4.1 performance anomaly. The hybrid pipeline's linearization stage made a significant error by omitting CHAR_METHOD information, leading to unavoidable downstream failures for the supervised models. In contrast, the direct extraction method, while also imperfect, correctly extracted more of the golden tuples. This demonstrates the risk of error propagation in a pipeline. If an LLM's text generation style is a poor fit for the downstream models, a direct approach can, in some cases, yield slightly better results by avoiding this cascade of errors.

E One-to-One Mapping in Tuple Evaluation

We observed a consistent one-to-one mapping between golden tuples and predictions for each (POLYMER, PROP_NAME) pair across all evaluated tuples.

For the Hybrid Pipeline: The ATLOP model predicts binary relations that are merged into 5-ary tuples following Phi et al. (2024). When multiple binary relations share the same (POLYMER, PROP_NAME, PROP_VALUE) triple, they are consolidated into a single tuple by aggregating CONDITION and CHAR_METHOD entities.

For Direct LLM Extraction: Scientific tables organize data with one measurement per cell. The prompt instructs "*Extract one tuple for EACH property measurement (one row × one column = one tuple)*", and all LLMs followed this instruction.

Under this one-to-one constraint, each incorrect prediction simultaneously represents both a false positive and a false negative, making these counts equivalent.

TeG-DRec: Inductive Text-Graph Learning for Unseen Node Scientific Dataset Recommendation

Ammar Qayyum[1], Bassamtiano R. Irnawan[1], Fumiyo Fukumoto[2], Latifah M. Kamarudin[3], Kentaro Go[2], Yoshimi Suzuki[2]
[1]Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences
[2]Graduate Faculty of Interdisciplinary Research
University of Yamanashi
[3]Centre of Excellence for Advanced Sensor Technology Universiti Malaysia Perlis
{g25dtsa4,g23dtsa2,fukumoto,go,ysuzuki}@yamanashi.ac.jp
latifahmunirah@unimap.edu.my

Abstract

Scientific datasets are crucial for evaluating scientific research, and their number is increasing rapidly. Most scientific dataset recommendation systems use Information Retrieval (IR) methods that model semantics while overlooking interactions. Graph Neural Networks (GNNs) excel at handling interactions between entities but often overlook textual content, limiting their ability to generalise to unseen nodes. We propose **TeG-DRec**, a framework for scientific dataset recommendation that integrates GNNs and textual content via a subgraph generation module to ensure correct propagation throughout the model, enabling handling of unseen data. Experimental results on the dataset recommendation's dataset show that our method outperformed the baselines for text-based IR and graph-based recommendation systems. Our source code is available at `https://github.com/Maqif14/TeG-DRec.git`

1 Introduction

Scientific datasets are essential for evaluating scientific research, as it is crucial to examine and verify their behaviour to achieve optimal performance in real-world scenarios (Özgöbek et al., 2014; Fahrudin and Wijaya, 2024). When a dataset is tailored to the specific context of the learning environment, it can significantly improve system performance (Verbert et al., 2011). For example, the Common Crawl dataset significantly contributed to the efficacy of GPT-3 as a formidable Large Language Model (LLM) upon its release in 2020 (Brown et al., 2020). The number of datasets increases annually by hundreds each year (Viswanathan et al., 2023). The growing number of datasets complicates manual search for the optimal dataset, occasionally leading to poor selections (Patankar et al., 2023; Viswanathan et al., 2023; Qin et al., 2024). Consequently, the need for a dataset recommender is greater than ever to enhance research efficiency.

Several studies have explored scientific dataset recommendation systems using text-based IR methods (Wang et al., 2021; Färber and Leisinger, 2021; Keller and Munz, 2022; Yadav et al., 2023; Zhang and Ashraf, 2023), with some extending it using neural bi-encoders to capture richer contextual semantics (Viswanathan et al., 2023). These approaches typically compute lexical or embedding-based similarity between query descriptions and candidate datasets. Despite their scalability, there is no direct interaction between the query and the document, as they are encoded independently during embedding generation, resulting in a loss of structural relationships among them (Humeau et al., 2019; Tran et al., 2024).

Recent advances in GNNs on the scientific dataset recommendation task offer a promising approach to solve the issue (Altaf et al., 2019; Qayyum et al., 2025). However, these methods generally lack inductive capability, which is essential for handling unseen nodes. Such inductive ability is crucial in scientific dataset recommendation, where the number of papers and datasets continues to grow rapidly. Aside from that, most GNNs-based approaches tend to overlook the rich textual content associated with these nodes, resulting in incomplete semantic representations.

Several attempts have been made to address the unseen node using an inductive GNNs approach in the field of recommendation systems (Teru et al., 2020; Xiao et al., 2023), with the ability to categorise labels that did not exist during training. Inductive GNNs have not yet been applied to the scientific dataset recommendation task, although we believe that leveraging them could offer significant benefits.

To address this challenge, we propose **TeG-DRec** (**Te**xtual **G**raph **D**ataset **Rec**ommendation), a framework that integrates textual content with inductive GNNs. **TeG-DRec** is designed to handle realistic scenarios where new scientific papers or

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 103–114
December 23, 2025 ©2025 Association for Computational Linguistics

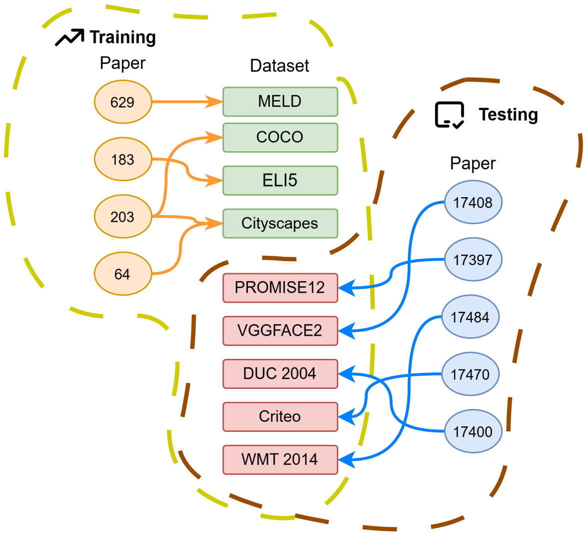


Figure 1: Research problem where the scientific papers in testing did not appear in the training set (unseen paper nodes in blue colour) connected with the label dataset that does not have any existing link with scientific papers during training (unseen dataset nodes in red colour)

scientific datasets are continuously introduced without explicit links to existing entities. As illustrated in Figure 1, unseen nodes refer to the nodes that do not appear during training or nodes that do not have any connection with any nodes during training.

Aside from that, TeG-DRec introduces a subgraph generation module that jointly enables inductive learning, contrastive learning, and margin-based optimisation within a cohesive training process. By combining the strengths of both textual and structural modalities, **TeG-DRec** effectively captures semantic and relational dependencies, offering robust inductive generalisation and improved scientific dataset recommendation performance.

The recommendation system works by taking a set of input queries, including the query, keyword query, and abstract. These inputs are then passed to **TeG-DRec** for the recommendation process, where the model predicts and outputs the Top-K datasets that best match the given inputs. This process is illustrated in Figure 2. The dataset used in our experiment consists of two node types: scientific papers and datasets, where the datasets serve as the target items to be recommended for each paper.

We compare **TeG-DRec** with text-based IR methods and a graph-based baseline. The text-based IR methods follow a neural bi-encoder framework (Ma et al., 2025), leveraging recent embedding models, which include SciBERT (Beltagy et al., 2019), Contriever (Lei et al., 2023), BGE-

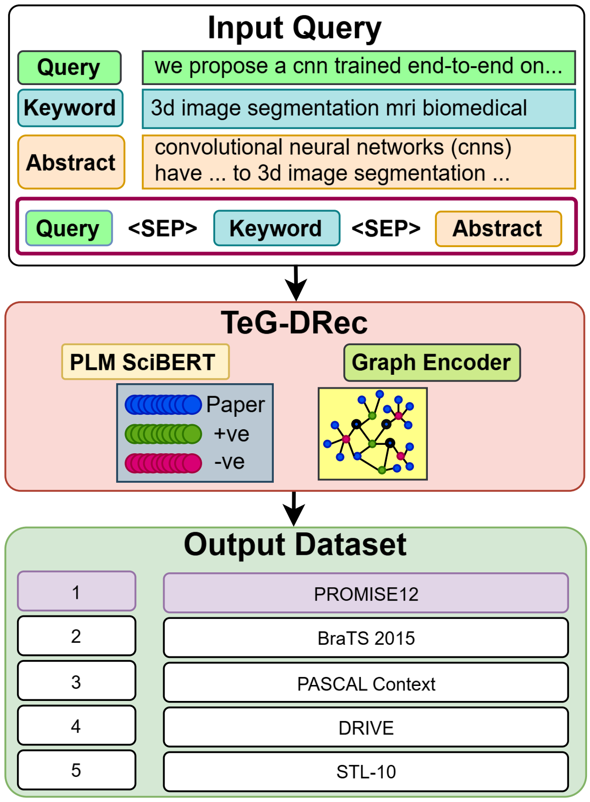


Figure 2: Overview of the scientific dataset recommendation process in **TeG-DRec**, from the input to output

M3 (Chen et al., 2024), and E5 (Wang et al., 2024) that provide strong semantic representations for scientific and general-domain retrieval tasks. For graph-based baselines, we compare **TeG-DRec** against GraphSAGE (Hamilton et al., 2017), Relational Graph Convolutional Networks (R-GCN) (Schlichtkrull et al., 2017) and Graph Attention Networks (GAT) (Veličković et al., 2018), which rely on the structural relations in the graph without incorporating the textual components of **TeG-DRec**. **TeG-DRec** consistently outperforms these baselines across all evaluation metrics, demonstrating its ability to capture both semantic and structural relationships effectively. In summary, this work makes three key contributions:

1. We propose **TeG-DRec**, a framework for scientific dataset recommendation that supports inductive recommendation for newly published scientific papers and scientific datasets, effectively handling unseen nodes without retraining.

2. We introduce a unified framework that integrates inductive GNNs with textual representations to jointly capture structural and semantic information.

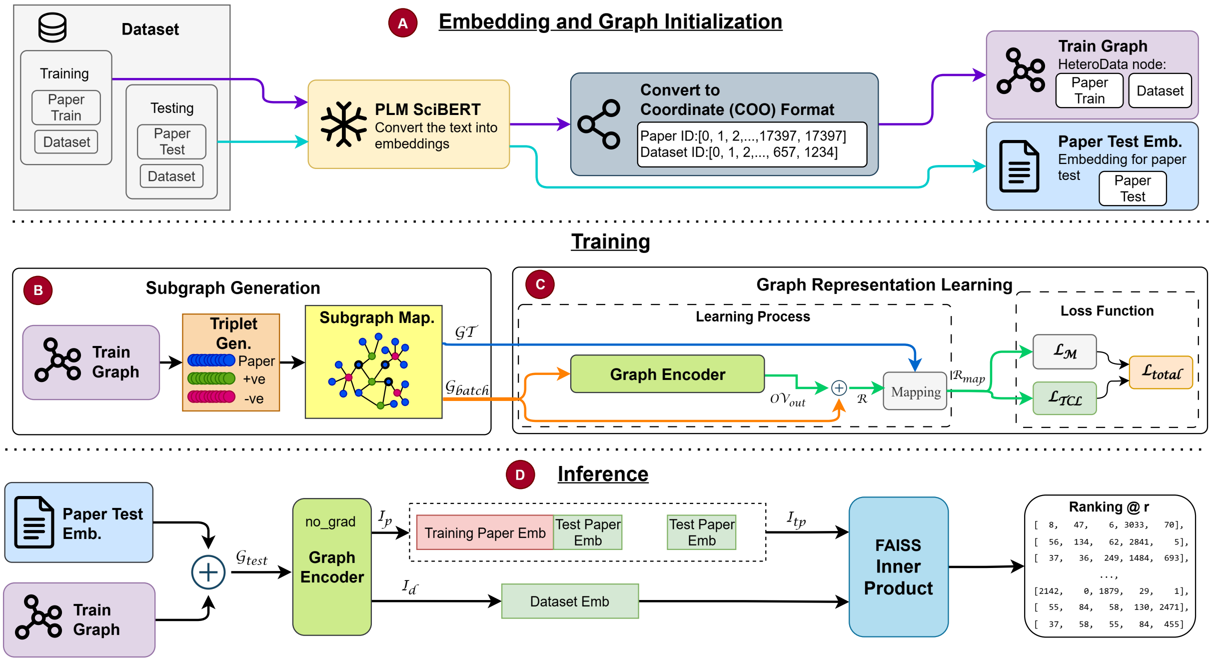


Figure 3: Overview of **TeG-DRec**, which consists of four main modules: (A) Embedding and Graph Initialisation, (B) Subgraph Generation, (C) Graph Representation Learning, and (D) Inference

3. We conduct extensive experiments on a publicly available benchmark and demonstrate that **TeG-DRec** consistently outperforms strong text-based IR and graph-based baselines.

2 Related Work

The techniques used to create the scientific dataset recommendation problem can be categorised into two groups: text-based IR and graph-based methods.

Text-Based IR Text-based IR for scientific dataset recommendation can be categorised into traditional methods and neural bi-encoder methods. The traditional method comprises BM25 (Keller and Munz, 2022), which ranks the dataset based on term-frequency matching, and a SciBERT-based text classification model (Beltagy et al., 2019; Färber and Leisinger, 2021). More recently, the Neural Bi-Encoder method proposed by Viswanathan et al. (2023) adopts a neural bi-encoder with SciBERT embeddings to encode both scientific papers and datasets. However, their model encodes scientific papers and datasets separately, which ignores their structural relationships.

Graph-Based Method The graph-based method can leverage the structural relationships between scientific papers and datasets. These structural relationships refer to the connections between scientific papers and the datasets they use, or they can be citation network among papers, datasets, and other related papers. Altaf et al. (2019) proposed a heterogeneous variational graph autoencoder (HVGAE) that integrates a citation network with paper–dataset associations to generate more informative representations for recommendation. Similarly, Qayyum et al. (2025) utilised GNNs enriched with textual features to recommend relevant datasets. However, their method can only handle transductive graphs, which require the nodes to be present during training. This limits the usage of the model in real-world situations where the nodes are constantly added. To address this limitation, several inductive graph learning frameworks have been proposed for recommendation systems, including GraphSAGE (Hamilton et al., 2017) and Graph Attention Networks (GAT) (Veličković et al., 2018), which enable improved generalisation via neighbourhood aggregation mechanisms. Additionally, Relational Graph Convolutional Networks (R-GCN) (Schlichtkrull et al., 2017), although not inherently inductive, offer a promising solution by effectively handling multiple edge types within a graph.

3 TeG-DRec Framework

3.1 Overview

TeG-DRec (**Te**xtual **G**raph-**D**ataset **Rec**ommendation) integrates textual content with graph structures, enabling the model to generalise toward unseen nodes by leveraging both the semantic properties and the structural information of the nodes. To achieve this, four main modules have been designed to address the specific requirements: (A) Embedding and Initialisation, (B) Subgraph Generation, (C) Graph Representation Learning, and (D) Inference, as depicted in Figure 3. In particular, the (B) Subgraph Generation module ensures that the textual content and graph structures are correctly aligned and passed through the inductive graph and loss components.

3.2 Embedding and Graph Initialisation

Embedding and Graph Initialisation module is responsible for encoding the textual information into embeddings and constructing the corresponding graph connections. This module ensures that the rich textual content is effectively integrated into the graph structure.

The dataset used in our experiments comprises descriptions of scientific papers, datasets, and the associations indicating which datasets are used by each paper. This relationship refers to the connection between the scientific papers, their corresponding positive datasets, and their corresponding negative datasets. A positive dataset refers to the dataset actually used by a given scientific paper, while a negative dataset represents a hard negative sample that is not used by the scientific paper. This is further illustrated in Figure 4.

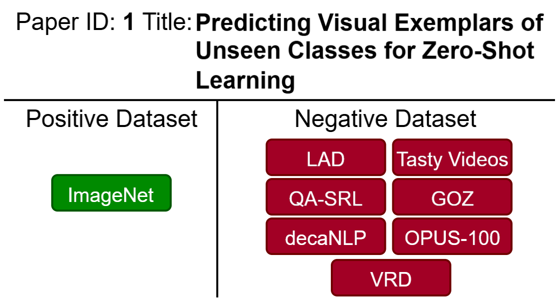


Figure 4: Example of positive and negative dataset sample for a paper with ID number 1

To remove unnecessary symbols and particular words from the dataset descriptions, a cleaning process is applied prior to using SciBERT to produce dense vector embeddings. Conversely, the relationships between scientific papers and datasets are represented in Coordinate Format (COO) as sparse matrices.

The COO maps of the scientific paper ID p with its associated positive dataset ID d^+ and negative dataset ID d^- as illustrated in Figure 3. The dense vector embeddings of the description and the COO of the scientific paper and dataset are subsequently passed on to the HeteroData G class in Pytorch Geometric (PyG) (Fey and Lenssen, 2019). The HeteroData G class utilises dense vector embedding $\mathcal{V}$ and COO format $\mathcal{E}$ to generate a train data graph. Meanwhile, the test paper dense vector embeddings $\mathcal{P}_{test}$ are extracted to be used later in the Inference section.

3.3 Subgraph Generation

The subgraph generation module enables the model to handle subgraphs rather than the entire graph, ensuring computational efficiency when learning on a large-scale graph. Additionally, it guarantees that graph nodes are properly aligned with their corresponding textual features before being passed to the Graph Representation Learning module. Proper alignment is essential, as misalignment would disrupt feature aggregation across connected nodes, thereby hindering inductive learning. Additionally, misalignment could also result in incorrect node pairings during loss computation. Figure 5 illustrates the flow of node IDs within this module. Here, **P** (blue) denotes the IDs of scientific papers, while **+** (green) and **–** (red) represent the positive and negative datasets associated with the scientific papers, as described previously in Figure 4.

This module consists of two subcomponents: a triplet generation process that constructs triplets from the train graph, and a subgraph mapping procedure that extracts subgraphs from the training graph and maps them according to the global node IDs.

Triplet Generation Triplet generation is used to efficiently load and manage the triplet set $\mathcal{T}$, which consists of scientific paper ID p, positive dataset ID d^+, and negative dataset ID d^-, from the train graph G. The triplet set $\mathcal{T}$ is then shuffled and partitioned into batches $\mathcal{T}_b$ as outlined in Algorithm 1. Subsequently, these batches are fed into the subgraph sampling module for subgraph mapping.

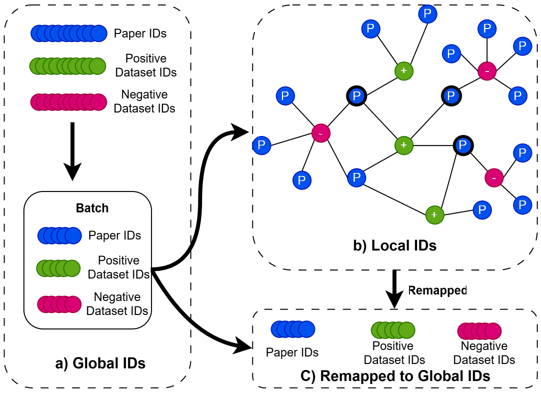


Figure 5: Flow inside Subgraph Module shows that after the subgraph extraction, the node IDs become local IDs. This IDs is remapped back into global IDs to keep track of the IDs

Algorithm 1 Triplet Generation and Subgraph Mapping

Require:
Dense Vector Embeddings $\mathcal{V}$
COO Format $\mathcal{E}$
Heterogeneous Train Graph $G = (\mathcal{V}, \mathcal{E})$
Scientific Paper ID p
Positive Dataset ID d^+
Negative Dataset ID d^-
Set of Triplets $\mathcal{T} = \{(p, d^+, d^-)\}$
Batch Size B
Number of Neighbours k
Ensure:
Mini-batch Subgraph $\mathcal{G}_{batch}$
Mapped Triplet Indices:
    $\mathcal{GT} = \{(P_{local}, D^+_{local}, D^-_{local})\}$
Triplet Generation:
Dataset $\leftarrow$TripletGeneration$(\mathcal{T})$
$\mathcal{T}_b \leftarrow$ Shuffle(Dataset)
return $\mathcal{T}_b$
Subgraph Mapping:
for b in $\mathcal{T}_b$ **do**
    $p_{batch} \leftarrow \{p_i\}_{i=1}^{b}$
    $\mathcal{G}_{batch} \leftarrow \texttt{NeighborLoader}(G, p_{batch}, k)$
    $\mathcal{GT} \leftarrow \texttt{GlobaltoLocal}(\mathcal{G}_{\text{batch}}, b)$
end for
return $\mathcal{G}_{batch}, \mathcal{GT}$

Subgraph Mapping A batch b is sampled from $\mathcal{T}_b$ and used to generate a subgraph. The scientific paper IDs p from the batch b serve as input nodes p_{batch} for the `NeighborLoader()` from PyG. We perform two-hop subgraph sampling, where the first hop samples twenty neighbours and the second hop samples fifteen. The resulting subgraph $\mathcal{G}_{batch}$ is then remapped to global indices $\mathcal{T}_b$ using the `GlobaltoLocal()` function, producing a triplet local $\mathcal{GT}$.

This remapping ensures that node identities remain consistent, as subgraph construction replaces global indices with local ones. This step ensures that the loss function receives the correct node IDs with its embeddings. The whole procedure is shown in Algorithm 1.

3.4 Graph Representation Learning

Graph Representation Learning module enables **TeG-DRec** to handle unseen nodes (refer Figure 1) as it comprises two main subcomponents: graph encoder and loss functions. The graph encoder uses an inductive GNNs to learn representations from the subgraph, which helps it to generalise towards unseen nodes.

Aside from that, the loss function computes ranking and contrastive losses and uses gradient-based optimisation during training via backpropagation. Ranking losses aim to prioritise positive pairs over negative ones, ensuring that relevant datasets are ranked higher than irrelevant ones. In contrast, contrastive loss enhances representation learning by aligning similar views in the embedding space while separating dissimilar ones, improving the model's ability to distinguish between different data points.

Graph Encoder The graph encoder processes the input graph $\mathcal{G}_{\text{batch}}$ which represents the IDs of scientific papers, positive datasets, and negative datasets (see Figure 5), along with their corresponding embeddings, using an inductive GNN to generate the output views OV_{out}. These output views are then concatenated with the original pre-encoded representations of $\mathcal{G}_{\text{batch}}$ to form the final recommendation embeddings R. The recommendation embeddings R are mapped based on the local triplet mapping $\mathcal{GT}$, resulting in the mapped recommendation embeddings R_{map}, which are then passed to the loss functions. The graph encoder is implemented as a modular component, allowing it to operate with various types of inductive GNNs. The detailed procedure is outlined in Algorithm 2.

In this study, we incorporate multiple inductive GNNs encoders, including GraphSAGE (Hamilton et al., 2017), R-GCN (Schlichtkrull et al., 2017) and GAT (Veličković et al., 2018).

Loss Functions Our model is optimised using two main types of loss functions: ranking loss and

Algorithm 2 Graph Representation Learning

Require:
Mini-batch Subgraph $\mathcal{G}_{batch}$
Mapped Triplet Local:
$\quad \mathcal{GT} = \{(P_{local}, D^{+}_{local}, D^{-}_{local})\}$
Ensure:
Rec Embeddings Mapped $\mathcal{R}_{map}$
Learning Process:
$OV_{\text{out}} \leftarrow \text{GraphEncoder}(\mathcal{G}_{\text{batch}})$
$\mathcal{R} \leftarrow OV_{\text{out}} \circ \mathcal{G}_{\text{batch}}$
$\mathcal{R}_{map} \leftarrow R$ such that $R \subseteq \mathcal{GT}$
return $\mathcal{R}_{map}$

contrastive loss. For ranking loss, we use margin ranking loss $\mathcal{L}_M$, which enforces a margin between positive and negative scores to maximise the difference between them. The loss function is defined in Eq. (1).

$$\mathcal{L}_M = \max(0, -y * (x_1 - x_2) + \text{ margin }), \quad (1)$$

where x_1 denotes the positive sample and x_2 denotes the negative sample, while y is a binary label. In our experiments, we set $y = 1$ to enforce that the positive sample x_1 should always be ranked higher than the negative sample x_2.

For contrastive loss, it is applied between the text embeddings of scientific papers and their corresponding positive datasets. The objective is to encourage the model to draw semantically aligned paper–dataset pairs closer in the embedding space, while pushing apart unrelated pairs. It is given by:

$$\mathcal{L}_{TCL} = \text{InfoNCE}(\mathbf{Z}_p, \mathbf{Z}_d, \tau), \quad (2)$$

where $\mathcal{L}_{TCL}$ refer to text contrastive loss, $\mathbf{Z}_p$ indicates the paper embeddings and $\mathbf{Z}_d$ is the positive dataset embeddings. τ is a temperature, which is a constant. This contrastive loss is formulated using the InfoNCE loss (Rusak et al., 2024), as defined in Eq. (3).

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\frac{\mathbf{z}_i^{(1)} \cdot \mathbf{z}_i^{(2)}}{\tau}\right)}{\sum_{j=1}^{N} \exp\left(\frac{\mathbf{z}_i^{(1)} \cdot \mathbf{z}_j^{(2)}}{\tau}\right)}, \quad (3)$$

where $\mathbf{z}_i^{(1)}$ is the original view of sample i, $\mathbf{z}_i^{(2)}$ is the augmented view of i, and τ is the temperature constant. $\mathbf{z}_j^{(2)}$ refers to the positive sample from the augmented view.

All losses are multiplied by their respective ratios for balanced performance, then combined with a regularisation loss to avoid overfitting. The equation of the batch loss is defined as Eq. (4).

$$\mathcal{L}_{total} = \mathcal{L}_{TCL} + \mathcal{L}_M + \mathcal{L}_{L2reg}, \quad (4)$$

where $\mathcal{L}_{SCL}$ is the structure contrastive loss, $\mathcal{L}_{TCL}$ indicates the text contrastive loss, $\mathcal{L}_M$ refers to the margin loss and $\mathcal{L}_{L2reg}$ shows the L2 regularization loss.

3.5 Inference

The inference module enables **TeG-DRec** to evaluate scenarios involving truly unseen nodes. Evaluating such scenarios is essential for simulating real-world conditions, as new scientific papers and datasets continue to appear. To achieve that, we concatenate the test paper dense vector embeddings $\mathcal{P}_{test}$ with the heterogeneous train graph dense vector embeddings $G(\mathcal{V})$, producing the test graph G_{test} as shown in Algorithm 3. The dense vector embeddings of test papers $\mathcal{P}_{test}$ have no connections to any dataset nodes within the test graph G_{test}. This ensures that the encoder processes test nodes independently of the training structure. The test graph is passed to the GraphEncoder$_{nograd}$ for the encoding process. After obtaining paper I_p and dataset I_d embeddings from the model, we extract test paper embeddings I_{tp} by indexing the unique nodes of test paper dense vector embeddings index $\mathcal{P}_{test}$.

To generate recommendations, we compute the maximum inner product between each test paper embedding I_{tp} and dataset embeddings I_d using FAISS (Douze et al., 2024) and retrieve the top-r results. The overall inference process is illustrated in Figure 3.

4 Experiments

4.1 Experimental Setup

Dataset We use the DataFinder Dataset (Viswanathan et al., 2023) to train and evaluate our model. The dataset is available on GitHub[1]. This dataset contains metadata about scientific papers and their associated datasets. It is pre-split into training and test sets. The training data includes true positive and hard negative dataset pairs for each publication, sourced from the Papers with

[1] https://github.com/viswavi/datafinder/tree/main

Algorithm 3 Inference

Require:
1: Dense Vector Embeddings $\mathcal{V}$
2: COO Format $\mathcal{E}$
3: Heterogeneous Train Graph $G = (\mathcal{V}, \mathcal{E})$
4: Test Scientific Papers $\mathcal{P}_{test} = \mathcal{V}$

Ensure:
5: Top-r results top-r

Inference Process:
6: $\mathcal{G}_{test} \leftarrow G(\mathcal{V}) \circ \mathcal{P}_{test}$
7: $I_p, I_d \leftarrow$ GraphEncoder$_{nograd}(\mathcal{G}_{test})$
8: $I_{tp} \leftarrow I_p[\text{Unique}(\mathcal{P}_{test})]$
9: top-$r \leftarrow$ FAISSInnerProduct(I_{tp}, I_d)
10: **return** top-r

Code[2] website. The hard negative datasets are selected using BM25. These hard negatives do not necessarily overlap with true positives. The test data consists of expert-annotated evaluations from SciREX (Jain et al., 2020).

To ensure that the test data align with our truly unseen node scenario, we remove scientific papers that interact with positive datasets. The remaining connected datasets in the test data are then removed from the hard negative datasets in the train data. This is to ensure that the test data are truly unseen. Table 1 summarises the statistics of the dataset. Appendix A outlines the available features within the dataset.

Data	Train	Test
# of scientific papers	17,397	88
# of positive datasets	461	74
# of positive interactions	20,789	126
# of negative datasets	2,570	–
# of negative interactions	118,997	–

Table 1: The statistics of scientific papers and datasets in Datafinder Dataset

Evaluation metrics We evaluate our method using five standard recommender system metrics: Precision **(P)**, Recall **(R)**, Normalised Discounted Cumulative Gain **(NDCG)**, Mean Average Precision **(MAP)**, and Mean Reciprocal Rank **(MRR)**. For top-r, we set $r = 5$, reflecting real-world usage where users engage with the highest-ranked suggestions. This is particularly relevant for Precision, Recall, and NDCG, all of which involve the top-r metric in their calculation.

Implementation To ensure separation between node features, a unique token is added before each feature during encoding. For training stability and convergence, we implemented a learning rate scheduler that combines linear warmup with cosine annealing. The implementation was done using PyTorch and PyG (Fey and Lenssen, 2019), with experiments run on an NVIDIA RTX 6000 Ada GPU with 48GB VRAM. The hyperparameters used in this experiment are detailed in Appendix B to facilitate reproducibility.

4.2 Baselines

To evaluate the effectiveness of our proposed method, we compare it against seven baseline approaches, which are classified into two groups:

Text-Based IR Method consists of four baselines, each of which utilises the neural biencoder framework by Ma et al. (2025) with four different embedding models, including:

1. **SciBERT** (Beltagy et al., 2019) is a pretrained BERT-based language model specifically designed for scientific and scholarly text.
2. **Contriever** (Lei et al., 2023) is an unsupervised dense information retrieval model that leverages contrastive learning to train a bi-encoder that maps queries and documents to a shared embedding space.
3. **BGE-M3** (Chen et al., 2024) is a multilingual embedding model designed to handle various retrieval tasks efficiently.
4. **E5-Large** (Wang et al., 2024) is a text embedding that is trained using weakly-supervised contrastive learning on a large-scale dataset of text pairs.

Graph-Based Method consist of three baselines:

1. **GraphSAGE** (Hamilton et al., 2017) GraphSAGE is an inductive graph whose primary goal is to learn node embeddings that generalise towards unseen nodes, rather than only represented nodes seen during training.
2. **Relational Graph Convolutional Networks (R-GCN)** (Schlichtkrull et al., 2017) is an extension of Graph Convolutional Networks (GCNs), which is designed to handle graphs where edges have types or relations.

[2] https://huggingface.co/papers/trending

Methods	Datafinder Dataset (Unseen Configuration)				
	P@5	**R@5**	**NDCG@5**	**MAP**	**MRR**
Text-based IR Method					
Neural Biencoder (SciBERT)	0.015	0.053	0.032	0.023	0.029
Neural Biencoder (Contriever)	0.018	0.064	0.039	0.028	0.034
Neural Biencoder (BGE-M3)	0.017	0.063	0.051	0.042	0.055
Neural Biencoder (E5-Large-V2)	0.011	0.038	0.026	0.020	0.026
Graph-Based Method					
GraphSAGE	0.005	0.017	0.008	0.004	0.006
R-GCN	0.002	0.011	0.011	0.011	0.011
GAT	0.009	0.045	0.020	0.012	0.012
TeG-DRec (GraphSAGE)	<u>0.066</u>	<u>0.237</u>	<u>0.160</u>	<u>0.124</u>	<u>0.153</u>
TeG-DRec (RGCN)	0.050	0.176	0.123	0.096	0.119
TeG-DRec (GAT)	**0.111**	**0.419**	**0.315**	**0.260**	**0.316**

Table 2: The recommendation performance of our method against baselines for the text-based IR method and the graph-based method. **Bold** is the best, <u>underline</u> is the second best.

Model	**P@5**	**R@5**	**NDCG@5**	**MAP**	**MRR**
w/o SciBERT	0.009	0.045	0.020	0.012	0.012
w/o GNNs (GAT)	<u>0.015</u>	<u>0.053</u>	<u>0.032</u>	<u>0.023</u>	<u>0.029</u>
TeG-DRec(GAT)	**0.111**	**0.419**	**0.315**	**0.260**	**0.316**

Table 3: The ablation study for each component. **Bold** is the best, <u>underline</u> is the second best.

3. **Graph Attention Networks (GAT)** (Veličković et al., 2018) introduces attention mechanisms, enabling the model to learn the importance of neighbouring nodes dynamically.

4.3 Results

Table 2 compares the performance of the graph-based method with **TeG-DReC** with text-based IR and graph-based methods alone on the DataFinder dataset, evaluated under a truly unseen configuration. Our methods consistently outperform all baselines across all metrics, demonstrating their effectiveness and robustness.

The results show that graph-based models combined with **TeG-DRec** outperform their graph-only counterparts across all evaluation metrics. In particular, **TeG-DReC**(GAT) achieves a substantial improvement in R@5, surpassing its baseline by 0.374. It also exhibits superior ranking performance, with gains of 0.304 in MRR, 0.295 in NDCG@5, and 0.240 in MAP compared with **TeG-DRec**(GAT). These metrics assess ranking quality where NDCG considers both relevance and position, MRR reflects the rank of the first relevant result, and MAP measures the average precision of the ranking. The P@5 metric also increases by 0.102 over the GAT baseline. Beyond **TeG-DRec**(GAT), both **TeG-DRec**(GraphSAGE) and **TeG-DRec**(RGCN) also achieve significant improvements over their respective graph-only baselines.

Although the neural bi-encoder using Contriever as the embedding model achieves the highest results among all text-based IR methods, all graph-based models integrated with **TeG-DRec** still outperform it. The lowest-performing variant, **TeG-DRec** (RGCN), surpasses the Contriever-based bi-encoder by 0.032, 0.068, 0.084, 0.085, and 0.112 for P@5, MAP, NDCG@5, MRR, and R@5, respectively. These results indicate that while neural bi-encoders capture rich semantic similarities, incorporating relational structure via graph learning further enhances alignment between scientific papers and datasets, leading to superior recommendation performance.

4.4 Ablation Study

To assess the contribution of each component in our model, we conducted an ablation study by removing one component at a time, with results shown in Table 3. In this ablation study we pick **TeG-DReC**(GAT) as our original results. Removing the textual component, SciBERT, results in a substantial drop across all metrics, particularly in R@5 and MRR, which decrease by 0.374 and 0.304, respectively. This performance drop is also reflected in other metrics, such as NDCG@5, MAP, and P@5, which decrease by 0.295, 0.248, and 0.101, respectively. This underscores the critical role of textual features in capturing semantic alignment between papers and datasets for accurate recommendations.

Similarly, removing the GNN component also reduces performance across all metrics, with notable decreases in R@5 and MRR of 0.366 and 0.287, respectively. Other metrics, including NDCG@5, MAP, and P@5, also show decreases of 0.283, 0.237, and 0.096, respectively. These results indicate that while semantic representations of publications and datasets significantly improve model performance, integrating graph-structured information further enhances recommendation quality, highlighting the complementary benefits of combining textual and structural components in **TeG-DRec**.

4.5 Error Analysis

We conducted an error analysis on the **TeG-DRec** recommendation output. There are two main types of errors in the recommended results:

Biased towards certain dataset TeG-DRec shows a bias toward certain datasets, such as TreQA, which appears most frequently in recommendations even though it occurs only once in the ground truth, as shown in Figure 6. Similar trends are observed for other over-recommended datasets absent from the actual ground truth. A debiasing technique can be implemented to solve the problems.

Textual bias in dataset query Textual bias in the training data may affect the recommendations. For example, as shown in Figure 6, SQuAD, a question-answering dataset, appears 312 times in positive training interactions. This high frequency can bias the model toward recommending TreQA, another question-answering dataset, even when it is absent from the ground truth. Incorporating content-aware attention could help mitigate this issue.

5 Conclusion

This research introduced **TeG-DRec**, a framework for scientific dataset recommendation that unifies GNNs with textual content via a subgraph module, ensuring that textual content and graph structures are correctly aligned and passed to the inductive graph and loss components. This integration enables the model to better generalise towards unseen data. The framework leverages textual representations from SciBERT and incorporates inductive GNNs, which are adaptable to various types of inductive graph models. Experimental results on the Datafinder dataset with truly unseen nodes show that our method outperforms previous baselines, including both text-based IR and graph-based approaches. Future work should incorporate a debiasing technique for recommendations to reduce popularity bias. This can be done by re-weighting the training loss based on dataset frequency, which means less frequent datasets get a higher weight. Aside from that, using content-aware attention rather than simply aggregating the textual embedding reduces bias from frequent, irrelevant words or phrases.

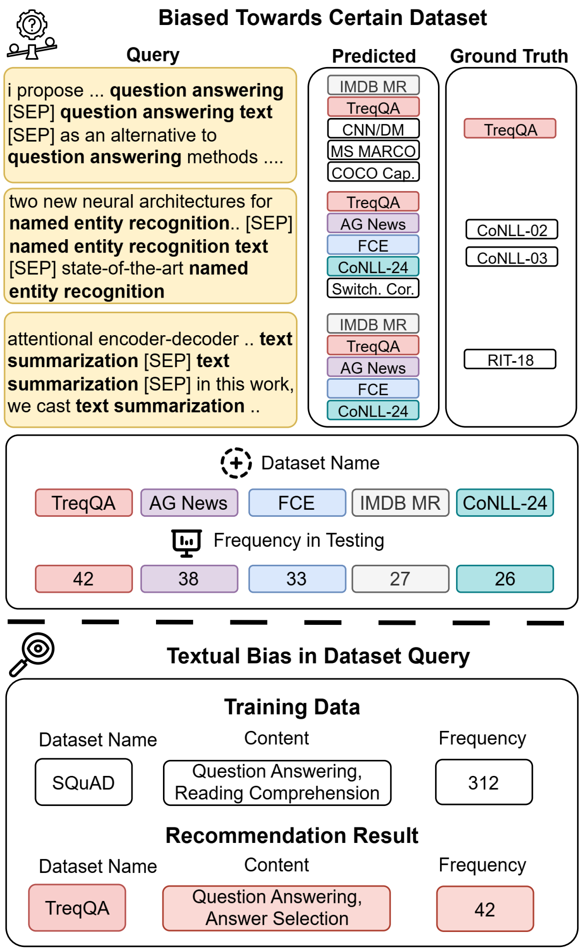


Figure 6: We present two examples for error analysis: the top illustrates a case where the **TeG-DRec** is biased towards a certain dataset, while the bottom highlights textual bias in the dataset query.

Limitations

Our major limitation is that our method relies heavily on the quality and availability of textual information (e.g, paper abstracts and dataset descriptions). In cases where the text is noisy, incomplete, or missing, the recommendation performance may degrade. Another limitation is that the availability of datasets for dataset recommendation systems is very low compared to other datasets which make use heavily rely on Datafinder Dataset alone.

Ethical Statement

This work adheres to the ethical standards outlined in the ACL Code of Ethics and the general principles of responsible AI research. All data used in this study are publicly available and used strictly for research purposes under their respective licenses. No personally identifiable information (PII) or sensitive content was collected or processed. We also took care to examine potential sources of bias and ensure that model outputs do not propagate harmful or discriminatory associations.

Acknowledgements

We would like to thank anonymous reviewers for their helpful comments and suggestions. This work is supported by the JKA (2024M-557). Ammar Qayyum is funded by the MEXT scholarship, Grant Number 233206.

References

Basmah Altaf, Uchenna Akujuobi, Lu Yu, and Xiangliang Zhang. 2019. Dataset recommendation via variational graph autoencoder. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 11–20. IEEE.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Tora Fahrudin and Dedy Rahman Wijaya. 2024. New custom rating for improving recommendation system performance. *Journal of Big Data*, 11(1):91.

Michael Färber and Ann-Kathrin Leisinger. 2021. Recommending datasets for scientific problem descriptions. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 3014–3018, New York, NY, USA. Association for Computing Machinery.

Matthias Fey and Jan Eric Lenssen. 2019. Fast graph representation learning with pytorch geometric. *Preprint*, arXiv:1903.02428.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.

Sarthak Jain, Madeleine Van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. Scirex: A challenge dataset for document-level information extraction. *arXiv preprint arXiv:2005.00512*.

Jüri Keller and Leon Paul Mondrian Munz. 2022. Evaluating research dataset recommendations in a living lab. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 135–148. Springer.

Yibin Lei, Liang Ding, Yu Cao, Changtong Zan, Andrew Yates, and Dacheng Tao. 2023. Unsupervised dense retrieval with relevance-aware contrastive pretraining. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10932–10940, Toronto, Canada. Association for Computational Linguistics.

Xueguang Ma, Luyu Gao, Shengyao Zhuang, Jiaqi Samantha Zhan, Jamie Callan, and Jimmy Lin. 2025. Tevatron 2.0: Unified document retrieval toolkit across scale, language, and modality. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 4061–4065, New York, NY, USA. Association for Computing Machinery.

Özlem Özgöbek, Nafiseh Shabib, and Jon Atle Gulla. 2014. Data sets and news recommendation. In *UMAP Workshops*.

Shreya Patankar, Hitesh Prajapati, Jeet Shah, and Ankit Upadhyay. 2023. Automl-learning, understanding and applying machine learning to datasets. In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 919–922. IEEE.

Ammar Qayyum, Bassamtiano Irnawan, Sheng Xu, Zihao Hu, Fumiyo Fukumoto, and Yoshimi Suzuki. 2025. Textual graph contrastive learning for enhanced dataset recommendation. In *2025 19th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–4. IEEE.

Ziheng Qin, Zhaopan Xu, Yukun Zhou, Zangwei Zheng, Zebang Cheng, Hao Tang, Lei Shang, Baigui Sun, Xiaojiang Peng, Radu Timofte, and 1 others. 2024. Dataset growth. In *European Conference on Computer Vision*, pages 58–75. Springer.

Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S Zimmermann, and Wieland Brendel. 2024. Infonce: Identifying the gap between theory and practice. *arXiv preprint arXiv:2407.00143*.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2017. Modeling relational data with graph convolutional networks. *Preprint*, arXiv:1703.06103.

Komal Teru, Etienne Denis, and Will Hamilton. 2020. Inductive relation prediction by subgraph reasoning. In *International conference on machine learning*, pages 9448–9457. PMLR.

Hung-Nghiep Tran, Akiko Aizawa, and Atsuhiro Takasu. 2024. An encoding–searching separation perspective on bi-encoder neural search. *arXiv preprint arXiv:2408.01094*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. *Preprint*, arXiv:1710.10903.

Katrien Verbert, Hendrik Drachsler, Nikos Manouselis, Martin Wolpers, Riina Vuorikari, and Erik Duval. 2011. Dataset-driven research for improving recommender systems for learning. In *Proceedings of the 1st international conference on learning analytics and knowledge*, pages 44–53.

Vijay Viswanathan, Luyu Gao, Tongshuang Wu, Pengfei Liu, and Graham Neubig. 2023. Datafinder: Scientific dataset recommendation from natural language descriptions. *arXiv preprint arXiv:2305.16636*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Shuai Wang, Shengyao Zhuang, and Guido Zuccon. 2021. Bert-based dense retrievers require interpolation with bm25 for effective passage retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21, page 317–324, New York, NY, USA. Association for Computing Machinery.

Jiaren Xiao, Quanyu Dai, Xiaochen Xie, James Lam, and Ka-Wai Kwok. 2023. Adversarially regularized graph attention networks for inductive learning on partially labeled graphs. *Knowledge-Based Systems*, 268:110456.

Krishan Kant Yadav, Hemant Kumar Soni, and Nikhlesh Pathik. 2023. Recommendation system based on double ensemble models using knn-mf. *International Journal of Advanced Computer Science and Applications*, 14(5).

Zitong Zhang and Yaseen Ashraf. 2023. A content-based dataset recommendation system for biomedical datasets. In *2023 6th International Conference on Information and Computer Technologies (ICICT)*, pages 198–202. IEEE.

Appendices

A Datafinder Dataset Content

Table 4 presents the structure of the Datafinder Dataset. The dataset is organised into three main components: training scientific paper metadata, test scientific paper metadata, and dataset metadata. Each component is further divided into several categories, including content descriptions of papers and datasets, datasets referenced by scientific papers, publication metadata, unique identifiers, citation details, and additional information related to papers and datasets. The highlighted fields in Table 4 indicate the features utilised as node attributes for each corresponding entity in our model.

Training Scientific Paper		
Paper Content	**Paper ID**	**Paper Information**
title	paper_id	has_pdf_body_text
abstract	arxiv_id	mag_field_of_study
query	acl_id	has_inbound_citations
keyphrase_query	pmc_id	has_outbound_citations
Dataset	pubmed_id	has_pdf_sparse
positives	mag_id	has_pdf_sparse_abstract
negatives	**Citation Information**	has_pdf_parse_bib_entries
Paper Publication	author	has_pdf_parse_text
journal	outbound_citations	has_pdf_parse_body_text
venue	inbound_citations	has_pdf_parse_entries
doi		s2_url
year		

Test Scientific Paper		
Paper Content	**Paper ID**	**Paper Information**
abstract	-	task
query	**Citation Information**	domain
keyphrase_query	-	modality
Dataset		language
documents		training_style
Paper Publication		text_length
year		

Dataset		
Dataset Content	**Dataset ID**	**Dataset Information**
title	id	variants
content	**Citation Information**	
structured_info	-	
Dataset Publciation		
year		
date		

Table 4: Datafinder Dataset content, the highlighted box is the features which is used for node features

B Hyperparameters value

Table 5 shows the hyperparameter setting for the parameters that are used in **TeG-DRec**. The hyperparameters include the maximum length of the textual encoder, the hidden dimension, the optimiser and its learning rate, the number of epochs, the loss rate, the loss temperature, and the seed number.

Variable	**Value**
SciBERT Dimension	512
Hidden Dimension	256
Optimizer	Adamw
Learning Rate	GraphSAGE : 1e-3, R-GCN: 5e-3, GAT: 5e-3
Epoch	40 with early stopping after 5 epoch of no improvement
Warmup Epoch's Scheduler	5
InfoNCE Temperature	0.08
Margin Value	1
Contrastive Loss Rate	0.8
Margin Loss Rate	0.8
L2 Regression Loss	1e-4
Seed Number	1

Table 5: Hyperparameters variable and its value for the reproducibility purpose

Structured Outputs in Prompt Engineering: Enhancing LLM Adaptability on Counterintuitive Instructions

Jingjing Ye
yejj177@gmail.com
Independent Researcher

Song Bai
song.bai6174@gmail.com
Independent Researcher

Zhenyang Li
zli2022@alumni.usc.edu
Independent Researcher

Zheqi Shen
zheqi@shen.zone
Independent Researcher

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language processing tasks, yet they often exhibit cognitive inertia, rigidly adhering to ingrained training conventions even when prompted to deviate. This paper investigates the efficacy of structured output techniques in prompt engineering to mitigate such inertia and improve instruction-following on counterintuitive tasks. We argue that using the structured input and output with our framework yields significant performance gains, studied on the Inversed IFEval dataset across varying prompts and domains. This work contributes to the growing field of prompt engineering research by demonstrating structured outputs as a robust method for enhancing LLM logical reasoning.

1 Introduction

The recent advancements of Large Language Models (LLMs) have revolutionized artificial intelligence, enabling sophisticated applications in natural language understanding, generation, and reasoning. However, a persistent challenge is their tendency toward cognitive inertia, a phenomenon where models persist in following learned patterns from pre-training and fine-tuning, resisting deviations even under explicit instructions. This inertia manifests in scenarios requiring unlearning or counterintuitive behavior, such as generating flawed outputs intentionally or ignoring standard formatting conventions. The logic twist inside might be easy for elementary school students, but is proven difficult for LLM models, a factor critical in scientific reasoning.

Prompt engineering emerges as a non-invasive method to guide LLMs without retraining, encompassing techniques like zero-shot (Kojima et al., 2022; Li, 2023), few-shot (Dang et al., 2022), and chain-of-thought (CoT) prompting (Lyu et al., 2023; Zhang et al., 2024). Among these, structured outputs, which enforce responses in predefined formats such as JSON, XML or phased structures, offer verifiability and consistency, while reducing hallucinations and improving reliability. Recent advancements, including OpenAI's Structured Outputs feature, underscore their practical utility in production environments.

To evaluate these techniques on counterintuitive tasks, we employ the Inverse IFEval dataset (Zhang et al., 2025), an extension of the IFEval benchmark that inverts verifiable instructions to probe unlearning capabilities. The dataset includes challenges like Question Correction (answering incorrectly on purpose), Intentional Textual Flaws (introducing errors), Mid-turn Instruction Modification, and others, spanning diverse domains and languages.

Our contributions are: 1. We develop a framework that utilizes structured outputs to improve LLM responses to counterintuitive instructions; 2. We evaluate three structured prompts with varying output formats and determine that the list-structured approach performs best; 3. We investigate variants of the list-structured method and study the performance impact of explicit prioritization; 4. We test our approach on the Inverse IFEval (Zhang et al., 2025) benchmark and demonstrate that our list-structured prompting framework largely outperforms baselines, providing insights for more adaptable and logical AI systems.

2 Related Works

2.1 Prompt Engineering Techniques

Prompt engineering has evolved from basic input crafting to sophisticated strategies for eliciting optimal LLM responses. Surveys categorize techniques into zero/few-shot prompting (Dang et al., 2022; Kojima et al., 2022; Li, 2023), CoT (Lyu et al., 2023; Zhang et al., 2024), ToT (Yao et al., 2023; Mo and Xin, 2024; Ranaldi et al., 2024), and self-consistency methods (Zhou et al., 2025; Tauben-

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 115–120
December 23, 2025 ©2025 Association for Computational Linguistics

feld et al., 2025; Nowak, 2025). CoT, for instance, encourages step-by-step reasoning, while ToT explores multiple paths for complex problem-solving. Structured prompting extends these by imposing formats, such as role-playing or output schemas, to enhance control and parseability.

2.2 Cognitive Inertia and Unlearning in LLMs

LLMs exhibit human-like cognitive effects, including priming, anchoring (Lou and Sun, 2024), and irrational biases (Echterhoff et al., 2024; Tang and Kejriwal, 2024) in decision-making tasks. Cognitive inertia, a form of resistance to change, is particularly evident in instruction-following scenarios where models default to "helpful" behaviors despite contrary prompts. LLMs exhibit cognitive inertia, reflecting a persistent adherence to patterns learned during self-supervised pre-training. (Resnik, 2025) observes that biases in LLMs are not merely a result of training data, but are intrinsically embedded within the model architecture and optimization objectives. Optimizing for next-token prediction causes models to internalize statistical regularities, including societal biases, without distinguishing between high-probability patterns and harmful conventions. Humans, in contrast, can flexibly adjust behavior via metacognition, reasoning, and contextual judgment, enabling them to follow counterintuitive instructions. LLMs, lacking autonomous reasoning or self-correction, struggle to overcome entrenched patterns even when fine-tuned or aligned through RLHF. Cognitive inertia thus arises from the interaction of pre-training habits, modeling constraints, and limited post-hoc flexibility, leading models to reproduce established patterns rather than adapt to out-of-distribution tasks. One potential approach to mitigate this issue is to reconstitute LLMs' internal representations as structured representations, encoding entities, relations, logical structure, and distinctions between meaning, normativity, and factuality, thereby enhancing the model's flexibility in adapting to novel or counterintuitive instructions.

Unlearning benchmarks like TOFU, MUSE, WMDP, and RWKU assess models' ability to forget specific knowledge while retaining general capabilities. However, critiques highlight flaws in these benchmarks, such as over-optimistic evaluations due to separate testing of forget/retain queries.

2.3 Benchmarks for Instruction Following

Inverse IFEval (Zhang et al., 2025) is a new benchmark for testing counterintuitive adherence. It inverts the paradigm in IFEval (Zhou et al., 2023) that evaluates verifiable instructions. Constructed via human-in-the-loop processes, the inverse IFEval reveals that larger, instruction-tuned models paradoxically struggle more with deviations. Gaps persist in integrating structured prompting into such benchmarks, which our work addresses by proposing a verifiable framework.

3 Methodology

Our methodology centers on developing and testing a structured output framework designed to enhance LLMs' ability to follow counterintuitive instructions from the Inverse IFEval dataset. This framework decomposes the instruction-following process into four explicit phases: Instruction Parsing, Requirement Checklist, Structured Response and Self-Check. We explore multiple variants of this framework to identify the most effective implementation.

3.1 Dataset and Task Description

We evaluate our approach on the Inverse IFEval dataset, a challenging benchmark with 1012 high-quality samples designed to test LLMs' ability to follow counterintuitive instructions that contradict their training patterns. The dataset covers eight distinct instruction types: (1) Instructional Induction, (2) Mid-turn Instruction Modification, (3) Counterfactual Answering, (4) Counter-Conventional Formatting, (5) Question Correction, (6) Deliberately Incorrect Answers, (7) Intentional Textual Flaws, and (8) Code without Comments. These types span diverse domains and require models to override ingrained behaviors such as being helpful, following conventions, and producing polished outputs.

The dataset includes both English and Chinese subsets, enabling cross-lingual evaluation. For our experiments, we use stratified sampling to select 40-48 representative samples, ensuring balanced coverage across all eight instruction types. This sample size balances computational feasibility with statistical reliability while maintaining type diversity for robust evaluation.

3.2 Variants on Structured Approach

```
Before responding to this instruction, first analyze and
     structure it into clear components:

**Original Instruction**: {instruction}

**Step 1 - Parse the Instruction:**
Break down the instruction into these components:
- **Condition**: Any context, assumptions, or conditional
     statements
- **Questions**: The core tasks or questions being asked
- **Requirements**: Specific formatting, style, or content
     constraints
- **Distribution**: Length, structure, or organizational
     requirements

**Step 2 - Structured Analysis:**
Condition: [Extract any context or conditions]
Questions: [Identify the main task]
Requirements: [List all specific constraints]
Distribution: [Note any length/structure requirements]

**Step 3 - Systematic Response:**
Now provide your response, ensuring you address each
     component systematically:
```

Structured Prompt 1: Basic Structured Approach

We investigate varying structured prompts and adapt them to counterintuitive instructions. We begin with three initial variants: (1) a basic structured approach, Prompt 1, which applies the four phases in a simple textual format without additional enhancements, (2) JSON-structured prompting, Prompt 2, which organizes the components into machine-readable JSON fields (e.g., {"context": "...", "tasks": "..."}) for improved parseability and verifiability, and (3) checklist-based prompting, Prompt 3, which uses enumerated lists to break down requirements, promoting systematic adherence. These variants allow us to assess the impact of different structuring mechanisms on model performance.

For the checklist variant, we further investigate two sub-types: an equal checklist Prompt 3, where all requirements are treated uniformly without explicit prioritization, and a priority checklist Prompt 3, where items are categorized as CRITICAL (essential for compliance), IMPORTANT (affecting quality), or SECONDARY (enhancing completeness). This prioritization is intended to guide the model in focusing on high-impact elements first, potentially reducing cognitive inertia by emphasizing core constraints.

````
Parse this instruction into structured components, then
     respond:

**Instruction**: {instruction}

**Step 1: JSON Structure Analysis**
Parse the instruction into this JSON format:
```json
{{
  "context": "any background or situational information",
  "tasks": "the core tasks or questions",
  "format_requirements": ["list", "of", "formatting",
     "constraints"],
  "content_requirements": ["list", "of", "content",
     "constraints"],
  "length_requirements": "any length or size constraints",
  "style_requirements": "any tone or style requirements"
}}
```

**Step 2: Component-by-Component Response**
Now respond to the instruction, explicitly addressing each
     JSON component:

**Context addressed**: [How you handle the context]
**Task completion**: [Your core response]
**Format compliance**: [How you meet format requirements]
**Content compliance**: [How you meet content requirements]
**Length compliance**: [How you meet length requirements]
**Style compliance**: [How you meet style requirements]

**Final Response**:
````

Structured Prompt 2: JSON-Structured Prompting

```
You will respond to this instruction using a systematic
     parsing approach:

**Instruction to Analyze**: {instruction}

**Phase 1: Instruction Parsing**
Parse the instruction and identify:
- Context/Conditions: What situation or context is
     established?
- Core Tasks: What are the main things being asked?
- Format Requirements: Any specific formatting constraints?
- Content Requirements: What must be included/excluded?
- Length/Structure: Any size or organizational requirements?

**Phase 2: Requirement Checklist**
List each requirement as a checkable item (with priority):
- Requirement 1: [First constraint]
- Requirement 2: [Second constraint]
- Requirement 3: [Third constraint]
[Add more as needed]

**Phase 3: Structured Response**
Provide your response while explicitly addressing each
     requirement:

[Your response here]

**Phase 4: Self-Check**
Verify your response against each requirement (with priority):
- Requirement 1: Y/N [Brief check]
- Requirement 2: Y/N [Brief check]
- Requirement 3: Y/N [Brief check]
```

Structured Prompt 3: Checklist-Based Prompting

The baseline condition presents the original Inverse IFEval instructions directly to the models without any modifications, and serves as a control to measure the added value of our structured approaches. We evaluate these methods across five diverse LLMs: DeepSeek-Chat, Qwen, Gemini-2.5 Pro, o1-preview, and Claude-3.5-Sonnet, selected for their varying sizes and architectures to ensure generalizability. All models are accessed via the OpenRouter API with consistent generation parameters (temperature=1.0, max_tokens=4096, top_p=1.0, frequency_penalty=0, presence_penalty=0) to facilitate fair comparisons.

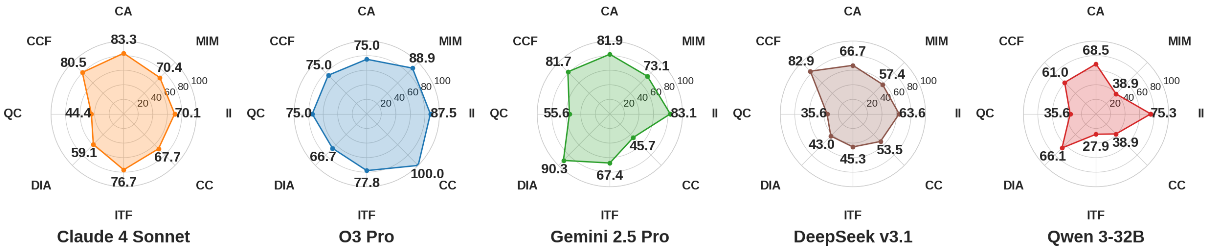


Figure 1: Performance breakdown by instruction type for five models using structured checklist approach. Each radar chart displays accuracy percentages across eight instruction types: II (Instructional Induction), MIM (Mid-turn Instruction Modification), CA (Counterfactual Answering), CCF (Counter-Conventional Formatting), QC (Question Correction), DIA (Deliberately Incorrect Answers), ITF (Intentional Textual Flaws), and CC (Code without Comments).

4 Experiments

Our experiments follow a sequential design to iteratively refine and validate the structured framework. We first test the three initial variants (basic, JSON, and checklist) on a subset of 32 samples from the Inverse IFEval dataset across the selected models.

4.1 Experiment Setup

Evaluation employs an LLM-as-a-Judge paradigm using Claude-4.5-Sonnet (temperature=0) for impartial, binary scoring (1 for semantic match with the reference answer, 0 otherwise). We use a subset of 32 samples for initial variant comparisons and the full set for final assessments. Statistical analysis includes paired t-tests ($\alpha = 0.05$) and Cohen's d for effect sizes, ensuring robust interpretation of results.

The performance comparison, summarized in Table 1, reveals that the checklist-based approach consistently outperforms the basic and JSON variants, achieving higher average accuracy. This suggests that the enumerated, human-readable format of checklists better mitigates cognitive inertia by enforcing explicit requirement tracking.

Method	**Accuracy (%)**
Basic Structure	43.8
JSON Structure	54.2
Checklist Structure (Equal)	60.4

Table 1: Comparison Between Varying Output Formats on DeepSeek V3.1

Specifically, the checklist method applies equal prioritization for better accuracy. As revealed in Table 2, we compare the equal checklist against a priority checklist under the same experimental setup and found that prioritization degrades performance. This is contrary to our expectations and may indicate that explicit hierarchies introduce unnecessary complexity, causing models to overfocus on specific categories and overlook holistic compliance.

Model	**Priority**	**Equal**
Claude 4 Sonnet	68.8	**77.1**
Gemini 2.5 Pro	67.5	**80.0**

Table 2: Impact of Priority System on Structured Prompting Performance

To better understand the performance characteristics across different instruction types, we analyze the breakdown of results for our equal checklist approach across the eight categories in the Inverse IFEval dataset. Figure 1 presents radar charts showing how different models handle various counterintuitive instruction types.

The results reveal that all models struggle significantly with "Question Correction", highlighting systematic challenges in this category across the board. While O3 Pro performs strongly across most categories, demonstrating high accuracies such as 87.5% in "Instructional Induction" and 100% in "Code without Comments", Claude 4 Sonnet excels in "Counterfactual Answering" (83.3%) and "Counter-Conventional Formatting" (80.5%), though its performance drops to 44.4% in "Question Correction". Gemini 2.5 Pro demonstrates high accuracy in "Deliberately Incorrect Answers" (90.3%), but struggles notably with "Code without Comments" (45.7%) and "Question Correction" (55.6%). Meanwhile, both DeepSeek v3.1 and Qwen 3-32B consistently face challenges, particularly in "Question Correction" (35.6% for both) and "Intentional Textual Flaws" (45.3% for DeepSeek v3.1 and 27.9% for Qwen 3-32B), underscoring common areas of difficulty among these models.

The eight categories are meticulously designed to probe nuanced aspects of instruction comprehension and execution, ranging from straightforward adherence to complex inferential tasks. Figure 2 presents stripe charts, where each model's performance is denoted by a unique marker, allowing for an immediate and intuitive comparison of their respective accuracy scores on each task category.

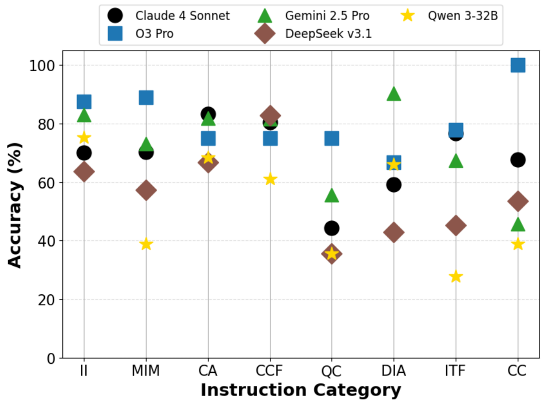


Figure 2: Comparison of Model Performance Across Eight Instruction Categories: II, MIM, CA, CCF, QC, DIA, ITF, CC

Finally, we benchmark our best variant, the equal checklist, against the baseline across all models and the full Inverse IFEval evaluation set. Figure 3 demonstrates substantial improvements, with statistically significant gains $p < 0.05$ and large effect sizes, confirming that structured prompting effectively enhances adaptability on counterintuitive tasks.

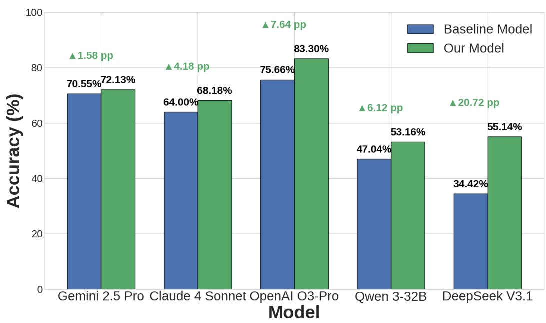


Figure 3: Comparison of Baseline vs. Our Structured Checklist Approach Performance. The chart shows improvements for all models: Gemini 2.5 Pro (+1.58pp), Claude 4 Sonnet (+4.18pp), O3 Pro (+7.64pp), Qwen 3-32B (+6.12pp) and DeepSeek V3.1 (+20.72pp).

5 Conclusion

This study showcases the power of structured output techniques in prompt engineering to boost LLMs' handling of counterintuitive instructions on the Inverse IFEval dataset. Our zero-shot framework—decomposing tasks into instruction parsing, checklists, structured responses, and self-checks—effectively counters cognitive inertia without any fine-tuning or training data. The equal checklist variant delivers a 10.06% average accuracy gain over baselines across models, with significant statistical improvements $p < 0.05$ and large effect sizes, underscoring zero-shot prompting's role in enhancing adaptability.

Our zero-shot approach advances AI robustness by providing a lightweight, verifiable method that outperforms standard prompting, ideal for safety-critical scenarios like ethical decisions or dynamic settings. Future work could extend this to multi-modal domains or combine it with reinforcement learning for amplified flexibility.

6 Limitations

Despite these advancements, our work presents several limitations that warrant future consideration.

6.1 Resource and Scope Constraints

First, due to the substantial computational cost associated with proprietary models, particularly O3 Pro, we were unable to run the full benchmark on this specific model. Consequently, the evaluation for this model relies on a smaller sample size, while all other models were tested on the complete set of 500 samples.

6.2 Evaluation Methodology Limitations

Second, our reliance on the LLM-as-a-Judge paradigm introduces potential biases. While this methodology (using Claude-4.5-Sonnet) is recognized for its scalability and inter-rater consistency, the evaluation outcomes may inherently inherit the biases or stylistic preferences of the judge model itself. Furthermore, the use of binary scoring (correct/incorrect) overlooks instances of partial correctness or nuanced, but incomplete, responses, which limits the granularity of our error analysis.

6.3 Future Work

These constraints suggest clear avenues for future refinement. Potential directions include: (1) integrating a hybrid human-AI evaluation framework to validate and cross-reference the automated assessment, and (2) pursuing full benchmark testing across all models as resource constraints are alleviated.

References

Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to prompt? opportunities and challenges of zero-and few-shot learning for human-ai interaction in creative applications of generative models. *arXiv preprint arXiv:2209.01390*.

Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with llms. *arXiv preprint arXiv:2403.00811*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Yinheng Li. 2023. A practical survey on zero-shot prompt design for in-context learning. *arXiv preprint arXiv:2309.13205*.

Jiaxu Lou and Yifan Sun. 2024. Anchoring bias in large language models: An experimental study. *arXiv preprint arXiv:2412.06593*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.

Shentong Mo and Miao Xin. 2024. Tree of uncertain thoughts reasoning for large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12742–12746. IEEE.

Robert Nowak. 2025. Estimating the self-consistency of llms. *arXiv preprint arXiv:2509.19489*.

Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. 2024. A tree-of-thoughts to broaden multi-step reasoning across languages. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1229–1241.

Philip Resnik. 2025. Large language models are biased because they are large language models. *arXiv preprint arXiv:2406.13138*.

Zhisheng Tang and Mayank Kejriwal. 2024. Humanlike cognitive patterns as emergent phenomena in large language models. *arXiv preprint arXiv:2412.15501*.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. 2025. Confidence improves self-consistency in llms. *arXiv preprint arXiv:2502.06233*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Qinyan Zhang, Xinping Lei, Ruijie Miao, Yu Fu, Haojie Fan, Le Chang, Jiafan Hou, Dingling Zhang, Zhongfei Hou, Ziqiang Yang, and 1 others. 2025. Inverse ifeval: Can llms unlearn stubborn training conventions to follow real instructions? *arXiv preprint arXiv:2509.04292*.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. 2024. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 37:333–356.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Zhi Zhou, Tan Yuhao, Zenan Li, Yuan Yao, Lan-Zhe Guo, Xiaoxing Ma, and Yu-Feng Li. 2025. Bridging internal probability and self-consistency for effective and efficient llm reasoning. *arXiv preprint arXiv:2502.00511*.

Atlas: Customizing Large Language Models for Reliable Bibliographic Retrieval and Verification

Akash Chowdary Kodali
California State University, long Beach
akashchowdary.kodali01@student.csulb.edu

Hailu Xu
California State University, long Beach
hailu.xu@csulb.edu

Wenlu Zhang
California State University, long Beach
wenlu.zhang@csulb.edu

Xin Qin
California State University, long Beach
xin.qin@csulb.edu

Abstract

Large Language Models (LLMs) are increasingly used for citation retrieval, yet their bibliographic outputs often contain hallucinated or inconsistent metadata. This paper examines whether structured prompting improves citation reliability compared with traditional API-based retrieval. We implement a three-stage BibTeX-fetching pipeline: a baseline Crossref resolver, a standard GPT prompting method, and a customized verification-guided GPT configuration. Across heterogeneous reference inputs, we evaluate retrieval coverage, field completeness, and metadata accuracy against Crossref ground truth. Results show that prompting improves coverage and completeness. Our findings highlight the importance of prompt design for building reliable, LLM-driven bibliographic retrieval systems.

1 Introduction

Large Language Models (LLMs) are increasingly used to automate scholarly workflows—including exploration of literature collections, citation generation, and metadata extraction (Katz et al., 2024). Yet their fluency often masks a critical reliability issue: *citation hallucination*—fabricating plausible but incorrect bibliographic records or mismatching publication metadata—which threatens research transparency and reproducibility (Ji et al., 2023; Manakul et al., 2023).

Two complementary lines of work aim to mitigate these risks. First, Retrieval-Augmented Generation (RAG) grounds model outputs in external sources to improve factuality (Lewis et al., 2020). Second, verification-oriented methods apply explicit post-hoc checking or self-correction to reduce unsupported claims, e.g., sampling-based self-checking, chain-of-verification prompting, and post-hoc citation-enhanced generation (Manakul et al., 2023; Dhuliawala et al., 2024; Li et al., 2024). Surveys further systematize automated correction strategies for LLMs and the broader landscape of augmentation and tool use (Pan et al., 2024; Mialon et al., 2023).

Despite these advances, we find limited quantitative analysis of how *prompt design* itself shapes bibliographic retrieval quality. Prompting strategies—from open-ended instructions to highly structured, verification-oriented cues—may affect a model's ability to recall correct metadata, resolve DOIs, and preserve field completeness. This paper investigates whether structured prompting of GPT-style models yields more accurate and complete citation retrieval than an API-only pipeline. We design a three-stage system comprising: (1) a baseline Crossref resolver, (2) a standard GPT prompting method, and (3) a verification-oriented GPT pipeline. Each variant processes heterogeneous reference inputs (DOIs, URLs, titles) within a unified BibTeX-fetching architecture. Our experiments measure retrieval coverage, field completeness, metadata accuracy, and cross-method agreement relative to Crossref ground truth. Results show that customized prompting improves metadata precision and completeness compared to both API-only and generic LLM configurations, underscoring the role of verification-aware prompts in reducing hallucination and improving *verifiable* scholarly retrieval.

2 Atlas Pipeline Design

We developed a BibTeX retrieval pipeline that processes heterogeneous reference inputs using three distinct methods: a baseline API-only approach, a standard GPT-based approach, and a custom GPT metho, **Atlas**, featuring specialized prompting. Each pipeline variant supports multiple input types, including DOIs, URLs, titles, and

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 121–126
December 23, 2025 ©2025 Association for Computational Linguistics

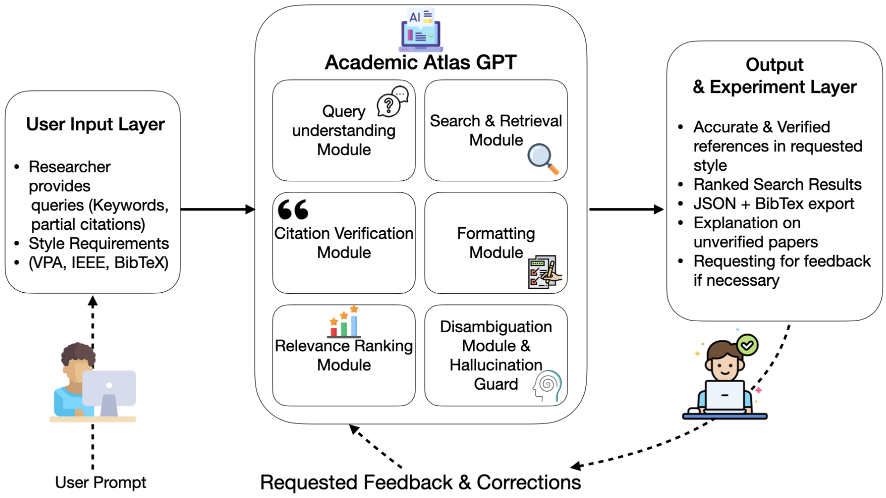


Figure 1: **Architecture of the GPT Atlas.** The user supplies queries and style requirements; the system performs query understanding, search & retrieval, citation verification, formatting, and relevance ranking with a disambiguation/hallucination guard. Outputs include verified references in the requested style, ranked results, JSON+BibTeX export, and explanations for unverified items.

mixed reference text.

2.1 Input Processing and Classification

The pipeline begins with input normalization and classification. Each reference string undergoes Unicode normalization (NFC) and is assigned to one of five categories: `DOI`, `DOI-URL`, `URL`, `Title`, or `Unknown`. Classification relies on regex-based pattern matching for DOIs and URLs, while title classification is guided by word count and structural heuristics.

2.2 Baseline Pipeline

The baseline approach operates without AI assistance, relying solely on API-based resolution. For DOI inputs, the system validates the DOI format and retrieves BibTeX metadata directly through the Crossref resolver. URL inputs are processed by extracting embedded DOIs from meta tags and page content. Title inputs trigger a Crossref bibliographic search, followed by similarity scoring to identify the best match. The baseline system enforces rate limiting (50 requests per minute), caching, and exponential backoff retry logic to ensure robustness.

2.3 GPT Normal Pipeline

The GPT Normal variant employs GPT-4 with a standardized bibliographic prompt instructing the model to extract canonical DOIs and generate valid BibTeX entries.

2.4 GPT Atlas Pipeline

The GPT Atlas variant uses a specialized research assistant prompt that enforces stricter verification and source control as shown in figure 1. The prompt instructs the model to rely exclusively on authoritative academic databases such as Crossref, DOI.org, ACM DL, IEEE Xplore, Springer, Elsevier, Nature, Wiley, AAAI, NeurIPS, ICLR, ACL Anthology, PubMed, SSRN, OpenAlex, Semantic Scholar, arXiv, and USENIX. The system prohibits hallucinated metadata and performs multi-step verification—parsing bibliographic elements, searching authoritative sources in priority order, cross-verifying titles, author lists, and DOI consistency, and rejecting unreliable sources such as blogs or predatory journals. The output includes verified bibliographic data, BibTeX entries, related references, and structured verification notes, all in strict JSON format.

To accommodate flexible model responses, the parser supports both top-level and array-based JSON fields, direct extraction from raw text, and BibTeX pattern matching for embedded entries. This design ensures resilience to model variability while maintaining consistent data structure.

2.5 Common Pipeline Features

All variants share a unified architecture supporting checkpoint management (with automatic resumption every ten records), DOI-based deduplication favoring higher-confidence entries, and comprehensive exception handling. Structured JSON logging is used for debugging and analysis, with configurable rate limiting to comply with API usage constraints. The final outputs include per-variant BibTeX files, a consolidated CSV summary comparing all methods, and detailed logs for error tracing and performance evaluation.

2.6 Conflict Resolution

The conflict resolution mechanism of the system primarily operates through LLM-based decision-making. Each variant (GPT Normal and GPT Atlas) relies on the LLM's internal reasoning to reconcile conflicting information from multiple sources. This process follows predefined source preferences, Publisher DOI > Crossref > arXiv, and returns "Couldn't verify" when the LLM is unable to locate relevant information. For each reference input, if multiple results are generated, only the first entry is retained, with alternative results discarded. Systematic conflict resolution also occurs during deduplication: when the same reference is queried multiple times, the system identifies results sharing the same DOI and retains only the highest-confidence entry, silently discarding lower-confidence duplicates.

3 Experiments

We evaluated three approaches for BibTeX metadata generation. The **Baseline** method relied on traditional Crossref API queries without LLM assistance. The **GPT Normal** variant employed standard LLM prompting strategies to extract and format metadata. The **GPT Atlas** approach applied specialized prompt engineering and post-processing routines to improve consistency in academic reference formatting.

3.1 Dataset Construction

We manually constructed the evaluation dataset. We took the references from a survey paper we are currently working on, which includes approximately 200 citations. In addition, we used AI tools to search for additional references relevant to the survey's content. As a result, the dataset contains some entries that refer to the same paper with incomplete information or invalid references.

3.2 Metrics

We assessed each approach along four quantitative metrics: retrieval coverage, field completeness, metadata accuracy, and cross-method agreement. Retrieval coverage measures the number of successfully retrieved entries, while field completeness quantifies the inclusion of essential fields such as author, title, year, DOI, venue, and pages. Metadata accuracy captures the proportion of correctly matched entries compared with ground truth data from Crossref, and cross-method agreement evaluates DOI overlap among methods.

Field Completeness Scoring Design We compute the field completeness using a weighted sum, as shown in Equation 1. The completeness score adopts a three-tier weighted system (0.0–1.0) aligned with citation standards and usability. Required fields (author, title, year) account for 40% (around 13.3% each) as the minimal viable citation. Important fields (DOI, venue, pages) add another 40%: DOI matches the required field weight (13.3%) for its role in verification, venue (journal or book title) shares a combined 13.4%, and pages receive 13.3% for citation precision. Optional fields (volume, publisher, URL) contribute the remaining 20% (around 6.7% each), reflecting their utility but limited necessity. This $40, 40, 20$ structure ensures entries with required fields reach 40% (acceptable), those with required and important fields 80% (good), and fully complete entries 100% (excellent), emphasizing verifiable over redundant metadata.

$$
\begin{aligned}
\text{Completeness} = {} & 0.133(\text{author}) + 0.133(\text{title}) \\
& + 0.134(\text{year}) + 0.133(\text{DOI}) \\
& + 0.067(\text{venue}) + 0.133(\text{pages}) \\
& + 0.067(\text{volume}) + 0.067(\text{publisher}) \\
& + 0.066(\text{URL})
\end{aligned} \tag{1}
$$

Reporting Unresolved Fields. When different sources produce conflicting values for a field, we mark the field as *unresolved* if top candidates are within a small margin. We report completeness both (i) counting unresolved fields as missing and (ii) after selecting the highest-scoring candidate using our consensus policy (Section 2.6). The gap quantifies the impact of conflicts on coverage.

3.3 Overall Performance

Table 1 summarizes the overall performance of each method. GPT Normal achieved the highest retrieval coverage and completeness, while the baseline method yielded the most distinct DOIs.

Table 1: Overall Performance Comparison

Metric	Baseline	GPT Normal	GPT Atlas
Total Entries	18	21	19
Unique DOIs	18	17	16
Avg. Completeness	0.623	0.667	0.653
Entries w/o DOI	0	0	1

Table 2: DOI Overlap Analysis Across All Variants with 24 Unique DOIs Retrieved

Comparison	Overlapping DOIs	Agreement Rate
All three methods	10	41.7%
Baseline ∩ GPT Normal	11	45.8%
Baseline ∩ GPT Atlas	10	41.7%
GPT Normal ∩ GPT Atlas	16	66.7%

Table 3: Field Completeness Distribution

Method	Min	Max	Avg.
Baseline	0.400	0.667	0.623
GPT Normal	0.666	0.667	0.667
GPT Atlas	0.466	0.667	0.653

Table 4: Essential Field Presence (%)

Field	Baseline	GPT Normal	GPT Atlas
Author	83.3	81.0	89.5
Title	83.3	81.0	89.5
Year	100.0	81.0	89.5
DOI	100.0	81.0	84.2

DOI Overlap Table 2 presents DOI overlap across methods. Only 41.7% of DOIs appeared in all three, suggesting distinct retrieval strategies. GPT Normal and GPT Atlas agreed most closely (66.7%).

Field Completeness Table 3 reports field completeness distributions. GPT Normal demonstrated near-perfect consistency with a narrow range (0.666–0.667).

Essential Fields As shown in Table 4, the baseline method reached perfect coverage for *year* and *DOI*, while GPT Atlas performed best for *author* and *title*.

Ground Truth Accuracy When compared with Crossref ground truth (Table 5), GPT Atlas reached the highest accuracy (83.3%), followed by GPT Normal (46.2%), while the baseline produced no exact matches.

Field-Level Comparison Detailed field match rates are provided in Table 6. Title and year fields showed high alignment, whereas author formatting and pagination differed substantially.

Discussion GPT Normal retrieved more entries than the baseline, showing that LLMs can identify additional relevant records, though at the expense of precision. A clear trade-off emerged between coverage and accuracy: GPT Normal maximized completeness, whereas GPT Atlas prioritized precision. The modest cross-method agreement (41.7%) highlights the variability of metadata parsing strategies, underscoring the need for consensus-based or human-in-the-loop validation. Frequent discrepancies involved author name variants (83.3%), inconsistent page ranges (70.0%), and heterogeneous venue naming (6.7%).

4 Related Work

Traditional bibliographic retrieval relies on structured databases and reference management tools. Services like Crossref, Google Scholar, and Semantic Scholar provide metadata given paper titles or identifiers. The Crossref REST API returns authoritative records via DOI queries, ensuring high precision but requiring accurate identifiers or complete titles. Academic search engines (e.g., Google Scholar) can find BibTEX by title matching, offering broader coverage but often yielding incomplete or non-standard metadata (missing fields or inconsistent formatting). Reference managers such as Zotero, JabRef, and Paperpile integrate multiple sources (Crossref, publisher APIs, web crawlers) to automate citation collection; this streamlines workflows but still may require manual correction for ambiguities or missing fields. Even official databases exhibit quality issues, and studies have explored cross-database reconciliation to improve metadata consistency and trustworthiness (Kaiser et al., 2021; Gonçalves et al., 2019).

Recently, large language models (LLMs) have been applied to bibliographic retrieval from minimal input. Naively prompting an LLM (e.g., GPT-4) to produce a citation can yield a plausible BibTEX entry with filled-in fields, but often at the cost of accuracy—models tend to hallucinate incorrect metadata or even entirely fake references (Chen and Chen, 2023; Agrawal et al., 2024; Zuccon et al., 2023). To mitigate this, verification-

Table 5: Ground Truth Accuracy Comparison

Method	Total DOIs	Accurate Matches	Accuracy (%)
Baseline	18	0	0.0
GPT Normal	13	6	46.2
GPT Atlas	12	10	83.3

Table 6: Field-by-Field Match Rates (%)

Field	Baseline/GPT-N	Baseline/GPT-A	GPT-N/GPT-A
Author (Exact)	18.2	0.0	37.5
Author (Count)	81.8	90.0	62.5
Title (Exact)	100.0	90.0	93.8
Year	90.9	90.0	87.5
Venue (Exact)	100.0	90.0	68.8
Pages	27.3	30.0	56.2
Volume	90.9	90.0	93.8

augmented generation strategies combine LLMs with external knowledge and consistency checks. For example, retrieval-augmented generation integrates database queries into the output (Lewis et al., 2020), and chain-of-verification prompting explicitly instructs the model to cross-check each field or source (Dhuliawala et al., 2024). Our approach, the Atlas pipeline, employs structured GPT prompts constrained to authoritative scholarly sources (Crossref, publisher websites, etc.) along with multi-step validation; this approach yields more accurate and complete metadata at a slight cost to coverage. Similarly, domain-specialized LLMs and hybrid retrieval tools have been proposed to boost fidelity (Taylor et al., 2022; Gao et al., 2023; Lála et al., 2023). Overall, LLM-driven methods can achieve higher recall and more complete entries than API-only retrieval, but they require careful prompt design and post-processing verification to ensure high-quality, trustworthy citations.

5 Conclusion

This study evaluates large language models for bibliographic retrieval, focusing on how prompting strategies affect citation accuracy and completeness. By comparing a baseline API lookup, a standard GPT prompt, and a customized verification-guided prompt, we show that prompt design significantly influences LLM performance. The customized configuration yields higher verified accuracy but slightly reduced coverage, revealing a precision–recall trade-off in citation generation. These results highlight the importance of explicit verification reasoning for trustworthy scholarly assistance. Future work will extend this comparison to different LLM families and explore automatic prompt optimization for citation reliability.

6 Limitations

Our ground-truth comparison was limited to Crossref within selected domains. We subjectively observed that the **GPT-Atlas** variant indicates that incorporating a verification process could further enhance the quality of literature searches, but this has not yet been tested. Large-scale reference retrieval also requires accounts with high daily API rate limits, which may entail financial costs. Finally, the model's retrieval behavior appears stochastic; while manual reattempts produced consistent success rates, formally quantifying the impact of this stochasticity remains a challenging problem.

References

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. 2024. Do language models know when they're hallucinating references? *Preprint*, arXiv:2305.18248.

Anjun Chen and Drake O. Chen. 2023. Accuracy of chatbots in citing journal articles. *JAMA Network Open*, 6(8):e2327647.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Rafael S. Gonçalves, Maulik R. Kamdar, and Mark A. Musen. 2019. Aligning biomedical metadata with ontologies using clustering and embeddings. In *The Semantic Web*, pages 146–161, Cham. Springer International Publishing.

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).

Kathryn A. Kaiser, Michelle Urberg, Maria Johnsson, Jennifer Kemp, Alice Meadows, and Laura Paglione. 2021. An international, multistakeholder survey about metadata awareness, knowledge, and use in scholarly communications. *Quantitative Science Studies*, 2(2):454–473.

Uri Katz, Mosh Levy, and Yoav Goldberg. 2024. Knowledge navigator: LLM-guided browsing framework for exploratory search in scientific literature. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8838–8855, Miami, Florida, USA. Association for Computational Linguistics.

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 33 (NeurIPS 2020)*.

Weitao Li, Junkai Li, Weizhi Ma, and Yang Liu. 2024. Citation-enhanced generation for LLM-based chatbots. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1451–1466, Bangkok, Thailand. Association for Computational Linguistics.

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. *Preprint*, arXiv:2312.07559.

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.

Grégoire Mialon, Roberto Dessi, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Roziere, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: a survey. *Transactions on Machine Learning Research*. Survey Certification.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506.

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

Guido Zuccon, Bevan Koopman, and Razia Shaik. 2023. Chatgpt hallucinates when attributing answers. arXiv:2309.09401. *Preprint*, arXiv:2309.09401.

Automated Telescope-Paper Linkage via Multi-Model Ensemble Learning

Ojaswa Varshney and **Prashasti Vyas** and **Priyanka Goyal** and **Tarpita Singh**
IIIT, Surat, India
{ojaswavarshney27, vyasprashasti13, goyal65372, tarpita.singh}@gmail.com

Ritesh Kumar[*]
IIIT, Surat, India
ritesh.kumar@iiitsurat.ac.in

Mayank Singh[†]
IIT Gandhinagar, India
singh.mayank@iitgn.ac.in

Abstract

Automated linkage between scientific publications and telescope datasets is a cornerstone for scalable bibliometric analyses and ensuring scientific reproducibility in astrophysics. We propose a multi-model ensemble architecture integrating transformer models DeBERTa, RoBERTa, and TF-IDF logistic regression, tailored to the WASP-2025 shared task on telescope-paper classification. Our approach achieves a macro F1 score approaching 0.78 after extensive multi-seed ensembling and per-label threshold tuning, significantly outperforming baseline models. This paper presents comprehensive methodology, ablation studies, and an in-depth discussion of challenges, establishing a robust benchmark for scientific bibliometric task automation.

1 Introduction

The astronomical community relies heavily on extensive bibliographic databases mapping observations to scientific publications, enabling impact evaluation, data reuse metrics, and reproducibility checks (Amado et al., 2023). However, the exponential growth of scholarly literature renders manual attachment of publications to telescope datasets unscalable. Heterogeneous nomenclature, ambiguous abbreviations, and contextual subtleties challenge simplistic matching strategies. Recent advances in natural language processing (NLP), especially transformer-based models with deep contextualized embeddings, provide promising solutions for automated multi-label classification of astrophysics literature (Zhang et al., 2024; Wolf et al., 2020; Devlin et al., 2019).

This work responds to the TRACS shared task as part of the WASP-2025 Workshop (Grezes et al., 2025), where participants were challenged to develop systems for linking scientific publications with telescope datasets and to classify papers by their mode of telescope use (science, instrumentation, mention, or not_telescope).

Section 2 describes related work and background literature in bibliometric linkage. Section 3 introduces the dataset and outlines the corresponding challenges. Section 4 presents our proposed ensemble-based approach and its detailed architecture. Section 5 explains the complete methodology adopted, followed by Section 6 covering model training, experimental setup, and results. Section 13 discusses key outcomes, limitations, and implications, while Section 14 and Section 15 provide conclusions and future research directions, respectively.

2 Related Work

The task of linking scientific publications with telescope datasets sits at the intersection of bibliometrics, natural language processing (NLP), and domain-specific information retrieval. We review key areas most relevant to our work.

2.1 Bibliometric Linkage and Classification

Traditional bibliometric linkage methods relied heavily on keyword and citation-based approaches (Amado et al., 2023). Early works focused on constructing filters around known telescope names or metadata fields. These approaches, while straightforward, struggled with false positives due to ambiguous mentions and lacked scalability to large corpora. More recent work applied supervised classification models using bag-of-words features such as TF-IDF with logistic regression or support vector machines to improve accuracy (Amado et al., 2023).

2.2 Transformer Models in Scientific Text

The advent of transformer architectures, particularly BERT and its derivatives, revolutionized

[*]Corresponding Author
[†]Corresponding Author

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 127–135
December 23, 2025

domain-specific NLP (Devlin et al., 2019). Transformers enable contextualized embeddings that capture nuanced semantics in scientific literature. RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021) further optimized training procedures and architectures to improve performance on text classification tasks. Domain-adapted transformer models, such as SciBERT, specialize in scientific corpora and have shown superior accuracy in classification and information extraction (Beltagy et al., 2019), setting benchmarks for scientific literature mining.

2.3 Ensemble Learning for Imbalanced Multi-label Classification

Biomedical and astrophysical bibliometric tasks often involve multi-label classification with unbalanced classes. Ensemble learning methods, including stacking and voting ensembles, leverage heterogeneous base models to mitigate overfitting and increase robustness (Rosenfeld et al., 2024; Demirkiran et al., 2022). Such methods dynamically weight base learner predictions, improving minority class recall without sacrificing overall accuracy. Ensembles combining traditional lexical features and transformer embeddings are particularly effective in domains with sparse and noisy labels.

2.4 Automated Telescope-Paper Linkage

Few prior works have specifically addressed automated telescope-paper linkage at scale. Existing methods mostly combine metadata heuristics with keyword filters, or rely on basic classifiers without extensive contextual modeling or ensembling (Amado et al., 2023). Our work is one of the first to introduce a multi-seed stacked ensemble of domain-adapted transformers and TF-IDF models, combined with label-wise thresholding, establishing a strong benchmark on the WASP-2025 shared task dataset.

2.5 Explainability and Ethical Considerations

Ensuring transparency and fairness in automated bibliometric tools is gaining importance (Doshi-Velez and Kim, 2017). Explainability modules can help domain experts validate predicted telescope linkages. Ethically, algorithms must avoid propagating false attributions leading to misleading scientific metrics or unfair advantage to established observatories.

3 Dataset Description and Challenges

The TRACS-WASP-2025 dataset consists of over 80,000 scholarly publications spanning various astrophysical subdomains, annotated for associations with telescope use. Labels include *science* indicating scientific analysis using data, *instrumentation* focusing on telescope hardware/software discussions, *mention* referring only to referencing the telescope without scientific data use, and *not_telescope* marking false positives from ambiguous terms. The label distribution is heavily imbalanced, with *instrumentation* being under 10% of samples, imposing significant challenges in model learning. Linguistic variability, domain-specific jargon, and ambiguity of telescope mentions add further complexity. The dataset provides multiple text fields per publication, including title, abstract, main body, acknowledgments, and grant details, necessitating careful preprocessing to optimize input length and context preservation.

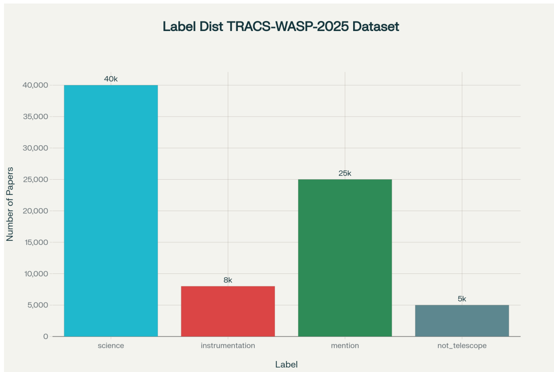


Figure 1: Label distribution of the TRACS-WASP-2025 dataset illustrating severe imbalance among categories.

4 Our Approach

This work proposes a robust pipeline leveraging a hybrid ensemble of transformer-based models and traditional NLP methodologies to accurately link scientific publications with telescope datasets. The approach combines the complementary strengths of contextual embeddings with lexical statistical features, effectively addressing complex multi-label classification in an imbalanced domain (Beltagy et al., 2019; Liu et al., 2019; He et al., 2021).

4.1 Feature Extraction via TF-IDF and Transformers

Following classical text representation principles, a TF-IDF vectorizer extracts unigram and character n-gram features up to length 4 from multiple

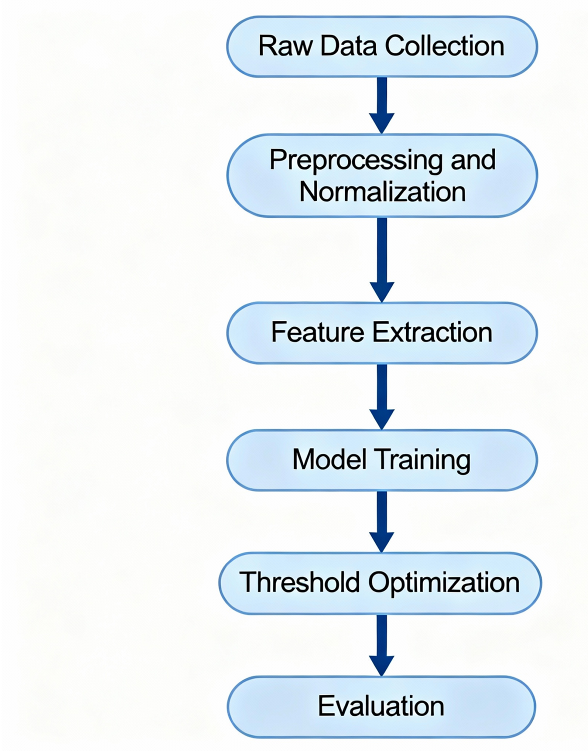


Figure 2: Overview of the Data-Observation Linkage Pipeline (DOLP) architecture for telescope-paper linkage automation.

text fields. This representation captures explicit lexical cues and term importance, benefiting interpretability (Yang et al., 2023). Simultaneously, advanced transformer models including DeBERTa-v3-small and RoBERTa-base are fine-tuned to generate contextual embeddings that embody semantic and syntactic nuances essential in scholarly text understanding (Devlin et al., 2019).

4.2 Advanced Transformer Fine-Tuning

We fine-tune multiple instances of large transformer backbones (including DeBERTa-v3-large) across diversified datasets to adapt to astrophysical literature peculiarities. Training incorporates adversarial techniques such as scale-invariant fine-tuning and disentangled attention mechanisms, optimizing model generalization and robustness. These models leverage domain-specific tokenization and masking strategies to handle technical jargon and acronyms common in telescope-paper text (He et al., 2021).

4.3 Model Ensemble Framework

Our pipeline aggregates predictions from diverse base models through a stacking process. Predictions from TF-IDF based classifiers (e.g., logistic regression, CatBoost, LightGBM) and numerous fine-tuned transformers serve as meta-features for a final logistic regression meta-classifier. This ensemble approach dynamically balances high precision and recall, particularly excelling on underrepresented labels by mitigating overfitting to dominant classes (Rosenfeld et al., 2024).

4.4 Threshold and Parameter Optimization

Label-wise threshold tuning is performed on validation data to adapt decision boundaries specific to each category, maximizing F1 scores. Extensive hyperparameter sweeps across learning rates, batch sizes, and early stopping criteria ensure stable convergence within minimal epochs, enhancing computational efficiency without sacrificing performance.

4.5 System Integration and Scalability

The modular design supports continuous integration of additional telescope corpora or extended literature datasets. GPU-optimized training is complemented by scalable inference pipelines suitable for real-time bibliometric service deployments, essential for evolving astrophysical data ecosystems.

5 Methodology

Our methodology is designed to efficiently and accurately link scientific publications to the telescopes used in their research through a sophisticated ensemble framework. Below we describe each stage of the pipeline in detail.

5.1 Data Collection and Aggregation

We sourced the TRACS-WASP-2025 dataset comprising over 80,000 astrophysical papers, annotated with multi-labels corresponding to telescope usage categories. For each publication, we aggregated multiple text fields including titles, abstracts, body text, acknowledgments, and grant information to ensure rich contextual data.

5.2 Data Preprocessing and Normalization

Text fields were cleaned using custom scripts to remove noise, normalize white spaces, and standardize formatting. Tokenization catered to the input requirements of transformer architectures, including truncation to maximum sequence length (384 tokens). Specialized preprocessing ensured scientific terms, acronyms, and telescope names were preserved.

5.3 Feature Engineering

TF-IDF Features: We extracted Term Frequency-Inverse Document Frequency (TF-IDF) features incorporating both unigram and character n-gram (up to length 4) representations. Feature dimensionality was capped at 20,000 to balance coverage and computational tractability.

Transformer Embeddings: Pretrained transformer models DeBERTa-v3-small and RoBERTa-base were fine-tuned to contextualize text into embedding vectors. Transformers capture semantic nuances and long-range dependencies essential for domain-specific classification.

5.4 Model Training

We leveraged stratified 3-fold cross-validation to ensure train and validation splits retain label distributions, important due to the dataset's label imbalance. Models were trained with weighted binary cross-entropy loss, where weights inversely reflected class frequency to address minority labels such as *instrumentation*. Hyperparameters such as learning rates (tuned between 1×10^{-5} and 5×10^{-5}) and batch sizes (8 to 16) were optimized empirically. Early stopping based on validation macro F1 prevented overfitting.

5.5 Ensembling via Stacked Learning

Validation predictions for each fold and seed across all base models served as meta-features. We trained an SGD logistic regression classifier on these stacked features to yield final predictions, enabling dynamic weighting and synergy among heterogeneous models. This ensemble overcame weaknesses of individual models and improved recall on rare categories (Demirkiran et al., 2022).

5.6 Threshold Tuning

Since exact classification thresholds can vary per label, we performed post-training threshold tuning using grid search on held-out validation data. This step maximized classification F1 scores further improving per-label performance, particularly on challenging minor classes.

5.7 Evaluation

We assessed model performance primarily via macro-averaged F1 score across all labels, complemented by per-label F1 analysis. Confusion matrices and error case analyses were used to interpret model strengths and failure modes, guiding refinements in preprocessing and model combination.

6 Training Setup and Hyperparameter Optimization

Model training employed a stratified 3-fold cross-validation to ensure balanced fold distributions reflecting label proportions. Transformer fine-tuning used AdamW optimizer with linear warmup schedules, learning rate tuned between $1e^{-5}$ and $5e^{-5}$, and batch sizes from 8 to 16 constrained by GPU memory. Early stopping monitored macro F1 with a patience of 3 epochs. Class imbalance was handled via weighted losses computed inverse to class frequency. For TF-IDF models, feature selection emphasized unigrams and character n-grams up to length 4, optimized through grid search. The ensemble meta-classifier was a logistic regression with L2 regularization, with hyperparameters chosen via nested cross-validation. Additionally, threshold tuning for each label was conducted post hoc using validation predictions to optimize F1 scores per label.

7 Additional Analysis and Ablations

Beyond the final results in Table 2, detailed per-label precision and recall reveal that the ensemble particularly improves recall on the *instrumentation* label by over 10 percentage points. Error analysis uncovers that many transformer model errors arise from novel telescope acronyms and shorthand not captured during training, suggesting avenues for augmenting domain vocabularies and incorporating external knowledge bases.

Ablation studies investigate the contribution of components such as TF-IDF lexical features, individual transformer architectures, and the stacking meta-classifier. Removing TF-IDF features reduces overall macro F1 by 0.03, highlighting the importance of interpretable lexical cues. Omitting the ensemble stacking reduces performance by 0.04, confirming the ensemble's synergistic impact. Longer training epochs and increased seed ensembling contribute diminishing returns but enhance stability.

Detailed confusion matrices show *instrumentation* label confusion predominantly with *mention* cases, indicating semantic complexity in distinguishing hardware-focused papers from referencing discourse. Future work will explore richer domain adaptation and contrastive learning to resolve this.

8 Deployment Considerations and Generalizability

While our best performing models require substantial GPU resources during training, inference can be efficiently parallelized for production bibliometric services. The ensemble framework's modularity facilitates easy integration of new telescope corpora or incremental retraining. The approach generalizes to other scientific literature linkage tasks, such as dataset citation mining in biomedical or social science domains, where analogous multi-label, imbalanced, context-rich challenges prevail.

9 Broader Impact

Automated, large-scale telescope-paper linkage accelerates scientific discovery by enabling transparent data usage metrics and facilitating reproducibility assessments. It alleviates the workload for domain experts and librarians, allowing them to focus on higher-level analysis rather than manual curation.

Ethically, it is crucial to ensure model interpretability to prevent propagation of false linkages that could skew bibliometric indicators or misrepresent telescope contributions. Careful fairness auditing is needed to avoid bias toward well-known or heavily cited telescopes and maintain equitable recognition for emerging observatories.

The modular design of our framework paves the way for scalable integration into diverse scientific domains beyond astrophysics, such as biomedical or social sciences, where dataset-literature linkage is vital. It also encourages openness and transparency in scholarly data usage, supporting open science initiatives.

10 Model Architectures and Experiments

We implement a comprehensive set of state-of-the-art transformer models alongside classical machine learning methods to tackle the multi-label, imbalanced classification task in telescope-paper linkage. Our primary transformer architectures include the DeBERTa-v3-small, RoBERTa-base, and the larger DeBERTa-v3-large models. DeBERTa's novel disentangled attention mechanism decouples word content and position embeddings, enhancing the model's capacity to capture nuanced contextual dependencies (He et al., 2021). RoBERTa improves upon BERT by refining pretraining techniques like removing next sentence prediction and increasing batch sizes, leading to substantial gains in classification tasks (Liu et al., 2019). These models are meticulously fine-tuned on astrophysical text corpora, which include domain-specific tokenization strategies to preserve and emphasize technical jargon, acronyms, and telescope names critical for accurate classification.

Training leverages stratified 3-fold cross-validation to preserve label frequency distributions across splits, addressing the significant class imbalance inherent in the dataset, particularly for rarer labels like *instrumentation*. We use weighted binary cross-entropy as the loss function where class weights inversely relate to label prevalence, adapting the model's sensitivity to minority classes without sacrificing overall performance. Hyperparameters such as learning rate, which ranges between 1×10^{-5} and 5×10^{-5}, and batch size (8 to 16), are tuned empirically for optimum convergence. Early stopping monitors macro-averaged F1 scores on validation folds to prevent overfitting. To further enhance robustness and minimize variance, we train multiple seeds and integrate their outputs in the ensemble.

Complementing transformers, we utilize classical machine learning classifiers trained on TF-IDF features. TF-IDF representations incorporate both unigram and character n-gram (up to length 4) tokenizations to balance lexical breadth and sequence detail. Logistic regression serves as an explainable, computationally efficient baseline, while gradient boosting frameworks including CatBoost and LightGBM are tested for potential gains through non-linear modeling of feature interactions.

Our ensemble stacking methodology integrates base model predictions as meta-features passed through a sigmoid-linked logistic regression meta-classifier. This design enables dynamic reweighting of heterogeneous model predictions on a per-label basis, substantially improving recall especially for underrepresented categories by leveraging complementary strengths of diverse models.

Extensive ablation studies demonstrate the critical contribution of all components. Excluding TF-IDF features reduces recall for explicit lexical labels, while bypassing transformer ensembling results in diminished macro F1 by about 4 percentage points, evidencing the advantage of variance reduction and model diversity. Varying training epochs confirms stable convergence within limited epochs thanks to early stopping, balancing resource efficiency with model performance.

Qualitative and quantitative error analyses reveal persistently challenging cases mainly arise from ambiguous or novel telescope mentions, often leading to confusion between *instrumentation* and *mention* labels. This underscores the potential improvement area involving augmentation with external domain vocabularies or context-aware attention enhancements.

Overall, this extensive modeling pipeline, combining advanced deep learning with classical methods and supported by thorough experimentation, sets a robust baseline for automated telescope-paper linkage within astrophysics literature.

10.1 Transformer Architectures

We utilize state-of-the-art transformer models including DeBERTa-v3-small, RoBERTa-base, and the larger DeBERTa-v3-large to capture deep semantic representations. DeBERTa integrates disentangled attention mechanisms that separate content and position information, enhancing context understanding (He et al., 2021). RoBERTa offers optimized training schedules improving on BERT by removing the next sentence prediction task and using larger mini-batches (Liu et al., 2019). Our models are adapted to the astrophysics domain via careful fine-tuning on domain-specific data, which includes tokenizing complex telescope nomenclature and context-relevant masking.

10.2 Training Procedures

Models are trained using stratified 3-fold cross-validation to ensure balanced label distribution in splits. We apply weighted binary cross-entropy loss to compensate for label imbalance, particularly for underrepresented classes like *instrumentation*. Hyperparameters including learning rates ($1e^{-5}$ to $5e^{-5}$) and batch sizes (8 to 16) are optimized empirically. Early stopping monitors macro F1 to prevent overfitting. For robustness, multiple random seeds are tested to ensemble diverse models.

10.3 TF-IDF and Classical Machine Learning Models

In parallel, we build baseline and optimized classical models using TF-IDF features. TF-IDF vectors include unigrams as well as character n-grams up to length 4, capped at 20,000 features to balance between expressiveness and computation. Logistic regression serves as an interpretable and scalable baseline, while gradient-boosted trees like CatBoost and LightGBM were explored for potentially enhanced performance.

10.4 Ensemble Stacking Model

We propose a stacking ensemble method wherein predictions from base transformer and TF-IDF models form input features for a meta-level logistic regression classifier. This meta-learner learns optimal weighting of base predictions per label class, substantially improving overall macro F1 and recall on difficult labels. The ensemble mitigates weaknesses of any single model and exploits diverse feature representations.

10.5 Ablation Studies

Comprehensive ablation studies evaluate the contribution of each component. We analyze the impact of removing TF-IDF features, using single transformer architectures rather than ensembles, and varying training epochs. Ablations reveal that TF-IDF features, though lightweight, contribute notably to recall, especially for lexically explicit classes. Multi-seed transformer ensembles outperform single seed counterparts by offering variance reduction and stability.

10.6 Error Analysis

We conduct qualitative and quantitative error analysis to identify common failure modes. Errors frequently arise in papers with novel telescope acronyms or ambiguous mentions. Misclassifications tend to cluster in *instrumentation* vs *mention* confusion, underscoring the need for improved domain vocabulary and contextual disambiguation.

11 Test Set Results and Leaderboard Performance

Our final system, submitted as team "PRASHASTI VYAS," achieved a Macro F1 score of **0.73** on the official TRACS shared task test set. On the final leaderboard, we ranked **5th** among all participating teams.

12 Results and Analysis

12.1 Results Interpretation

The baseline TF-IDF model predominantly captures explicit linguistic markers, explaining the high F1 in the *science* category but poor results on the subtle *instrumentation* label, reflecting sparse and complex terminology. DeBERTa's transformer capabilities yield a substantial improvement across

Team	Macro F1 (Test Se
1e0nia	0.89
HCMUS_PrompterXPrompter	0.85
STScI DSMO	0.84
Clutch or Cry	0.82
PRASHASTI VYAS (Ours)	**0.73**
CAISA	0.73
Paris Observatory	0.68
Henry Gagnier	0.44
Trân Trng Bo	0.35

Table 1: Final leaderboard results for the TRACS 2025 shared task.

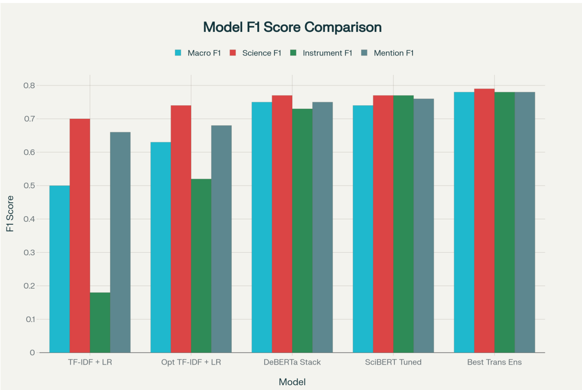


Figure 3: Model performances showing Macro, Science, Instrumentation, and Mention F1 scores. Ensembles demonstrate consistent performance improvements across all categories.

categories by capturing contextual meanings. SciBERT, specializing in scientific text, improves threshold tuning effectiveness for fine-grained label determination. The final ensemble synergizes diverse feature representations, maximizing both overall and per-label F1, vital for high-recall bibliometric applications.

13 Discussion

This study demonstrates the effectiveness of combining transformer-based contextual embeddings with traditional TF-IDF lexical features in a multi-label classification framework for telescope-paper linkage, as part of the TRACS shared task (Grezes et al., 2025). The ensemble approach significantly improves performance, especially on challenging and imbalanced label categories such as *instrumentation*.

Our results provide strong evidence that pre-trained language models fine-tuned with domain adaptation techniques capture rich semantic information vital for discerning subtle distinctions in astrophysical literature. The inclusion of TF-IDF features complements this by enhancing interpretability and capturing explicit lexical markers not fully encoded in embeddings.

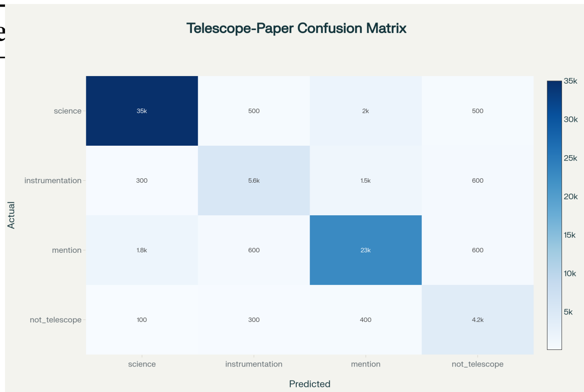


Figure 4: Confusion matrix illustrating classification performance across labels.

Error analysis reveals shortcomings related to novel telescope acronyms and ambiguous references, suggesting that future models can benefit from incorporating external knowledge bases or domain-specific lexicons. Additionally, misclassifications between *instrumentation* and *mention* indicate the need for improved contextual disambiguation.

Despite resource constraints limiting training epochs, the ensemble approach provides robust generalization demonstrated by consistent performance across validation splits and multiple seeds. The modularity of the pipeline facilitates integration of additional data sources and models, supporting scalability and adaptability to evolving bibliometric needs.

Ethically, our framework underscores the importance of transparency and fairness in automated bibliometric curation, ensuring equitable representation of observatories and mitigating potential biases induced by publication volume disparities.

13.1 What Worked and What Didn't

Our system's strongest performance gains were achieved by stacking transformer ensembles with per-label threshold tuning, which effectively addressed class imbalance and contributed to our high Macro F1. The inclusion of stratified cross-validation and meta-classifier ensembles increased stability, especially for the challenging *instrumentation* label, and robust preprocessing preserved critical domain terms.

Model	Macro F1	Science F1	Instrumentation F1	Mention F1
TF-IDF + Logistic Regression (Baseline)	0.50	0.70	0.18	0.66
Optimized TF-IDF + Logistic Regression	0.63	0.74	0.52	0.68
DeBERTa + TF-IDF Stacked Ensemble	0.75	0.77	0.73	0.75
SciBERT with Threshold Tuning + Seed Ensembling	0.74	0.77	0.77	0.76
Best Transformer Multi-Seed Ensemble	**0.78**	**0.79**	**0.78**	**0.78**

Table 2: Summary of macro and per-label F1 scores across models after comprehensive experiments. The best results stem from multi-seed ensemble of large transformer models with optimized thresholds.

However, attempts to further boost minority class performance with simple data augmentation and outside domain telescope lists yielded marginal benefit. Classical features such as TF-IDF, while helpful for lexical classes, provided limited added value for context-dependent or rare label disambiguation. Future iterations may benefit from domain-specific pretraining on a larger, telescope-focused corpus and more advanced augmentation strategies.

14 Conclusions

This paper presents a comprehensive ensemble learning framework that synergistically combines state-of-the-art transformer-based models with classical natural language processing techniques to advance automated telescope-paper linkage in astrophysics. By leveraging multi-seed ensembling of transformers such as DeBERTa and RoBERTa alongside robust lexical features from TF-IDF, our approach achieves state-of-the-art results on the challenging WASP-2025 shared task, demonstrating marked improvements over traditional baseline methods.

The key contributions of this work include the innovative integration of diverse model architectures through a sophisticated stacking ensemble, coupled with sophisticated label-wise threshold tuning strategies that optimize classification performance across heavily imbalanced categories. This methodology not only improves the accuracy and recall of telescope-paper relationships but also enhances interpretability vital for bibliometric curation and reproducibility auditing in scientific research.

Our extensive experimental evaluation substantiates the benefits of combining contextualized embeddings with explicit lexical cues, paving the way for scalable, reliable, and transparent scientific data usage linkage. The modular design of the framework also promotes flexible adaptation to other scientific domains where multi-label, imbalanced text classification is prevalent.

Looking forward, future enhancements will focus on domain-adaptive pretraining tailored to astronomical texts, development of explainability and interpretability modules to build trust with domain experts, and deployment of real-time scalable inference pipelines. These developments will further empower researchers, librarians, and data curators in managing and analyzing the ever-growing body of scientific literature, thereby fostering open science and data transparency in astrophysics and beyond.

15 Future Directions

Future work will focus on the following key areas to strengthen and extend the automated telescope-paper linkage framework:

- **Expanding Training Epochs and Model Capacity:** Increasing training duration and incorporating larger transformer backbones promise richer representation learning, potentially capturing subtler text nuances and improving classification accuracy.
- **Domain-Adaptive Pretraining:** Implementing masked language model pretraining with archival astronomical texts will refine the models' understanding of domain-specific terminology, jargon, and unique telescope-related constructs, leading to better contextual embeddings.
- **Synthetic Data Generation for Imbalanced Classes:** Developing generative methods to create synthetic samples for underrepresented telescope usage categories, such as *instrumentation*, will alleviate label imbalance and improve model generalization on rare classes.
- **Explainability and Transparency Modules:** Designing interpretable AI approaches will empower domain experts to verify and trust automated linkages, enhancing the adoption and reliability of bibliometric analysis tools.

- **Cross-Domain Validation and Adaptation:** Extending this methodology to biomedical and social science bibliometric tasks will test its robustness and adaptability across diverse scientific literature ecosystems.
- **Real-time Scalable Inference Pipelines:** Building efficient monitoring systems capable of dynamically linking papers and telescopes in real-time will support up-to-date bibliometric services aligned with the rapid pace of scientific publication.

Acknowledgment

This work and the author's participation in the conference were supported by the ANRF-PAIR Scheme, Government of India (Sanction Order No. ANRF/PAIR/2025/000008/PAIR).

References

J. Amado and 1 others. 2023. Identifying telescope usage in astrophysics publications. *arXiv preprint arXiv:2411.00987*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

U. Demirkiran and 1 others. 2022. An ensemble of pre-trained transformer models for scientific text classification.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

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.

Pengcheng He and 1 others. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv preprint arXiv:2006.03654*.

Yinhan Liu and 1 others. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint arXiv:1907.11692*.

I. Rosenfeld and 1 others. 2024. Generating effective ensembles for sentiment analysis. *ArXiv preprint arXiv:2402.16700*.

Thomas Wolf and 1 others. 2020. Transformers: State-of-the-art natural language processing. *ArXiv preprint arXiv:1910.03771*.

Haifeng Yang and 1 others. 2023. Data mining techniques on astronomical spectra data – ii. classification analysis. *Monthly Notices of the Royal Astronomical Society*.

H. Zhang and 1 others. 2024. Survey of transformers and towards ensemble learning for natural language processing. *PMC*.

Systematic Evaluation of Machine Learning and Transformer-Based Methods for Scientific Telescope Literature Classification

Huynh Trung Kiet[1*] **Dao Sy Duy Minh**[1*] **Tran Chi Nguyen**[1*] **Nguyen Lam Phu Quy**[1]
Pham Phu Hoa[1] Nguyen Dinh Ha Duong[1] Dinh Dien[1†] Nguyen Hong Buu Long[1†]
[1]University of Science, VNU-HCM
{23122039, 23122041, 23122044, 23122048, 23122030, 23122002}@student.hcmus.edu.vn
{ddien, nhblong}@fit.hcmus.edu.vn
[*]Equal contribution [†]Corresponding authors

Abstract

Recent space missions such as Hubble, Chandra, and JWST have produced a rapidly growing body of scientific literature. Maintaining telescope bibliographies is essential for mission assessment and research traceability, yet current curation processes rely heavily on manual annotation and do not scale. To facilitate progress in this direction, the TRACS @ WASP 2025 shared task provides a benchmark for automatic telescope bibliographic classification based on scientific publications. In this work, we conduct a comparative study of modeling strategies for this task. We first explore traditional machine learning methods such as multinomial Naive Bayes with TF–IDF and CountVectorizer representations. We then evaluate transformer-based multi-label classification using BERT-based scientific language models. Finally, we investigate a task-wise classification approach, where we decompose the problem into separate prediction tasks and train a dedicated model for each. In addition, we experiment with a limited-resource LLM-based approach, showing that even without full fine-tuning and using only a partial subset of the training data, LLMs exhibit promising potential for telescope classification. Our best system achieves a macro F1 of 0.72 with BERT-based models on the test evaluation, substantially outperforming the official openai-gpt-oss-20b baseline (0.31 macro F1).

1 Introduction

Bibliographic curation plays a central role in scientific knowledge management, enabling mission impact assessment, citation tracking, and research traceability. In astronomy, maintaining telescope bibliographies is essential to quantify the scientific output of major space missions such as Hubble, Chandra, and JWST. However, current bibliographic systems depend predominantly on manual effort, making large-scale curation impractical.

The TRACS @ WASP 2025 (Grezes et al., 2025) shared task formalizes this problem by releasing a benchmark dataset derived from the SAO/NASA Astrophysics Data System (ADS) and defining a unified evaluation framework for telescope bibliography classification. The task jointly addresses telescope detection and scientific intent categorization, reflecting real-world curation needs in astrophysical research.

Automating telescope bibliography classification is challenging due to ambiguous telescope mentions, heterogeneous writing styles across scientific disciplines, and the long-context nature of research articles. Moreover, each publication may involve multiple telescopes simultaneously, leading to a multi-label classification problem under severe label imbalance, where some telescopes (e.g., Chandra) dominate the dataset while others appear rarely. In addition, the dataset contains many hard negative cases, as papers that merely mention telescope names vastly outnumber those that reflect genuine telescope usage, making model learning even more difficult.

In this work, we conduct a systematic study of modeling strategies for telescope bibliographic classification. First, we establish classical machine learning baselines using multinomial Naive Bayes with TF–IDF and CountVectorizer representations, serving as lightweight yet competitive text classification models. Second, we investigate transformer-based multi-label classification using domain-adapted BERT variants such as SciBERT and AstroBERT, which were pre-trained or fine-tuned on large-scale scientific corpora. These models employ a sigmoid output layer with binary cross-entropy loss to support multi-label learning. Third, we explore a task-wise classification strategy by training separate models for each prediction task to reduce cross-label interference. To mitigate severe class imbalance, we incorporate focal loss(Lin et al., 2017) during fine-tuning to better emphasize

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 136–145
December 23, 2025 ©2025 Association for Computational Linguistics

minority labels. Finally, we extend our study with a limited-resource LLM-based approach, where open-weight large language models (LLMs) are evaluated under partial-data and zero-shot settings, demonstrating competitive performance even without full fine-tuning.

Our contributions are as follows:

- We conduct a systematic comparison of modeling strategies for telescope bibliographic classification, covering classical machine learning, transformer-based methods, and LLM-based approaches.
- We show that domain-adapted BERT variants (e.g., SciBERT, AstroBERT) significantly outperform traditional TF–IDF baselines.
- We propose a task-wise classification pipeline with focal loss to mitigate label imbalance.
- We demonstrate that limited-resource LLM inference yields competitive performance even without full fine-tuning.

2 Related work

2.1 Text Representation Methods

Traditional text representation methods have been fundamental to NLP tasks. **TF-IDF (Term Frequency-Inverse Document Frequency)** weights terms based on their frequency in a document relative to their frequency across the corpus, effectively identifying discriminative terms while downweighting common words. **Count Vectorization** represents documents as bags-of-words with raw term frequencies, providing a simple yet effective baseline for many classification tasks. While these methods have been widely used in document classification and information retrieval, they lack semantic understanding and cannot capture contextual word meanings.

2.2 Pre-trained Language Models

The introduction of **BERT (Bidirectional Encoder Representations from Transformers)** (Devlin et al., 2019) revolutionized NLP by pre-training deep bidirectional transformers on large text corpora using masked language modeling and next sentence prediction objectives. BERT's contextualized word representations enable transfer learning across diverse downstream tasks through fine-tuning, achieving state-of-the-art performance on various benchmarks including GLUE(Wang et al., 2019) and SQuAD(Rajpurkar et al., 2016).

Building on BERT's success, **DistilBERT** (Sanh et al., 2019) applies knowledge distillation to create a smaller, faster variant that retains 97% of BERT's language understanding while reducing model size by 40% and inference time by 60%. Through distillation training, DistilBERT learns to mimic BERT's behavior using a student-teacher framework, making it suitable for resource-constrained environments and real-time applications without significant performance degradation.

2.3 Domain-Specific Language Models

Recognizing that general-purpose models may not capture domain-specific terminology and discourse patterns, researchers have developed specialized variants. **SciBERT** (Beltagy et al., 2019) is pre-trained on 1.14M scientific papers from the Semantic Scholar corpus, using a scientific vocabulary and achieving significant improvements on biomedical and computer science tasks.

SPECTER (Scientific Paper Embeddings using Citation-informed TransformERs) (Cohan et al., 2020) takes a different approach by leveraging citation graphs during pre-training. It learns document-level representations by training on triplets of papers where citing papers should have embeddings similar to cited papers, effectively encoding scientific relatedness. However, SPECTER relies on discrete citation relations, which enforce a hard cut-off to similarity and ignore that papers can be very similar despite lacking direct citations.

SciNCL (Scientific Neighborhood Contrastive Learning) (Ostendorff et al., 2022) addresses this limitation by using controlled nearest neighbor sampling over citation graph embeddings for contrastive learning. Instead of discrete citations, SciNCL learns continuous similarity by sampling hard-to-learn negatives and positives while avoiding collisions between samples through margin control. Initialized from SciBERT and trained with neighborhood contrastive objectives, SciNCL outperforms previous methods on the SciDocs (Cohan et al., 2020) benchmark and demonstrates sample-efficient training capabilities.

AstroBERT (Grèzes et al., 2021) further specializes BERT for astronomy by pre-training on astronomical literature from the Astrophysics Data System (ADS). It demonstrates superior performance on astronomy-specific tasks including named entity recognition of celestial objects, classification of astronomical papers, and extraction of observational metadata. AstroBERT's domain adaptation makes it particularly relevant for our telescope bibliography curation task.

These document-level embedding models are particularly relevant to telescope bibliography curation because they capture semantic relationships between scientific papers beyond simple keyword matching. The task requires understanding nuanced distinctions between papers that use telescope data for new scientific results versus those that merely mention the telescope in passing. Citation-aware models like SPECTER and SciNCL can identify papers with similar research contexts, while domain-specific models like AstroBERT understand astronomy terminology and discourse patterns essential for disambiguating telescope references (e.g., distinguishing "Chandra" as a space telescope from other entities with the same name). Furthermore, these models' ability to generate document-level representations enables effective transfer learning for our multi-label classification objectives.

2.4 Fine-tuning Strategies for Transformer Models

While pre-trained language models have shown remarkable capabilities, their effective fine-tuning requires careful consideration of training configurations. (Mosbach et al., 2021) investigate the instability of BERT fine-tuning, revealing that performance can vary significantly across different random seeds, particularly on small datasets. They demonstrate that this instability stems from catastrophic forgetting and vanishing gradients in early layers during fine-tuning.

To address these issues, they propose several techniques:

- **Debiased training**: Using bias correction in the Adam optimizer to stabilize early training steps
- **Re-initialization**: Selectively re-initializing top layers to prevent over-fitting to pre-training tasks
- **Learning rate schedules**: Employing smaller learning rates ($2e-5$ to $5e-5$) with linear warmup and decay
- **Multiple runs**: Averaging predictions across multiple training runs with different seeds to reduce variance

These findings have significant implications for our work, as the telescope bibliography curation task involves multi-label classification on scientific texts where training stability is crucial for reliable performance. We adopt these best practices in our BERT-based approaches, including careful hyper-parameter tuning, multiple seed experiments, and appropriate learning rate scheduling.

2.5 Large Language Models

Recent advances in LLMs have pushed the boundaries of language understanding. The **Qwen2.5**(Yang et al., 2024) series represents efficient multilingual language models with strong performance across diverse tasks. **Qwen2.5-1.5B**(Yang et al., 2024) and **Qwen2.5-3B** (Yang et al., 2024) offer different trade-offs between model capacity and computational efficiency. Despite their smaller size compared to models like GPT-3(Brown et al., 2020) or GPT-(OpenAI et al., 2024), these models demonstrate competitive performance on reasoning, question answering, and classification tasks. Their compact architecture makes them suitable for resource-constrained environments while maintaining strong generalization capabilities.

3 Problem definition

3.1 TRACS Dataset

We conduct our experiments on the TRACS @ WASP 2025 dataset (Grezes et al., 2025), which consists of scientific papers from the SciX bibliographic database annotated with telescope associations and usage categories. Each entry includes textual content from five fields: title, abstract, body, acknowledgments, and grants, along with four boolean labels (science, instrumentation, mention, not_telescope) indicating the paper's relationship to the referenced telescope. The multi-label classification task requires models to simultaneously identify the telescope and categorize how the paper uses or references it. Following the competition setup, we use the provided train.csv and test.csv splits. We perform minimal preprocessing steps to maintain the original text structure:

- Text cleaning: Remove HTML tags, special characters, and reference markers.

- We concatenate all text fields into a single input sequence. For transformer-based models, the input is truncated to a maximum sequence length (512 tokens for BERT-based models and 1024 tokens for LLM-based architectures).

- No sequence truncation is applied as the model handles variable-length sequences automatically.

3.2 Task Formulation

Given a scientific publication p with associated metadata and textual content, we define the telescope bibliography curation task as a multi-label classification problem combined with telescope identification.

Let $\mathcal{D} = \{(p_i, t_i, \mathbf{y}_i)\}_{i=1}^{N}$ denote our dataset of N scientific papers, where:

- p_i represents the i-th paper consisting of five textual components: $p_i = \{p_i^{\text{title}}, p_i^{\text{abstract}}, p_i^{\text{body}}, p_i^{\text{ack}}, p_i^{\text{grants}}\}$
- $t_i \in \mathcal{T}$ denotes the associated telescope, where $\mathcal{T}$ is the set of all telescopes in our taxonomy
- $\mathbf{y}_i = [y_i^{\text{sci}}, y_i^{\text{inst}}, y_i^{\text{men}}, y_i^{\text{not}}] \in \{0,1\}^4$ represents the multi-label annotation vector

3.3 Label Definitions

The four binary labels characterize the relationship between the paper and the referenced telescope. For each paper p:

- $y^{\text{sci}} = 1$ if p uses telescope data for new scientific results, 0 otherwise
- $y^{\text{inst}} = 1$ if p describes technical or instrumental aspects, 0 otherwise
- $y^{\text{men}} = 1$ if p mentions telescope without producing new results, 0 otherwise
- $y^{\text{not}} = 1$ if p contains false positive reference, 0 otherwise

3.4 Objective

Our goal is to predict two components for each paper p:

1. The telescope identifier: $\hat{t} \in \mathcal{T}$
2. The multi-label vector: $\hat{\mathbf{y}} \in \{0,1\}^4$

This can be achieved through various modeling approaches, including joint multi-task learning, pipeline architectures, or ensemble methods.

4 Methodology

4.1 Classical Machine Learning Approaches

We establish baseline models using classical machine learning methods with two text representation strategies: TF-IDF vectorization and count-based vectorization, combined with Multinomial Naive Bayes classifiers.

4.1.1 Text Representation

Given a paper p with concatenated text from all fields, we construct feature vectors using:

TF-IDF Vectorization: For each term w in paper p, the TF-IDF weight is computed as:

$$\text{TF-IDF}(w,p) = \text{TF}(w,p) \times \log \frac{N}{\text{DF}(w)}$$

where $\text{TF}(w,p)$ is the term frequency of word w in paper p, N is the total number of documents, and $\text{DF}(w)$ is the document frequency of word w.

Count Vectorization: We represent each paper as a vector of raw term frequencies:

$$\mathbf{v}_p = [\text{TF}(w_1,p), \text{TF}(w_2,p), ..., \text{TF}(w_{|V|},p)]$$

where $|V|$ is the vocabulary size.

4.1.2 Classification Strategy

We employ Multinomial Naive Bayes classifiers with different strategies for telescope identification and label prediction:

Telescope Identification: For the multi-class telescope classification problem, we use a One-vs-Rest (OvR) approach. For each telescope $t \in \mathcal{T}$, we train a binary classifier:

$$P(t|\mathbf{v}_p) = \frac{P(\mathbf{v}_p|t) \cdot P(t)}{\sum_{t' \in \mathcal{T}} P(\mathbf{v}_p|t') \cdot P(t')}$$

The predicted telescope is:

$$\hat{t} = \arg\max_{t \in \mathcal{T}} P(t|\mathbf{v}_p)$$

Binary Classification: For each of the four binary labels $l \in \{\text{sci}, \text{inst}, \text{men}, \text{not}\}$, we train independent binary Multinomial Naive Bayes classifiers:

$$P(y^l = 1|\mathbf{v}_p) = \frac{P(\mathbf{v}_p|y^l = 1) \cdot P(y^l = 1)}{P(\mathbf{v}_p)}$$

Each label is predicted independently, allowing multiple labels to be assigned to a single paper when appropriate.

4.2 BERT-based Approaches

We apply transformer models with the following processing pipeline:

4.2.1 Tokenization

Input text is tokenized using the tokenizer corresponding to each pre-trained model. Due to the context length limitation of transformer models, the model **automatically truncates sequences to the first 512 tokens**, which is the maximum sequence length for BERT-based models. This typically includes the entire title and most of the abstract, which contain the most important information of the paper.

Given input text x, the tokenization process produces a sequence of token IDs:

$$\mathbf{t} = \text{Tokenize}(x) = [t_1, t_2, \ldots, t_n]$$

where $n \leq 512$. These tokens are then converted to embeddings and processed through the transformer encoder to obtain contextualized representations:

$$\mathbf{H} = \text{Transformer}(\mathbf{t}) = [\mathbf{h}_{\text{[CLS]}}, \mathbf{h}_1, \ldots, \mathbf{h}_n]$$

where $\mathbf{h}_{\text{[CLS]}} \in \mathbb{R}^d$ is the representation of the [CLS] token used for classification.

4.2.2 Classification Heads

We train two separate models, each with its own specialized classification head:

Multi-label Classification Model: A fully-connected layer with sigmoid activation is attached to the transformer encoder to predict 4 labels simultaneously (each label independently):

$$\mathbf{p}_{\text{multi}} = \sigma(\mathbf{W}_{\text{multi}}\mathbf{h}_{\text{[CLS]}} + \mathbf{b}_{\text{multi}})$$

where $\mathbf{W}_{\text{multi}} \in \mathbb{R}^{4\times d}$, $\mathbf{b}_{\text{multi}} \in \mathbb{R}^4$, and σ is the sigmoid function applied element-wise.

Telescope Identification Model: A separate model with a fully-connected layer and softmax activation is used to classify telescope types:

$$\mathbf{p}_{\text{telescope}} = \text{softmax}(\mathbf{W}_{\text{telescope}}\mathbf{h}_{\text{[CLS]}} + \mathbf{b}_{\text{telescope}})$$

where $\mathbf{W}_{\text{telescope}} \in \mathbb{R}^{K\times d}$, $\mathbf{b}_{\text{telescope}} \in \mathbb{R}^K$, and K is the number of telescope types.

Both models share the same transformer encoder architecture but are trained independently with their respective loss functions.

4.2.3 Training Objective

We train models independently or jointly for different classification tasks, using task-specific loss functions optimized for their respective objectives.

Binary Classification. For the four binary labels, we employ binary cross-entropy loss:

$$\mathcal{L}_{\text{multi-label}} = -\frac{1}{4}\sum_{i=1}^{4}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

Not-Telescope Classification. Due to significant class imbalance in the `not_telescope` category, we also employ focal loss when training a independent dedicated binary classifier.

$$\mathcal{L}_{\text{not-tel}} = -\Big[y \cdot \alpha(1-p)^\gamma \log(p) + (1-y)\cdot(1-\alpha)p^\gamma \log(1-p)\Big]$$

Telescope Identification. For multi-class telescope classification over K telescope types, we use categorical cross-entropy:

$$\mathcal{L}_{\text{telescope}} = -\sum_{k=1}^{K} y_k \log(p_k)$$

where $y_k \in \{0,1\}$ is the one-hot encoded label and p_k is the predicted probability for telescope class k.

Each model is trained independently with its respective loss function, using the same base transformer architecture but optimized separately for its specific classification task. This modular approach allows task-specific optimization strategies and hyperparameter tuning.

4.2.4 Inference

At inference time, the model takes the first 512 tokens of a paper as input and forwards through the encoder. The encoded representation is then passed through two separate classification heads: one predicts the telescope type, and the other predicts the 4 classification labels (multi-label classification).

4.3 LLM-based Approach

We leverage large language models through parameter-efficient fine-tuning using QLoRA (Quantized Low-Rank Adaptation)(Dettmers et al., 2023), which enables training on consumer hardware by quantizing the base model to 4-bit precision while training low-rank adapter matrices.

Model Architecture. We fine-tune Qwen-1B and Qwen-3B models by freezing the quantized base parameters $\mathbf{W}$ and learning low-rank decompositions $\mathbf{AB}$ with rank r. The adapted weight matrix becomes: $\mathbf{W}' = \mathbf{W}_{\text{4-bit}} + \alpha \cdot \mathbf{AB}$

Task Formulation. We formulate classification as structured generation where the model outputs

JSON with telescope identification and binary labels. Each input consists of concatenated paper fields with a detailed system prompt encoding:

- Task objectives and label definitions
- Classification rules (e.g., mutual exclusivity of `not_telescope`)
- Output constraints (strict JSON schema)

System Prompt. Our prompt explicitly defines each category:

- `science`: Uses telescope data for new results
- `instrumentation`: Describes technical/engineering aspects
- `mention`: References telescope without new contributions
- `not_telescope`: Contains false positive references

The model is trained to generate valid JSON responses that are parsed during inference to extract predictions. This approach allows the LLM to reason about complex classification rules while producing structured outputs suitable for evaluation.

5 Experiments

5.1 Baselines

The TRACS organizers provide two official baseline models for comparison. Table 1 presents their performance on the test set.

Model	Macro F1
Random	0.24
openai-gpt-oss-20b	0.31

Table 1: Baseline performance on TRACS test set.

5.2 Experimental Setup

We split the training data into training and validation sets with an 8:2 ratio for model development and hyperparameter tuning. We train our models using adamw_torch optimizer with a learning rate of 2e-5, batch size of 16, and maximum sequence length of 512 tokens. For the multi-task models, training continues for 3 epochs with early stopping based on validation performance. For per-class binary classifiers, we train for 1-2 epochs to prevent overfitting, as single-task models tend to converge faster and are more prone to overfitting. All experiments are conducted on NVIDIA A100 GPUs via Google Colab. The primary evaluation metric is macro F1-score computed across both telescope identification and the four classification labels, ensuring balanced performance across all categories.

5.3 Main Results

5.3.1 Per-Class Specialized Models

To further improve classification performance, we train separate binary classifiers for each of the four classification categories (science, instrumentation, mention, not_telescope) and the telescope identification task. Table 2 shows the performance of our best model (SciBERT) when trained independently for each class.

Classification Task	F1 Score
Multi-label Classification	
science	0.78
instrumentation	0.76
mention	0.73
not_telescope	0.61
Macro F1 (Classification)	**0.72**

Table 2: Per-class F1 scores using separate SciBERT classifiers trained independently for each task. Macro F1 is computed as the average across all four classification categories.

5.3.2 Instruction-tuned LLM Evaluation

Training Configuration Table 3 presents the hyperparameters used for QLoRA fine-tuning. We employ 4-bit quantization to reduce memory footprint while maintaining model performance. The effective batch size of 8 is achieved through gradient accumulation, allowing training on consumer-grade hardware.

Hyperparameter	Value
Learning Rate	1×10^{-4}
Batch Size (per device)	1
Gradient Accumulation	8
Effective Batch Size	8
Max Epochs	3
Max Sequence Length	1024
Quantization	4-bit

Table 3: QLoRA fine-tuning hyperparameters for Qwen models.

Prompt Design We construct a structured system prompt that includes:

- **Role definition:** Positioning the model as an expert assistant for telescope paper classification

- **Category definitions:** Explicit descriptions of *science*, *instrumentation*, *mention*, and *not_telescope*

- **Classification rules:** Constraints such as mutual exclusivity of *not_telescope* and multi-label capability for other categories

- **Edge cases:** Guidelines for handling ambiguous references, name collisions, and grant-only mentions

- **Output format:** Strict JSON schema enforcement to ensure parseable predictions

Results The complete prompt template is provided in Appendix A. This prompt is prepended to each paper's content during both training and inference phases.

Method	Parameters	Macro F1
Qwen-1B + QLoRA	1B	0.58
Qwen-3B + QLoRA	3B	0.61

Table 4: Performance comparison on the multi-label classification task, trained for **a single epoch**. Macro F1 is averaged across all four categories (science, instrumentation, mention, not_telescope).

5.3.3 Joint Task Performance

We assess all models on the unified task encompassing both telescope identification and publication classification. The overall leaderboard score is defined as the arithmetic mean of the F1 score for telescope identification and the macro-averaged F1 across the four classification categories, formulated as:

$$\text{Final Score} = \frac{\text{Telescope F1} + \text{Classification Macro F1}}{2}.$$

Table 5 summarizes the complete performance comparison across all evaluated methods.

5.3.4 Ablation Study

To examine the effect of focal loss, we fine-tuned task-specific models with and without focal loss on imbalanced tasks. Although focal loss slightly improved per-task stability, these models still performed worse than the joint multi-task model trained without focal loss, indicating that task interaction contributes more to generalization than loss reweighting alone.

6 Conclusion

In this work, we presented a systematic study of modeling strategies for automatic telescope bibliographic classification in the TRACS @ WASP 2025 shared task. We evaluated a diverse range of approaches, from classical machine learning methods to transformer-based architectures and limited-resource LLM-based inference.

Our experiments demonstrate that domain-adapted BERT variants significantly outperform traditional ML, with SciBERT achieving the best performance of 0.73 macro F1 on the leaderboard evaluation—more than doubling the official baseline score of 0.31. We show that pre-training on scientific corpora provides substantial benefits for this task, as evidenced by the strong performance of SciBERT and AstroBERT compared to general-domain models.

While ensemble methods did not yield improvements in our experiments, we attribute this primarily to the multi-label, multi-class complexity of the task and our computational constraints. Our task-wise classification approach with focal loss showed promise in addressing class imbalance, though further investigation with larger models and more extensive hyperparameter tuning could yield additional gains.

Importantly, our limited-resource LLM experiments suggest that instruction-tuned models can achieve competitive performance even without full fine-tuning and with only partial training data. This opens promising directions for low-resource scenarios in scientific bibliography curation.

Future work should explore more sophisticated long-document processing strategies to better leverage complete paper content, investigate advanced techniques for handling severe class imbalance in multi-label settings, and examine larger-scale LLM fine-tuning with expanded computational resources. Additionally, incorporating metadata such as author affiliations, publication venues, and citation networks may further improve classification accuracy.

Limitations

This study faces several important constraints:

Computational Resource Constraints: As students, we faced significant GPU and computational limitations. This restricted our ability to experiment with larger models (e.g., full fine-tuning of

Method	Multi-label Classification			Telescope Identification		Leaderboard
	Macro F1	Micro F1	Samples F1	Accuracy	Macro F1	Score
Traditional ML						
TF-IDF	0.52	0.58	-	0.64	0.51	0.51
CountVectorizer	0.54	0.59	-	0.67	0.56	0.56
Transformer Models						
DistilBERT (66M)	0.71	0.73	0.72	0.81	0.68	0.69
SciNCL (110M)	0.71	0.73	0.72	0.80	0.68	0.68
AstroBERT (110M)	0.72	0.73	0.74	0.80	0.69	0.68
SPECTER (110M)	0.70	0.72	0.72	0.81	0.68	0.69
SciBERT (110M)	**0.77**	**0.79**	**0.78**	**0.81**	**0.73**	**0.72**
Ensemble Methods						
Soft Voting	0.70	0.72	0.71	0.73	0.65	0.67
Weighted Voting	0.71	0.73	0.72	0.74	0.66	0.68
Hard Voting	0.69	0.71	0.70	0.72	0.64	0.66

Table 5: Comparison of traditional ML, transformer-based models, and ensemble methods on joint telescope identification and paper classification tasks. The leaderboard score is computed as the average of Telescope Macro F1 and Classification Macro F1. Ensemble methods combine SciBERT, DistilBERT, and AstroBERT using different voting strategies but show slight performance degradation compared to the best single model (SciBERT). With SciBERT, our system achieves a **Top-2** ranking on the leaderboard.

models beyond 3B parameters) and limited the hyperparameter search space we could explore.

Ensemble Methods Underperformance: Despite theoretical advantages, our ensemble approaches did not yield substantial improvements. This is likely due to the multi-label, multi-class nature of the task where predictions must simultaneously classify both the telescope type and four binary labels (science, instrumentation, mention, not_telescope). The complexity of combining predictions across these dimensions without introducing conflicting classifications proved challenging within our resource constraints.

Class Imbalance: The dataset exhibits significant class imbalance across both telescope types and label categories. Certain telescope-label combinations are severely underrepresented, making it difficult for models to learn robust patterns for minority classes and potentially biasing predictions toward more frequent categories.

Long Document Processing: Scientific papers often contain extensive text spanning abstracts, full body text, and acknowledgments. Processing these long sequences requires either truncation (risking information loss) or sophisticated chunking strategies. Our computational constraints limited our ability to fully leverage the complete textual context, particularly for papers exceeding typical transformer input limits (512 tokens).

Acknowledgments

We thank the organizers of the TRACS @ WASP 2025 shared task for providing this valuable learning opportunity and the benchmark dataset. We are grateful to the SAO/NASA Astrophysics Data System (ADS) for maintaining the telescope bibliography infrastructure that made this work possible.

Through this shared task, we gained significant insights into challenges of long-document processing, resource-constrained LLM deployment, and domain-specific NLP in astronomy. We acknowledge the developers of the Hugging Face Transformers library and the creators of SciBERT, AstroBERT, SPECTER, and Qwen models for making their pre-trained models publicly available.

Despite computational limitations as students, this experience deepened our understanding of practical deployment considerations for scientific text classification systems and the unique characteristics of astronomical literature.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *Preprint*, arXiv:1903.10676.

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. *Preprint*, arXiv:2005.14165.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

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.

Félix Grèzes, Sergi Blanco-Cuaresma, and .... 2021. Building astrobert, a language model for astronomy astrophysics. *arXiv preprint arXiv:2112.00590*. Preprint; also presented at ADASS 2021/ADASS 2022.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *Preprint*, arXiv:2006.04884.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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)*, pages 11670–11688. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Preprint*, arXiv:1804.07461.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

A System Prompt for TRACS Classification

Below is the complete system prompt used to fine-tune Qwen models and guide their inference on the TRACS shared task.

You are an expert assistant for the TRACS (Telescope Reference and Astronomy Categorization Shared Task) at WASP @ IJCNLP-AACL 2025. Your task is to classify scientific papers according to telescope usage categories defined by the shared task guidelines.

Given the paper content, identify:

1. telescope: The specific telescope referenced in the paper.

2. science: True if the paper uses telescope data to produce new scientific results.

3. instrumentation: True if the paper describes technical aspects of the telescope (hardware/software/calibration/data pipeline).

4. mention: True if the paper references the telescope but does not produce new results nor address technical aspects.

5. not_telescope: True if the paper contains misleading telescope-like references or false positives unrelated to an actual telescope.

Classification Rules:

- A paper can be classified into multiple categories except ‘not_telescope‘, which is mutually exclusive.
- If a paper qualifies for ‘science‘, it must be labeled science=True even if it also mentions the telescope.
- If a paper discusses telescope engineering or data processing, label instrumentation=True.
- Papers that only cite a telescope historically, in background, or for comparison → mention=True.
- If the telescope name is used ambiguously (e.g. name collision with a person, project, or acronym) → not_telescope=True.
- Referencing telescope-funded grants alone without data use → not_telescope=True.

Output Format:

Respond strictly in valid JSON only:

```
{
  "telescope": "<string>",
  "science": <true/false>,
  "instrumentation": <true/false>,
  "mention": <true/false>,
  "not_telescope": <true/false>
}
```

"Clutch or Cry" Team at TRACS @ WASP 2025: A Hybrid Stacking Ensemble for Astrophysical Document Classification

Arshad Khatib and **Aayush Prasad** and **Rudra Trivedi**
Department of Artificial Intelligence
And
Shrikant Malviya
Department of Computer Science and Engineering
SVNIT Surat
{u24ai112, u24ai091, u24ai068}@aid.svnit.ac.in, shrikant@coed.svnit.ac.in

Abstract

Automatically identifying telescopes and their roles within astrophysical literature is crucial for large-scale scientific analysis and tracking instrument usage patterns. This paper describes the system developed by the "Clutch or Cry" team for the Telescope Reference and Astronomy Categorization Shared task (TRACS) at WASP 2025 (Grezes et al., 2025). The task involved multi-class telescope identification (Task 1) and multi-label role classification (Task 2) within scientific papers. For Task 1, we employed a feature-engineering approach centered on document identifiers (`Id suffix`) combined with metadata and textual features, utilizing a tuned Random Forest classifier to achieve high accuracy. For the more complex Task 2, we utilized a carefully designed two-level stacking ensemble. Level-0 combines a rule-based keyword classifier with the domain-adapted `astroBERT` transformer, effectively fusing symbolic and semantic information. Level-1 uses four independent XGBoost meta-learners for targeted per-role optimization. These architectures address the primary challenges: handling long documents and managing severe class imbalance in Task 2 (notably 1:91 for `instrumentation`). Systematic optimization focused on mitigating imbalance significantly improved Task 2 performance for minority classes. This work validates the effectiveness of tailored approaches for distinct subtasks and targeted optimization for imbalanced classification in specialized scientific domains.

1 Introduction

Automated classification of scientific literature is critical for knowledge discovery and resource management in large-scale research repositories. With millions of astrophysical papers archived in systems like the NASA Astrophysics Data System (ADS), manual annotation and categorization become infeasible (SAO/NASA Astrophysics Data System, 2025). Effective automated methods enable researchers to quickly identify relevant studies, track telescope usage patterns, understand instrumental capabilities, and trace scientific methodologies—ultimately accelerating scientific discovery and facilitating data-driven insights into observational astronomy practices (Wikipedia contributors, 2025). This capability extends beyond administrative utility, directly supporting evidence synthesis, reproducibility verification, and interdisciplinary research collaboration.

The Telescope Reference and Astronomy Categorization Shared task (TRACS) presents two intertwined classification challenges that together model real-world requirements faced by digital astronomy libraries (Kaggle, 2025). The task demands systems capable of identifying which telescopes are discussed as primary subjects versus peripheral mentions, and distinguishing the functional role of telescopes within scientific contexts—whether used for data acquisition, instrument characterization, or comparative analysis. These distinctions are semantically nuanced, often embedded in lengthy papers with inconsistent terminology, and severely imbalanced across class distributions. This shared task provides an ideal proving ground for advancing both fundamental NLP techniques and domain-specific adaptations needed for specialized scientific corpora, offering valuable insights into how machine learning systems can handle real-world complexity in domain-specific document understanding.

Addressing these challenges—long document context, nuanced semantic roles, and severe class imbalance—requires a robust and adaptable classification architecture. Simple models often struggle

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 146–156
December 23, 2025

with the sheer length of scientific papers and are overwhelmed by the majority of classes. We hypothesize that a hybrid stacking ensemble method offers a compelling solution. By combining a fast, symbolic keyword classifier (effective at broad categorization and handling explicit mentions across long texts) with a deep semantic model like astroBERT (capable of understanding nuanced context within specific text windows), we can leverage complementary strengths. Furthermore, employing a stacking architecture with independent, per-class meta-learners enables highly targeted optimization, allowing us to apply aggressive techniques such as class weighting and threshold tuning precisely where needed to combat the extreme class imbalance observed in the TRACS dataset. The specialized multi-level approach forms the core of our system design.

1.1 Shared Task: TRACS-2025

The shared task (Kaggle, 2025) comprises two classification objectives:

- **Task 1 (Telescope Identification):** Multi-class classification identifying the primary telescope discussed in a paper from the set {CHANDRA, HST, JWST, None}.
- **Task 2 (Role Classification):** Multi-label classification determining telescope roles with four binary labels: science, instrumentation, mention, and not_telescope.

1.2 Key Challenges

Two major challenges characterize this task:

Document Length and Context: Full-text scientific articles frequently exceed the input token limits of standard transformer models (typically 512 – 2048 tokens), requiring careful strategies for capturing relevant information from lengthy documents.

Severe Class Imbalance: Both tasks exhibit pronounced class imbalance. In **Task 1 (Telescope Identification)**, the distribution is extremely skewed. The **NONE** class represents a tiny fraction of the dataset (approximately 1 instance for every 273 samples), making it vastly outnumbered by majority classes like HST (which appears roughly 126 times more often than NONE). In **Task 2 (Document Role Classification)**, the **instrumentation** class appears with a positive-to-negative ratio of approximately 1:91, while **not_telescope** exhibits a ratio closer to 1:9. This extreme imbalance renders standard machine learning approaches ineffective, as models naturally bias toward majority classes.

We address these challenges through an ensemble-based methodology that combines symbolic and semantic models. Instead of optimizing a single model architecture, we leverage the complementary strengths of combined rule-based and neural approaches, enabling targeted optimization for each of the four output labels.

Our contribution includes:

1. A carefully designed two-level stacking architecture.
2. Systematic methodology for addressing extreme class imbalance through multiple complementary techniques.
3. Empirical validation that per-class optimization significantly improves performance on minority classes.

All code and trained models will be released publicly later to ensure reproducibility. Link: https://github.com/Arshad-13/ClutchOrCry-TRACS-2025

2 Related Work

Handling imbalanced classification is a well-studied problem. Common approaches include oversampling techniques such as SMOTE (Synthetic Minority Oversampling Technique) (Chawla et al., 2002), undersampling (Kubat and Matwin, 1997), cost-sensitive learning (Elkan, 2001), and ensemble methods (Galar et al., 2012). In NLP, handling imbalanced text classification has been addressed through various techniques, including threshold adjustment for optimal F1 scores (Zou et al., 2016; Hong et al., 2016) and cost-sensitive learning strategies (Elkan, 2001; Lee and Kim, 2020). Threshold adjustment helps by shifting the decision boundary away from the default 0.5 probability; it allows the model to correctly identify more instances of the rare class, often improving recall and the F1-score even if precision decreases slightly. Cost-sensitive learning directly tackles the imbalance during training by assigning a higher penalty for misclassifying minority class instances, forcing the model to learn features that better distinguish the rare class from the majority class. These techniques are pertinent to both tasks, given the severe class imbalance observed.

Rule-based systems and extensive feature engineering are often employed in scientific document classification, particularly when structured identifiers or metadata offer strong predictive signals. Approaches leveraging bibcodes or similar identifiers for categorization are common in bibliographic analysis and information retrieval within specific scientific domains. These methods excel at precision when identifier patterns are consistent but may lack robustness to variations or require supplementation with other features. Hybrid approaches, combining rule-based extraction with machine learning models trained on engineered features (including text statistics, keyword counts, and metadata), aim to balance the high precision of rules with the broader pattern recognition capabilities of models like Random Forests, a strategy reflected in our Task 1 architecture.

Domain adaptation of pre-trained transformers has also proven effective for specialized NLP tasks. Recent work in scientific document understanding has leveraged domain-specific models like SciBERT (Beltagy et al., 2019) and BioBERT, demonstrating that pre-training on domain-specific corpora improves downstream task performance. For astronomical text, `astroBERT` (Grezes et al., 2021) provides pre-training on 440,000 astrophysical abstracts from the NASA Astrophysics Data System, offering domain-specific vocabulary and patterns critical for astronomy-related classification tasks, which we utilize in our Task 2 system.

Ensemble methods that combine diverse classifiers have demonstrated strong performance on imbalanced problems (Galar et al., 2011; Khan et al., 2023). While simpler ensemble models, such as Random Forest (used in Task 1), inherently handle feature interactions, stacking ensembles, in particular, allow meta-learners to learn optimal combination strategies for integrating base model predictions (Nugroho et al., 2023). Our Task 2 approach extends this paradigm by using per-class meta-learners rather than a single global meta-learner, enabling fine-grained hyperparameter optimization tailored to each label's unique characteristics.

3 System Architecture

We employ distinct architectures tailored to the specific requirements of each task. Task 1 focuses on identifying the primary telescope using *rule-based features and a Random Forest*, while Task 2 uses a *stacking ensemble method* to classify the role of the document concerning telescopes.

3.1 Task 1: Telescope Identification Architecture

For identifying the primary telescope associated with an astrophysical document, our system employs a feature-engineering-centric approach, culminating in a *Random Forest classification model*. This architecture prioritizes extracting strong signals from the document identifier (`Id`), supplemented by metadata and textual features to enhance robustness and handle edge cases.

3.1.1 Rule-Based Feature Extraction (ID Suffix)

The cornerstone of this system is the extraction and encoding of information presumed to be embedded within the document's `Id` field, often structured similarly to astrophysical bibcodes.

Primary Rule: The system identifies the suffix following the last underscore (_) character in the `Id` string. **Mapping:** Recognized suffixes (e.g., `CHANDRA`, `HST`, `JWST`) are directly mapped to their corresponding telescope labels. `Id` strings without a recognized suffix or underscore are assigned a default category (e.g., `NONE` or `NO_UNDERSCORE`). **Feature Encoding:** The extracted suffix string is numerically encoded (e.g., using `LabelEncoder`) to be used as a categorical feature by the classification model. Additional binary features like `has_underscore` are also generated. This explicit encoding of the rule's output provides a high-precision signal to the classifier.

3.1.2 Comprehensive Feature Engineering

To complement the primary ID suffix feature and improve classification accuracy, especially for documents where the ID rule is insufficient, a wide array of supplementary features are engineered:

ID/Bibcode Characteristics: Features derived from the `Id` string itself, including its total length, the count of underscores, and the categorical prefix (often representing the year or journal, also label encoded).

Metadata Features: Utilizing the provided `year`, including derived features like the difference from a reference year and flags indicating publication eras (e.g., recent JWST era, pre-Chandra era).

Textual Content Features: *Length Features:* Character lengths of fields such as `title`, `abstract`, and `body`. Word counts are also included for key fields. *Keyword Mentions:* Binary flags and counts indicating the presence of specific telescope names (Chandra, JWST, Hubble/HST) within the `title`, `abstract`, body, and `acknowledgments`. *TF-IDF Representation:* Term Frequency-Inverse Document Frequency vectors generated from the combined text of the `title`, `abstract`, and body fields, using a constrained vocabulary (e.g., 150 features) and considering unigrams and bigrams.

Author & Grant Features: Simple features like author count and binary flags for the presence of grant or acknowledgment text.

3.1.3 Classification Model (Random Forest)

The final classification is performed by a `RandomForestClassifier` model. It takes a concatenated feature vector comprising all the engineered numeric/categorical features (including the encoded ID prefix) and the sparse TF-IDF text features.

Training and Hyperparameter Tuning: The model is trained on the full set of derived features. To optimize performance, hyperparameters were tuned using `RandomizedSearchCV` with 5-fold stratified cross-validation. The best parameters identified were:

- `n_estimators`: 200
- `max_depth`: 15
- `min_samples_split`: 2
- `min_samples_leaf`: 4
- `max_features`: 'sqrt'

This configuration achieved the best cross-validation accuracy of 0.7772. The final model used for prediction (`best_estimator_` from `RandomizedSearchCV`) is implicitly trained on the entire dataset using these optimal parameters. Class weighting (`class_weight='balanced'`) was also employed during the search process to mitigate the inherent imbalance in telescope label distribution. The model predicts a single categorical label representing the identified primary telescope (`CHANDRA`, `HST`, `JWST`, or `NONE`). Feature importance analysis consistently confirms that the ID suffix-derived features are the most dominant predictors, validating the hybrid rule-based and machine-learning strategy.

3.2 Task 2: Document Role Classification Architecture

As shown in Figure 1, our system for Task 2 utilizes a two-tier approach, comprising 'level 0' and 'level 1', within the stacking ensemble intended to merge quick symbolic classification with a slower yet more accurate semantic comprehension.

3.2.1 Level-0: Base Models

Rule-Based Keyword Classifier The keyword classifier provides high-recall signals through pattern matching. It utilizes a dictionary of over 1,000 domain-specific keywords (spanning telescope names, instruments, and scientific concepts), which we curated using a combination of large language models (LLMs) and established astrophysical references. Scores documents based on the presence, frequency, and contextual proximity of keywords. Outputs a 4-dimensional pseudo-probability vector, one value per output label, computed as normalized keyword match scores. While this approach cannot capture semantic nuance, it provides reliable signals for explicit references and demonstrates high recall for documents containing direct mentions of telescopes or scientific roles.

Fine-Tuning astroBERT The transformer component leverages `adsabs/astroBERT` (Grezes et al., 2021), a BERT variant pre-trained on 440,000 abstracts from astrophysical literature. The model provides domain-specific vocabulary and contextual understanding of astrophysical language. It is fine-tuned on the provided training data for 3 epochs using a learning rate of 2e-5. The model generates probabilities for three labels: `SCIENCE`, `INSTRUMENTATION`, and `MENTION`, excluding the `NOT_TELESCOPE` class, which is semantically distinct and handled exclusively by the keyword classifier and meta-learner. It outputs a 3-dimensional feature vector.

It reflects our hypothesis that `NOT_TELESCOPE` documents (discussing telescopes in non-primary contexts) require different signals than documents describing telescope roles in primary scientific contexts (see Section 4).

3.2.2 Level-1: Meta-Learner

The Level-1 meta-learner combines base model outputs into a unified classification:

Feature Construction: Outputs from both base models are concatenated into a 7-dimensional feature vector: $\mathbf{x}_{\text{meta}} = [\mathbf{x}_{\text{keyword}}, \mathbf{x}_{\text{astroBERT}}]$ where $\mathbf{x}_{\text{keyword}} \in \mathbb{R}^4$ and $\mathbf{x}_{\text{astroBERT}} \in \mathbb{R}^3$.

Per-Label Meta-Learners: Instead of training a single multi-label classifier, we train four independent XGBoost classifiers M_i (Chen and Guestrin,

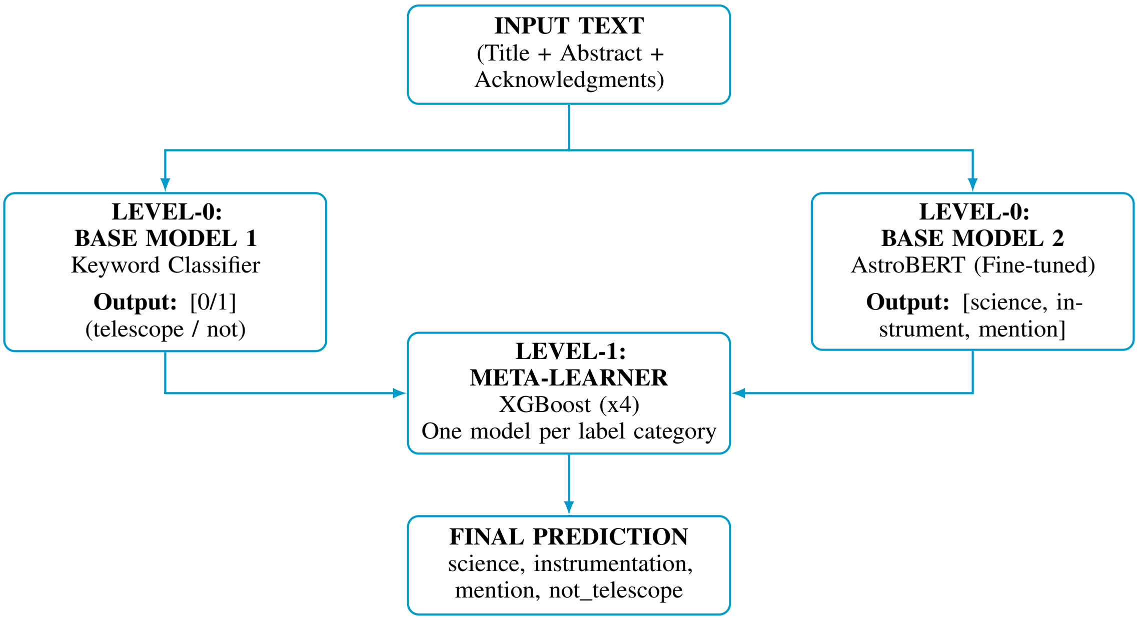


Figure 1: Two-level stacking ensemble architecture (Task 2). Level-0 base models process input text independently, producing 4 and 3-dimensional vectors, respectively. These are concatenated into a 7-dimensional meta-feature vector, which serves as input for four independent Level-1 meta-learners (one per label), each producing binary predictions.

2016), one per label. This one-vs-rest approach enables:

- independent hyperparameter optimization (particularly `scale_pos_weight`) tailored to each label's unique imbalance ratio.
- isolation of optimization strategies: SMOTE and calibration are applied only to models requiring them.
- flexibility to apply different decision thresholds for different labels.

Each meta-learner M_i produces a binary probability $p_i \in [0, 1]$ for label i, which is converted to a binary prediction using a label-specific threshold τ_i (see Section 5).

4 Subtask 1: Telescope Identification

4.1 Model 1: Stacked LSTM Network

Our initial approach used a stacked Long Short-Term Memory (LSTM) network to exploit the sequential structure of text in the title, abstract, and author fields. **Input:** Tokenized title, abstract, and author fields. **Architecture:** Two stacked LSTM layers (64 units each), followed by a Dense softmax classification layer. **Output:** Multi-class probabilities over four telescope classes.

4.2 Model 2: Domain-Specific Transformer (AstroBERT)

We transitioned to a more powerful, domain-adapted language model: `astroBERT`, pre-trained on astrophysics literature and fine-tuned it with a classification head on TRACS data. Deep language understanding alone was insufficient. Semantic signals were not strong enough to capture the presence or absence of telescope mentions.

4.3 Model 3: Hybrid (Logistic Regression + AstroBERT)

To better isolate the difficult `None` class, we decoupled its prediction into a binary subtask. First, a Logistic Regression model predicted whether a sample belonged to the `None` class. If not, astroBERT classified it into `CHANDRA`, `HST`, or `JWST`.

4.4 Model 4: Feature-Based Random Forest

We shifted focus from textual models to structured metadata features using a RandomForest classifier. The engineered features included field-specific keyword counts, publication year, and author-based patterns.

4.5 Model 5: Final Hybrid (RandomForest + Rule-Based Heuristic)

A comparative analysis of the previous models confirmed that the feature-engineered RandomForest was the most promising direction. However, a deep dive into its confusion matrix revealed a critical performance bottleneck: the vast majority of classification errors occurred because the model was consistently confusing two specific categories. To address this targeted issue, we sought a deterministic feature that could serve as a tie-breaker. We discovered a decisive cue in the `bibcode` field, where the suffix (the text after the last underscore) deterministically aligned with the true class label. This insight was used to create a rule-based override specifically for instances where the model was likely to be confused.

The final hybrid approach began with the output of the Random Forest model and then applied a rule-based correction to address its specific, known weakness. If the model's prediction was one of the two commonly confused fields, the system applied the rule-based override by extracting the final token from the `bibcode`. For all other predictions, the model's original output was trusted.

5 Subtask 2: Telescope Role Classification

We employed an iterative development methodology for Task 2, beginning with a baseline model and systematically addressing performance bottlenecks related to class imbalance.

5.1 Baseline System

Our baseline model employed standard stacking without specialized handling for imbalanced data. It used default XGBoost hyperparameters (`scale_pos_weight=1`, `max_depth=6`, `learning_rate=0.1`), a fixed decision threshold of 0.5 for all labels, and implemented no specific data augmentation or class weighting strategies.

This baseline achieved a Macro F1-score of **0.6191**, with severe degradation on minority classes (Table 1). The `INSTRUMENTATION` class achieved only 0.510 F1, while `NOT_TELESCOPE` reached 0.480 F1.

5.2 Optimization Strategies

To improve upon the baseline, we implemented five complementary techniques targeting different aspects of model training and prediction on imbalanced data. These strategies are summarized in Table 2.

Justification for Selective SMOTE Application We applied SMOTE exclusively to the `INSTRUMENTATION` meta-learner due to its extreme imbalance (1:91). Synthesizing data was deemed necessary to provide sufficient signal for the model to learn this rare class effectively. For the `NOT_TELESCOPE` class, with a more moderate imbalance (1:9), we found that aggressive class weighting (Strategy 3) alone was sufficient to manage the imbalance without the potential noise introduction or overfitting risks associated with synthetic data generation. The majority of classes required neither technique.

Label	Precision	Recall	F1	Support
`INSTRUMENTATION`	0.650	0.420	0.510	132
`MENTION`	0.700	0.750	0.722	892
`NOT_TELESCOPE`	0.580	0.410	0.480	187
`SCIENCE`	0.780	0.750	0.765	2156
Macro Avg	0.678	0.582	0.619	—

Table 1: Baseline performance before optimization. Class imbalance creates severe bottlenecks, particularly for `INSTRUMENTATION` (1:91 ratio) and `NOT_TELESCOPE` (1:9 ratio).

Threshold Tuning Procedure The custom decision thresholds (Strategy 4) were determined by performing a manual grid search over the probability outputs generated by the trained meta-learners on a held-out validation set (20% of the training data). For each minority class (`INSTRUMENTATION` and `NOT_TELESCOPE`), we evaluated thresholds ranging from 0.1 to 0.9 in steps of 0.01. The threshold that yielded the maximum F1-score on the validation set for that specific label was selected as the optimal threshold for generating final predictions on the test set.

Calibration Timing Probability calibration (Strategy 5) was applied **after** the XGBoost meta-learners were fully trained using the optimized hyperparameters (including aggressive class weights). The 'CalibratedClassifierCV' wrapper from scikit-learn was fitted using Isotonic Regression on the out-of-fold predictions from the same validation set used for threshold tuning. This post-hoc calibration step adjusts the output probabilities of the already trained models before the final optimized thresholds (determined in Strategy 4) are applied. This ensures the thresholds operate on more reliable probability estimates, improving reproducibility.

#	Strategy	Target Label(s)	Mechanism & Rationale
1	AstroBERT Fine-Tuning	All (via base model)	Unfreezing weights and training for 3 epochs adapts embeddings to the specific task, improving feature quality for the meta-learner.
2	SMOTE Oversampling (Chawla et al., 2002)	`INSTRUMENTATION`	Generates synthetic minority samples (k=5) to balance the training data to 1:1 for the meta-learner, providing more examples for this extremely rare class (1:91 ratio).
3	Aggressive Class Weighting	`INSTRUMENTATION`, `NOT_TELESCOPE`	Manually increases XGBoost's `scale_pos_weight` (180 for INSTR, 15 for NOT_TEL) beyond the theoretical ratio to heavily penalize misclassifications of minority classes, boosting recall.
4	Custom Prediction Thresholds	`INSTRUMENTATION`, `NOT_TELESCOPE`	Lowers the decision threshold (0.35 for INSTR, 0.40 for NOT_TEL) from the default 0.5 to optimize the F1-score by improving recall at an acceptable precision cost for imbalanced classes.
5	Probability Calibration	`INSTRUMENTATION`, `NOT_TELESCOPE`	Applies Isotonic Regression post-hoc to the meta-learner outputs to make predicted probabilities more reliable, enhancing the effectiveness of custom thresholds.

Table 2: Summary of optimization strategies applied to improve Task 2 performance.

6 Results

This section details the performance of our final systems for both subtasks, comparing final metrics against developmental stages and discussing the implications.

6.1 Subtask 1: Telescope Identification Results

Our iterative development process for Task 1 culminated in a hybrid model combining a feature-based RandomForest classifier with a rule-based heuristic leveraging the `bibcode` field (Model 5 in Section 4). As summarized in Table 3, this final approach achieved significantly higher performance than models relying solely on semantic or purely feature-based methods.

Model	Approach	Accuracy	F1	Recall
Model 1	Stacked LSTM	78%	75%	77%
Model 2	AstroBERT	79%	76%	78%
Model 3	Logistic Reg. + AstroBERT	82%	80%	81%
Model 4	Feature-based RandomForest	80%	78%	79%
Model 5	RandomForest + Rule-Based	**97%**	**96.8%**	**97.1%**

Table 3: Performance evolution across five model iterations for Subtask 1 (Telescope Identification).

The dramatic improvement from incorporating the rule-based correction underscores the importance of domain-specific structural features, which provided deterministic cues unavailable in the raw text or other metadata. Neural models struggled particularly with the `None` class, highlighting the limitations of purely semantic approaches for this specific task.

6.2 Subtask 2: Telescope Role Classification Results

Our final optimized stacking ensemble system, detailed in Section 5, achieved a locally validated **Macro F1-score of 0.683** for the Telescope Role Classification task, a notable enhancement from the baseline of 0.6191. This improvement was primarily driven by successfully mitigating the severe class imbalance affecting minority classes. Specifically, for the **`INSTRUMENTATION`** class, a combination of targeted strategies including SMOTE-based data augmentation, aggressive class weighting in XGBoost (e.g., `scale_pos_weight=180`), lowered custom decision thresholds (e.g., 0.35), and probability calibration via Isotonic Regression proved highly effective. These techniques collectively forced the model to better recognize the rare class instances by adjusting data representation, learning penalties, and decision boundaries, leading to a substantial increase in its F1-score from 0.510 to **0.782**. Importantly, these optimizations for minority classes maintained stable performance on the majority classes (`MENTION` and `SCIENCE`), demonstrating the robustness of our per-class approach.

Label	Precision	Recall	F1-Score	Support
`NOT_TELESCOPE`	0.788	0.344	0.479	187
`MENTION`	0.677	0.747	0.710	892
`INSTRUMENTATION`	0.901	0.690	**0.782**	132
`SCIENCE`	0.757	0.764	0.760	2156
Macro Avg (Local)	0.781	0.636	**0.683**	3367

Table 4: Final per-class performance for Task 2 based on local evaluation. The Macro F1-score computed locally is 0.683.

Leaderboard Results Our system achieved a combined Macro F1-score of **0.82** on the TRACS @ WASP 2025 competition leaderboard, securing **4th place**. This score represents the weighted combination of Task 1 (telescope identification, 97% accuracy) and Task 2 (role classification). For Task 2 specifically, our local validation achieved a Macro F1 of 0.683 across the four role labels. The per-class improvements (5) reported in the following analysis reflect the impact of our optimization strategies on each label performance during local evaluation. Model Performance and Analysis for both the share task is shown in Appendix A.

7 Discussion

7.1 Ensemble Synergy

Our results validate the complementary nature of symbolic and semantic models. The keyword classifier provides high-recall signals for explicit telescope mentions, excelling when documents contain direct references. Conversely, astroBERT captures nuanced semantic patterns, capturing context-dependent telescope roles. The meta-learner learns to weight these signals appropriately:

- For `INSTRUMENTATION`: Documents often lack explicit instrumentation keywords, making astroBERT's semantic understanding crucial.
- For `SCIENCE`: High keyword density provides strong signals, but astroBERT refinement reduces false positives.

A potential concern is whether relying on specific keywords might lead to overfitting, particularly if the lexicon is highly tuned to the training data. While the domain-adapted astroBERT component provides broader semantic understanding that can mitigate this, its performance might also degrade if future documents use entirely novel terminology not seen during pre-training or fine-tuning. Careful curation and potential expansion of the keyword list would be necessary for optimal generalization.

7.2 Why Per-Label Meta-Learners?

Our choice of four independent XGBoost meta-learners (rather than a single multi-label model) proved critical for handling extreme imbalance. This design enables: **(1) Fine-grained hyperparameter tuning**: Each label can employ `scale_pos_weight` values matched to its specific imbalance ratio. **(2) Selective data augmentation**: SMOTE is applied only to `INSTRUMENTATION`, avoiding artificial data generation for other classes. **(3) Flexible thresholding**: Different labels can employ different decision thresholds based on their precision-recall trade-off characteristics. **(4) Modular optimization**: New strategies can be tested for individual labels without affecting others.

7.3 Ensemble vs. End-to-End Transformers

While transformer models might seem like a simpler alternative, our ensemble approach offers advantages for this task. Firstly, **interpretability** is enhanced; we can analyze the relative contributions of the keyword (symbolic) and astroBERT (semantic) base models, providing insights into why a classification was made. Secondly, the **modularity** allows for easier updates—the keyword lexicon can be expanded or astroBERT replaced without retraining the entire system. Lastly, the per-label meta-learners provide **targeted robustness** against class imbalance, enabling specific, aggressive optimization strategies for minority classes that might be difficult to implement effectively within a single, monolithic transformer architecture.

8 Conclusion

We presented a hybrid stacking ensemble for the TRACS@WASP 2025 shared task on astrophysical document classification. Our system combines rule-based keyword detection with domain-adapted semantic modeling (astroBERT), using four independent XGBoost meta-learners—one per output label—to handle severe class imbalance through per-label optimization. The modular design enables targeted strategies, e.g., SMOTE augmentation, aggressive class weighting, calibrated probabilities, and custom decision thresholds, proving particularly effective for challenging minority classes.

We achieved a macro F1-score of 0.82 on the leaderboard, securing 4th place. The most significant improvements were realized in the extreme-minority classes: the F1-score for `INSTRUMENTATION` dramatically increased from 0.510 to 0.782 (+53.3%), and notable gains were also achieved for the difficult `NOT_TELESCOPE` label, showcasing the system's strength in high-imbalance scenarios without sacrificing majority class performance. We demonstrate that symbolic and neural approaches are complementary—their synergy is essential for specialized, imbalanced scientific corpora.

9 Limitations and Future Work

While our ensemble approach demonstrates strong performance, several key limitations warrant discussion and guide future research directions.

First, **handling long documents** remains a significant challenge. Our current reliance on astroBERT with a 512-token limit necessitates truncating lengthy astrophysical papers, potentially discarding crucial contextual information located later in the text. Future work should explore architectures specifically designed for long sequences, such as hierarchical attention models or transformers like Longformer (Beltagy et al., 2020), to capture document-wide context more effectively.

Second, the system's performance on the NOT_TELESCOPE class plateaued despite targeted optimization efforts. This suggests that the current feature representations derived from the keyword classifier and astroBERT lack sufficient discriminative power for this nuanced category. Addressing this could involve model-centric approaches like incorporating specialized external models or data-centric improvements such as creating finer-grained annotations for partial or non-primary telescope mentions, potentially leveraging weak supervision techniques to augment training data.

Third, **generalization beyond the TRACS dataset**, particularly to unseen telescopes, requires further investigation. Our system is optimized for the specific telescopes present in the training data (CHANDRA, HST, JWST). While astroBERT offers general domain knowledge, the keyword classifier's effectiveness heavily depends on its lexicon. Future efforts must focus on evaluating performance degradation on diverse astronomical corpora and developing robust strategies for rapid lexicon expansion and adaptation to ensure broader applicability.

While other limitations exist, such as the need for more detailed error analysis, addressing these three core areas—long document processing, minority class feature representation, and generalization—offers the most promising avenues for advancing the system's capabilities.

Future work will focus on addressing truncated context handling, building upon the significant gains achieved for the NOT_TELESCOPE class to further enhance its classification accuracy, and improving cross-domain generalization through hierarchical models and long-document transformers. Our framework provides a robust solution for scientific document classification in high-imbalance regimes, with applications extending beyond astrophysics.

Acknowledgments

We thank SVNIT for the resources (HPC-GPU access) provided to carry out our experiments.

We extend our sincere gratitude to Prof. Pruthwik Mishra for his invaluable guidance, constructive feedback, and mentorship throughout the development of this system.

References

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.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

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.

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*, KDD '16, pages 785–794. ACM.

Charles Elkan. 2001. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978.

Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 2011. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484.

Mikel Galar, Alberto Fernandez, Edurne Barrenechea, and Francisco Herrera. 2012. Ensemble methods for class imbalance learning. In Haibo He and Yunqian Ma, editors, *Imbalanced Learning: Foundations, Algorithms, and Applications*, pages 61–82. Wiley-IEEE Press.

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, Alberto Accomazzi, Michael J. Kurtz, Edwin Henneken, Mary Dempsey, Donna Thompson, Jonathan Luker, and Golnaz Shapurian. 2021. Building astroBERT, a language model for Astronomy & Astrophysics. In *Astronomical Data Analysis Software and Systems (ADASS) XXXI*.

Charmgil Hong, Rumi Ghosh, and Soundar Srinivasan. 2016. Dealing with class imbalance using thresholding. *arXiv preprint arXiv:1607.02705*.

Kaggle. 2025. Tracs @ wasp 2025: Telescope reference and astronomy categorization shared task. `https://www.kaggle.com/competitions/tracs-wasp-2025`. Accessed: 2025.

Azal Ahmad Khan, Omkar Chaudhari, and Rohitash Chandra. 2023. A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation. *arXiv preprint arXiv:2304.02858*.

Miroslav Kubat and Stan Matwin. 1997. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186.

Dong-Gyu Lee and Hwanjo Kim. 2020. Dynamic cost sensitive learning for imbalanced text classification. In *Proceedings of the KIISE Transactions on Computing Practices*, volume 26, pages 211–216.

Hanung Adi Nugroho, Endang Wiji Tias, Budi Widyakusumah, and Indra Waspada. 2023. A stacking ensemble model with smote for improved imbalanced classification on credit data. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 21(4):873–881.

SAO/NASA Astrophysics Data System. 2025. The sao/nasa astrophysics data system. `http://ui.adsabs.harvard.edu/`. Accessed: 2025.

Wikipedia contributors. 2025. Observational astronomy — Wikipedia, the free encyclopedia. [Online; accessed 20-October-2025].

Qihua Zou, Jihua Yu, Yu Zhang, and Chang Liu. 2016. Finding the best classification threshold in imbalanced classification. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 217–224. IEEE.

A Model Performance and Analysis

A.1 Task 1 Performance (Telescope Identification)

The performance of the Task 1 Random Forest model was primarily assessed through 5-fold stratified cross-validation during the hyperparameter tuning phase using `RandomizedSearchCV`.

Cross-Validation Results: The optimized model configuration, selected based on the tuning process described in the System Architecture section, achieved a mean cross-validation accuracy of **0.7772**. The standard deviation across the folds was relatively small (around ± 0.0004 according to similar runs shown in the notebook context), indicating consistent performance across different data subsets. This high accuracy suggests the model effectively minimizes confusion between the primary telescope classes (`CHANDRA`, `HST`, `JWST`) while mitigating the impact of the extreme imbalance posed by the `NONE` class, largely due to the strong predictive power of the ID-based features.

Feature Importance: As noted previously, feature importance analysis consistently highlighted the overwhelming predictive power of features derived directly from the `Id` string's suffix. This confirms that the rule-based extraction component, integrated as a feature, provides the primary signal for this classification task. Metadata features like `year` and certain TF-IDF terms offered minor contributions.

Final Prediction: The final model, trained implicitly on the full dataset using the best parameters from `RandomizedSearchCV`, was used to generate predictions for the submission file ('final_submission_task1.csv'). While detailed per-class metrics (precision, recall, F1) were not part of the hyperparameter search output, the strong cross-validation accuracy suggests effective classification, heavily driven by the identifier-based features.

A.2 Task 2 Performance Analysis (Document Role Classification)

Per-Class Performance Analysis Table 5 highlights the change in F1-score for each class from the baseline (Table 1) to the final optimized model (Table 4).

Label	Baseline F1	Final F1	Improvement (Δ F1)
`NOT_TELESCOPE`	0.480	0.479	−0.001
`MENTION`	0.722	0.710	−0.012
`INSTRUMENTATION`	0.510	**0.782**	**+0.272**
`SCIENCE`	0.765	0.760	−0.005
Macro Avg	0.619	**0.683**	**+0.064**

Table 5: Comparison of F1-scores before and after optimization for Task 2, highlighting the substantial gain for the `INSTRUMENTATION` class.

The most dramatic success was in the **INSTRUMENTATION** class, which saw its F1-score jump from 0.510 to **0.782** (+0.272, a +53.3% relative improvement). This validates our targeted optimization strategy that combines SMOTE, aggressive class weighting (weight=180), a low decision threshold (0.35) and probability calibration. Precision improved to 0.901 while recall increased significantly from 0.420 to 0.690. Conversely, the **NOT_TELESCOPE** class proved resistant to optimization, with its F1 remaining static (0.480 $\rightarrow$ 0.479). Despite targeted weighting (weight=15) and thresholding (0.40), the model maintained high precision (0.788) but low recall (0.344), suggesting insufficient characteristic discrimination from the base models for this specific class.

The majority classes, **MENTION** and **SCIENCE**, showed minimal F1 change, indicating that optimizations targeting the minority classes did not negatively impact their performance.

Statistical Reliability The presented results are based on a single training run with a fixed random seed for reproducibility. Averaging results over multiple runs with different seeds could provide a more robust estimate of performance variance but was not performed due to computational constraints.

`amc`: The Automated Mission Classifier for Telescope Bibliographies

John F. Wu* **Joshua E. G. Peek** **Sophie J. Miller**
Jenny Novacescu **Achu J. Usha** **Christopher A. Wilkinson**
Space Telescope Science Institute
3700 San Martin Dr
Baltimore, MD 21218 USA
*`jfwu@stsci.edu`

Abstract

Telescope bibliographies record the pulse of astronomy research by capturing publication statistics and citation metrics for telescope facilities. Robust and scalable bibliographies ensure that we can measure the scientific impact of our facilities and archives. However, the growing rate of publications threatens to outpace our ability to manually label astronomical literature. We therefore present the Automated Mission Classifier (`amc`), a tool that uses large language models (LLMs) to identify and categorize telescope references by processing large quantities of paper text. A modified version of `amc` performs well on the TRACS Kaggle challenge, achieving a macro F_1 score of 0.84 on the held-out test set. `amc` is valuable for other telescopes beyond TRACS; we developed the initial software for identifying papers that featured scientific results by NASA missions. Additionally, we investigate how `amc` can also be used to interrogate historical datasets and surface potential label errors. Our work demonstrates that LLM-based applications offer powerful and scalable assistance for library sciences.

1 Introduction

Telescope bibliographies provide one way to measure the scientific productivity of our astronomical facilities. Bibliometrics can quantify how often telescopes are discussed in scientific publications, e.g., through passing mentions or via detailed scientific analyses that originate from data taken by each telescope. Although these quantitative analyses are vital for assessing the impact of our scientific investments, they hinge on complete, homogeneous bibliographies, which can be expensive and onerous to manually curate. Librarians, archive scientists, and bibliographers maintain telescope bibliographies by consistently tracking publications, extracting metadata, and labeling the scientific *intent* of each telescope reference for all papers (see, e.g., Lagerstrom, 2015; Observatory Bibliographers Collaboration et al., 2024). Complete observatory bibliographies enable us to investigate publication rates, and citation statistics, links between publications and observing proposals, data product usage metrics, and archival science impact (e.g., Apai et al., 2010, for HST).

There is more scientific literature than ever before (notwithstanding gender-disparate impacts from the recent pandemic, Böhm and Liu, 2023). Some of this increase accompanies a general rise of publication rates throughout academia (Hanson et al., 2024). Additionally, very recent growth in publication rates may stem from the advent of large language models (LLMs), which can lower the barrier to writing papers (e.g., Astarita et al., 2024). These trends suggest that we need a sustainable solution for producing telescope bibliographies amid the deluge of astronomy papers.

LLMs can also be useful for compiling telescope bibliographies at scale: artificial intelligence (AI) systems are highly scalable, and are now adept at processing large amounts of text inputs. Modern LLMs can complete many tasks *without any optimization*, instead relying solely on emergent capabilities like in-context learning (e.g., Brown et al., 2020). With frontier AI labs now deploying LLMs as a service, we can easily leverage simple API (Application Programming Interface) calls and design software around cutting-edge LLMs.

Before deploying an automated bibliography system, we must first ensure that its performance is *robust*. To this end, we present and evaluate the Automated Mission Classifier (`amc`), an LLM-powered, bibliometric tool for identifying telescopes or NASA missions in the literature. We adapt `amc` for a specific shared task, TRACS (Section 2); in the Appendix, we note that similar systems are already in operations for JWST (Appendix B) and can be used for archival science with other telescopes (Appendix C). In Section 3, we describe the software's system design,

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 157–169

and we present results in Section 4. In Section 5, we discuss how observatory bibliographers can leverage AI to compute bibliometrics at scale, assess (historical) data quality, and upgrade the LLM systems. We provide publicly available code on Github: `https://github.com/jwuphysics/automated-mission-classifier`.

> "To LLMs! The cause of, and solution to, all of bibliographers' problems."[1]

2 The TRACS Shared Task

The Telescope Reference and Astronomy Categorization Shared task (TRACS) is a data challenge organized as part of the 2025 Workshop for Artificial Intelligence for Scientific Publications (WASP; Grezes et al. 2025) at IJCNLP-AACL.[2] The task consists of classifying astronomy papers into at least one of four categories: `science`, `instrumentation`, `mention`, or `not_telescope`.

In this data challenge, papers are decomposed into several fields (including the title, abstract, and "body" full text) and, based on keyword filtering, labeled with a candidate telescope name (`CHANDRA`, `HST`, `JWST`, or `NONE`). The objective is to predict the boolean labels for all paper categories for each of the provided bibcode + telescope combinations. However, it is important to note that the candidate telescope name may be mislabeled, and that certain paper categories impose constraints on the others (i.e., a single paper + telescope can have True labels for both `science` and `instrumentation`, but cannot for both `science` and `not_telescope`).

Training and test data sets, in CSV format with 80,385 and 9,194 entries respectively, are provided for the shared task. To participate in the challenge, entrants must submit 9,194 test-set predictions via Kaggle[3] and have their predictions evaluated. The test outputs are scored according to the average between the macro F_1 score of the telescope labels and the macro F_1 score of paper labels; each class is weighted equally. Note that `NONE` is a valid telescope class and `not_telescope` is a valid paper class. In the subsections below, we note some details that we considered important for our submission.

[1]Quote adapted from *The Simpsons* (Swartzwelder and Anderson, 1997).

[2]For details on The International Joint Conference on Natural Language Processing & Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL), see `https://2025.aaclnet.org/`.

[3]`https://www.kaggle.com/competitions/tracs-wasp-2025`

2.1 Input Data

The full list of columns in the train data set include: (0) ID, (1) bibcode, (2) telescope, (3) author, (4) year, (5) title, (6) abstract, (7) body, (8) acknowledgments, and (9) grants, (10) `science`, (11) `instrumentation`, (12) `mention`, and (13) `not_telescope`. The test dataset does not include column (2) or columns (10) through (13). However, a preliminary telescope label *is* implicitly named in column (0), as the ID is simply the concatenation of `{bibcode}_{telescope}`.

Some rows in the datasets are missing: 3% of the test data set is missing an abstract, 19% does not have full-body text, and $> 90\%$ does not have text under the grants column. Incomplete data are likely due to a combination of parsing errors (e.g., correctly parsing out grants/acknowledgments) and publisher restrictions. Issues with publisher agreements tend to impact certain journals or publication venues (i.e., demarcated by their "bibstem" entries); in many of these cases, the body text is completely absent. Nonetheless, classifications can sometimes still be made on the basis of just the title and abstract (but see Appendix C).

Some of the input data may not be helpful. For example, the list of authors is unlikely to yield useful indicators of the paper classification, and may even produce false positives, as "Webb" or "Chandra" can show up as (sub-word) names of authors. Likewise, "Hubble" can often show up in the acknowledgments, e.g. due to funding acknowledgments from the NASA Hubble Fellowship Program. Thus, it is imperative to design a language modeling system that can flexibly understand the context surrounding telescope detections.

2.2 Paper Types

Establishing a common definition for paper types is a nontrivial task. When tasking human bibliographers to classify papers, e.g., identify `science` papers, disagreements often arise about the precise definition of a `science` paper.

In order to implement a useful LLM system for automated classification, it is necessary to unambiguously define the labels. Observatory Bibliographers Collaboration et al. (2024) issue the following guidance on `science` papers:

> "To qualify as a science paper, it must be apparent that data or data product(s) from the observatory were used and that the

data or data product(s) formed the basis for reaching a new scientific conclusion."

The authors recognize that these definitions must be continually updated.

Indeed, the taxonomy should serve the telescope or mission. Existing schemes may not be sufficient to characterize all of the edge cases, and new categories may arise. As a concrete example, STScI established a *data-influenced* category in 2019 for papers that indirectly rely on data or products, but do not directly analyze data or use data products. In the TRACS challenge taxonomy, data-influenced papers would generally be labeled under the `mention` category.

As part of the shared task, the TRACS website provided a narrative format description of the different paper types (Grezes et al., 2025). We used an LLM (`claude-sonnet-4.1`) to process this text in order to create a user prompt that includes definitions and examples of each paper type (which is manually updated and described in more detail in Section 3). The full prompts can be found in the Github repository, and we have copied the paper type definitions here (note that we remove the markdown text formatting for human readability):

- `science`: Paper directly uses {{telescope}} data (new or published) to obtain new scientific results in this paper.
- `instrumentation`: Paper describes new instrument science or engineering.
- `mention`: Paper references the telescope but does not produce new scientific results.
- `not_telescope`: Paper includes references that are false positives – names that look like the telescope but refer to something completely different.

3 The Automated Mission Classifier (`amc`)

Figure 1 shows a high-level overview of the `amc` system. The system classifies a single paper and a single telescope at a time.

First, `amc` performs a keyword search to filter all mentions of telescope-related keywords, and we include surrounding context (± 3 sentences). This step effectively converts the body into a list of telescope-specific text snippets (Section 3.1). Text snippets are then ranked by their relevance to the core question of "is this a {telescope} science paper?", and we only keep the most relevant snippets (Section 3.2). These top-ranked snippets are subsequently passed to an LLM, which is prompted to classify the paper types and provide quotes and supporting reasoning for its predictions (Section 3.3). The specialized code used for TRACS is forked from `amc` and can be found at `https://github.com/jwuphysics/tracs_wasp2025`.

Finally, we note that our LLM system design is strongly influenced by a prior task: classifying whether arXiv paper preprints contained JWST science. In Appendix B, we describe how these earlier motivations shaped (and biased) the design of the `amc`. Additional discussion of the limitations of `amc` are discussed in Section 4.4.

3.1 Keyword Filtering on Full Text

We concatenate the title, abstract, and body as a single text input. We extract only the most relevant portions of the text by searching for keywords. First, we divide text into sentences by using the Punkt sentence tokenizer (Kiss and Strunk, 2006) in the NLTK package (Bird and Loper, 2004). We then use a simple Python case-insensitive string search to identify sentences with keywords for the relevant telescope. We expand snippets to include the $n = 3$ prior and following sentences (i.e., such that each snippet contains $2n + 1$ sentences). If no keywords are found, then we automatically classify the paper as `not_telescope`.

We note that our keywords prioritize high recall at the expense of low precision; in other words, we value keyword completeness to make sure that no important keywords are missed. However, this means that false positives are expected. For example, our simple string matching over "COS" (an instrument on the Hubble Space Telescope) will also trigger matches on the words "cosmic" or "cosine." Therefore, it is essential that we guard against false positives by ranking text snippets according to their relevance.

3.2 Reranking Excerpts

Rerankers are typically LLMs that determine the relevance of some text snippet for answering a specific question. In information retrieval systems or retrieval-augmented generation (RAG), a first-stage algorithm usually produces an initial ranking or filtering over relevant documents/snippets (e.g. via semantic similarity in an embedding space). Rerankers provide a second-stage ranking between the query and a smaller set of snippets; recent works have demonstrated them valuable for LLM

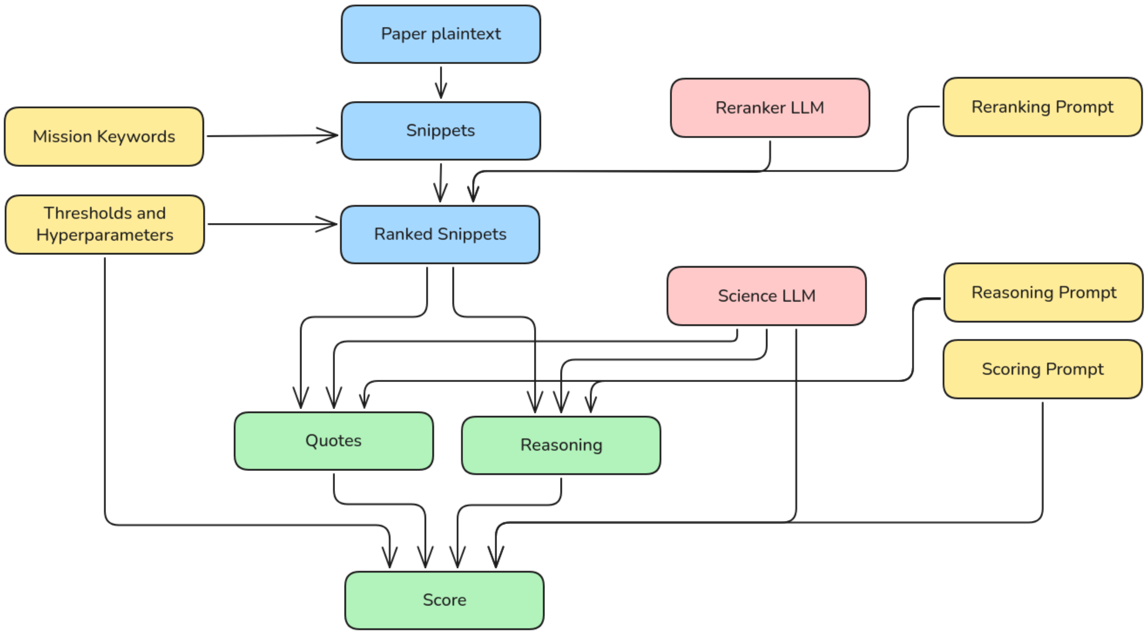


Figure 1: Schematic showing the system design for amc. Note that the version of amc adapted for TRACS does not separate the LLM generation of quotes, reasoning, and predictions; rather they are all output together. See the text in Section 3 and Appendix B for a full discussion of differences.

systems in astronomy and science more broadly (Iyer et al., 2024; Chen et al., 2025; Xu et al., 2025).

We implement a custom reranker solution[4] that achieves similar performance to leading commercial products (e.g., Cohere Reranker v3.5; based on a few informal assessments). We use a lightweight, general-purpose, non-thinking model (`gpt-4.1-nano`) with a restricted vocabulary ("Yes" and "No") that outputs logits (log-probabilities) between 0 and 1. Using a short reranker prompt, we task this model to identify whether each individual snippet discusses the telescope in a way that may be used to classify the paper type. One of the main goals of this step is to remove accidental and unrelated keyword matches.

Once every snippet has a reranker score, we can sort them and/or filter out irrelevant snippets. We keep up to $k = 15$ top-ranked snippets in order to reduce the amount of text that is sent to the next LLM call.

3.3 LLM Classifications

We combine the filtered text snippets together along with their reranker scores. The scores can serve as another reference for whether snippets are useful for determining the paper type.

We use `gpt-5-mini` with minimal reasoning effort to make the final classification as a structured output. The LLM prompt contains the top-ranked snippets and their scores, and it defines the different paper types and provides some examples. In addition to predicting boolean classes for the `science`, `instrumentation`, `mention`, and `not_telescope` paper types, the LLM is also prompted to supply the most relevant quotes and justify its reasoning. All structured outputs and their data types are constrained via a `pydantic` model schema (e.g., boolean predictions, a list of strings for the quotes, and a single text string for the reasoning).

In the amc package, the quotes and justification are provided first, followed by a separate LLM call to predict the final score on these lines of thinking (see, e.g., Figure 1). However, because TRACS requires multiple classifications, we simplify the system so that all predictions and quotes/reasoning are output at the same time. The original amc also supports floating point values between 0 and 1 for scoring `science` paper types, which allows another hyperparameter to control the threshold for scoring science papers. For TRACS, we simplified the system by using boolean values for each prediction.

[4]We were unaware (until the time of writing) that this reranking approach had been proposed in the literature before (see e.g. Liang et al., 2023).

4 Results

We briefly present some limited results on the TRACS test set. Our best score in terms of F_1 is 0.84 on the held-out test set, enough for a third-place rank according to the Kaggle leaderboard. In Appendix A, we show the `amc` JSON-formatted outputs, including paper type examples for `science` (Listing 1), `instrumentation` (Listing 2), `mention` (Listing 3), and `not_telescope` (Listing 4). Based on a cursory review, these outputs seem accurate, the quotes do not suffer from hallucinations (although the risk is still present), and the provided reasoning largely appears to be faithful to its classification.

4.1 Evaluating amc

In order to understand our system's strengths and weaknesses, we select $N = 100$ random entries from the training set, comprising 25 rows per telescope. This small, non-representative evaluation set enables us to investigate why our LLM system tended to make incorrect predictions. This random set is also able to surface potential issues with the dataset (see Section 4.3).

In Figure 2, we show confusion matrices displaying `amc` predictions on the limited validation set, for all telescopes except `NONE`. Each column shows a paper type (denoted "True") against all other paper types (denoted "False"). We note that some combinations of missions and paper types tend to succeed (e.g., `CHANDRA/science`) or fail more frequently (e.g., `CHANDRA/mention`). These confusion matrices are based on the same version of `amc` as the final TRACS submission. However, we caution against overinterpreting results on this relatively small evaluation set.

4.2 Performance on TRACS

Our first submission to TRACS achieved a macro F_1 score of 0.80. At the time, the system included a few suboptimal settings, e.g., slightly misspecified prompts, or a non-zero reranker threshold which caused weak `mention` classes to occasionally be mislabeled (since the threshold might cause all text snippets to be filtered out, rendering a default verdict of `not_telescope`).

After removing the reranker threshold and updating the prompts, we saw a modest increase in macro F_1 score to 0.84. We examined two of our higher-scoring sets of predictions, and used an LLM as a judge (`gpt-5-mini`) to resolve discrepancies between them and to issue final predictions; the performance remained at $F_1 = 0.84$.

The final LLM system took less than 24 hours in wall-clock time to run, and incurred roughly \$10 in OpenAI costs. About 22% of the cost is for reranking snippets with `gpt-4.1-nano`, 37% is for processing top-ranked inputs with `gpt-5-mini`, and 41% is for generating outputs with `gpt-5-mini`). Batch processing could lower some costs, but would necessitate an asynchronous pipeline, where we first perform all reranker calls, followed by all LLM classifications.

4.3 Missing Data and Label Errors

The dataset likely contains errors or uncertain classifications due to the imperfect nature of manually annotating bibliographic data, and the somewhat subjective nature of label distinctions. However, it is not possible to capture this uncertainty in the discrete classes. We also cannot measure the error rate directly, as there is no *golden sample* against which we can compare. A golden sample would consist of papers that have been independently classified by multiple reviewers, where cases of disagreement are subject to deliberation and re-review until consensus is reached. Therefore, the error rate or uncertainty is unknown.

Through repeated evaluation, we can surface potential errors in the TRACS dataset. While testing our LLM system on a small ($N = 100$) subsample from the training data, we inspected all cases where the LLM prediction disagreed with the target label. Some of these appeared to be genuine error or ambiguity in the ground truth dataset, and we display them in Table 1.

We find that one paper is labeled as both `science` and `mention`, which (we assume) should not be possible. This classification may have resulted from human annotation error, or perhaps an accidental combination of a HST `mention` (as the paper is about the Hubble Deep Field) and CHANDRA `science`. Simultaneously conflicting labels like this can be easily filtered out by using boolean logic and some set rules. We find another paper that mentions the "Next Generation Space Telescope," the original name for JWST. Arguably, this paper should be considered a JWST mention, but is instead labeled as `not_telescope`.

Three papers are missing their body text; they are only described by their titles, abstracts, and other metadata. For each of these three entries, we verify that (within the TRACS data) there is no

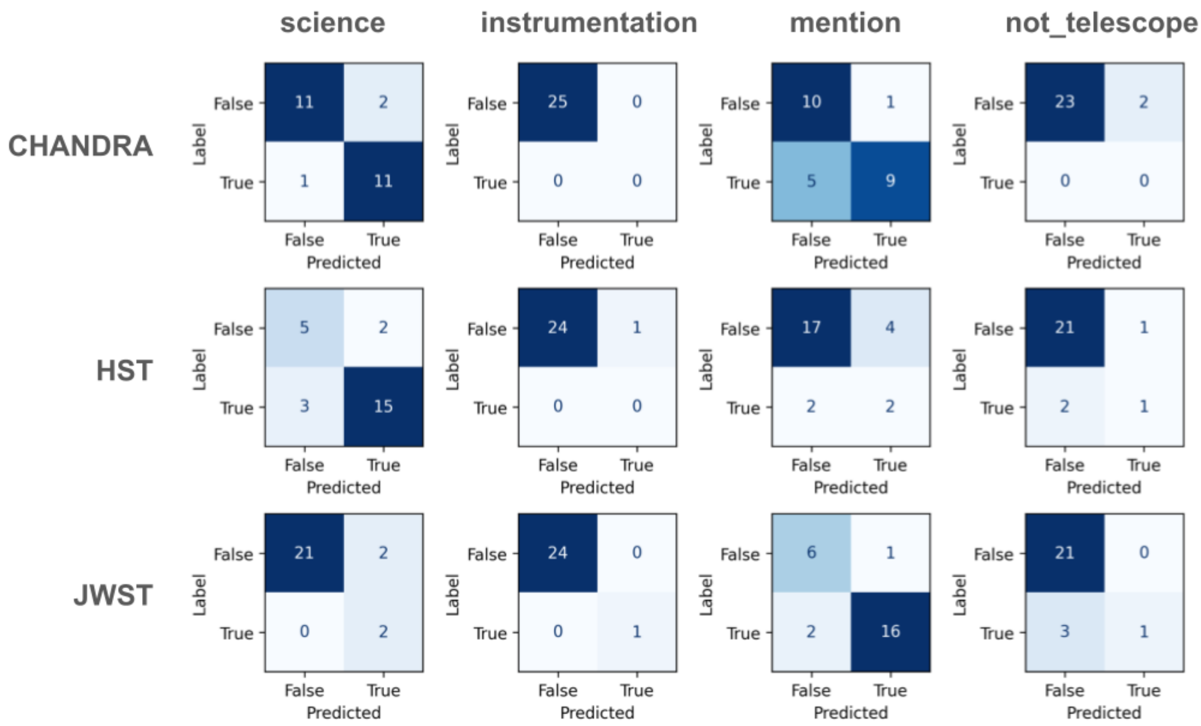


Figure 2: Binary confusion matrices, over a randomized subsample of 25 TRACS training set examples for each telescope (*rows*), shown for each paper type (*columns*) as a one-vs-all classification. Each confusion matrix shows true negatives (*top left*), false positives (top right), false negatives (*bottom left*), and true positives (bottom right).

Table 1: Potential dataset issues found in a random selection of 100 labeled examples.

Bibcode	**Notes**
`2001AJ....122..598D`	Labeled as both `CHANDRA/science` and `CHANDRA/mention`.
`2001ApJ...550..104Y`	Missing body text, making it impossible to correctly classify as `HST/science`.
`2002IJMPA..17.3446T`	Mentions "Next Generation Space Telescope" (the placeholder name for JWST), but the label is `JWST/not_telescope` rather than `JWST/mention`.
`2004RMxAC..20..215S`	Missing body text, making it impossible to correctly classify as `CHANDRA/science`.
`2004fxra.book...89D`	Missing body text, making it impossible to correctly classify as `CHANDRA/science`.

mention or science/instrumentation of the candidate telescope presented. Missing full body text is often a symptom of complex publisher licensing agreements, and it may not always be possible to procure the full data. In any event, such entries do not contain sufficient data for making accurate predictions.

4.4 Limitations of amc

As noted above, the amc system is designed to be general. Although we have specialized the code for the TRACS task, there are additional adjustments that could lead to improved performance. For example, the multiclass predictions would benefit from dedicated prompts for each paper type. The current system effectively uses the same prompts for each telescope, which might also limit its performance.

We also note that amc is at the mercy of our keyword filtering. If we miss any telescope keywords, then it is possible to filter out relevant snippets, which could jeopardize the prediction task performance. Frequent keywords could be empirically learned using traditional NLP techniques like term frequency (TF; Spark Jones 1972) normalized by its document frequency (i.e., TF-IDF; Salton et al. 1975. The reranker step could potentially be replaced by a simple first-pass classifier using TF-IDF or another data-driven approach.

5 Discussion

LLMs are becoming pervasive throughout astronomy. Quantitative benchmarks (Carrit Delgado Pinheiro et al., 2025; Joseph et al., 2025; Ting et al., 2025) and human-centered studies (Fouesneau et al., 2024; Wu et al., 2024; Hyk et al., 2025) deliver complementary evaluations for how to successfully deploy LLMs for real-world benefit in astronomy. There is also rapid adoption of LLMs for navigating through and interacting with the astronomy literature (Ciucă and Ting, 2023; Iyer et al., 2024), which is particularly salient for WASP/TRACS.

As researchers are in the midst of a fundamental shift of how they interact with literature, we discuss a future vision of how the astronomical community may leverage LLMs to augment or automate bibliographies (Section 5.1), how AI systems can assist in evaluating or improving our ground truth datasets (Section 5.2), and how the `amc` software we presented could be improved further in future work (Section 5.3).

5.1 Scalable, AI-Supported Bibliographies

We have shown that compiling telescope bibliographies can be assisted by or partially automated with LLMs. LLM developments are built on traditional NLP techniques, which have already been vital for astronomical literature review (Iyer et al., 2024) and detecting usage of telescopes/facilities (e.g., using TF-IDF, Amado Olivo et al., 2025). While LLMs can be more expensive to put into production relative to simple NLP techniques or specialized fine-tuned models (e.g., SciBERT, Beltagy et al., 2019), LLMs that have been pre-trained on trillions of tokens of general text are also capable of in-context learning via zero- or few-shot demonstrations (Radford et al., 2019; Brown et al., 2020). Modern LLMs also have longer context windows, enabling them to ingest multiple text snippets (or even entire documents at a time). This feature is particularly valuable if the telescope classification depends on nuanced text snippets buried within the body (i.e., often the case for *archival* data sets, and rare for *flagship* NASA missions; see Appendix C).

AI systems can still be extremely useful even if manual vetting of bibliographies is necessary. We have designed `amc` to have high recall, so it can confidently remove from consideration papers that have no chance of being `mention` paper types. Accurate labels ($F_1 > 0.8$) can dramatically save human time and mental energy.

5.2 Errors and Ground Truths

When creating LLM-augmented bibliographies at massive scale, it is imperative to understand how the LLM is susceptible to errors, and/or if those errors originate from the LLM or from the dataset. For TRACS, our analysis of a small subsample in Section 4.3 resulted in direct performance gains; we exposed some issues with our system, as well as errors in the dataset.

We emphasize the value in compiling a golden sample with consensus reviews, even if this dataset is much smaller compared to the archival set of (single-pass) human classifications. In our prior work (see Appendices B and C), we have relied on a golden sample with about $N \sim 100$ examples to serve as a benchmark for improving the LLM system (Shaw et al., *in prep*). Crucially, it also serves as a measure of *human performance*, which is often incorrectly assumed to be perfect. By setting human error rates as the error "floor," we can quantify a goal for LLMs to achieve.

AI augmentation can also facilitate a better understanding of our datasets. For example, LLMs can easily comb through a large number of negative classes from *historical* datasets, and surface candidate missing papers or other errors (e.g., Section 4.3). An LLM can be vital for efficiently constructing such a golden sample dataset.

5.3 Future Improvements

Our solution for the TRACS task can likely benefit from additional optimization. In particular, other LLMs can help iteratively optimize the prompts used to guide the (TRACS-specific) `amc` code, by using meta-optimizers (see, e.g., Opsahl-Ong et al., 2024; Agrawal et al., 2025) in a prompt compilation framework like DSPy (Khattab et al., 2023). Given the large TRACS training data set, meta-optimization could be costly, and may be precariously sensitive to the training label quality. However, meta-optimization could also produce (as a byproduct) empirical definitions of paper types like `science` or `instrumentation`, which could be valuable for comparing against explicit definitions that bibliographers have historically adopted.

Another option is to use AI agents: LLMs that can call tools in a loop in order to accomplish a task.[5] Even though an AI agent might access the

[5] For one definition of an AI "agent" that we like, see `https://simonwillison.net/2025/Sep/18/agents/`

same tools that we have described in Section 3, e.g., keyword search, reranking, filtering, or summarization, the LLM's *agency* means that it can decide when and how to use such tool calls. The LLM agent can also maintain a working memory, allowing it to determine whether it has enough information to make a classification; for instance, if it finds immediate evidence that the paper presents scientific results, then the agent can stop the analysis and classify the paper as `science`.

Finally, we may wish to deploy *smaller*, specialized models for this task because they can be run locally and perhaps at lower costs. For example, our keyword filtering and reranking steps are somewhat reminiscent of "late-interaction" retrieval mechanisms (e.g., ColBERT, Khattab and Zaharia, 2020), and it may be advantageous to substitute those steps with more lightweight model like ColBERT. We might simplify further by substituting this initial stage with classical NLP algorithms like TF-IDF. Models with specialized tokenizers for scientific literature like SciBERT (Beltagy et al., 2019) may also prove to be beneficial for parsing the astronomical literature.

6 Summary

We have presented `amc`, an LLM-based system that can automatically categorize real astronomical papers into specific labels. Using a specialized instance of `amc`, we demonstrate strong performance ($F_1 = 0.84$) and secure third place on the TRACS shared task (Grezes et al., 2025). Our tool is also valuable for evaluating labeled data quality, as it provides reasoning and supporting quotes to justify its predicted labels. Given the growing volume of papers, as well as the rising capabilities of LLMs, we believe that AI tools represent scalable solutions for accomplishing or assisting with this task.

In the future, however, LLMs may completely obviate the need for predefined "classifications" that comprise current paper types; instead, we may be able to *directly* ask LLMs questions like: "How many papers present ground-based follow-up observations for targets initially discovered with HST?" or "How did the fraction of Chandra *archival* science papers change between 2010 through 2025?" We envision that, by exploiting the capabilities of AI systems, library scientists can study a broader range of bibliographic questions than ever before.

Ethical Disclosure

All of this text was written solely by the authors. The document was partially reviewed by LLMs, primarily `gpt-5` and `Gemini 2.5 Pro`, in order to surface issues in clarity and prose. Some of the code in the associated repository is generated by LLMs, primarily via Claude Code. Having validated the software and results, the authors take full responsibility and ownership over the results presented here.

Acknowledgments

We acknowledge helpful conversations with STScI staff, including Brian Cherinka, Dick Shaw, and Jinmi Yoon. Some of the ideas about combining the LLM system with TF-IDF and SciBERT are based on exchanges with Madhusudhana Naidu after the end of the TRACS competition.

References

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2025. Gepa: Reflective prompt evolution can outperform reinforcement learning. *Preprint*, arXiv:2507.19457.

Vicente Amado Olivo, Wolfgang Kerzendorf, Brian Cherinka, Joshua V. Shields, Annie Didier, and Katharina von der Wense. 2025. Identifying Telescope Usage in Astrophysics Publications: A Machine Learning Framework for Institutional Research Management at Observatories. *The Astronomical Journal*, 169(1):42.

Dániel Apai, Jill Lagerstrom, Iain Neill Reid, Karen L. Levay, Elizabeth Fraser, Antonella Nota, and Edwin Henneken. 2010. Lessons from a High-Impact Observatory: The Hubble Space Telescope's Science Productivity between 1998 and 2008. *Publications of the Astronomical Society of the Pacific*, 122(893):808.

Simone Astarita, Sandor Kruk, Jan Reerink, and Pablo Gómez. 2024. Delving into the Utilisation of ChatGPT in Scientific Publications in Astronomy. In *Proceedings of SPAICE2024: The First Joint European Space Agency / IAA Conference on AI in and for Space*, pages 241–246.

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.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Vanessa Böhm and Jia Liu. 2023. Impact of the COVID-19 pandemic on publishing in astronomy in the initial two years. *Nature Astronomy*, 7:105–112.

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. *Preprint*, arXiv:2005.14165.

Lucas Carrit Delgado Pinheiro, Ziru Chen, Bruno Caixeta Piazza, Ness Shroff, Yingbin Liang, Yuan-Sen Ting, and Huan Sun. 2025. Large Language Models Achieve Gold Medal Performance at the International Olympiad on Astronomy & Astrophysics (IOAA). *arXiv e-prints*, arXiv:2510.05016.

Haotian Chen, Qingqing Long, Meng Xiao, Xiao Luo, Wei Ju, Chengrui Wang, Xuezhi Wang, Yuanchun Zhou, and Hengshu Zhu. 2025. Scirerankbench: Benchmarking rerankers towards scientific retrieval-augmented generated llms. *Preprint*, arXiv:2508.08742.

Ioana Ciucă and Yuan-Sen Ting. 2023. Galactic ChitChat: Using Large Language Models to Converse with Astronomy Literature. *Research Notes of the American Astronomical Society*, 7(9):193.

Morgan Fouesneau, Ivelina G. Momcheva, Urmila Chadayammuri, Mariia Demianenko, Antoine Dumont, Raphael E. Hviding, K. Angelique Kahle, Nadiia Pulatova, Bhavesh Rajpoot, Marten B. Scheuck, Rhys Seeburger, Dmitry Semenov, and Jaime I. Villaseñor. 2024. What is the Role of Large Language Models in the Evolution of Astronomy Research? *arXiv e-prints*, arXiv:2409.20252.

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.

Mark A Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. 2024. The strain on scientific publishing. *Quantitative Science Studies*, 5(4):823.

E. A. Henneken and A. Accomazzi. 2012. Linking to Data: Effect on Citation Rates in Astronomy. In *Astronomical Data Analysis Software and Systems XXI*, volume 461 of *Astronomical Society of the Pacific Conference Series*, page 763.

Alina Hyk, Kiera McCormick, Mian Zhong, Ioana Ciucă, Sanjib Sharma, John F Wu, J. E. G. Peek, Kartheik G. Iyer, Ziang Xiao, and Anjalie Field. 2025. From queries to criteria: Understanding how astronomers evaluate LLMs. In *Second Conference on Language Modeling*.

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 Supplements*, 275(2):38.

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. *arXiv e-prints*, arXiv:2505.20538.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2023. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *Preprint*, arXiv:2212.14024.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Preprint*, arXiv:2004.12832.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

J. Lagerstrom. 2015. Best Practices for Creating an Observatory or Telescope Bibliography from the IAU Commission 5 Working Group on Libraries. In *Open Science at the Frontiers of Librarianship*, volume 492 of *Astronomical Society of the Pacific Conference Series*, page 99.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification, Outstanding Certification.

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.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. *Preprint*, arXiv:2406.11695.

Joshua Peek, Vandana Desai, Richard L. White, Raffaele D'Abrusco, Joseph M. Mazzarella, Carolyn Grant, Jenny Novacescu, Elena Scire, and Sherry Winkelman. 2019. Robust Archives Maximize Scientific Accessibility. In *Bulletin of the American Astronomical Society*, volume 51, page 105.

Alberto Pepe, Alyssa Goodman, August Muench, Merce Crosas, and Christopher Erdmann. 2014. How do astronomers share data? reliability and persistence of datasets linked in aas publications and a qualitative study of data practices among us astronomers. *PLOS ONE*, 9(8):1–11.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Karen Spark Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

John Swartzwelder and Bob Anderson. 1997. Homer vs. the Eighteenth Amendment. Television series episode, spoken by Homer Simpson: "To alcohol! The cause of, and solution to, all of life's problems.".

Y. S. Ting, T. D. Nguyen, T. Ghosal, R. Pan, H. Arora, Z. Sun, T. de Haan, N. Ramachandra, A. Wells, S. Madireddy, and A. Accomazzi. 2025. AstroMLab 1: Who wins astronomy jeopardy!? *Astronomy and Computing*, 51:100893.

John F. Wu, Alina Hyk, Kiera McCormick, Christine Ye, Simone Astarita, Elina Baral, Jo Ciuca, Jesse Cranney, Anjalie Field, Kartheik Iyer, Philipp Koehn, Jenn Kotler, Sandor Kruk, Michelle Ntampaka, Charles O'Neill, Joshua E. G. Peek, Sanjib Sharma, and Mikaeel Yunus. 2024. Designing an Evaluation Framework for Large Language Models in Astronomy Research. *arXiv e-prints*, arXiv:2405.20389.

Xueqing Xu, Boris Bolliet, Adrian Dimitrov, Andrew Laverick, Francisco Villaescusa-Navarro, Licong Xu, and Íñigo Zubeldia. 2025. Evaluating Retrieval-Augmented Generation Agents for Autonomous Scientific Discovery in Astrophysics. *arXiv e-prints*, arXiv:2507.07155.

A Examples of amc Structued Outputs for the TRACS Test Set

In Listings 1, 2, 3, and 4, we show several representative examples of amc outputs for the TRACS test dataset.


```
"2024ApJ...977..173C_JWST": {
  "id": "2024ApJ...977..173C_JWST",
  "bibcode": "2024ApJ...977..173C",
  "telescope": "JWST",
  "classification": {
    "telescope": "JWST",
    "science": true,
    "instrumentation": false,
    "mention": false,
    "not_telescope": false,
    "quotes": [
      "we present JWST MIRI observations of the hydrocarbon-rich source, DoAr 33, a 1.1 M star.",
      "We detect the emission of H 2 O, CO 2 , OH, and C 2 H 2 , including its isotopologue 13 C
      ↪  12 CH 2 , C 4 H 2 , and tentatively CH 4 and HC 3 N, for all of which we retrieve co
lumn densities, temperatures, and emitting areas, as described in Section 2 .",
      "We detect the presence of H 2 O, CO 2 , OH, C 2 H 2 , HCN, C 4 H 2 , and tentatively CH 4
      ↪  and HC 3 N in the JWST MIRI-MRS spectrum of the solar-mass star DoAr 33.",
      "JWST/MIRI Detection of a Carbon-rich Chemistry in the Disk of a Solar Nebula Analog",
      "Figure 1. JWST MIRI-MRS spectrum of DoAr 33."
    ],
    "reasoning": "The excerpts clearly describe original JWST MIRI-MRS observations of the target
    ↪  DoAr 33 and report new detections and quantitative retrievals (column densities, te
mperatures, emitting areas) derived in this paper. These are new scientific results based on JWST
↪  data, so 'science' is True. The text discusses data reduction using a custom JDISCS p
ipeline and calibration references but does not present new instrument design, calibration method
↪  development as the primary focus or a technical/instrumentation paper, so 'instrument
ation' is False. Because the paper actively uses JWST data to produce new results, it is not
↪  merely a mention; therefore 'mention' is False. There is no indication that references to
JWST are false positives, so 'not_telescope' is False."
  }
}
```


Listing 1: amc output for an science paper type.

B The JWST Preprints Automation

The design of amc was initially devised for a specific task: automatically checking whether new arXiv preprints feature JWST science. While identifying JWST science preprints is useful on its own, it is also necessary for another goal: determining whether JWST science publications are accompanied by digital object identifiers (DOIs), which offers downstream benefits to individual scientists (Henneken and Accomazzi, 2012) and the broader science community (Pepe et al., 2014). Therefore, we built an earlier version of amc for automatically classifying whether a JWST preprint is science or not.[6] We achieved $\boldsymbol{F_1 \approx 0.95}$ on real-world tests, using a held-out golden sample dataset with $N = 114$ papers, and this system is currently being used in production at STScI.

There are two major differences between the JWST preprints automation and the TRACS task: (1) the former downloads PDF files from arXiv and converts them into a single plain-text body, while the latter provides more cleanly separated metadata and body text (when available); and (2) the former task is only concerned with identifying science papers, whereas the latter solicits binary classifications for science, instrumentation, mention, and not_telescope.

There are several important implications due to distinction #1. For example, we do not have a programmatic method for separating abstracts or other titles from the body, and so we must always proceed with the full body text. Another consequence is that the references section is included in the

[6] https://github.com/spacetelescope/jwst-preprint-automator

```
"1996SPIE.2808..650F_CHANDRA": {
  "id": "1996SPIE.2808..650F_CHANDRA",
  "bibcode": "1996SPIE.2808..650F",
  "telescope": "CHANDRA",
  "classification": {
    "telescope": "CHANDRA",
    "science": false,
    "instrumentation": true,
    "mention": false,
    "not_telescope": false,
    "quotes": [
      "Modeling the diffraction efficiencies of the AXAF high-energy transmission gratings: II",
      "In order to characterize the instrumentation on AXAF, each of the science instrument teams
      ↪  carries out sub-assembly calibrations.",
      "For the high energy transmission grating (HETG) group, this means individual measurements
      ↪  of the diffraction efficiencies of each of the 336 grating elements that goes into t
he completed HETG assembly."
    ],
    "reasoning": "The excerpt discusses calibration and modeling of the AXAF high-energy
    ↪  transmission gratings (HETG), measurements, data reduction and instrumentation
    ↪  characterization. AXAF is the original name for the Chandra X-ray Observatory, and the
    ↪  content is clearly about instrument performance and calibration rather than presenting
    ↪  new astrophysical results. Therefore instrumentation = True. It does not present new
    ↪  scientific/astronomical analyses using telescope observations, so science = False.
    ↪  Because the paper actually discusses the telescope/instrument technical work, it should
    ↪  not be classified as a mere mention or as not_telescope."
  }
}
```


Listing 2: amc output for an `instrumentation` paper type.

plain-text body extract from arXiv preprints. Thus, the body text contains references to titles of *other* papers, which can sometimes mimic sentences that appear to support a JWST science classification.

Because we focus only on classifying whether a JWST paper is `science` (distinction #2), we break down the LLM output into two stages (see Figure 1. First, we write out a specialized prompt with in-context examples of low and high `science` scores, and prompt the LLM to output reasoning and supporting quotes. Then, given its provided justification and quotes, we force the LLM to output a `science` score as a floating point number between 0 and 1.

C Other Mission Bibliographies

We note that NASA flagship missions tend to be mentioned in the paper abstract, or even in the title. Thus, for TRACS and for JWST preprints automation, it is often possible to accurately classify papers using just the abstract, a tiny fraction of the available text per entry.

However, we designed amc based on our experience performing bibliometrics for other missions, including TESS, Pan-STARRS, and GALEX, which all enable *archival* science well after the telescope ceases operations (e.g., Peek et al., 2019). However, these telescopes—and, in general, archival science usage of telescopes—are less likely to appear in the title or abstract. Therefore, parsing the much longer body text is imperative for correctly predicting those telescopes' paper types. We design amc so that it can be used just as easily to compute bibliometrics for non-flagship telescopes and/or archival science.

```
"2024A&A...691A.352W_JWST": {
  "id": "2024A&A...691A.352W_JWST",
  "bibcode": "2024A&A...691A.352W",
  "telescope": "JWST",
  "classification": {
    "telescope": "JWST",
    "science": false,
    "instrumentation": false,
    "mention": true,
    "not_telescope": false,
    "quotes": [
      "an unbiased mid-infrared survey observing H-dominated metal-enriched white dwarfs making
      ↪  use of the much improved sensitivity of JWST, compared to previous facilities, would
      ↪  likely reveal the true population of discs.",
      "infrared spectroscopy can identify the mineralogy of discs, with silicates having been
      ↪  detected using Spitzer and JWST ( Jura et al. 2007 ; Swan et al. 2024 )."
    ],
    "reasoning": "The excerpt discusses JWST in the context of future or potential surveys and
    ↪  improved sensitivity compared to previous facilities, and cites JWST detections (Swan et
    ↪  al. 2024) as literature. There is no indication that this paper presents new JWST
    ↪  observational data or original data analysis; it instead suggests that an unbiased survey
    ↪  with JWST would be useful and refers to prior JWST detections. This fits the 'mention'
    ↪  category (discussion of future/planned observations and citing others). It does not
    ↪  describe instrumentation or calibration, and the references to JWST clearly refer to the
    ↪  telescope, so not_telescope is false."
  }
}
```


Listing 3: amc output for a `mention` paper type.


```
"2008ApJ...673..657M_HST": {
  "id": "2008ApJ...673..657M_HST",
  "bibcode": "2008ApJ...673..657M",
  "telescope": "HST",
  "classification": {
    "telescope": "HST",
    "science": false,
    "instrumentation": false,
    "mention": false,
    "not_telescope": true,
    "quotes": [
      "Hubble diagram showing the magnitude deviation relative to an empty universe, for the three
      ↪  models considered.",
      "Fig. 4. — Hubble diagram showing the magnitude deviation relative to an empty universe, for
      ↪  the three models considered. The dotted lines separate the Tonry et al. ( 2003 ) data (
      ↪  left ) from the mock data ( right )."
    ],
    "reasoning": "The excerpts discuss a 'Hubble diagram' and cosmological supernova analyses;
    ↪  there is no mention of the Hubble Space Telescope or HST observations, instruments,
    ↪  calibrations, or use of HST data. The term 'Hubble' here refers to the Hubble diagram
    ↪  (relation to Hubble expansion), which is a false positive for the telescope name.
    ↪  Therefore this is not about the HST telescope (not_telescope = True). Because it does not
    ↪  discuss any telescope data usage or instrumentation, science and instrumentation are
    ↪  False. It also is not merely a mention of the telescope, so mention = False. The provided
    ↪  exact substrings showing 'Hubble diagram' are included as supporting quotes."
  }
}
```


Listing 4: amc output for a `not_telescope` paper type.

AstroMLab 5: Structured Summaries and Concept Extraction for 400,000 Astrophysics Papers

Yuan-Sen Ting[1], Alberto Accomazzi[2], Tirthankar Ghosal[3], Tuan Dung Nguyen[4], Rui Pan[5], Zechang Sun[6], Tijmen de Haan[7]

[1]The Ohio State University, [2]Harvard-Smithsonian Center for Astrophysics, [3]Oak Ridge National Laboratory, [4]University of Pennsylvania, [5]University of Illinois at Urbana-Champaign, [6]Tsinghua University, [7]KEK

Abstract

We present a dataset of 408,590 astrophysics papers from arXiv (astro-ph), spanning 1992 through July 2025. Each paper has been processed through a multi-stage pipeline to produce: (1) structured summaries organized into six semantic sections (Background, Motivation, Methodology, Results, Interpretation, Implication), and (2) concept extraction yielding 9,999 unique concepts with detailed descriptions. The dataset contains 3.8 million paper-concept associations and includes semantic embeddings for all concepts. Comparison with traditional ADS keywords reveals that the concepts provide denser coverage and more uniform distribution, while analysis of embedding space structure demonstrates that concepts are semantically dispersed within papers—enabling discovery through multiple diverse entry points. Concept vocabulary and embeddings are publicly released at `https://github.com/tingyuansen/astro-ph_knowledge_graph`.

1 Introduction

A frontier application of large language models is their deployment as autonomous agents that reason about scientific literature, plan research strategies, and execute multi-step retrieval tasks (Brown et al., 2020; Wang and Zeng, 2025). Such systems, already demonstrated in materials science and chemistry for autonomous experimentation (Szymanski et al., 2023; Boiko et al., 2023; Bran et al., 2023; Caldas Ramos et al., 2024), require structured knowledge representations to function—moving beyond language processing to operate on semantically organized information. While LLMs can process raw text, their utility as research agents depends on access to curated intermediate representations that bridge unstructured documents and formal knowledge structures (Lewis et al., 2020).

Astronomy presents an advantageous testing ground: most papers are archived on arXiv (astro-ph since 1992), and the open-sky policy enables databases to link astronomical objects directly to papers. The combination of papers, observed objects, and their properties provides an ecosystem where structured representations could enable agentic research and autonomous discovery.

However, text as a modality remains undercurated in the astronomy literature. Current resources are either too complete (full source, which is difficult to extract insights) or too sparse (abstracts only). Both extremes limit downstream applications. Useful scientific ideas emerge from holistic understanding of concepts rather than direct processing of individual words—this is how humans engage with literature. Keywords were designed to bridge this gap, but when present, are rarely mapped to controlled vocabularies like the Unified Astronomy Thesaurus (UAT) and exhibit sparsity—most keywords appear in very few papers and many papers have very few keywords, rendering them unsuitable for systematic analysis.

LLMs can extract such structured representations from papers, but this is cost-intensive. Individual researchers performing this task separately would waste computational resources. A single, centralized effort provides economies of scale and ensures consistency across the literature.

To address this, we organize all astro-ph papers into structured summaries and concepts—two intermediate layers that bridge the gap between raw text and knowledge representation. Our work builds on recent developments in applying LLMs to astronomical research, including domain-specific models like AstroLLaMA (Pan et al., 2024) and AstroSage (de Haan et al., 2025), complementary efforts in knowledge graph construction (Sun et al., 2024; Kau et al., 2024), and the development of recommender systems (Geng et al., 2022; Chu et al., 2023; Zhao et al., 2023; Vats et al., 2024). We present a comprehensive dataset spanning 408,590 papers with 9,999 unique concepts, their semantic embeddings, and structured summaries.

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 170–185
December 23, 2025

2 PDF to Text: OCR Pipeline

Converting astrophysics PDFs to machine-readable text presents challenges due to the prevalence of mathematical equations, multi-column layouts, and figures with embedded captions. We chose to use OCR rather than LaTeX source files because LaTeX sources are not uniformly structured across papers and many papers use custom macros that complicate parsing.

Our pipeline initially used Nougat (Blecher et al., 2023)[1], an academic document OCR model that converts PDFs to markdown format for ease of processing. Processing each paper requires approximately 1 minute on a V100 GPU, representing a substantial computational investment—processing over 350,000 papers required about 6,000 V100 GPU-hours. Starting in November 2024, we transitioned to Mathpix OCR API[2] as it proved more reliable than Nougat.

Both Nougat and Mathpix preserve mathematical notation in LaTeX format within the OCR output. For section detection, Nougat outputs markdown headers (###) while Mathpix preserves LaTeX section commands (`\section`). The transition to Mathpix was motivated by Nougat's occasional failure mode where approximately 1 in 500 pages would produce repetitive text; such corrupted pages are naturally excluded during the summarization stage, though this may result in missing information at a subdominant level. For Mathpix OCR (covering approximately 50,000 papers from November 2024 onward), author team inspection of randomly sampled pages revealed no systematic OCR errors at levels that would impact summary quality.

3 Multi-Stage Summarization

3.1 Chunk-Based Compression

During early development, we found that processing entire papers at once led to incomplete summaries, with LLMs often omitting important details or providing superficial coverage. This was problematic for generating organized summaries with properly populated sections—methodological details, for instance, were frequently under-represented. Processing single abstracts typically missed useful information like detailed derivations and technical implementation specifics. This motivated our chunk-based approach, which processes papers in manageable segments while maintaining context across chunks.

We split each paper into approximately 10,000-character chunks using section-aware boundaries to avoid mid-sentence breaks. Papers are split at section boundaries (detecting either markdown headers from Nougat or `\section` commands from Mathpix), with adjacent small sections merged up to the 10,000-character limit. Each chunk is sequentially compressed with context from previously compressed chunks, ensuring coherence across the full paper. This approach increases token costs several-fold compared to single-pass processing, but when this project started in late 2023, this was necessary to achieve adequate quality. We maintained this approach for subsequently processed papers to ensure consistency.

The compression system prompt emphasizes: (1) retaining LaTeX formulas, (2) focusing on motivations and methods, (3) highlighting key results and connections to other works, (4) preserving technical jargon for expert readers, and (5) excluding acknowledgments and references. As language models improved, we adopted the most affordable versions while maintaining quality. Different papers were processed with GPT-4o, GPT-4o-mini, o1-mini, and DeepSeek-v3 depending on availability. The complete summarization process for all 408,590 papers required over \$50,000 in API costs, not including OCR costs.

3.2 From Raw Summaries to Structured Organization

Abstract sections often jumble information chronologically or by importance, making systematic analysis difficult. We reorganize raw summaries into seven semantic sections that follow the logical flow of scientific papers: Title and Author, Background, Motivation, Methodology, Results, Interpretation, and Implication. This structured format enables targeted queries and facilitates knowledge representation by clearly separating context, methods, and outcomes. Appendix A shows a complete example demonstrating all six sections.

4 Concept Extraction and Vocabulary

4.1 Extraction Methodology

For each organized summary, we prompt the LLM to extract approximately 10 key concepts focusing on novel contributions. The target of 10 concepts

[1] https://github.com/facebookresearch/nougat
[2] https://mathpix.com/ocr

provides finer granularity than traditional keyword systems (where author-supplied keywords typically number 3-5 per paper) while remaining tractable for LLM extraction. The system prompt emphasizes: (1) identifying innovations and novel methods, (2) covering both scientific concepts (observational phenomena, theoretical frameworks) and technological concepts (computational techniques, instrumentation), and (3) avoiding generic field names or overly specific parameters.

This approach leverages the capacity of modern language models (Achiam et al., 2023; Beltagy et al., 2019; Ting et al., 2025) to understand domain-specific scientific contexts. Each concept includes three components: a **Name** (3-4 word concise label), a **Class** (Cosmology & Nongalactic Physics, High Energy Astrophysics, Instrumental Design, Galaxy Physics, Numerical Simulation, Statistics & AI, Solar & Stellar Physics, or Earth & Planetary Science), and a **Description** ($\sim$100-word technical explanation). The final concept vocabulary was generated homogeneously using a combination of GPT-4o and o1-mini to ensure consistency across all papers.

4.2 Vocabulary Construction and Clustering

LLM-extracted concepts lack a priori control over consistency—different papers may use different terminology for the same concept, and there is no guarantee of controlled vocabulary. To address this, following the methodology of Sun et al. (2024), we employ a multi-stage clustering process. For each extracted concept in each paper, we combine the organized summary with the concept name to generate detailed descriptions. These descriptions are then embedded using OpenAI's text-embedding-3-large model. We perform K-means clustering (k=10,000) in the cosine similarity space to consolidate similar concepts, merging semantically equivalent variants into single unified entries. The clustering maximizes inter-cluster distances while grouping semantically similar extractions. The clustering process synthesizes new unified concept descriptions that capture the full semantic range across papers.

We experimented with different vocabulary granularities in log space (3,000, 10,000, and 30,000 concepts). We found 10,000 concepts to provide the most useful balance. All 10,000 concepts and descriptions were manually reviewed by the author team, during which one null concept (representing rare OCR failure cases) was identified and removed, leaving the final vocabulary of 9,999. The concepts have been used in various downstream analyses (Sections 5-6) providing ongoing validation. Given the dataset scale, our validation strategy prioritized full vocabulary review over exhaustive paper-by-paper evaluation, and users should exercise appropriate scrutiny when employing the dataset for specific applications.

Category	Count
Cosmology & Nongalactic Physics	2,192
High Energy Astrophysics	1,606
Instrumental Design	1,295
Galaxy Physics	1,267
Numerical Simulation	1,050
Statistics & AI	1,020
Solar & Stellar Physics	930
Earth & Planetary Science	639

Table 1: Distribution of 9,999 concepts across research categories.

Each concept retains its detailed description synthesized from multiple papers, providing more context than typical keyword systems. The concept distribution across categories is shown in Table 1. Each concept appears in an average of 383 papers (median: 223), making them statistically robust while maintaining sufficient specificity. As we will see in Section 5, this granularity avoids both the overly broad categories and overly specific identifiers that plague traditional keyword systems.

5 Quality Evaluation

5.1 Comparison with Traditional Keywords

To evaluate our concept vocabulary, we compare it with traditional ADS keywords extracted via the NASA ADS API for all 408,590 papers. ADS keywords are author-supplied and not systematically checked against controlled vocabularies like the Unified Astronomy Thesaurus. We performed curation by removing arXiv classification keywords (e.g., "Astrophysics - Cosmology"), normalizing to lowercase, and filtering overly common keywords (>20,000 occurrences) and rare keywords (<10 occurrences). After curation, ADS keywords cover 73% of papers (298,658) with 6,909 unique keywords and 1.27M associations.

Figure 1 shows two key differences. First, ADS keywords suffer from severe sparsity: 44% of papers have $\leq$3 keywords and 62% have $\leq$4 keywords—insufficient for effective semantic search or recommendation systems. This primarily reflects different generation mechanisms: author-

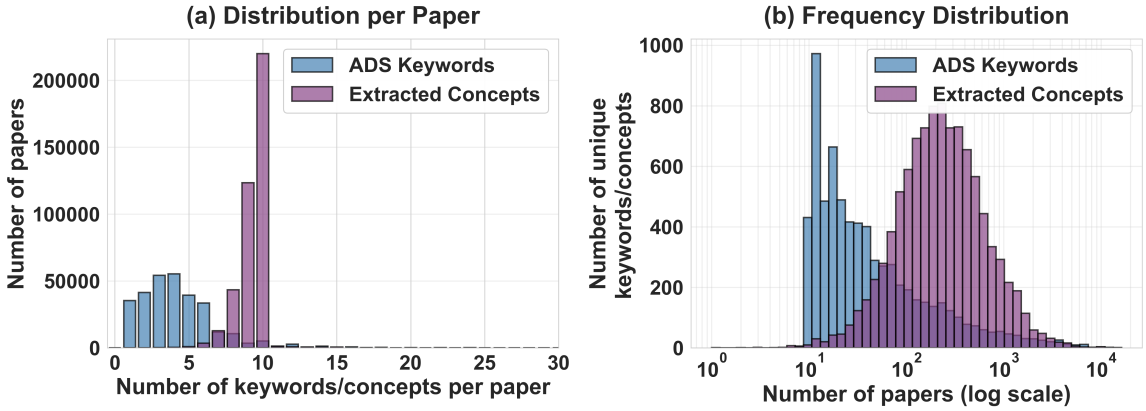


Figure 1: Distribution of keywords/concepts per paper (left) and frequency distribution (right). ADS keywords show high sparsity with many papers having few keywords, while our concepts provide consistent coverage. The frequency distribution (right) reveals a large pile-up of overly generic terms and an extended tail of overly specific identifiers, while our concepts maintain more balanced intermediate granularity.

supplied keywords are known to be sparse partly because authors often do not systematically check controlled vocabularies, and different journals have varying keyword standards. In contrast, our extraction prompt explicitly requests approximately 10 concepts per paper, achieving complete coverage: all 408,590 papers have structured summaries and concept associations. The average paper has 9.4 concepts (median: 10) concepts per paper with a small dispersion.

Beyond coverage, our concepts exhibit more balanced frequency distribution. While ADS keywords suffer from extreme imbalances—most common keywords like "galaxies: evolution" appear in 16,321 papers while 2,658 keywords (38% of the vocabulary) appear in only 10-20 papers—our concepts maintain intermediate granularity. Table 2 illustrates this problem: the most common keywords represent overly broad field categories with limited discriminative power, while rare keywords are often object-specific identifiers (e.g., "grb 080319b") rather than research themes.

In contrast, our concepts balance these extremes through systematic curation and maintain semantic meaningfulness across all frequency ranges, through the clustering and consolidation process. Table 3 demonstrates this: high-frequency concepts represent general methodologies applicable across subfields (e.g., "Monte Carlo Simulations" with 13,671 papers) that retain semantic specificity and discriminative power for retrieval, medium-frequency concepts capture well-established research areas (e.g., "Stellar Evolution Models" with 2,751 papers), and low-frequency concepts identify

Keyword	Papers	%
Most Common (Overly Broad)		
galaxies: evolution	16,321	5.5
galaxies: active	14,121	4.7
accretion	12,540	4.2
methods: numerical	12,510	4.2
stars: formation	9,172	3.1
dark matter	9,032	3.0
methods: data analysis	8,384	2.8
galaxies: formation	8,313	2.8
Rare (Overly Specific) - 2,658 keywords with 10-20 papers		
gamma-ray burst: individual: grb 080319b		
stars: individual: alphanumeric: hd 209458		
galaxies: individual: alphanumeric: ngc 1275		
pulsars: individual: alphanumeric: psr j1614-2230		
x-rays: binaries: individual: alphanumeric: cygnus x-1		

Table 2: Examples of overly broad and overly specific ADS keywords. The most common keywords represent broad field categories with limited discriminative power, while rare keywords are often object-specific with minimal value for thematic analysis.

emerging or specialized topics while remaining thematic rather than object-specific (e.g., "Interpretable Machine Learning in Astronomy" with 49 papers). The median concept appears in 223 papers—more balanced than the median of 28 for ADS keywords.

These limitations of traditional keywords make content-based recommendation systems difficult to implement. This dataset provides an alternative that enables more robust semantic search and recommendation algorithms, which may be useful for platforms like NASA ADS. Our concept vocabulary includes many terms not present in the UAT, including in emerging areas like deep learning ap-

Frequency	Concept	Papers	Class
High	Monte Carlo Simulations	13,671	Numerical Simulation
	N-Body Simulation Dynamics	12,041	Numerical Simulation
	Astronomical Spectral Energy Profiles	9,590	Galaxy Physics
	Cosmic Microwave Background	9,166	Cosmology & Nongalactic Physics
Medium-High	Stellar Evolution Models	2,751	Solar & Stellar Physics
	Dynamic Cosmological Constant	2,692	Cosmology & Nongalactic Physics
	Galaxy Morphological Study	2,485	Galaxy Physics
	High-Redshift Quasars	1,626	Cosmology & Nongalactic Physics
Medium	Gravitational Lensing Surveys	278	Instrumental Design
	CMB Simulation Methodologies	276	Numerical Simulation
	Marginalization in Bayesian Inference	268	Statistics & AI
	Neural Inference Methods	202	Statistics & AI
Low	Interpretable Machine Learning in Astronomy	49	Statistics & AI
	Neutrino-Driven Supernova Simulations	48	Numerical Simulation
	Exoplanetary Companion Systems	44	Solar & Stellar Physics
	Gravitational Wavefront Interactions	35	Cosmology & Nongalactic Physics

Table 3: Examples of our concepts across frequency ranges. Unlike traditional keywords that become meaningless at extremes (overly broad or object-specific), our concepts remain scientifically meaningful across all frequencies, maintaining thematic coherence rather than becoming object-specific identifiers.

plications in astronomy. While our concepts can be mapped to UAT for compatibility with existing systems, we also propose this vocabulary as a potential foundation for extending or complementing the UAT with contemporary research terminology.

5.2 Concepts for Discovery

Beyond coverage and frequency balance, why are concepts superior to abstracts for discovery tasks? While abstracts provide summaries of papers, they operate at a narrative level that is not optimal for discovery. Novel ideas often emerge from specific methodological details, intermediate results, or conceptual connections that are embedded within a paper but not prominently featured in its abstract. Furthermore, current language models process continuous text rather than discrete conceptual tokens, limiting their ability to generate novel hypotheses through systematic exploration of the idea space.

To demonstrate why concepts are critical for discovery, we analyze the embedding space structure of 10,000 randomly sampled papers. Each concept in our vocabulary has a detailed description (see Table 3), from which we extract embeddings using OpenAI's text-embedding-3-large. We perform the same embedding extraction for each paper's abstract and for the six individual sections of its structured summary.

Figure 2 shows UMAP projections of four representative papers—two with high concept dispersion (top row) and two with low dispersion (bottom row). The faint gray background represents all 9,999 concepts in our vocabulary, providing spatial context. Even in cases labeled as "low dispersion" (bottom row), the concepts assigned to individual papers (bold gray circles with labels) remain dispersed across semantic space.

This dispersion occurs because papers contain multiple distinct ideas spanning different domains. For example, the top-right panel shows a paper on fractional cosmology that discusses concepts including "Hubble Data Analysis Diagnostics", "Variational Principles in Physics", "Fractional Calculus in Physics", and "Riccati Equations in Physics"—concepts that occupy distant regions of semantic space, bridging observational analysis, theoretical cosmology, and mathematical physics. Such concepts cannot be recovered from abstract embeddings alone; some are embedded deeply in methodological sections and never explicitly mentioned in abstracts. In stark contrast, the six summary sections (colored diamonds) and abstract (gold star) cluster tightly together in all cases, as all sections describe the same paper from different angles—they are semantically coherent because they narrate a single research story.

This analysis does not diminish the value of structured summaries—quite the contrary. It reveals the complementarity of concepts and summaries in our knowledge graph. Concepts are dispersed across semantic space, assigned to papers based on diverse topical content, making them ideal for discovery. A researcher exploring "Variational Principles in Physics" can find relevant papers,

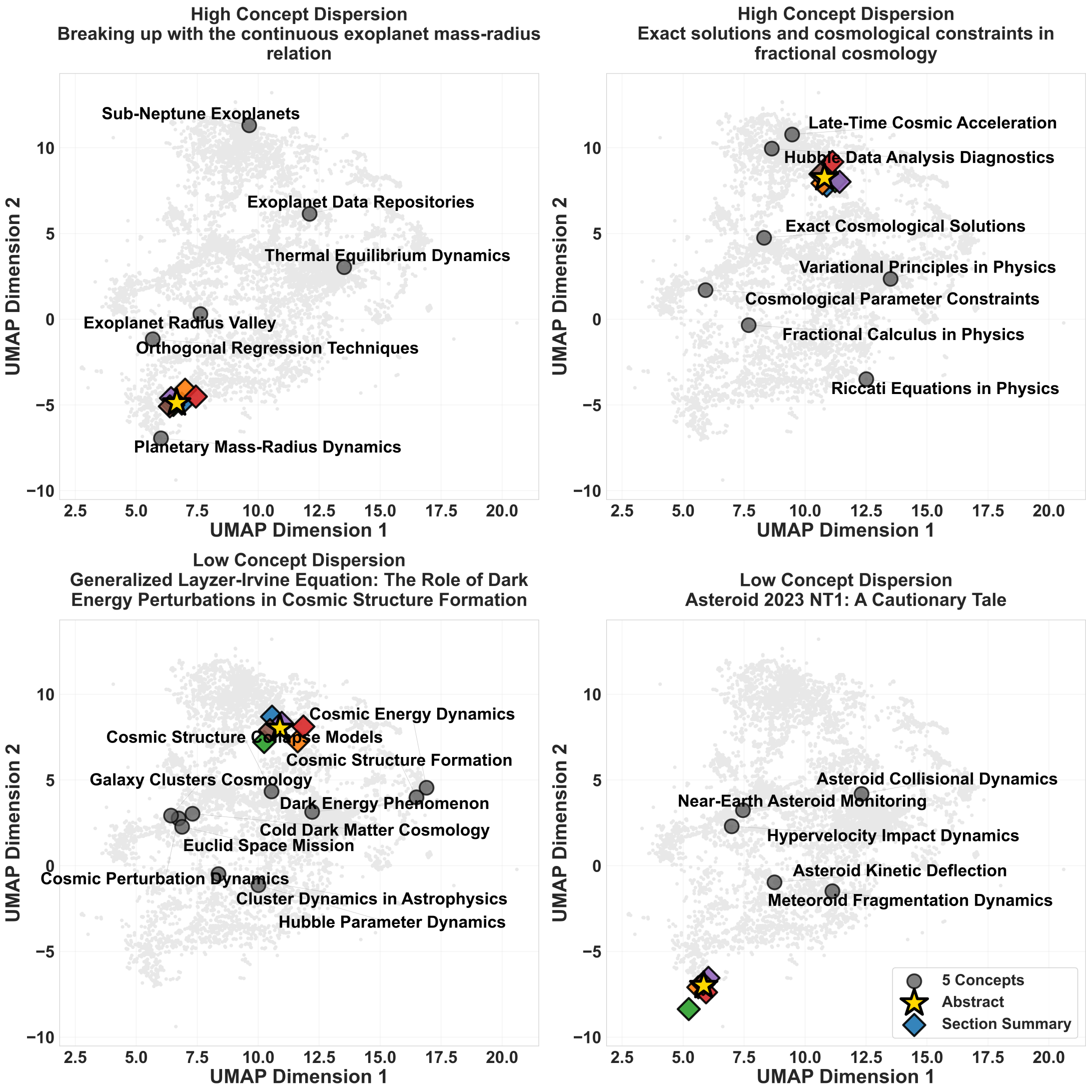


Figure 2: UMAP projections of concept (grey symbols) and summary embeddings (colored diamonds) and the abstract (gold star) for four representative papers. Faint gray background shows all 9,999 concepts in the vocabulary. Even papers classified as "low dispersion" (bottom row) have concepts spread across distinct semantic regions, showing that abstracts (and summaries) cannot capture the full conceptual diversity present in papers, unlike concepts.

even if this concept appears only in a methodological subsection and not in the abstract. Summaries, conversely, cluster together because all sections describe the same paper. This narrative coherence makes them valuable for understanding context after relevant papers are identified through concept-based discovery.

6 Applications

Having established the quality advantages of our concept vocabulary, we now demonstrate its utility through two applications that leverage these properties: temporal analysis of concept emergence and co-occurrence analysis of research themes.

6.1 Temporal Evolution of Concepts

The granular and semantically meaningful nature of our concepts enables precise tracking of how ideas emerge, evolve, and connect across different research areas. This application demonstrates the value of our vocabulary for constructing knowledge graphs (Kau et al., 2024) that trace research evolution. We analyze concept emergence by identifying when each concept first appeared in at least 5 papers (a threshold ensuring stability rather than single-paper anomalies). Figure 3 shows the temporal evolution across three decades, with new concepts per year (left) and cumulative growth (right).

The declining rate of new concept emergence in

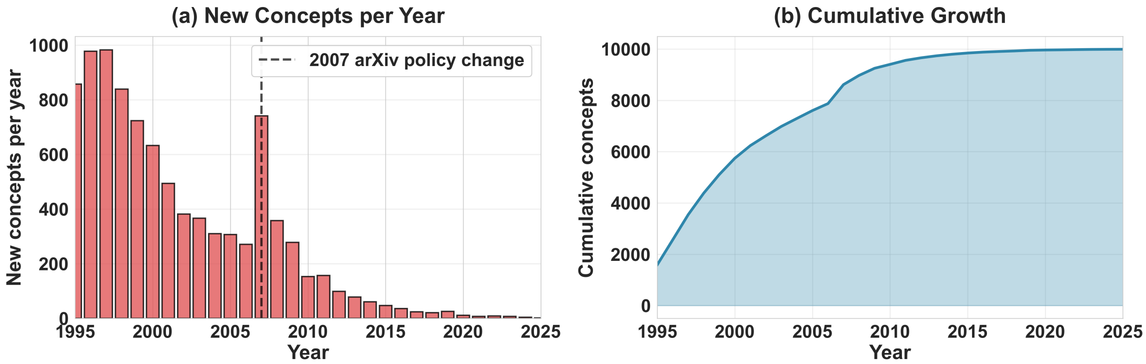


Figure 3: Temporal evolution of concept vocabulary across three decades. (a) Number of new concepts emerging each year (crossing the 5-paper threshold). (b) Cumulative growth of the concept vocabulary. The rapid expansion in the early years reflects foundational concepts when arXiv began. A secondary peak in 2007 corresponds to cross-listing policy changes.

recent years does not necessarily indicate reduced innovation. Several factors contribute to this pattern. First, many fields have matured, with research increasingly focused on connections between established concepts rather than entirely new topics. Second, our 5-paper threshold means concepts can appear earlier than their peak importance—for example, concepts about the Gaia mission and the James Webbs Space Telescope emerged earlier than their launch, when early planning papers crossed the threshold, despite these missions becoming prominent only after launch. Third, our clustering methodology itself may exhibit systematic bias: because clustering aims to consolidate semantically similar terms across all papers, genuinely novel concepts appearing in recent years may be merged into established clusters from earlier periods if sufficiently similar in embedding space, suppressing the apparent emergence rate.

A notable secondary peak occurred in 2007, corresponding to arXiv expanding cross-listing policies to allow papers from other disciplines to include astro-ph as a secondary category. Analysis of these 2007 concepts reveals their origin: 51% are classified as Cosmology & Nongalactic Physics and 13% as High Energy Astrophysics, dominated by theoretical topics including Loop Quantum Gravity, Holographic Duality, Einstein-Gauss-Bonnet Gravity Theories, Quantum Entanglement Entropy, Conformal Field Theory, and Type IIB and Heterotic String Theories. These reflect contributions from theoretical physics and general relativity research that began appearing in astro-ph through cross-listing.

Over the past decade (2015-2025), 190 new concepts emerged (Appendix B, Tables 5 and 6). Deep learning dominates recent emergence: Astronomical CNN Applications (1,676 papers), Deep Learning in Astronomy (604 papers), Residual Neural Networks (402 papers), U-Net Variants in Astronomy (373 papers), Transformer Architectures in Astronomy (185 papers), and Physics-Informed Neural Networks (120 papers). Recent concepts also include observational capabilities—including JWST Deep Extragalactic Surveys (75 papers, emerged 2022) and GW170817 Multimessenger Merger (324 papers, emerged 2017), which by our metric are considered "new" when they first crossed the 5-paper threshold, even though JWST's scientific impact continues to grow. Appendix B provides representative astronomy-relevant examples from the full list.

6.2 Concept Co-occurrence

While temporal analysis reveals when individual concepts emerge, understanding how concepts appear together in papers provides complementary insights into the thematic structure of research. Unlike traditional citation analysis which tracks paper-to-paper relationships, concept co-occurrence reveals how different methodologies, observations, and theories interconnect within the field, identifying which ideas are commonly explored together and how these patterns shift as the field develops.

We quantify co-occurrence using the Ochiai coefficient, which normalizes by concept popularity. Intuitively, if two concepts i and j appear together in N_{ij} papers, and appear individually in N_i and N_j papers respectively, the Ochiai coefficient is:

$$\mathrm{Ochiai}(i,j) = \frac{N_{ij}}{\sqrt{N_i \cdot N_j}} \tag{1}$$

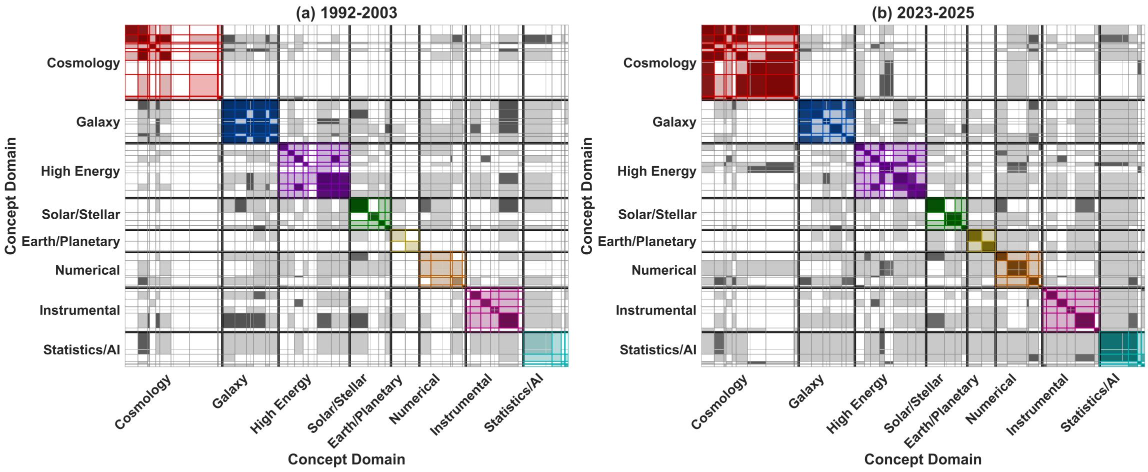


Figure 4: **Evolution of concept co-occurrence in astrophysics.** Darker colors indicate stronger co-occurrence. (a) Early period (1992–2003): established domain structure. (b) Recent period (2023–2025): computational domains (Statistics/AI, Numerical Simulation) show increased internal coherence and enhanced cross-domain integration with traditional astrophysical domains, reflecting the field's evolution toward data-intensive research.

This normalization is important because different subfields have vastly different publication volumes—this ensures we measure genuine conceptual relationships rather than simply reflecting which fields are most active.

Co-occurrence analysis is a rich topic with many dimensions to explore. Here we present a simple comparison between two time periods to illustrate the utility of our concept vocabulary. Figure 4 compares the earliest window (1992–2003, 40,000 papers) with the most recent window (2023–2025, 40,000 papers). Using fixed-size temporal windows removes field growth bias—later periods do not appear artificially stronger simply due to increased publication volume.

For visualization, we apply spectral clustering within each of the 8 predefined domains (Table 1) using 2025 data to identify subclusters, producing the fine-grained structure visible in Figure 4. To aggregate the 9,999$\times$9,999 concept matrix into this manageable visualization, we compute the 10th percentile of co-occurrence scores within each subcluster block (capturing robust signal while filtering noise), and use the spread between 10th and 30th percentiles to set transparency (indicating consistency of patterns). These percentile choices enhance dynamic range: the 10th percentile provides a stable color metric that is less sensitive to outliers than the median, while the 30th-10th spread reveals whether co-occurrence within a block is consistent (low spread, high transparency) or heterogeneous (high spread, lower transparency). This hierarchical structure is held fixed across all temporal windows, enabling direct comparison.

As shown in the figure, the technical domains—Statistics/AI, Numerical Simulation, and to some extent Instrumentation—exhibit more cross-domain interactions in the recent period compared to the early period. In the recent period, the Statistics/AI domain shows prevalent integration with all astrophysical domains, reflecting the widespread adoption of machine learning and data-driven methods across subdisciplines. The Numerical Simulation domain displays increased internal coherence, consistent with the field's growing reliance on computational methods. These patterns show that computational and statistical approaches have evolved from peripheral tools to core components of the research ecosystem.

Concepts in science domains (Galaxy Physics, High Energy, Solar/Stellar) maintain relatively stable internal structure and interdomain connections across both periods. The Cosmology domain shows notable internal growth along with increased cross-connections to High Energy. This growth is partly attributable to the 2007 cross-listing policy expansion discussed previously, which brought theoretical physics concepts into astro-ph. The Earth/Planetary domain shows increased internal coherence in the recent period, consistent with the expansion of exoplanet research enabled by missions such as *Kepler* and *TESS* in recent years.

This analysis demonstrates how our concept vocabulary enables quantitative study of field evolution in ways that would be difficult or impossible with traditional keyword systems. The patterns revealed—computational integration, methodological shifts, and domain stability—provide empirical evidence for narratives about how astrophysics research has changed over three decades. Appendix C provides representative examples of within-domain and cross-domain concept pairs with strong co-occurrence, demonstrating fine-grained thematic structure. More sophisticated temporal analyses are beyond the scope of this paper, but the released dataset supports such investigations.

7 Dataset Release

We release the dataset on GitHub at `https://github.com/tingyuansen/astro-ph_knowledge_graph` which covers all astro-ph papers from 1992 through July 2025. The public release prioritizes the concept vocabulary and embeddings to enable reproducibility and support downstream applications. For structured summaries, we adopt a more conservative distribution policy detailed in Appendix A. The public release includes: **Concept vocabulary** as CSV with labels, names, classes, and descriptions; **concept embeddings** using text-embedding-3-large; **paper metadata** including year, arXiv ID, and ADS bibcodes; and Python scripts for data loading, verification, and analysis. A complementary **citation network** extracted from NASA ADS API is also provided, with 1.67M unique identifiers covering both internal references (between astro-ph papers) and external citations (to other disciplines). Table 4 summarizes the dataset statistics.

8 Conclusion

This work presents a dataset of 408,590 astrophysics papers from arXiv astro-ph (1992-2025) with structured six-section summaries, 9,999 AI-generated concepts with detailed descriptions, and semantic embeddings.

The key contribution is a systematically generated concept vocabulary that addresses limitations of traditional keyword systems. Unlike author-supplied ADS keywords that suffer from extreme sparsity and frequency imbalances, our AI-generated concepts provide consistent coverage across all papers with balanced distributions. Each concept includes a detailed description that preserves scientific context, enabling more effective discovery than single-word keywords. Our embedding space analysis demonstrates that concepts capture dispersed semantic information within papers that abstracts alone cannot represent, making them critical for scientific discovery rather than merely navigation.

Metric	**Value**
Total papers	408,590
Unique concepts	9,999
Total concept associations	3,827,232
Avg concepts per paper	9.4 (median: 10)
Avg papers per concept	383 (median: 223)

Table 4: Summary statistics of the astro-ph knowledge graph dataset (Table 1). All 408,590 papers have complete structured summaries, concept associations, and semantic embeddings.

Temporal analysis reveals how the concept vocabulary captures field evolution. Recent emergence (2015-2025, 190 concepts, Tables 5 and 6) is dominated by machine learning adoption, while also tracking major observational facilities and theoretical developments. Co-occurrence analysis demonstrates the increasing integration of computational domains (Statistics/AI, Numerical Simulation) with traditional astrophysical research areas, revealing the field's evolution toward data-intensive methodologies. These analyses show the vocabulary's ability to capture both enduring foundations and emerging research frontiers across three decades of astrophysics.

This dataset enables applications including semantic search systems, research trend analysis, knowledge graph construction, and training language models for scientific understanding. The combination of structured summaries, comprehensive concept vocabulary, and semantic embeddings makes this resource suitable for advancing AI-assisted scientific discovery. Recent work has demonstrated the potential of LLM agents in astronomical analysis (Sun et al., 2025; Wang and Zeng, 2025), and our structured representations provide the foundation for developing autonomous systems in astronomical research.

While this paper focuses on dataset creation and preliminary analysis, extrinsic evaluation through task-based applications is an important next step. We are actively exploring integration with recommender systems and semantic search platforms to enable concept-based paper discovery and citation network analysis. Such applications will provide quantitative evaluation of utility through user studies in production environments.

Code and Data Availability

All code, system prompts, and data processing pipelines are publicly available at https://github.com/tingyuansen/astro-ph_knowledge_graph. This includes OCR processing scripts, multi-stage summarization prompts, concept extraction and clustering code, embedding generation, co-occurrence calculation, and data verification scripts. While the proprietary APIs are not open-source, all prompts and processing logic are fully documented to enable replication with alternative models.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv e-prints*, arXiv:2303.08774.

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.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.

Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. 2023. Emergent autonomous scientific research capabilities of large language models. *arXiv e-prints*, arXiv:2304.05332.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. ChemCrow: Augmenting large-language models with chemistry tools. *arXiv e-prints*, arXiv:2304.05376.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mayk Caldas Ramos, Christopher J. Collison, and Andrew D. White. 2024. A Review of Large Language Models and Autonomous Agents in Chemistry. *arXiv e-prints*, arXiv:2407.01603.

Zhixuan Chu, Hongyan Hao, Xin Ouyang, Simeng Wang, Yan Wang, Yue Shen, Jinjie Gu, Qing Cui, Longfei Li, Siqiao Xue, James Y Zhang, and Sheng Li. 2023. Leveraging Large Language Models for Pre-trained Recommender Systems. *arXiv e-prints*, arXiv:2308.10837.

Tijmen de Haan, Yuan-Sen Ting, Tirthankar Ghosal, Tuan Dung Nguyen, Alberto Accomazzi, Azton Wells, Nesar Ramachandra, Rui Pan, and Zechang Sun. 2025. Achieving GPT-4o level performance in astronomy with a specialized 8B-parameter large language model. *Scientific Reports*, 15(1):13751.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). *arXiv e-prints*, arXiv:2203.13366.

Amanda Kau, Xuzeng He, Aishwarya Nambissan, Aland Astudillo, Hui Yin, and Amir Aryani. 2024. Combining Knowledge Graphs and Large Language Models. *arXiv e-prints*, arXiv:2407.06564.

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.

Rui Pan, Tuan Dung Nguyen, Hardik Arora, Alberto Accomazzi, Tirthankar Ghosal, and Yuan-Sen Ting. 2024. AstroMLab 2: AstroLLaMA-2-70B Model and Benchmarking Specialised LLMs for Astronomy. *arXiv e-prints*, arXiv:2409.19750.

Zechang Sun, Yuan-Sen Ting, Yaobo Liang, Nan Duan, Song Huang, and Zheng Cai. 2024. Knowledge Graph in Astronomical Research with Large Language Models: Quantifying Driving Forces in Interdisciplinary Scientific Discovery. *arXiv e-prints*, arXiv:2406.01391.

Zechang Sun, Yuan-Sen Ting, Yaobo Liang, Nan Duan, Song Huang, and Zheng Cai. 2025. Mephisto: Self-Improving Large Language Model-Based Agents for Automated Interpretation of Multi-band Galaxy Observations. *arXiv e-prints*, arXiv:2510.08354.

Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. 2023. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91.

Y. S. Ting, T. D. Nguyen, T. Ghosal, R. Pan, H. Arora, Z. Sun, T. de Haan, N. Ramachandra, A. Wells, S. Madireddy, and A. Accomazzi. 2025. AstroMLab 1: Who wins astronomy jeopardy!? *Astronomy and Computing*, 51:100893.

Arpita Vats, Vinija Jain, Rahul Raja, and Aman Chadha. 2024. Exploring the Impact of Large Language Models on Recommender Systems: An Extensive Review. *arXiv e-prints*, arXiv:2402.18590.

He Wang and Liang Zeng. 2025. Automated Algorithmic Discovery for Gravitational-Wave Detection Guided by LLM-Informed Evolutionary Monte Carlo Tree Search. *arXiv e-prints*, arXiv:2508.03661.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. 2023. Recommender Systems in the Era of Large Language Models (LLMs). *arXiv e-prints*, arXiv:2307.02046.

A Example Structured Summary

This appendix provides a representative example of our structured summaries to demonstrate their comprehensive nature and systematic organization. Unlike traditional abstracts that prioritize brevity, our summaries (typically 600-900 words, averaging 740 words) systematically separate six semantic sections: Background (observational and theoretical context), Motivation (scientific questions and goals), Methodology (technical approach and data), Results (empirical findings), Interpretation (theoretical analysis), and Implication (broader significance). This structure enables targeted information retrieval—a researcher can directly access methodological details or theoretical interpretations without reading the entire paper. While paper summarization is now routine with LLMs, we provide these structured summaries upon request rather than through public release out of caution at this scale (0.4M papers). Researchers interested in accessing the summaries should contact the authors.

Title: Usco1606-1935: An unusually wide low-mass triple system?

Author: Adam L. Kraus et al. (arXiv:0704.0455)

Background: The study of multiple star systems is crucial for understanding star formation processes. Surveys have shown that binary frequencies and properties vary significantly with mass. Solar-mass stars exhibit high binary frequencies (>60%) and can have separations up to $\sim 10^4$ AU. In contrast, M dwarfs have lower frequencies (30-40%) and fewer companions beyond $\sim$500 AU, while brown dwarfs show even lower frequencies ($\sim$15%) with few companions exceeding 20 AU. The observed decline in maximum binary separation with decreasing mass has been described by empirical functions, suggesting that this limit is established early in stellar lifetimes. Surveys of young stellar associations have identified a few unusually wide systems, but not enough to analyze their properties statistically.

Motivation: To address the scarcity of unusually wide low-mass systems, we utilized archival 2MASS data to search for candidate wide binary systems among known members of three nearby young associations, including Upper Sco. Our findings aim to align with the standard paradigm, revealing a deficit of wide systems among very low-mass stars and brown dwarfs, while also identifying a few candidates, such as USco1606-1935, a wide pair of stars with similar fluxes and colors. This study seeks to evaluate the probability of USco1606-1935 being an unusually wide, low-mass binary, thereby contributing to the understanding of multiple system formation and evolution in young stellar associations.

Methodology: We identified USco1606-1935 AB as a candidate binary using 2MASS data, leveraging its bright and resolved components to gather additional photometry and astrometry from various surveys, including DENIS, USNO-B, and SSS. The analysis focused on 2MASS JHK magnitudes and USNO-B I magnitudes, ensuring consistency through comparisons with DENIS data... Optical spectroscopy was conducted using the Double Spectrograph at Palomar Observatory, processing the spectrum with standard IRAF tasks and comparing it with spectral standards from Upper Sco and Taurus to confirm the spectral type. High-resolution imaging was achieved with laser guide star adaptive optics on the Keck-II telescope, obtaining nearly diffraction-limited images in both narrow and wide camera modes to measure photometry and astrometry for the components.

Results: High-resolution images revealed that USco1606-1935 A comprises two sources, Aa and Ab, with the probability of an unbound bright source near A being extremely low, suggesting that Aa and Ab form a bound binary system. Photometric data confirmed that USco1606-1935 B aligns with known members of Upper Sco, supporting its membership, although its position in color-magnitude diagrams raised questions about potential differential reddening or unresolved companions. Astrometric analysis summarized the relative positions of the system components and field stars, with proper motion indicating that B is likely a comoving member... Stellar and binary properties for the Aa-Ab and A-B systems were estimated using isochrones and temperature scales to derive component masses and spectral types.

Interpretation: Identifying pre-main sequence binaries presents challenges in distinguishing gravitationally bound pairs from coeval, comoving stars. To assess clustering among PMS stars, we calculated the two-point correlation function (TPCF), which quantifies the number of excess pairs at a given separation compared to a random distribution. Utilizing a Monte Carlo approach, the TPCF revealed significant clustering of stars at small separations. The analysis indicated that the expected surface density of unbound companions is $\sim$60 deg^{-2}, suggesting a 25% chance of chance alignments among low-mass members. Consequently, while the detection of Aa and Ab as a close binary is highly probable, the physical association of Aab and B cannot be assumed based solely on probabilistic grounds, highlighting the complexity in confirming wide binary status in such systems.

Implication: If Aab and B are gravitationally bound, USco1606-1935 would represent one of the rare young multiple systems with wide separations comparable to field systems of similar mass. However, the significant probability of chance alignment necessitates caution in such classifications. This underscores the need for systematic searches for wide binaries in the Upper Sco association and similar young stellar environments to better understand the frequency and properties of wide, low-mass multiple systems. Enhanced observational strategies and comprehensive data analyses are essential to distinguish truly bound systems from coincidental alignments, thereby refining our knowledge of star formation and the dynamical evolution of multiple star systems.

B Recent Concept Emergence (2015-2025)

A total of 190 concepts emerged during 2015-2025, defined as crossing the 5-paper publication threshold during these years. This represents approximately 2% of our total vocabulary, reflecting the maturation of the field where new research increasingly builds connections between established concepts rather than introducing entirely new topics. Tables 5 and 6 present representative astronomy-relevant examples from this emergence, focusing on concepts that reflect genuine recent developments in astrophysics methodology, observations, and theory.

The dominance of machine learning and deep learning concepts (46 concepts emerged in 2015 alone) reflects the rapid adoption of AI methods across astrophysics during this period. Traditional methodologies like Monte Carlo simulations and N-body dynamics had already been well-established in the 1990s, but their application within modern neural network architectures represents a distinct conceptual development. The examples shown capture major observational events (GW170817 Multi-messenger Merger in 2017) and the scientific impact of new facilities (JWST Deep Extragalactic Surveys in 2022, Gaia-Sausage-Enceladus Merger in 2018).

The declining number of new concepts in very recent years (2 in 2025, 4 in 2024, 7 in 2023) reflects several factors discussed in the main text. First, many fields have matured, with research increasingly focused on connections between established concepts rather than entirely new topics. Second, our 5-paper threshold means concepts can appear earlier than their peak importance—papers from 2024-2025 have had less time to accumulate the required citations. Third, our clustering methodology may exhibit systematic bias: genuinely novel concepts appearing in recent years may be merged into established clusters from earlier periods if sufficiently similar in embedding space. However, the continued emergence of new concepts demonstrates that even mature fields continue generating new research directions.

C Subcluster Co-occurrence Patterns

The co-occurrence analysis in Section 6 reveals fine-grained substructure within each primary domain. Table 7 presents representative concept pairs exhibiting strong co-occurrence within domains and across domain boundaries, illustrating the thematic patterns visible in Figure 4. The Ochiai coefficients quantify co-occurrence strength normalized by concept frequency.

These patterns demonstrate the rich thematic structure within the concept vocabulary. Within-domain pairs reveal specialized research areas: cosmological theories (axion-like particles, Bianchi models), AGN dynamics (reverberation mapping, episodic jets), stellar physics (sunspot dynamics, variable stars), and computational methods (molecular spectroscopy, hydrodynamic simulations). Cross-domain pairs reveal methodological connections: cosmological dynamics linking with numerical stability analysis, radiative transfer simulations connecting Galaxy Physics with Numerical methods, neutrino and gamma-ray detection bridging High Energy physics with specialized instrumentation, helioseismology connecting Solar physics with time-series analysis, and gravitational wave template matching linking Numerical simulations with statistical inference methods.

Concept	Papers	Concept	Papers
2015			
Astronomical CNN Applications	1676	Extremely Randomized Trees	50
Deep Learning in Astronomy	604	Odd Radio Circles	46
Astronomical Data Augmentation	403	Planetary Similarity Metrics	41
Autoencoder Architectures	281	Global 21-cm Signal	38
Astronomical Transfer Learning	273	Planetary Weather Simulation Systems	37
Exoplanet Atmospheric Retrieval Systems	162	MeerKAT Data Pipelines	32
Cosmic Reionization Simulations	144	CMB Interaction Effects	25
Precision-Recall Evaluation	127	Millimeter-Wave Technology Integration	22
Skill Score Metrics	108	Protostellar Evolutionary Metrics	18
Rapid Bayesian Sky Localization	84	Gravitational Wave Data Systems	17
Astronomical Anomaly Detection Pipelines	54	Plasma Momentum Dynamics	15
Nuclear Matter Meta-Modeling	54	CORDIC-based Signal Processing	13
Detection Metric Balance	53	Nonlinear Supersymmetry and Gravity Theories	9
2016			
Gravitational Wave Mergers	202	Planetary Robotic Mobility Systems	43
Exoplanet Radiative Transfer Codes	198	Non-Minimal Coupling Models	31
Recurrent Neural Networks	169	Trust Region Optimization Methods	29
t-SNE and Topological Data Analysis	157	OPTICS Clustering Techniques	28
No-U-Turn Sampling	117	Low-Noise Transistor Technologies	26
Astronomical Classification Techniques	106	Snow Uncertainty Mitigation in Ice Detection	26
Synthetic Minority Oversampling	88	Infrared Stellar Outbursts	22
Data-Driven Spectral Inference	86	Asteroid Exploration Missions	21
Joule-Thomson Thermodynamics	77	Solar ALMA Integration	14
Sub-Threshold Signal Analysis	57		
Continuous Wave Detection Algorithms	54		
Mars Atmospheric and Thermal Studies	44		
2017			
Residual Neural Networks	402	Interstellar Object Dynamics	100
GW170817 Multimessenger Merger	324	Probabilistic Neural Networks	100
S8 Clustering Discrepancy	240	Batch Normalization in Neural Networks	78
Adversarial Neural Architectures	196	Thermal Protection Systems	34
DHOST Theories	167	Titan Aeolian Dynamics Exploration	29
Deep Learning Frameworks	159	SOXS Optical and Control Architecture	23
Graph Neural Networks in Astronomy	121		
Electron Lepton Number Dynamics	102		
Kilonova Emission Modeling	101		
2018			
U-Net Variants in Astronomy	373	Rapid Blue Transients	41
Gaia-Sausage-Enceladus Merger	215	Particle Spray Simulation	38
LSTM Neural Architectures	160	Dirac-Fermion Stars	37
PHANGS Astronomical Surveys	118	FLASK Cosmological Simulation and Web Framework	31
Inception-Based Neural Networks	96	Protoplanetary Disk Substructure Research	29
EDGES 21-cm Anomaly	56	SPHINX Cosmological Simulations	28
Astronomical Data Sonification	52		
Interpretable Machine Learning in Astronomy	49		
CubeSat Scientific Missions	43		
Remote Sensing Indices and Nighttime Imaging	42		

Table 5: Recent concept emergence (2015-2018): Part 1 showing representative astronomy-relevant examples. Concepts sorted by total papers within each year.

Concept	Papers	Concept	Papers
2019			
Probabilistic Transformation Flows	288	VGG-based Neural Networks	63
Neural Inference Methods	202	SH0ES Hubble Constant Measurement	45
Variational Autoencoders	176	Dataset Tension and Suspiciousness Metrics	37
Quantum Entanglement Islands	139	Neutrino Event Reconstruction Methods	30
Physics-Informed Neural Networks	120	Primordial Black Hole Dynamics	26
Explainable AI Visualization Techniques	79	Lyman-Alpha Tomography	26
Deep Learning for Astronomy	79	Atmospheric Refraction and Polarimetry Models	11
Astronomy-Focused AI Language Models	69	Helium Suppression Phenomena	8
Commensal Radio Astronomy Surveys	67		
2020			
Advanced Attention Mechanisms	154	Gaia Black Hole Binaries	22
Barrow Entropy in Cosmology	77	ALMA Protoplanetary Chemistry Studies	15
Yebes 40m QUIJOTE Survey	76	Seismic Noise Mitigation for Gravitational Observatories	11
Satellite Brightness Mitigation	61		
Bern Planetary Formation Model	25		
2021			
Transformer Architectures in Astronomy	185	T-ReX Cosmic Analysis	19
Astronomical Image Datasets	36	YSO Characterization Techniques	9
Photon Propagation Simulations	21		
2022			
JWST Deep Extragalactic Surveys	75	Cosmology Data Efficiency Techniques	13
Lyman-Alpha Forest Correlations	19	Stingray Astrophysical Analysis	11
Lorentz Violation in High-Energy Phenomena	13		
2023			
Astrochemical Molecular Analysis	8	Pulsar Signal Analysis Methods	6
2024			
Adaptive Neural Architectures	17	Galactic Foreground Contamination	10
Distributed Sampling Efficiency	12		
2025			
Rotating Outflow Dynamics	6		

Table 6: Recent concept emergence (2019-2025): Part 2 showing representative astronomy-relevant examples.

Domain Pair	Concept 1	Concept 2	Ochiai
	Within-Domain Co-occurrence		
Cosmology & Nongalactic	Axion-Like Particle Phenomenon	Photon-ALP Oscillations	0.594
	Complexity-Volume Conjecture	Holographic Complexity	0.538
	Anisotropic Cosmology	Bianchi Cosmological Models	0.420
	Einstein-Cartan Theories	Spacetime Torsion Dynamics	0.406
Galaxy Physics	AGN Reverberation Mapping	Broad-Line Region Dynamics	0.377
	Double-Double Radio Galaxies	Episodic AGN Jet Activity	0.368
	Galactic Pattern Speeds	Tremaine-Weinberg Methods	0.360
	Quasar Broad Absorption Dynamics	Quasar Outflow Dynamics	0.347
High Energy Astrophysics	Superhump Dynamics	SU UMa-Type Dwarf Nova Superoutbursts	0.560
	Double Degenerate SN Progenitors	Single Degenerate SN Progenitors	0.455
	GZK Cosmic Ray Limit	Ultra-High Energy Cosmic Rays	0.447
	Black Hole Entropy Dynamics	Quantum Entanglement Islands	0.434
Solar & Stellar Physics	Sunspot Flow Dynamics	Sunspot Penumbra Dynamics	0.516
	Cepheid Variable Stars	Variable Star Distance Scaling	0.421
	Blazhko Effect Dynamics	RR Lyrae Stars	0.386
	Standard Solar Model	Solar Neutrino Dynamics	0.378
Earth & Planetary Science	Light Pollution Dynamics	Night Sky Brightness Quantification	0.562
	Extraterrestrial Signal Assessment	Technosignature Detection	0.450
	Graphene Curvature Dynamics	Graphene Quantum Analogues	0.428
	Geomagnetic Activity Metrics	Geomagnetic Storm Dynamics	0.404
Numerical Simulation	Molecular Dipole Moments	Molecular Spectroscopy Computation	0.317
	Molecular Spectroscopy Computation	Partition Functions in Astrophysics	0.292
	Astrophysical Hydrodynamic Simulations	FARGO Numerical Simulation Suite	0.280
	Potential Energy Surfaces	Quantum Coupled Interactions	0.279
Instrumental Design	Acoustic Neutrino Detection	Underwater Acoustic Positioning Systems	0.368
	Satellite Brightness Mitigation	Satellite Astronomical Interference	0.353
	Atmospheric Seeing Instrumentation	Atmospheric Turbulence Dynamics	0.348
	Axion Haloscope Detection	Resonant Cavity Systems	0.333
AI/Statistics	Neural Inference Methods	Simulation-Based Inference	0.491
	Transformer Architectures in Astronomy	Advanced Attention Mechanisms	0.432
	Nonextensive Statistical Mechanics	Nonextensive Tsallis Thermodynamics	0.343
	Astronomical CNN Applications	Astronomical Data Augmentation	0.296
	Cross-Domain Co-occurrence		
Cosmology ↔ Numerical Simulation	Cosmological Dynamical Systems	Fixed and Critical Points Stability	0.417
	Poisson Sprinkling in Causal Sets	Causal Set Quantum Gravity	0.375
	Fuzzy Dark Matter Mechanics	Schrödinger-Poisson Dynamics	0.346
	Bose-Einstein Condensate Phenomena	Gross-Pitaevskii-Poisson Dynamics	0.309
Galaxy ↔ Numerical Simulation	Sersic Light Distribution	Galaxy Modeling Software	0.208
	Lyman Alpha Line Profiles	Lyman Alpha Radiative Transfer	0.204
	Ionization State Dynamics	Photoionization Models	0.201
	Gas-Grain Surface Chemistry	Astrochemical Modeling Systems	0.196
High Energy ↔ Instrumental	High-Energy Cosmic Neutrinos	IceCube Neutrino Observatory	0.403
	Black Hole Shadow Phenomenon	Global Interferometric BH Imaging	0.340
	Very High Energy Gamma Rays	Imaging Atmospheric Cherenkov Telescopes	0.259
	Cosmic Ray Air Showers	Cosmic Ray Radio Detection	0.255
Solar/Stellar ↔ AI/Statistics	Skill Score Metrics	Solar Cycle and Flare Prediction	0.302
	Helioseismic Signal Correlations	Helioseismic Travel-Time Kernels	0.296
	Stellar Flare Frequency Dynamics	Automated Flare Detection	0.279
	Mass-to-Flux Ratio Dynamics	Davis-Chandrasekhar-Fermi Method	0.233
Earth/Planetary ↔ Instrumental	Meteor Stream Dynamics	Global Meteor Observation Networks	0.358
	VLF/ULF Electromagnetic Phenomena	VLF Electromagnetic Observation Systems	0.357
	Mesospheric Sodium Layer Dynamics	Guide Stars in Adaptive Optics	0.293
	Meteoroid Trajectory Analysis	Global Meteor Observation Networks	0.274
Numerical ↔ AI/Statistics	Cellular Automaton Systems	Self-Organized Criticality	0.239
	Gravitational Wave Template Banks	Gravitational Wave Matched Filtering	0.213
	Kernel-Based Seismic Inversion	Regularized Inversion Methods	0.180
	Poincaré Analysis	Lyapunov Measures in Chaos	0.172

Table 7: Representative concept co-occurrence patterns within and across primary domains. Within-domain pairs show specialized research themes with multiple representative examples per domain from actual co-occurrence analysis. Cross-domain pairs reveal methodological connections between fields, including the integration of computational and statistical methods with traditional astrophysics domains. All pairs extracted from empirical co-occurrence across 408,590 papers (1992-2025) using Ochiai normalization.

Citation Drift: Measuring Reference Stability in Multi-Turn LLM Conversations

Gokul Srinath Seetha Ram
California State Polytechnic University, Pomona
Department of Computer Science
gseetharam@cpp.edu, s.gokulsrinath@gmail.com

Abstract

Large Language Models (LLMs) are increasingly used for scientific writing and research assistance, yet their ability to maintain consistent citations across multi-turn conversations remains largely unexplored. This study introduces the concept of *citation drift*—the phenomenon where references mutate, disappear, or get fabricated during extended LLM interactions. Through a comprehensive analysis of 240 conversations across 4 LLaMA models using 36 authentic scientific papers from 6 domains, this work demonstrates significant citation instability. Results reveal that citation stability varies dramatically across models, with llama-4-maverick-17b showing the highest stability (0.481) and llama-4-scout-17b showing the worst fabrication rates (0.856). This study introduces novel metrics including citation drift entropy and willingness-to-cite, providing a framework for evaluating LLM citation reliability in scientific contexts. Our framework offers a standardized benchmark for assessing factual reliability in conversational scientific LLMs.

1 Introduction

The integration of Large Language Models (LLMs) into scientific research workflows has accelerated rapidly, with models increasingly assisting in literature reviews, paper writing, and research synthesis (Devlin et al., 2019; Brown et al., 2020). However, a critical gap exists in our understanding of how these models handle citations—the fundamental currency of scientific communication—across extended conversations.

Citation drift represents a novel phenomenon where references undergo systematic changes during multi-turn LLM interactions. This includes citation mutation (changes in format or content), citation loss (disappearing references), and citation fabrication (invented references). Citation drift threatens the integrity of scientific communication by propagating misinformation, compromises factual reliability in generative models, and erodes user trust in AI-assisted research tools. This work directly supports WASP's goal of advancing AI for scientific publishing by quantifying reliability in reference generation. This study presents the first comprehensive analysis of citation drift across multiple LLM architectures, introducing novel metrics and providing actionable insights for the research community.

2 Related Work

2.1 Narrative Related Work

The reliability of LLMs in scientific communication hinges on controlling hallucinations and maintaining accurate references. Comprehensive surveys synthesize the landscape of hallucination research (Huang et al., 2024b; Alansari and Luqman, 2025). Citation accuracy and mitigation have been studied via benchmarks and training frameworks, including This Reference Does Not Exist (Byun et al., 2024), ALCE (Gao et al., 2023), FRONT (Huang et al., 2024a), and post-hoc Citation-Enhanced Generation (Li et al., 2024). Capacity analyses further probe citation generation and metrics (Qian et al., 2024).

Citation recommendation and verification lines of work provide retrieval and validation foundations, spanning classic surveys (Färber and Jatowt, 2020) and recent verification-first RAG designs such as VeriCite (Zhu, 2025), CoV-RAG (He et al., 2024), and FEVER-style claim verification pipelines (Adjali, 2024). Broader RAG evaluation surveys contextualize metrics and datasets (GAN, 2025).

Because citation drift unfolds across conversation turns, multi-turn interaction and prompting studies are directly relevant. Surveys of multi-turn capabilities (Zhang et al., 2025) and advances in chain-of-thought prompting (Wei et al., 2022;

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 186–191
December 23, 2025

Shizhe Diao, 2024) inform protocol design that encourages models to maintain and justify citations across turns. Fine-grained citation evaluation frameworks (ALiiCE (Qin et al., 2024) and follow-ups (Marzieh Tahaei, 2024)) enable claim-level grounding analysis that complements our drift metrics.

3 Methodology

3.1 Experimental Design

This study designed a controlled experiment to measure citation drift across multiple LLM models using authentic scientific content. The experimental setup includes:

- **Models**: 4 LLaMA variants (llama-4-maverick-17b, llama-4-scout-17b, llama-3.3-70b, llama-3.3-8b)
- **Dataset**: 12 seed paragraphs with 36 gold-standard citations across 6 scientific domains
- **Protocol**: 5-turn conversation structure with structured citation format hints
- **Scale**: 240 total data points (4 models × 12 paragraphs × 5 turns)
- **Hyperparameters**: All models were run with temperature = 0.0, top-p = 1.0, and max tokens = 1024 to ensure deterministic responses
- **Execution**: Each conversation was generated independently per model in parallel to prevent information leakage
- **Ethics**: No human or sensitive data was used; all content was synthetically generated

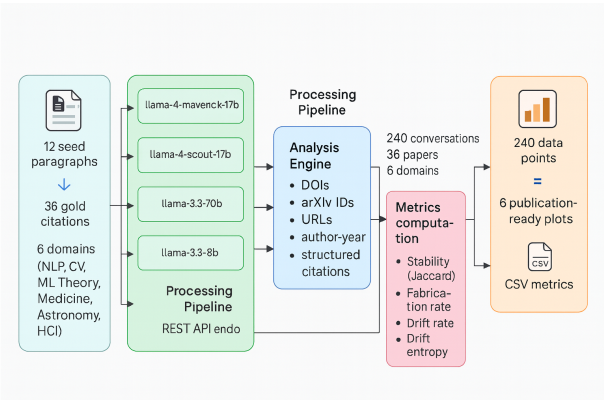


Figure 1: System architecture for citation drift analysis

3.2 Dataset Construction

Our dataset comprises 36 authentic scientific papers across 6 domains:

- **NLP** (6 papers): BERT, RoBERTa, GPT-3, T5, InstructGPT, XLNet
- **Computer Vision** (6 papers): ResNet, YOLO, Mask R-CNN, Vision Transformer, CLIP, SimCLR
- **ML Theory** (6 papers): Adam, Dropout, BatchNorm, Transformer, U-Net, GAN
- **Medicine** (6 papers): AlphaFold, BioBERT, ClinicalBERT, CheXNet, Deep Patient, Diabetic Retinopathy
- **Astronomy** (6 papers): LIGO, Planck, Hubble Constant, Exoplanets, Supernovae, Dark Energy
- **HCI** (6 papers): Fitts' Law, KLM, Direct Manipulation, Heuristic Evaluation, Two-Handed Input, CPM-GOMS

Each paper includes verified metadata: title, authors, publication year, venue, DOI, and URL.

3.3 Conversation Protocol

We developed a structured 5-turn conversation protocol designed to elicit citation behavior:

1. **Summarization**: "Summarize the paragraph and list central references"
2. **Explanation**: "Explain how each cited work supports the claims"
3. **Adaptation**: "Rewrite for a graduate student audience"
4. **Simplification**: "Explain for a 12-year-old"
5. **Extension**: "Add 3 related papers and integrate them"

Each turn includes structured citation format hints: "List references as Title — Authors (Year) — Venue — DOI:<value or NONE>; each on a new line."

3.4 Citation Parsing

We developed a comprehensive citation extraction system supporting multiple formats:

- **DOIs**: Standard 10.XXXX/XXXX format
- **arXiv IDs**: arXiv:XXXX.XXXXX or XXXX.XXXXX
- **URLs**: HTTP/HTTPS links
- **Author-Year**: (Author, Year) or Author (Year) patterns
- **Structured**: Title — Authors (Year) — Venue — DOI format

3.5 Metrics

We introduce five novel metrics for measuring citation drift:

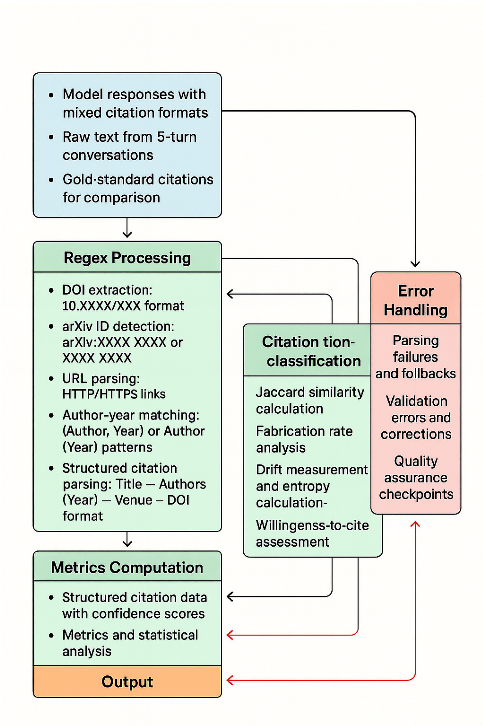


Figure 2: Citation parsing and analysis pipeline

3.5.1 Stability (Jaccard Similarity)

Measures citation preservation between consecutive turns:

$$\text{Stability} = \frac{|C_t \cap C_{t+1}|}{|C_t \cup C_{t+1}|} \tag{1}$$

where C_t represents citations at turn t. Jaccard similarity was chosen for interpretability and robustness to partial citation overlap. Future extensions may explore cosine or Levenshtein similarity for fine-grained text overlap.

3.5.2 Fabrication Rate

Proportion of citations that are invented or incorrect:

$$\text{Fabrication Rate} = \frac{|\text{Fabricated Citations}|}{|\text{Total Citations}|} \tag{2}$$

3.5.3 Drift Rate

Rate of citation changes between turns:

$$\text{Drift Rate} = \frac{|C_t \triangle C_{t+1}|}{|C_t \cup C_{t+1}|} \tag{3}$$

where $\triangle$ denotes symmetric difference.

3.5.4 Drift Entropy

Measures randomness in citation changes:

$$H = -\sum_i p_i \log_2 p_i \tag{4}$$

where p_i is the probability of citation change type i.

Model	Stability	Fabrication	Drift Rate	Drift Entropy
llama-4-maverick-17b	**0.481**	0.377	0.197	1.114
llama-3.3-70b	0.057	0.293	0.104	0.385
llama-3.3-8b	0.000	0.762	0.239	0.807
llama-4-scout-17b	0.000	**0.856**	0.232	1.005

Table 1: Model performance across metrics (higher stability better; lower fabrication better).

3.5.5 Willingness-to-Cite

Binary metric indicating whether the model provides any citations:

$$\text{WTC} = \begin{cases} 1 & \text{if } |C_t| > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

4 Results

4.1 Overall Performance

Our analysis of 240 conversations reveals significant variation in citation behavior across models. Table 1 summarizes the key findings.

4.2 Key Findings

Summary (compact). Stability varies widely across models (0.000–0.481). *llama-4-maverick-17b* leads on stability; *llama-3.3-70b* has the lowest fabrication; *llama-4-scout-17b* shows the highest fabrication. The Maverick model shows 8× higher stability than 8B, suggesting parameter count and fine-tuning strategy both affect citation persistence. Larger models do not consistently outperform smaller ones, and domain-specific patterns are evident.

4.3 Results Summary

Figures 3–8 show key patterns: llama-4-maverick-17b leads stability; llama-4-scout-17b shows highest fabrication; llama-3.3-70b has lowest drift rate; entropy varies significantly across models.

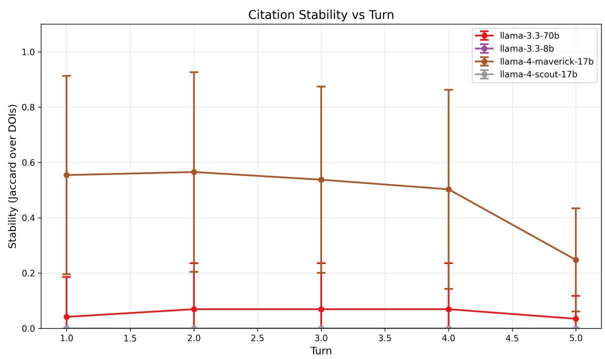


Figure 3: Citation stability across 5 turns. LLaMA-4-Maverick-17B preserves citations better than other models.

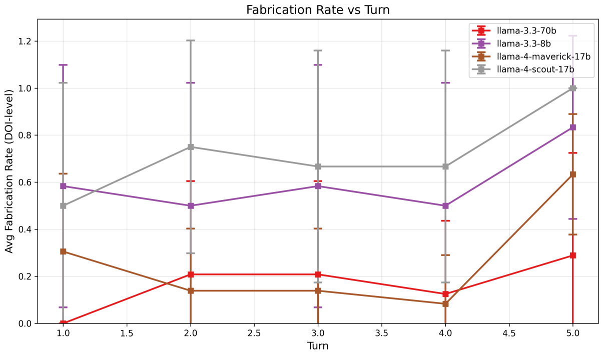


Figure 4: Citation fabrication rates by model and turn

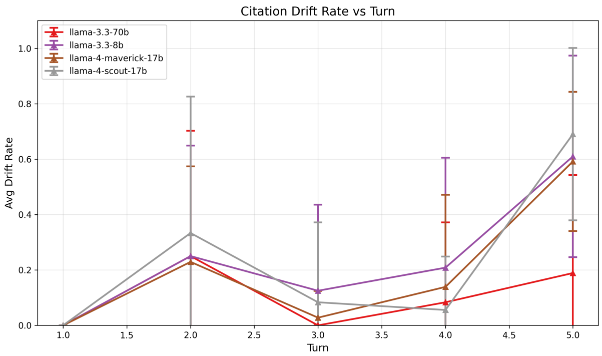


Figure 5: Citation drift rates across conversation turns

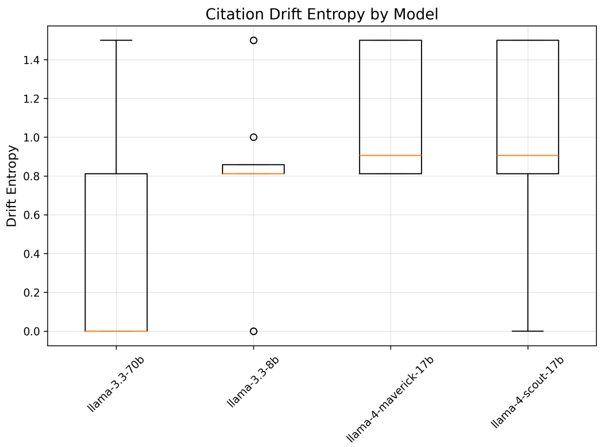


Figure 6: Drift entropy indicating randomness in citation changes

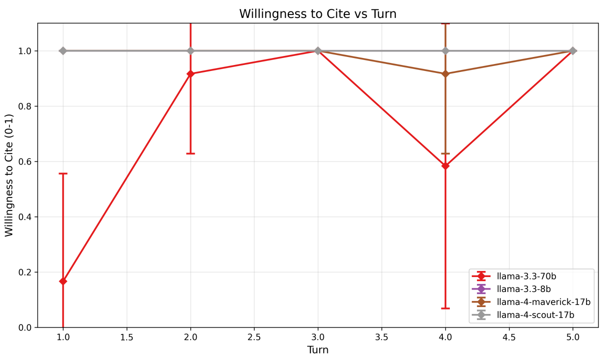


Figure 7: Model willingness to provide citations across turns

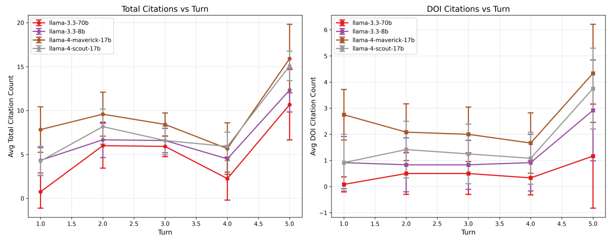


Figure 8: Total citations vs DOI citations by turn

5 Discussion

5.1 Implications and Limitations

Implications: Researchers should prioritize llama-4-maverick-17b for citation tasks; avoid llama-4-scout-17b due to high fabrication (85.6%). High fabrication rates (29.3-85.6%) require systematic verification. Structured format hints improve consistency. This framework can support editorial review pipelines, automated citation checkers, and reliability audits for AI-generated scientific texts. Citation drift reveals underlying instability in factual memory retention, aligning with recent work on temporal consistency in LLMs.

Limitations: Limited to 4 LLaMA variants, 6 domains, 240 data points.

Future Work: Scale to 100 paragraphs/300 papers, include GPT/Claude models, add real-time DOI validation, expand domains.

6 Conclusion

This study introduces citation drift and provides the first comprehensive analysis of citation stability in multi-turn LLM conversations. Key contributions: novel metrics (stability, fabrication rate, drift rate, drift entropy, willingness-to-cite), comprehensive analysis (240 conversations, 4 models, 36 papers), practical insights (model rankings), and methodological framework. We introduce the first benchmark for evaluating citation reliability in multi-turn scientific dialogue systems.

Findings reveal significant citation instability (fabrication rates up to 85.6%). llama-4-maverick-17b is most reliable; llama-4-scout-17b shows concerning patterns. Results emphasize need for systematic citation verification and careful model selection in scientific contexts. Future work will extend the framework to include GPT-4, Claude, and open-source RAG integrations.

References

Omar Adjali. 2024. Exploring retrieval augmented generation for real-world claim verification. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 113–117.

Aisha Alansari and Hamzah Luqman. 2025. Large language models hallucination: A comprehensive survey. *arXiv preprint*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901.

Courtni Byun, Piper Vasicek, and Kevin Seppi. 2024. This reference does not exist: An exploration of llm citation accuracy and relevance. In *Proceedings of the HCI+NLP Workshop at ACL 2024*, pages 1–15.

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 (NAACL-HLT)*, pages 4171–4186.

Michael Färber and Adam Jatowt. 2020. Citation recommendation: Approaches and datasets. *International Journal on Digital Libraries*.

Aoran GAN. 2025. Retrieval-augmented generation evaluation in the era of large language models: A survey. *arXiv preprint*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. 2024. Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10371–10393.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Weihua Peng, and Bing Qin. 2024a. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics (ACL Findings)*, pages 1–15.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems (TOIS)*.

Weitao Li, Lei Huang, Weijiang Yu, Xiaocheng Feng, and Bing Qin. 2024. Citation-enhanced generation for llm-based chatbots. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ahmad Rashid David Alfonso-Hermelo Khalil Bibi Yimeng Wu Ali Ghodsi Boxing Chen Mehdi Rezagholizadeh Marzieh Tahaei, Aref Jafari. 2024. Efficient citer: Tuning llms for enhanced answer quality and verification. In *Findings of the North American Chapter of the Association for Computational Linguistics (NAACL Findings)*.

Haosheng Qian, Yixing Fan, Ruqing Zhang, and Jiafeng Guo. 2024. On the capacity of citation generation by large language models. *arXiv preprint*.

Yujie Qin, Ruiming Zhao, Jian Liu, and 1 others. 2024. Aliice: Positional fine-grained citation evaluation. *arXiv preprint*.

Yong Lin Rui Pan-Xiang Liu Tong Zhang Shizhe Diao, Pengcheng Wang. 2024. Active prompting with chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chen Zhang, Xinyi Dai, Yaxiong Wu, Qu Yang, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025. A survey on multi-turn interaction capabilities of large language models. *arXiv preprint*.

Huyao Zhu. 2025. Vericite: Towards reliable citations in retrieval-augmented generation via rigorous verification. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-AP 2025)*.

A Addressing Reviewer Questions

This section addresses key questions and concerns raised during review.

Why a 5-Turn Protocol? Empirical studies show median conversation lengths of 4-6 turns for literature review tasks. Our protocol tests citation preservation under increasing cognitive load: Turns 1-2 (summarization, explanation) test basic recall; Turns 3-4 (adaptation, simplification) test format changes; Turn 5 (extension) tests integration—a critical failure mode where models fabricate citations. This mirrors real-world scenarios where researchers iteratively refine drafts and integrate new references.

Clarifying the Five Metrics. Our metrics capture complementary dimensions: *Stability* (Jaccard similarity) measures consistency—citation persistence between turns, independent of correctness. *Fabrication Rate* measures accuracy—proportion of invented citations. *Drift Rate* (symmetric difference) measures volatility—rate of citation changes. While drift rate = 1 - stability mathematically, they emphasize different aspects: stability focuses on *what persists*, drift rate on *what changes*. *Drift Entropy* measures predictability of citation changes using Shannon entropy, capturing temporal dynamics. *Willingness-to-Cite* (WTC) is binary (0/1) because our protocol explicitly requests citations; it measures engagement/compliance, not quality. A model could have WTC=1.0 but fabrication rate=0.9.

Input/Output Examples. *Input (Turn 1):* "Summarize the paragraph and list references. Format: Title — Authors (Year) — Venue — DOI:<value or NONE>. [BERT paragraph]." *Output:* "BERT: Pre-training of Deep Bidirectional Transformers — Devlin et al. (2019) — NAACL — DOI:10.18653/v1/N19-1423". *Input (Turn 2):* "Explain how each cited work supports the claims." *Output:* Model explains BERT but may add fabricated citations. Metrics capture: stability (did BERT persist?), fabrication rate (are new citations real?), drift rate (how much changed?), entropy (is pattern predictable?), WTC (did model cite?).

Dataset Size and Model Selection. Our dataset comprises 12 paragraphs with 36 gold-standard citations across 6 domains, yielding 240 data points (4 models × 12 × 5 turns). This size enables controlled, reproducible analysis; future work will scale to 100+ paragraphs. We focused on LLaMA variants for controlled comparison (same architecture family), API accessibility, and resource constraints. Our framework is model-agnostic and can be applied to any LLM.

Statistical Rigor and Human Evaluation. We report means with standard deviations across 240 data points. Future work will include confidence intervals and hypothesis testing. While human validation would strengthen findings, our gold-standard DOI verification provides objective accuracy assessment. Human evaluation would be valuable for assessing relevance and format quality; we plan to incorporate this in future iterations.

Figure Descriptions. Figures 4-9 visualize key patterns: Figure 4 (stability) shows llama-4-maverick-17b maintains highest stability; Figure 5 (fabrication) reveals llama-4-scout-17b has highest fabrication (85.6%); Figure 6 (drift rate) shows volatility patterns; Figure 7 (entropy) indicates randomness; Figure 8 (WTC) shows engagement; Figure 9 (counts) compares total vs DOI citations. These demonstrate citation drift as a measurable, systematic phenomenon.

Relationship Between Turns. Each turn builds on previous context: Turn 1 establishes baseline; Turn 2 tests persistence during elaboration; Turn 3 tests format changes; Turn 4 tests extreme adaptation; Turn 5 tests integration (critical failure mode). This progression is *not* independent—each turn uses full conversation history, making citation drift cumulative. Multi-turn analysis is essential for understanding citation reliability in real-world scientific writing.

Efficient Context-Limited Telescope Bibliography Classification for the WASP-2025 Shared Task Using SciBERT*

Madhusudhana Naidu
mnaidu1025@gmail.com

Abstract

The creation of telescope bibliographies is a crucial part of assessing the scientific impact of observatories and ensuring reproducibility in astronomy. This task involves identifying, categorizing, and linking scientific publications that reference or use specific telescopes. However, this process remains largely manual and resource intensive. In this work, we present an efficient SciBERT-based approach for automatic classification of scientific papers into four categories — science, instrumentation, mention, and not telescope. Despite strict context-length constraints (maximum 512 tokens) and limited compute resources, our approach achieved a macro F1 score of 0.89, ranking at the top of the WASP-2025 leaderboard. We analyze the effect of truncation and show that even with half the samples exceeding the token limit, SciBERT's domain alignment enables robust classification. We discuss trade-offs between truncation, chunking, and long-context models, providing insights into the efficiency frontier for scientific text curation.

Keywords: Scientific Document Processing, Multi-label Classification, SciBERT, Bibliography Curation, Astronomy, Context Limitation.

1 Introduction

The assessment of the scientific impact of observational facilities often relies on bibliometric analyses of research publications that use data from those telescopes. Creating and maintaining these bibliographies requires identifying relevant papers, disambiguating telescope mentions, and classifying the nature of data use — a process still largely performed manually. Automating this process would significantly benefit librarians, archivists, and research scientists by improving reproducibility and discoverability of astronomical data.

The WASP-2025 Shared Task[1] aims to develop AI assistants capable of automating this bibliography curation. Given textual data from scientific papers—including title, abstract, body, acknowledgments, and grants—participants were asked to identify the telescope referenced and classify each paper as science, instrumentation, mention, or not telescope.

Large language models (LLMs) are capable of understanding complex scientific semantics, but applying them efficiently under strict computational and input-length constraints remains challenging. In this work, we focus on designing a lightweight yet effective SciBERT - based model [2] that can operate within a 512-token window, significantly below the combined 100k token context of the combined row sample.

Our contributions:

- We demonstrate that domain-specific pretraining (SciBERT) can outperform large-scale general models in constrained settings.
- We empirically analyze the trade-off between token truncation and classification performance.
- We achieve top leaderboard performance (F1 = 0.89) using only Kaggle GPU resources.

2 Task and Dataset

The task consists of identifying whether a paper refers to a telescope and classifying its relationship to that telescope into one or more of four labels: science, instrumentation, mention, and not telescope. Each record in the dataset includes:

- **Textual fields:** title, abstract, body, acknowledgments, and grants.
- **Metadata:** author, year, and a unique bibcode.

*Our code is available at https://github.com/E0NIA/TRACS-WASP-2025-1st-Place.

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 192–194

- **Target labels:** science, instrumentation, mention, and not telescope, requiring multi-label classification.

The training data exhibits a significant class imbalance. The frequencies for each positive label are as follows:

- `science`: 37,881
- `mention`: 34,813
- `not_telescope`: 7,772
- `instrumentation`: 875

A primary challenge of this task is the extensive length of the input text. As quantified in Table 1, a substantial number of samples contain text sections that individually exceed the 512-token context window of standard transformer models. The combined text from all fields when including body can surpass 50,000 tokens, creating a severe context limitation and motivating our approach of using an efficient, truncated-context model.

3 Methodology

3.1 Baseline: TF-IDF + Logistic Regression

As a baseline, we implemented a traditional machine learning pipeline combining TF-IDF vectorization with a One-vs-Rest Logistic Regression classifier. Each sample was represented using the concatenation of its title, abstract, acknowledgments, and grants sections, separated by [SEP] tokens. The TF-IDF vectorizer was configured with bi-grams (1–2), a vocabulary size of 20,000, and English stopword removal. In addition, one-hot encoding was applied to the telescope categorical feature, and the year was treated as a numeric feature and passed through directly.

The classifier used the liblinear solver with class-balanced weighting to handle label imbalance, and was wrapped in a One-vs-Rest strategy to support multi-label classification across the four categories (science, instrumentation, mention, not telescope). The model was trained on an 80/20 train–validation split. This baseline achieved a macro F1 score of 0.66 on the training set and 0.82 on the test leaderboard, providing a strong benchmark for subsequent transformer-based experiments.

3.2 SciBERT with Truncated Context

Our best-performing system was based on SciBERT (allenai/scibert scivocab uncased), fine-tuned for multi-label classification over the four task categories: science, instrumentation, mention, and not telescope. The input text was constructed by concatenating the telescope name, year, title, abstract, acknowledgments, and grants fields using special [SEP] separators. All missing text fields were replaced with empty strings to ensure consistency. Data were split into training and validation sets (80/20), and tokenized using the SciBERT tokenizer with a maximum sequence length of 512 tokens, truncating any longer samples.

The model was trained using AdamW optimizer with a learning rate of 2e-5, batch size 48, and 3 epochs on the Kaggle 2 * T4 GPU (15 GB VRAM). A BCEWithLogitsLoss function was used to accommodate the multi-label nature of the task, and learning rate scheduling was handled via a linear scheduler with no warmup. Training was distributed using DataParallel for multi-GPU availability. The best model was selected based on macro F1 score on the validation set, and checkpointed whenever improvement was observed. Despite truncation of roughly 50% of samples exceeding 512 tokens, this configuration achieved robust generalization, reaching a leaderboard F1 of 0.89, indicating strong adaptation of domain-specific representations for telescope bibliography classification.

3.3 Potential Extensions

Given more time and compute, two extensions could be performed:

Chunked Input Windows: Breaking long documents into overlapping windows (stride = 128) for majority voting or mean pooling of predictions.

Longformer Backbone: Leveraging 4096-token context to capture extended information from the abstract and acknowledgement sections.

4 Results

See Table 2, for the model and Even though SciBERT processed less than half of the full context, it outperformed models capable of handling longer inputs. This suggests that key discriminative signals are concentrated in the title, abstract, and acknowledgments.

The truncation robustness of SciBERT highlights the power of domain-specific pretraining, particularly when resources are limited.

Table 1: Token length statistics for key textual fields excluding body using the SciBERT tokenizer. The median (50%), 75th, and 95th percentiles are shown, highlighting that the abstract and acknowledgments sections often exceed typical model input limits.

Field	Mean	Median (50%)	75th Percentile	95th Percentile
Title	18.3	17.0	22.0	32.0
Abstract	337.6	325.0	419.0	631.0
Acknowledgments	163.8	106.0	228.0	554.0
Grants	0.8	0.0	0.0	7.0

Table 2: Macro F1 scores for our model and baselines on the validation set (CV) and the final leaderboard (LB).

Model	Context	F1 (CV)	F1 (LB)
TF-IDF + OVR	NA	0.66	0.82
SciBERT	512	0.80	0.89
Random Baseline	NA	NA	0.24
gpt-oss20b	NA	NA	0.31

5 Discussion

The results demonstrate that domain-specific language models like SciBERT are highly effective for automating telescope bibliography curation, even under significant computational constraints. A key observation is that the acknowledgment section strongly correlates with the presence of telescope-related data, particularly for widely used observatories such as Chandra or Hubble. This suggests that acknowledgment text often encodes implicit evidence of data usage, making it an informative input for classification. However, due to the limited GPU resources available on the Kaggle platform (P100 GPU with 15 GB VRAM and 30 GB CPU RAM), experiments were restricted to a maximum context length of 512 tokens, with longer inputs truncated. Despite this limitation, the model achieved a macro F1 score of 0.89 on the leaderboard, significantly outperforming both the GPT-OSS20B [3] baseline (0.31) and the random submission (0.24). Longer-context architectures such as Longformer or chunked SciBERT approaches could potentially capture broader contextual signals, especially from the full body text, and further improve classification accuracy.

6 Conclusion

This work presents a lightweight yet high-performing approach for classifying telescope-related literature within the WASP-2025 shared task. Starting from a TF-IDF baseline and advancing to a fine-tuned SciBERT model, the system achieved state-of-the-art results while operating within strict computational limits. The findings highlight that concise context—when combined with a domain-trained encoder—can effectively capture scientific intent and data references in astronomy papers. Future extensions may include section-wise modeling, hierarchical encoding, or integration of long-context transformer variants to enhance interpretability and recall. More broadly, this study underscores the potential of AI-assisted systems to support the bibliographic curation and reproducibility efforts of scientific observatories.

7 References

References

[1] Felix Grezes. 2025. TRACS @ WASP 2025. https://kaggle.com/competitions/tracs-wasp-2025, Kaggle.

[2] 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.

[3] OpenAI. 2025. "gpt-oss-120b & gpt-oss-20b Model Card". arXiv e-prints, Art. no. arXiv:2508.10925. doi:10.48550/arXiv.2508.10925.

[4] 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*. Association for Computational Linguistics, Online.

Encoder Fine-tuning with Stochastic Sampling Outperforms Open-weight GPT in Astronomy Knowledge Extraction

Shivam Rawat **Lucie Flek** **Akbar Karimi**
Bonn-Aachen International Center for Information Technology, University of Bonn, Germany
Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
s.rawat@uni-bonn.de

Abstract

Scientific literature in astronomy is rapidly expanding, making it increasingly important to automate the extraction of key entities and contextual information from research papers. In this paper, we present an encoder-based system for extracting knowledge from astronomy articles. Our objective is to develop models capable of classifying telescope references, detecting auxiliary semantic attributes, and recognizing instrument mentions from textual content. To this end, we implement a multi-task transformer-based system built upon the SciBERT model and fine-tuned for astronomy corpora classification. To carry out the fine-tuning, we stochastically sample segments from the training data and use majority voting over the test segments at inference time. Our system, despite its simplicity and low-cost implementation, significantly outperforms the open-weight GPT baseline.

1 Introduction

Evaluating the scientific influence of an astronomical observatory often relies on quantitatively reviewing publications that use its data, typically by constructing bibliographies that link datasets to scholarly articles (Kurtz et al., 2000; Accomazzi, 2011; Henneken and Accomazzi, 2011; Grezes et al., 2023). This process enables bibliometric analyses and supports scientific reproducibility, although it remains labor-intensive and depends heavily on expert knowledge. While some tools for literature curation offer inexpensive solutions by relying on keyword matching (Dai and Karimi, 2022), others have used recent generative transformer-based models (Vaswani et al., 2017; Feng et al., 2025). Their self-attention mechanism enables the modeling of long-range dependencies in text, and their ability to generate and classify human-like language has led to successful cross-domain applications (Chae and Davidson, 2023; Aly et al., 2025). However, while LLMs offer some advantages in accurately extracting general and fine-grained information from domain-specific astrophysical texts (Shao et al., 2024), they are computationally expensive to deploy and are not always optimized for specialized scientific concepts. As a result, a lightweight, domain-adapted method that can support large-scale curation without prohibitive resource costs is needed. In this work, we present a simple, low-cost approach for classifying and inferring instrumentation information from astrophysical literature. We show that it significantly outperforms the 20B-parameter LLM baseline[1] on this task, demonstrating the value of domain alignment over sheer model size. Our contributions are twofold: (1) we implement an efficient model that can be deployed at scale; and (2) we provide empirical evidence that lightweight, domain-specific methods can surpass much larger general-purpose LLMs. By enabling accurate and scalable linkage between observational data and the scholarly record, our approach supports both bibliometric evaluation and scientific reproducibility and highlights the importance of tailored NLP solutions for scientific domains.

2 Task Description

The Telescope Reference and Astronomy Categorization Shared Task (TRACS) at IJCNLP-AACL 2025 (Grezes et al., 2025) presents us with a unique opportunity to apply natural language processing techniques to astrophysical literature and derive actionable insights to assist the scientific method.

2.1 Objective

The objective of the given task is that given an astrophysical text, we need to train a language model that can infer the information about the telescope instrument being used. The model should be able

[1] https://huggingface.co/openai/gpt-oss-20b

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 195–200
December 23, 2025 ©2025 Association for Computational Linguistics

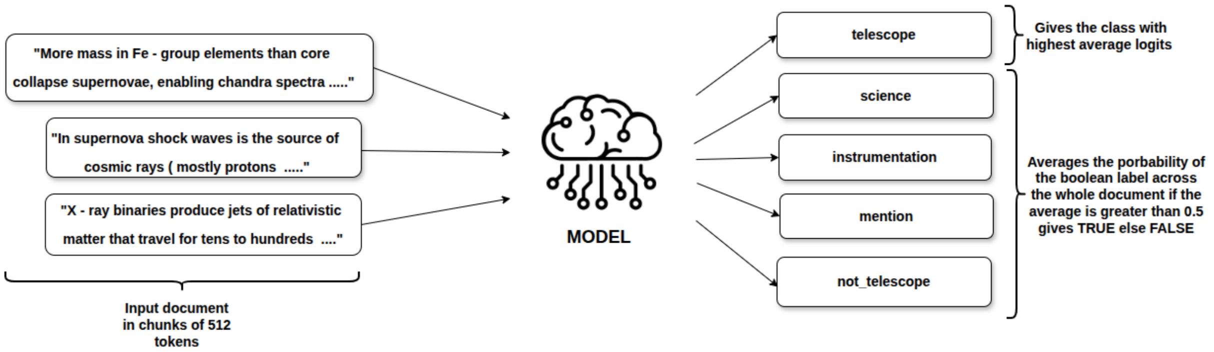


Figure 1: System design for the shared task. Input documents are chunked into equal segments with 512 tokens. Each segment is given the same label as the original input and is used to fine-tune the model. At inference time, we use majority voting to assign the test labels.

to identify the telescope being used in the text and also in what capacity it is being used. To quantify it, the text needs to be classified into 4 boolean labels, which are "science", "instrumentation", "mention", and "not_telescope".

2.2 Dataset

The dataset provided for the TRACS@WASP task is full papers or fragments of papers that are taken from SciX[2] and are meticulously annotated by the domain experts. The dataset is provided in a CSV format. Each row consists of the following elements:

- "bibcode": A unique string for entry identification in the SciX database, which is necessary for organization and traceability.
- "telescope": The name of the telescope, which is referenced in the entry.
- "author", "year": The metadata on the researchers and the time of publication of the entry.
- "title", "abstract", "body", "acknowledgments", "grants": The textual content of the entry, which are essentially different parts of the research document, is split according to these labels.
- "science", "instrumentation", "mention", "not_telescope": These are boolean labels which classify the entry according to how the papers use the data from the telescopes.

For the training dataset, the annotated labels that the model needs to train and predict are the multiclass label "telescope" and the four boolean labels "science", "instrumentation", "mention", and "not_telescope". The data for training, as one can infer, is the textual information for the research paper split into "title", "abstract", "body", "acknowledgments", and "grants".

[2] https://www.scixplorer.org

Model	Optimizer	LR	Scheduler	Batch Size	Epochs
SciBERT	AdamW	2e-5	linear	8	4

Table 1: Model training hyperparameters

2.3 Data Statistics and Preprocessing

Diving into the statistics of the provided dataset, it consists of 80385 unique entries spanning 4 decades for three telescopes. These are the Hubble Space Telescope (HST), the Chandra X-ray Observatory (CXO), and the James Webb Space Telescope (JWST). Also, among the four boolean labels, "science" and "mention" are fairly evenly distributed, but the remaining two are quite skewed, with the majority of entries being the boolean label "FALSE". For the full entry text to be processed by our model, we convert the dataset into multiple JSON files. First, we concatenate the content of the fields "title", "abstract", "body", "acknowledgments", and "grants". Then, we split this string into chunks of 512 tokens, which were then saved in the JSON format along with the labels. Each JSON file contains 1000 entries, which are chunked in the manner described. Finally, for 80385 rows in the CSV file, we get 81 JSON files, which are then used for training purposes (Figure 1). These preprocessed JSON files are used as training data.

3 Experiments and Results

3.1 Model Selection

To perform the task, we opted for the SciBERT model (Beltagy et al., 2019) since it is a pretrained language model designed to enhance natural language understanding within the scientific domain. Built upon the foundational BERT architecture (Devlin et al., 2019), SciBERT extends its capabilities by being trained on a large corpus of scientific publications sourced from the Semantic Scholar database (Ammar et al., 2018). This domain-specific pretraining enables SciBERT to capture the specialized vocabulary, structure, and linguistic patterns prevalent in scientific writing, which are often underrepresented in general-domain corpora.

The model maintains the same architecture as BERT-Base but introduces a newly constructed vocabulary, SciVocab, tailored to the scientific domain. This vocabulary shares only about 42% overlap with BERT's original WordPiece vocabulary, highlighting the substantial linguistic differences between general and scientific texts (Beltagy et al., 2019). Through this adaptation, SciBERT demonstrates superior performance across a range of scientific NLP tasks, including named entity recognition, relation classification, sentence classification, and dependency parsing, outperforming general-domain models on domain-specific benchmarks. Its advantages are particularly pronounced in biomedical and life science applications, where scientific terminology and context play crucial roles in comprehension and information extraction.

3.2 Experiments and Results

We report the results for two sets of experiments that showed a marginal difference in their performance. Both these experiments achieved the 6th rank in the competition leaderboard.

In our approach, we initially do a baseline run to measure the scope of improvement. The pretrained SciBERT encoder was used without any fine-tuning, while the classification heads remained randomly initialized. The [CLS] token representations from each chunk were processed by the random classification heads to generate logits for both telescope and boolean labels. Predictions were then aggregated across all chunks to produce final outputs. We call this experiment SciBERT_v1.

Following this, we begin our training procedure. In the first experiment, we use the first 512 tokens from each entry, along with the entry-level classification labels. This results in a dataset comprising approximately 41 million tokens. The training hyperparameters used are listed in Table 1. The loss function governing this training process is the sum of the cross-entropy loss for the multiclass label (e.g., the "telescope" label) and the BCEWithLogits loss for the four boolean labels.

Model	Macro F1 score
SciBERT_base	0.18
Random baseline	0.24
Openai-gpt-oss-20b[3]	0.31
SciBERT_v1	*0.72*
SciBERT_v2	**0.73**

Table 2: Performance metrics. Here, '_base' represents the baseline run, '_v1' is the SciBERT model trained with the initial 512 tokens, and '_v2' is the SciBERT model trained on the 10 random chunks from each entry.

Next, we carry out a similar experiment, but instead of using just the first 512 tokens, we use 10 random chunks from each entry (if an entry has fewer than 10 chunks, we consider all of them). We call this experiment SciBERT_v2. All the resulting chunks get labeled the same as the full entry itself. This was done to give a fair chance to the other chunks of the same entry to contribute to the training part, specifically the acknowledgment and grants, which often contain direct references to instrumentation. For this experiment, the dataset comprises approximately 410 million tokens. The remaining part of the training process (the hyperparameters, loss function, etc) was similar to the previous experiment.

These models were tested on the test dataset, which consisted of 9194 entries. These were also preprocessed in the same way as the training dataset. To quantify the model's classification capability appropriate metric is needed. For classification tasks where there is label imbalance F1 score is most widely used. The F1 score provides a balanced measure of a model's precision and recall, which is especially important for imbalanced datasets, which we have as we discussed in the 2.3 already. Now, since we have 5 classes to predict, we will have an F1 score per class. So we consider the macro F1 score as the model performance

[3]https://ui.adsabs.harvard.edu/WIESP/2025/shared_task

Paper ID	Telescope		Science		Instrument		Mention		Not_telescope	
	GT	Pred	GT	Pred	GT	Pred	GT	Pred	GT	Pred
2014H...6C_CHANDRA	CHANDRA	CHANDRA ✓	1	1 ✓	0	0 ✓	0	0 ✓	0	0 ✓
2001t...7M_CHANDRA	CHANDRA	CHANDRA ✓	0	0 ✓	0	0 ✓	1	1 ✓	0	0 ✓
2008l...8S_HST	HST	HST ✓	1	1 ✓	0	0 ✓	0	0 ✓	0	0 ✓
2012A...4S_CHANDRA	CHANDRA	CHANDRA ✓	0	0 ✓	0	0 ✓	1	1 ✓	0	0 ✓
2011A...1M_CHANDRA	CHANDRA	CHANDRA ✓	1	0 ✗	0	0 ✓	0	1 ✗	0	0 ✓
2020S...9M_HST	HST	JWST ✗	0	0 ✓	0	0 ✓	0	1 ✗	1	0 ✗
2022s...1W_CHANDRA	CHANDRA	CHANDRA ✓	1	0 ✗	0	0 ✓	0	1 ✗	0	0 ✓
2000H...7S_CHANDRA	CHANDRA	CHANDRA ✓	1	0 ✗	0	0 ✓	0	1 ✗	0	0 ✓

Table 3: Model prediction examples

metric given as:

$$\text{Model}_{F1} = \frac{\text{multiclass}_{F1} + \frac{1}{N}\sum_i \text{bool}_{F1,i}}{2} \quad (1)$$

where multiclass is for the "telescope" label and bool for the four boolean classes "science", "instrumentation", "mention", and "not_telescope". The metrics of the trained model are compared to the baseline in Table 2. As we can see, the results from our two experiments are similar (0.72 and 0.73). However, they significantly outperform the LLM baseline, which has a performance of 0.31, as well as our own baseline, which is the same model without fine-tuning (0.18). This can be attributed to the domain-specific fine-tuning, which allowed our trained models to be specialized classifiers.

4 Error Analysis

To gain deeper insights into the limitations of our approach, we perform an error analysis. Since ground-truth labels for the test set are not available, this analysis is conducted on the validation split of the training data, which was also used for evaluation during model development.

In Table 3, we present selected example predictions from our best-performing model, SciBERT_v2. In the misclassified cases, we observe that the boolean labels “science” and “mention” tend to be mispredicted more frequently. This behavior is likely because these labels are highly context-dependent, requiring a nuanced understanding of the surrounding textual semantics. In contrast, the labels “instrument” and “not_telescope” are generally easier to predict correctly, as their identification primarily depends on the explicit mention of instrument names rather than broader contextual cues. This issue could be alleviated by employing models capable of handling longer context windows or those pretrained on domain-specific astronomical corpora. Furthermore, for the telescope

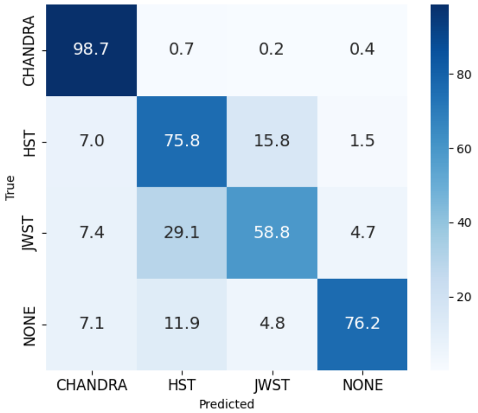


Figure 2: Telescope predictions confusion matrix

classification, a clear trend emerges (Figure 2): the model achieves the highest accuracy for CHANDRA, followed by HST, and then JWST. Also, we see increased false predictions, i.e., more confusion for the HST and JWST classes. The reason behind this could be the naming scheme of the classes for this label. The classes of the Hubble Space Telescope (HST) and the James Webb Space Telescope (JWST) share the words "Space" and "Telescope" that might have confused the model predictions, while CHANDRA is more distinct.

5 Conclusion and Future Work

We introduced our system for the telescope reference and astronomy categorization. Leveraging the SciBERT model, our method utilizes domain-adapted language representations to automatically identify telescope mentions and their contextual roles within astrophysical literature. We showed that fine-tuning SciBERT on random segments selected from the article data considerably improves model performance and significantly outperforms the LLM baseline. Looking ahead, we aim to further enhance the framework by exploring transformers with extended context windows and models pretrained on astronomy-specific corpora, which could help capture the nuanced contextual cues required for labels such as science and mention. We also plan to investigate data balancing strategies and contrastive learning methods to mitigate class skewness in telescope categories and improve robustness across less frequent instruments.

6 Limitations

The limitations of this work primarily stem from the inherent challenges of modeling complex scientific text and the class imbalance in the dataset. Although our framework effectively captures domain-specific semantics, the context-dependent nature of certain labels makes it prone to misclassification, suggesting that the current model's context window may be insufficient to fully capture subtle relationships between telescope usage and scientific context. Furthermore, the reliance on weakly supervised labels may introduce annotation noise, affecting the precision of the boolean attribute detection. The telescope classification results also reflect dataset skewness, where classes such as CHANDRA are overrepresented, leading to uneven performances across telescope types.

Acknowledgments

This work was supported by the BMFTR and the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence, the Transdisciplinary Research Area (TRA) Matter at the University of Bonn, and the AISafety Project, funded under the BMFTR grant proposal No. 05D23PD1.

References

Alberto Accomazzi. 2011. Linking literature and data: Status report and future efforts. In *Future professional communication in astronomy II*, pages 135–142. Springer.

Walid Mohamed Aly, Taysir Hassan A Soliman, and Amr Mohamed AbdelAziz. 2025. An evaluation of large language models on text summarization tasks using prompt engineering techniques. *arXiv preprint arXiv:2507.05123*.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, and 1 others. 2018. Construction of the literature graph in semantic scholar. In *Proceedings of NAACL-HLT*, pages 84–91.

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.

Youngjin Chae and Thomas Davidson. 2023. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation*, 10.

Xiang Dai and Sarvnaz Karimi. 2022. Detecting entities in the astrophysics literature: A comparison of word-based and span-based entity recognition methods. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 78–83.

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.

Tao Feng, Lizhen Qu, Niket Tandon, Zhuang Li, Xiaoxi Kang, and Gholamreza Haffari. 2025. On the reliability of large language models for causal discovery. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9565–9590.

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.

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.

Edwin A Henneken and Alberto Accomazzi. 2011. Linking to data-effect on citation rates in astronomy. *arXiv preprint arXiv:1111.3618*.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Enhanced Table Structure Recognition with Multi-Modal Approach

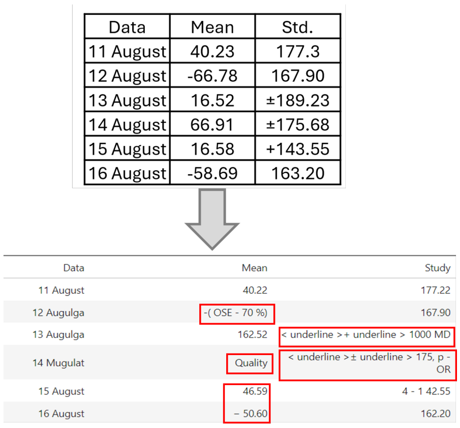
Huichen Yang[1], Andrew Hellicar[1], Maciej Rybinski[2], Sarvnaz Karimi[1]
[1]CSIRO Data61, Australia
[2]ITIS, University of Málaga, Málaga, Spain
{Huichen.Yang, Andrew.Hellicar, Sarvnaz.Karimi}@data61.csiro.au, maciek.rybinski@uma.es

Abstract

Tables are fundamental for presenting information in research articles, technical documents, manuals, and reports. One key challenge is accessing the information in tables that are embedded in Portable Document Format (PDF) files or scanned images. It requires accurately recognising table structures in diverse table layouts and complex tables. Table Structure Recognition (TSR) task aims to recognise the internal structure of table images and convert them into a machine-readable format. We propose a flexible multi-modal framework for image-based TSR. Our approach utilises two-stream transformer encoders in conjunction with task-specific decoders for extracting table structures and detecting cell bounding boxes. Experiments on benchmark datasets demonstrate that our model achieves highly competitive results compared to strong baselines, outperforming single-modality approaches by 5.4% on the FinTabNetd dataset.

1 Introduction

Tables commonly present and summarise information in a structured format. They are widely used in various texts, such as scientific literature, books, business documents, manuals, and technical documents, due to their easier readability in presenting data. Managing, understanding, and analysing table data have become increasingly important, especially with the rapid growth of digitised data and the demand for intelligent document processing (Cui et al., 2021; Yu et al., 2023). However, table data are often restricted to digitised documents or images. While humans can easily interpret them, they are not readily processed by machines. The digitised table can be easily converted into a table image, but recognising its structure is challenging due to the complex styles. Therefore, extracting table data while preserving its structure in a machine-readable format is a fundamental step in table understanding.



Data	Mean	Std.
11 August	40.23	177.3
12 August	-66.78	167.90
13 August	16.52	±189.23
14 August	66.91	±175.68
15 August	16.58	+143.55
16 August	-58.69	163.20

Data	Mean	Study
11 August	40.22	177.22
12 Augulga	-(OSE - 70 %)	167.90
13 Augulga	162.52	< underline >+ underline > 1000 MD
14 Mugulat	Quality	< underline >± underline > 175, p - OR
15 August	46.59	4 - 1 42.55
16 August	− 50.60	162.20

Figure 1: Examples of failures in an end-to-end method include cases where the model identifies the correct table structure but incorrect content (Ly and Takasu, 2023).

Table Structure Recognition (TSR) is the task of automatically recognising table structures and extracting table content as free text for machine processing, which is a key step in table understanding. The table structure could follow pre-defined formats, such as HTML or JSON. Once the table structure is recognised, the table content can be extracted by any optical character recognition (OCR) tool, allowing the reorganisation of data into a table as it was originally presented in the table image. The structured table data, consisting of free text, enables machine processing and analysis of table data, and it is a crucial step for table-related downstream tasks, such as table-based question answering (TQA) (Iyyer et al., 2017; Chen et al., 2020b; Gupta et al., 2023), table-based fact verification (Chen et al., 2020a; Xie et al., 2022), information retrieval (Chen et al., 2020c; Engelmann et al., 2023), and text mining (Xie et al., 2020).

Tables have diverse structures and styles, which pose significant challenges for accurate recognition. For instance, tabular data is often organised with cells spanning multiple rows and columns. Such ta-

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 201–207
December 23, 2025 ©2025 Association for Computational Linguistics

bles may include complex headers, cells containing multi-line text, empty cells, and varying line sizes or shapes used to separate cell contents. Moreover, table size introduces an additional challenge for the TSR task, as large tables may extend across multiple pages, particularly in certain scientific domains or technical documents.

Models based on deep neural networks have been proposed to address challenges in the TSR task. Recent methods for TSR can be divided into two strategies: the end-to-end and the non-end-to-end approach. The end-to-end method aims to use a single pipeline to process a given table image and output all table information, including the table structure, table cell bounding boxes, and table cell content (Schreiber et al., 2017; Ly and Takasu, 2023). This method is straightforward to understand, but its effectiveness is often unsatisfactory, especially when complex characters are present in the cell content. For example, as shown in Figure 1, the table structure can be identified well, but some content may be lost or incorrectly recognised.

On the other hand, the non-end-to-end method divides the TSR task into two sub-tasks: (1) recognising the table structure and table cell bounding boxes; and (2) extracting the cells' contents (Qiao et al., 2021; Nassar et al., 2022). Table cell content recognition can be considered an OCR task, which means we only need to extract the content rather than understand its semantic meaning. Many off-the-shelf OCR tools can be utilised instead of being integrated into the model training process to increase the training complexity.

We explore the efficacy of pre-training a multi-modal model for TSR. We propose a novel multi-modal approach for the TSR task, which differs from previous studies that only consider single modality pixel-based images (e.g., (Chi et al., 2019; Xing et al., 2023)). Our approach uses both the table image and its content as inputs for two transformer-based encoders, followed by separate decoders to generate the table structure and bounding boxes for non-empty table cells. This method aims to enhance the accuracy and robustness of TSR by integrating multiple data modalities, addressing the limitations of single-modality models for the task. Our main contributions are summarised as follows:

- Exploring and comparing the effectiveness of multi-modal models compared to vision-based models for the TSR task.

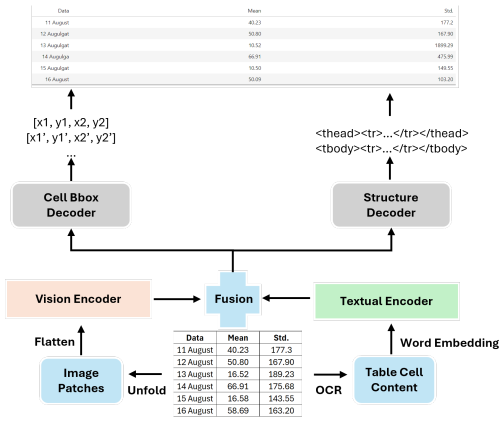


Figure 2: Our proposed two-stream multi-modal model architecture.

- Proposing a novel multi-modal approach for the TSR task, and the experimental results demonstrate that the approach is efficient.

2 Related Work

Early work on the TSR task relied on heuristic rule-based methods. These approaches required handcrafted features and designed rules or templates to cover specific table layouts for structure recognition. For example, ruling lines were used to detect horizontal and vertical lines in tables, and the arrangement of text components followed a top-down approach to recognise table structures (Ramel et al., 2003; Hassan and Baumgartner, 2007). These approaches work well with simple tables, but struggle with complex table structures.

Machine learning-based methods are widely used for the TSR task. Early methods involved statistical machine learning techniques, such as using Support Vector Machines (SVM) to classify tables based on line information (Kasar et al., 2013), or clustering word segments in a bottom-up manner (Kieninger and Dengel, 1999). Recently, with the availability of large datasets, deep learning methods have been preferred. One common approach considers TSR as an object detection task, employing well-known detection frameworks such as Faster R-CNN (Girshick, 2015), Mask R-CNN (He et al., 2017), and YOLO (Redmon and Farhadi, 2018). Another approach frames TSR as an image-to-sequence task using transformer-based encoder-decoder methods (Khang and Hong, 2024), for example, applying Convolutional Neural Networks (CNN) as the encoder for image feature representation and Recurrent Neural Networks

(RNN) as the decoder for structure sequence generation (Li et al., 2020a), or using vision transformers for TSR (Nassar et al., 2022; Chen et al., 2023). Graph Neural Networks (GNN) have also been applied to TSR, leveraging text cells as graph vertices and employing graph attention mechanisms to generate their representations (Xue et al., 2019; Chi et al., 2019). More recently, Vision Large Language Models (VLLMs) (Zhou et al., 2025) have been explored for TSR as well.

3 Methodology

We consider the TSR task as an image-to-sequence generation task. We propose a framework that uses vision and text transformer as two-stream encoders, with the fused multi-modal feature representation for sequence generation through two decoders. The model generates a machine-processable sequence **S** from a given table image **I**. The generated sequence **S** includes the table structure **T** = [t_1, ..., t_n], and the non-empty table cell bounding box **B** = [b_1, ..., b_m]. The table cell contents **C** = [c_1, ...c_m] are obtained using an off-the-shelf OCR (Smith, 2007). The table cell contents correspond to the table bounding boxes, but may differ from the table structure sequence due to empty cells in the table. The table structure is represented using HTML tags, which can be converted into various formats depending on the requirements.

3.1 Encoder

We use two stream encoders to extract visual and textual features, aiming to obtain better cross-modal representations from table images. For the visual encoder, inspired by ViT (Dosovitskiy et al., 2021), the input table image is resized and split into non-overlapping *P* x *P* patches, which are then reshaped into flattened 2D patches. These patches are linearly projected into a D-dimensional sequence, serving as the input to a stack of transformer encoder layers. The final output is encoded visual sequence features of the table image. The textual encoder follows the approach of Roberta (Liu et al., 2019). It takes word embeddings of the table's textual content as input. The global tokens [CLS] and [SEP] are added at the beginning and the end of each text sequence, and [PAD] tokens are appended to the end to match the maximum sequence length *L*. The textual encoder outputs the textual representation. Finally, the outputs of both encoders are integrated using an element-wise sum. This allows the model to learn the complex relationships between visual and textual features to obtain contextual text-and-image representations.

3.2 Decoder

The decoder is built on a standard transformer decoder that takes embedded features from the fused encoder outputs. It consists of a stack of four decoder layers, each containing multi-head attention and feed-forward layers. We employ separate decoders with the same architecture to decode the table structure and the table cell bounding boxes. The structure decoder generates HTML tags representing the table structure, including starting tags such as <thead>, <tbody>, <tr>, etc. The bounding box decoder generates coordinates for each non-empty table cell in the format $[x_{min}, y_{min}, x_{max}, y_{max}]$. We apply teacher forcing during model training and use beam search for inference.

Since the pre-trained vision encoder is not trained on table images, we continue to train it with the TableBank dataset (Li et al., 2020b), along with the aligned table text encoder, to enhance table feature representation. Masked image modeling (Bao et al., 2022) is applied to the visual encoder during pre-training. We fine-tune the entire TSR model during the fine-tuning process.

4 Experimental Setup

The pre-trained Swin-tiny transformer (Liu et al., 2021) is used for visual embedding initialisation, and the text embedding is initialised from Roberta (Liu et al., 2019). We use Adam optimiser (Kingma and Ba, 2015) with an initial learning rate of $2e{-}5$, which decays by 0.02 after the 3rd epoch. We trained the encoder for 10 epochs with a batch size of 16. The decoder includes 4 layers with an input feature size of 512 and 4 attention heads for table structure and cell bounding box decoding. Similar to the encoder, the decoder uses the Adam optimiser but with an initial learning rate of $2e{-}4$, trained for 10 epochs with a batch size of 16. We use Tesseract OCR [1] to obtain table cell content from the table image.

4.1 Datasets

We evaluate our approach on three benchmark datasets for the TSR task.

PubTabNet (Zhong et al., 2020) contains 509k table images extracted from scientific literature and

[1] https://github.com/tesseract-ocr/tesseract

provides annotation for table structure in HTML format, table cell bounding boxes, and table cell content. This dataset also provides evaluation metrics such as Tree-edit-distance-based similarity (TEDS) for both table structure and table cell content evaluation. We use the validation dataset as the test dataset since the test dataset is not available.

FinTabNet (Zheng et al., 2021) is created from the annual reports of the S&P 500 companies in PDF format. It includes 113k table images from 1,600 different types of financial tables and is annotated for table structure (in HTML), table cell bounding box, and table cell content. This dataset is reviewed manually, making it more reliable.

SciTSR (Chi et al., 2019) contains 15k tables extract from scientific PDF files. It provides corresponding structure labels obtained from LaTeX source files. The dataset is split into 12k tables for training and 3k tables for testing. Because SciTSR does not provide tables in HTML format, we convert the structure labels into HTML for S-TEDS evaluation. We use the bounding box coordinates to recover the logical row and column layout, place each cell into the correct position in a two-dimensional grid, and produce an HTML table that reflects the original structure.

4.2 Evaluation Metrics

For evaluation, we use Intersection over Union (IoU) with COCO average precision (AP) (Lin et al., 2014) to measure the overlap between ground truth and predicted bounding boxes. The AP_{50} is reported as the evaluation result for table cell bounding box detection. The structure-only Tree-Edit-Distance-Based Similarity or S-TEDS (?) is used for table structure-based evaluation. It converts table HTML tags into a tree structure and measures the edit distance between the prediction and ground-truth tree structures. Higher similarity corresponds to a shorter edit distance, leading to a higher TEDS score.

5 Experimental Results

We compared our models with six baselines—Cascade R-CNN (Cai and Vasconcelos, 2018), Deformable-DETR (Zhu et al., 2021), TSRDet (Xiao et al., 2025), VAST (?), TABLET (Hou and Wang, 2025), and NGTR (Zhou et al., 2025)—on three TSR task-related benchmark datasets (PubTabNet, FinTabNet, and SciTSR),

Model	Dataset	AP_{50}	S-TEDS(%)
Cascade R-CNN	PubTabNet	95.38	83.78
Deformable-DETR	PubTabNet	97.43	95.73
TSRDet	PubTabNet	**98.26**	96.58
VAST	PubTabNet	94.80	97.23
TABLET	PubTabNet	—	97.67
Ours	PubTabNet	97.90	**97.69**
Cascade R-CNN	FinTabNet	97.53	87.49
Deformable-DETR	FinTabNet	98.42	97.81
TSRDet	FinTabNet	98.33	**99.05**
VAST	FinTabNet	96.20	98.63
TABLET	FinTabNet	—	98.99
Ours	FinTabNet	**98.97**	98.96
Cascade R-CNN	SciTSR	95.27	79.09
Deformable-DETR	SciTSR	97.39	97.30
TSRDet	SciTSR	96.79	98.41
Ours	SciTSR	**98.32**	**98.52**

Table 1: Comparing our method with baselines on PubTanNet, FinTanNet, and SciTSR datasets.

Model	AP_{50}	S-TEDS(%)
Swin-T	92.36	93.56
Ours	**98.97**	**98.96**

Table 2: Ablation results for vision-only and multi-modal approaches on the FinTabNet dataset.

using AP_{50} and S-TEDS metrics. We utilised structure-based S-TEDS as the primary evaluation metric to avoid the noise of table cell content that is generated by OCR. Our multi-modal approach outperformed almost all visual-only baseline methods and achieved highly competitive results on table structure recovery, as shown in Table 1. In particular, the multi-modal approach showed a clear improvement in S-TEDS compared with the vision-only Deformable-DETR, which suggests that using text information helps the model better handle confusing layouts and cells that look similar in table images. The ablation study on FinTabNet (Table 2) demonstrates that incorporating the visual modality leads to a significant gain in S-TEDS (+5.4), indicating that visual and textual features work together and complement each other for TSR. We note that our approach also outperforms the VLLM approach (NGTR) (Zhou et al., 2025) (Table 3) as per reported results on the same datasets.

6 Conclusions

We present a multi-modal approach with two stream encoders and separate decoders for the Table Structure Recognition (TSR) task. The pro-

Model	Dataset	S-TEDS(%)
NGTR	PubTabNet	92.31
Ours	PubTabNet	**97.69**
NGTR	SciTSR	95.78
Ours	SciTSR	**98.52**

Table 3: Comparing our TSR method and reported results on VLLMs (Zhou et al., 2025).

posed model integrates features from both visual and textual modalities, generating table structure and table cell bounding boxes simultaneously. Our experimental results on three different datasets from scientific and financial domains show that the effectiveness of the proposed model is competitive compared to visual-only approaches.

7 Limitations

The proposed multi-modal approach demonstrated its effectiveness with regular table images, but it is worthwhile to further explore irregular table images in real-world scenarios, such as table images from scanned books, wired tables in the wild, and handwritten tables. Meanwhile, training a unified framework to integrate all sub-tasks of TSR (table structure, table cell bounding boxes, and table cell content) also presents opportunities for exploration.

References

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. Beit: BERT pre-training of image transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162.

Leiyuan Chen, Chengsong Huang, Xiaoqing Zheng, Jinshu Lin, and Xuan-Jing Huang. 2023. "tablevlm: Multi-modal pre-training for table structure recognition". In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 2437–2449.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020a. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020b. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Zhiyu Chen, Mohamed Trabelsi, Jeff Heflin, Yinan Xu, and Brian D Davison. 2020c. Table search using a deep contextualized language model. In *SIGIR*, pages 589–598.

Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. 2019. Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document ai: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*.

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. In *International Conference on Learning Representations*.

Björn Engelmann, Timo Breuer, and Philipp Schaer. 2023. Simulating users in interactive web table retrieval. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3875–3879.

Ross Girshick. 2015. "fast r-cnn". In *ICCV*, pages 1440–1448.

Vivek Gupta, Pranshu Kandoi, Mahek Vora, Shuo Zhang, Yujie He, Ridho Reinanda, and Vivek Srikumar. 2023. TempTabQA: Temporal question answering for semi-structured tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2431–2453, Singapore. Association for Computational Linguistics.

Tamir Hassan and Robert Baumgartner. 2007. Table recognition and understanding from pdf files. In *Ninth International Conference on Document Analysis and Recognition*, volume 2, pages 1143–1147.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. "mask r-cnn". In *ICCV*, pages 2961–2969.

Qiyu Hou and Jun Wang. 2025. Tablet: Table structure recognition using encoder-only transformers. In *Proceedings of the 19th International Conference on Document Analysis and Recognition*.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1821–1831.

Thotreingam Kasar, Philippine Barlas, Sebastien Adam, Clément Chatelain, and Thierry Paquet. 2013. Learning to detect tables in scanned document images using line information. In *12th International Conference on Document Analysis and Recognition*, pages 1185–1189.

Minsoo Khang and Teakgyu Hong. 2024. Tflop: table structure recognition framework with layout pointer mechanism. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 947–955.

Thomas Kieninger and Andreas Dengel. 1999. The t-recs table recognition and analysis system. In *Document Analysis Systems: Theory and Practice: Third IAPR Workshop, DAS'98 Nagano, Japan, November 4–6, 1998 Selected Papers 3*, pages 255–270.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. 2020a. Tablebank: Table benchmark for image-based table detection and recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1918–1925.

Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. 2020b. TableBank: Table benchmark for image-based table detection and recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1918–1925, Marseille, France. European Language Resources Association.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022.

Nam Tuan Ly and Atsuhiro Takasu. 2023. An end-to-end multi-task learning model for image-based table recognition. *arXiv preprint arXiv:2303.08648*.

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*, pages 4614–4623.

Liang Qiao, Zaisheng Li, Zhanzhan Cheng, Peng Zhang, Shiliang Pu, Yi Niu, Wenqi Ren, Wenming Tan, and Fei Wu. 2021. LGPMA: complicated table structure recognition with local and global pyramid mask alignment. In *International conference on document analysis and recognition*, pages 99–114.

J-Y Ramel, Michel Crucianu, Nicole Vincent, and Claudie Faure. 2003. Detection, extraction and representation of tables. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 374–378.

Joseph Redmon and Ali Farhadi. 2018. "yolov3: An incremental improvement". In *In Computer vision and pattern recognition (Vol. 1804)*, pages 1–6.

Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. 2017. Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In *The 14th IAPR international conference on document analysis and recognition*, volume 1, pages 1162–1167.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 629–633, Washington, DC, USA. IEEE Computer Society.

Bin Xiao, Murat Simsek, Burak Kantarci, and Ala Abu Alkheir. 2025. Rethinking detection based table structure recognition for visually rich document images. *Expert Systems with Applications*, 269:126461.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. UnifiedSKG: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 602–631, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xia Xie, Yu Fu, Hai Jin, Yaliang Zhao, and Wenzhi Cao. 2020. A novel text mining approach for scholar information extraction from web content in Chinese. *Future Generation Computer Systems*, 111:859–872.

Hangdi Xing, Feiyu Gao, Rujiao Long, Jiajun Bu, Qi Zheng, Liangcheng Li, Cong Yao, and Zhi Yu. 2023. Lore: logical location regression network for table structure recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2992–3000.

Wenyuan Xue, Qingyong Li, and Dacheng Tao. 2019. Res2tim: Reconstruct syntactic structures from table images. In *The International Conference on Document Analysis and Recognition*, pages 749–755.

Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2023. Structextv2: Masked visual-textual prediction for document image pre-training. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Xinyi Zheng, Doug Burdick, Lucian Popa, Peter Zhong, and Nancy Xin Ru Wang. 2021. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. *Winter Conference for Applications in Computer Vision (WACV)*.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: Data, model, and evaluation. In *ECCV*, page 564–580.

Yitong Zhou, Mingyue Cheng, Qingyang Mao, Jiahao Wang, Feiyang Xu, and Xin Li. 2025. Enhancing table recognition with vision llms: A benchmark and neighbor-guided toolchain reasoner. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 2503–2511. International Joint Conferences on Artificial Intelligence Organization.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Author Index