astrophysics data system

https://ui.adsabs.harvard.edu/ | https://scixplorer.org

Partner

CENTER FOR ASTROPHYSICS HARVARD & SMITHSONIAN

Identifying Instruments in the ADS/SciX Corpus with LLM Agents

Jean-Claude Paquin[1] and the ADS team

[1] Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA.

Introduction

Bibliographies are essential for observatories to assess the impact of their instruments and facilities. However, identifying which papers reference specific instruments is traditionally a manual, time-consuming process. The Astrophysics Data System (ADS/SciX) contains over 3 million astronomy records, with **1.5 million full-text articles**. To enhance the usefulness of this corpus, we aim to automatically extract and index references to astronomical instruments. Simple keyword matching is insufficient due to name ambiguities and underspecified references. Our goal is to develop an LLM-augmented pipeline that identifies real-world instruments within the full SciX corpus efficiently and accurately.

Methods

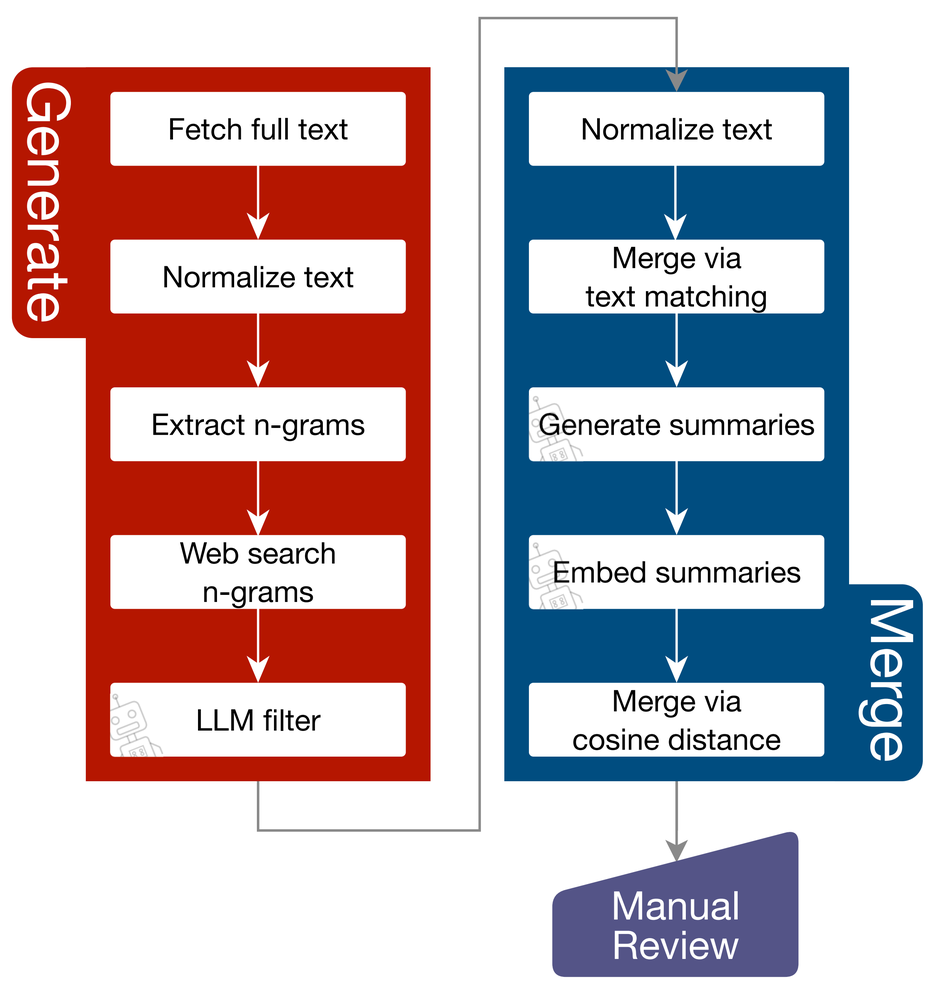
Our identification pipeline operates in two main phases:

1 Candidate Generation

- Normalize full-text records and extract n-grams triggered by keywords (e.g., “telescope,” “spectrograph”).
- Keep n-grams appearing in ≥20 records to reduce noise.
- Use web search to ground each n-gram, and an LLM filter to determine whether it refers to a real instrument.
- Accept candidates supported by a majority of LLM responses to reduce hallucinations.

2 Candidate Merging

- Normalize and expand instrument names (e.g., “4m” →“4 meters”), remove parentheticals.
- Merge duplicates using text and semantic similarity metrics. Use web search summaries and embeddings to group semantically equivalent names (cosine similarity > 0.93).
- Manually review the final instrument groupings for validation.



Results

- Our initial pipeline run processed millions of records, generating 1,740 distinct instrument names, of which only 56 required manual review.
- The LLM agents grounded with web search significantly increased the number of true-positive instrument identifications compared to unguided text matching.
- Tasks that would traditionally take a human curator a week of dedicated work were completed within hours, demonstrating the scalability and practical efficiency of this LLM-assisted approach.

Next Steps

- Improve the reliability of the automatic merging step to eliminate the need for manual review.
- Enhance detection of lesser-known and newly introduced instruments using contextual LLM reasoning.
- Integrate the instrument index into ADS/SciX search capabilities, enabling users to query by instrument and facility.
- Run the pipeline incrementally as new papers are added to the ADS/SciX corpus, maintaining an up-to-date instrument catalog.

Jean-Claude Paquin
jean-claude.paquin@cfa.harvard.edu

Accomazzi, Alberto; Kurtz, Michael J.; Protopapas, Pavlos; Grant, Carolyn S.; Henneken, Edwin; Lockhart, Kelly; Hostetler, Timothy W.; Templeton, Matthew R.; Chen, Shinyi (Jennifer); Grezes, Felix; Koch, Jennifer; Allen, Tom; Jacovich, Taylor; Blom, Harry; Chivvis, Daniel; de Macedo Alves, Fernanda; Paquin, Jean-Claude; Bartlett, Jennifer; Polimera, Mugdha; Kelbert, Anna; Di Re, Paul; Pomerantz, Jeffrey; Alkan, Atilla Kaan; Kundu, Suze; Blanco-Cuaresma, Sergi; Jarmak, Stephanie.