astrophysics data system

Machine Learning & AI

Felix Grezes and the ADS Team

ADS Users Group Meeting, 20 Nov. 2025

Smithsonian Institution

CENTER FOR ASTROPHYSICS HARVARD & SMITHSONIAN

SciX

NASA Partner

Machine Learning Initiatives at SciX

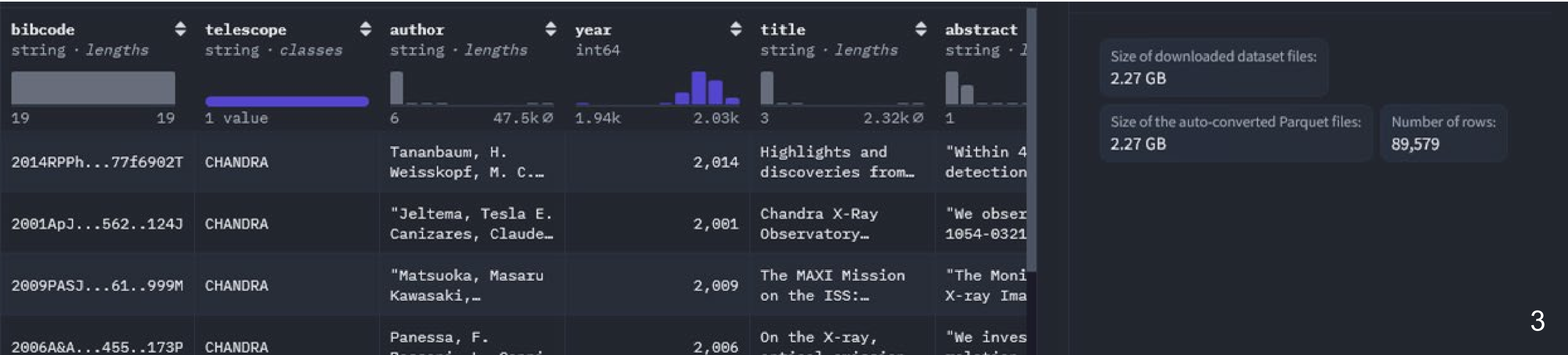
- **Machine learning datasets and models**
    - Expert-curated, publicly available, permissively licensed

- **R&D efforts focused on data enrichment via ML + AI pipelines**
    - Automated categorization
    - Automated keyword labeling
    - (Planetary features detection)
        - On-hold due to developer, project scientist resignations

- **Internal and external collaborations**
    - Exploratory projects to help us keep current with emerging technologies

Datasets: TRACS

- **Telescope Reference and Astronomy Categorization**
    - over 89K samples labeled by space telescope and usage
        - currently: CHANDRA, HST & JWST (VLT in the future)
        - Usage categories: science, instrumentation, mention
    - curated in conjunction with major telescope bibliography curators
    - guide LLMs toward bibliography curation
    - publicly available https://huggingface.co/datasets/adsabs/TRACS



bibcode string · *lengths* 19 19	telescope string · *classes* 1 value	author string · *lengths* 6 47.5k ⌀	year int64 1.94k 2.03k	title string · *lengths* 3 2.32k ⌀	abstract string · *l* 1
2014RPPh...77f6902T	CHANDRA	Tananbaum, H. Weisskopf, M. C.…	2,014	Highlights and discoveries from…	"Within 4 detection
2001ApJ...562..124J	CHANDRA	"Jeltema, Tesla E. Canizares, Claude…	2,001	Chandra X-Ray Observatory…	"We obser 1054-0321
2009PASJ...61..999M	CHANDRA	"Matsuoka, Masaru Kawasaki,…	2,009	The MAXI Mission on the ISS:…	"The Moni X-ray Ima
2006A&A...455..173P	CHANDRA	Panessa, F.	2,006	On the X-ray,	"We inves

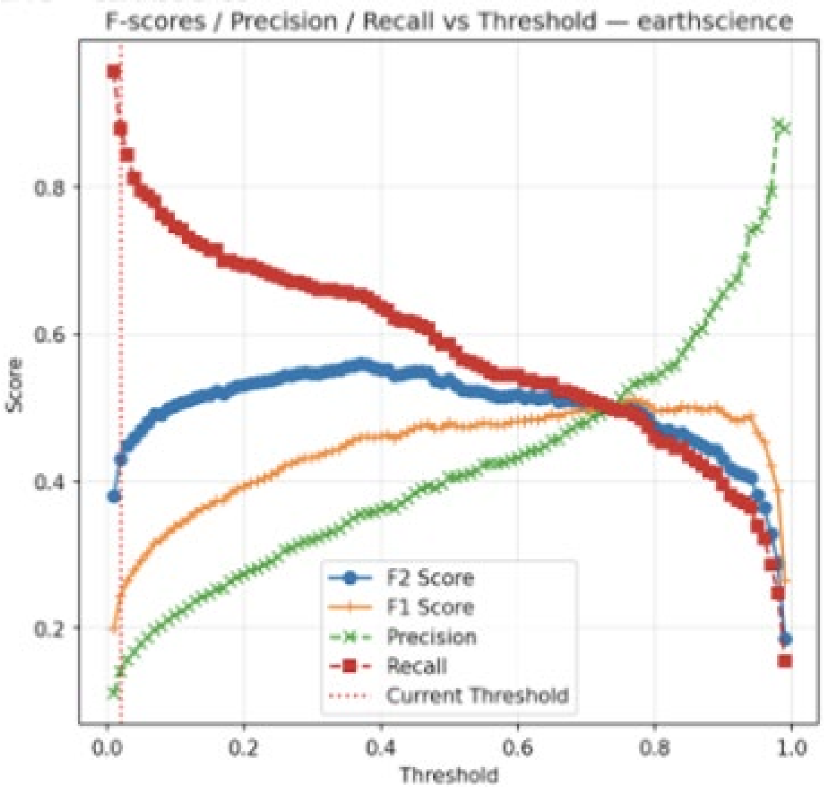
Size of downloaded dataset files:
2.27 GB

Size of the auto-converted Parquet files:
2.27 GB

Number of rows:
89,579

SciX Classifier improvements

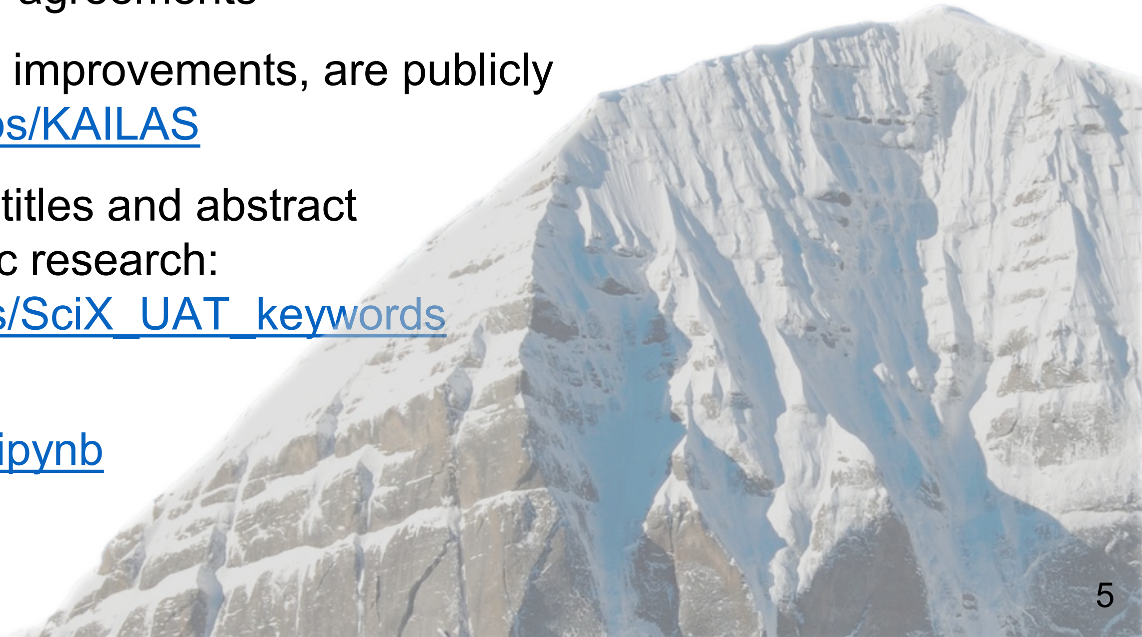
- **Planned: Focus on Earth Science**
    - move away from astroBERT and retrain the SciX Classifier using INDUS, the cross disciplinary, foundational LLM by NASA IMPACT and IBM
    - Improve the training dataset with more Earth Science examples



KAILAS: Keyword AI Labeler At SciX

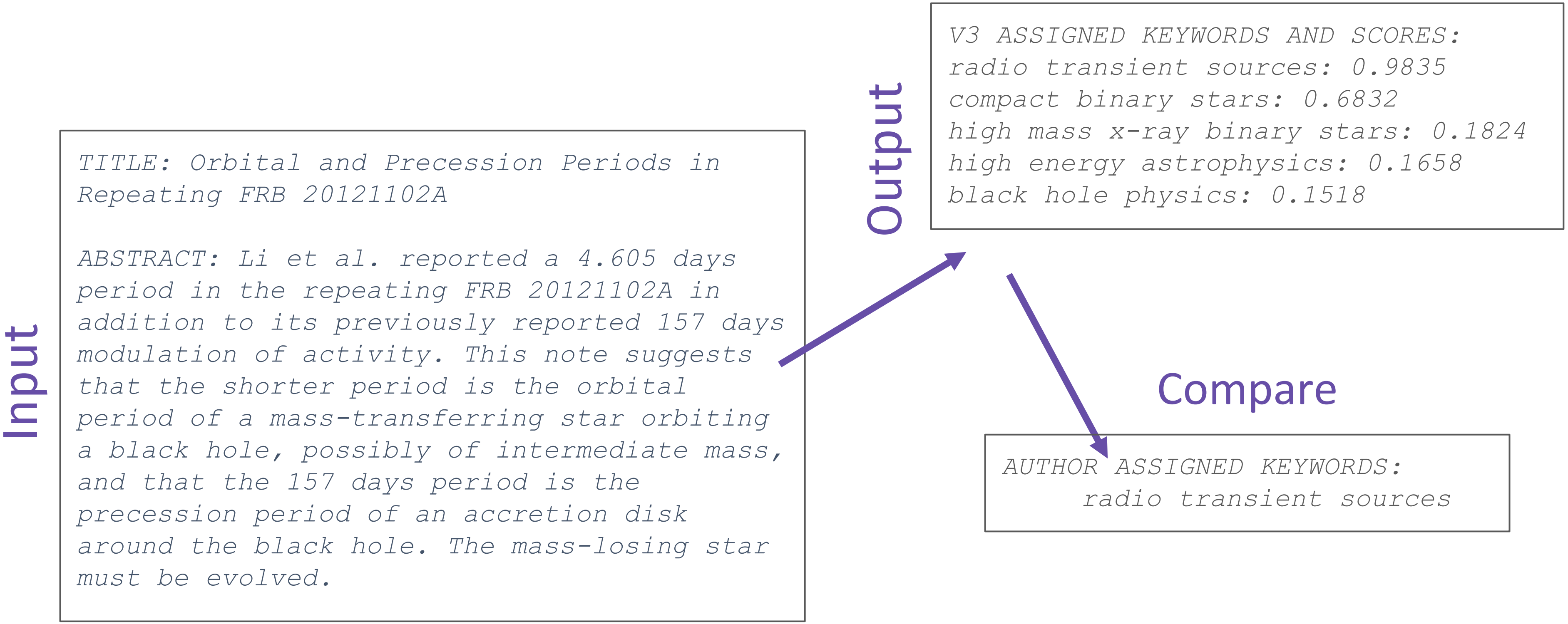
KAILAS automatically assigns UAT keywords to astronomy papers.

- KAILAS is a RoBERTa-based model trained on the full text of ~425,00 astronomy papers (3+ million tokens) in our database that have been tagged by authors/publishers with UAT keywords
- Development respects publisher agreements
- Trained model, as well as future improvements, are publicly available: huggingface.co/adsabs/KAILAS
- Version of the dataset with only titles and abstract is publicly available for academic research: huggingface.co/datasets/adsabs/SciX_UAT_keywords
- Try model yourself
KAILAS Demo of UAT labelling.ipynb



mount Kailas[h] from the Himalayas

KAILAS: Evaluation Process



Input

TITLE: Orbital and Precession Periods in Repeating FRB 20121102A

ABSTRACT: Li et al. reported a 4.605 days period in the repeating FRB 20121102A in addition to its previously reported 157 days modulation of activity. This note suggests that the shorter period is the orbital period of a mass-transferring star orbiting a black hole, possibly of intermediate mass, and that the 157 days period is the precession period of an accretion disk around the black hole. The mass-losing star must be evolved.

Output

V3 ASSIGNED KEYWORDS AND SCORES:
radio transient sources: 0.9835
compact binary stars: 0.6832
high mass x-ray binary stars: 0.1824
high energy astrophysics: 0.1658
black hole physics: 0.1518

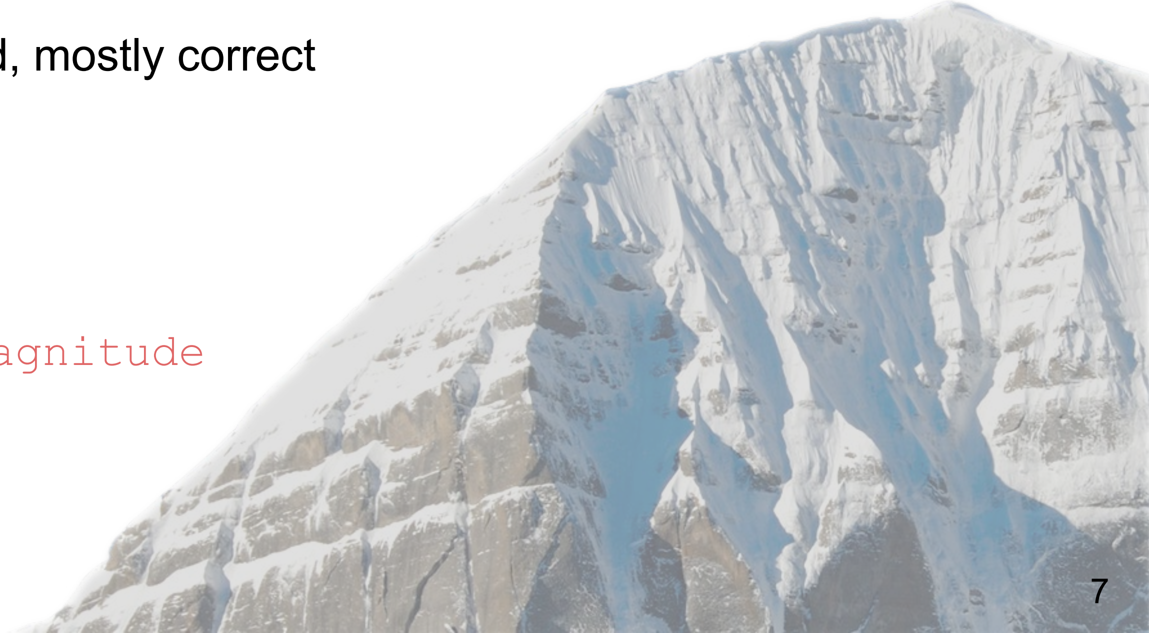
Compare

AUTHOR ASSIGNED KEYWORDS:
radio transient sources

KAILAS: Current Evaluation

Feedback from AAS Working Group on the UAT

- KAILAS v2 total of 77 papers + 1 control reviewed by 8 experts
- Keywords relevant 158 times (68%), potentially relevant 36 times, & wrong 39 times (17%).
- Keywords assigned are varied, mostly correct but not better than human yet
  - Author practices erratic
  - Uneven coverage
  - Poisoned training set - `magnitude`
- More feedback desired!
- Assessment of v3 begun



KAILAS: Evolution

- **Past**
  - v1 trained on ~18K astronomy papers in our database that have been tagged by authors/publishers with UAT keywords
- **Present**
  - v2 trained on >500K papers: older publications, synonyms and crosswalks from other thesauri
  - evaluated by the WG UAT to be as good as authors labelling their papers (but not better)
  - beta functionality present on the SciX user interface (on select papers)
- **Future**
  - v3+ increase coverage by improving training dataset further
  - exploring new methods to

Post Doctoral Research: Atilla Kaan Alkan

- **About Me (before my postdoc)**
  - **PhD in Natural Language Processing for Astrophysics (CEA-Irfu & LISN, Université Paris-Saclay)**
    - Building corpora (TDAC, astroECR) and information extraction systems (named entities and relation extraction, coreference resolution).
- **Now at ADS (Postdoc)**
  - **Domain Adaptation and Modeling**
    - Developing multilabel text classification models to automatically assign Unified Astronomy Thesaurus (UAT) concepts;
    - Designing and evaluating domain-specific language models.
  - **Information Extraction and Disambiguation**
    - Creating pipelines for software name recognition in scientific publications (geoscience);
    - Designing disambiguation approaches to link extracted software mentions to their canonical entities.
  - **Language Resource Development**
    - Building annotated datasets for named entity recognition and disambiguation.

Internal Collaborations: Pavlos Protopapas

- **Faculty at Harvard School of Engineering and Applied Sciences**
  - Leads student projects in collaboration with ADS staff
  - Focus on exploratory pilot projects
- **Citation Predictor**
  - Led by master student Benjamin Basseri
  - Suggests citations supporting scientific claims
- **SciX Labs**
  - Led by students Ashish Kumar, Karthik Rathod, Swarnava Bhattacharjee
  - Framework and UI to host pilot projects by internal and external collaborators
  - ADS UI staff will eventually move this project into production readiness

External Collaborations: 3rd WASP 2025

- **Co-organized WASP 2025**
  - Workshop for Artificial Intelligence for Scientific Publications, part of IJCNLP-AACL 2025
  - partnership with Dr. Tirthankar Ghosal of Oak Ridge National Lab
  - 37 submissions
  - Keynote speakers:
    - Karin Verspoor - RMIT University
    - Kartheik Iyer - NASA Hubble Fellow, Columbia University
  - TRACS Shared Task
    - TRACS @ WASP 2025 | Kaggle

External Collaborations: uTBD

- **Distributed research collaboration**
    - Atilla collaborating on the HypoGen for Science project
    - Mugdha collaborating on AstroCoder and AstroTalks/TalkFinder projects + member of Executive Board
- **Co-organizing the Language AI in the Space Sciences workshop**
    - Main workshop site
    - To be held at STScI in March 2026