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

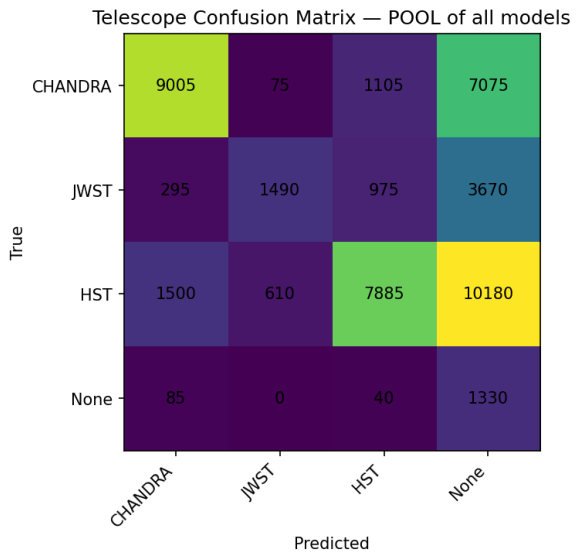

Figure 1: **Confusion matrix across all models: telescope accuracy.**

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 pre-training, 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 fine-tuned 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.