TRACS: Telescope Reference & Astronomy Categorization Dataset and Shared Task

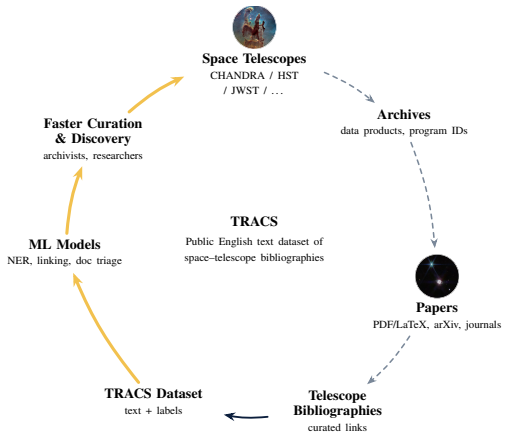
Felix Grezes[1], **Kelly Lockhart**[1], **Alberto Accomazzi**[1], **Jennifer Lynn Bartlett**[1], **Ethan Seefried**[2], **Anjali Pandiri**[3], **Tirthankar Ghosal**[2]
[1]Center for Astrophysics | Harvard & Smithsonian, USA
[2]Oak Ridge National Laboratory, USA
[3]Florida State University, USA
[1]{felix.grezes,kelly.lockhart,aaccomazzi,jennifer.bartlett}@cfa.har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

Abstract

To assess the scientific impact of observational facilities, Astronomers analyze publications which have made use of the data by the facility. This requires bibliographies annotating and linking data products with the literature, allowing bibliometrics to measure impact of the data. Creating bibliographies is laborious, involves specialists searching the literature for names, acronyms and identifiers, then determining how/if observations were used in those publications. These bibliographies are useful beyond metrics: from the perspective of a research scientist, the data-literature connections provide a critical path to data discovery and access. Thus, by leveraging the efforts of librarians and archivists, we can use telescope bibliographies to support the scientific inquiry process. To this end, we introduce **TRACS**: **T**elescope **R**eference & **A**stronomy **C**ategorization dataset and **S**hared task, consisting of over 89K public English texts from space telescope bibliographies, categorized using a new, simple taxonomy based on discussion with established bibliographers. We hope TRACS will be used to train modern Machine Learning and Artificial Intelligence models that in turn will help data curators build bibliographies. We evaluate how off-the-shelf Large Language Models perform on automatic bibliography curation as a baseline analysis. Results show the difficulty of the task and the need for custom tools.

1 Introduction

Astronomical instruments generate a wealth of data, not just directly with measurements, but indirectly as well, in the form publications that make use of these measurements or describe software created to handle them. To properly credit the teams behind the telescopes, bibliographies linking the software and research back to the telescope are needed.



Since its launch as the NASA Astrophysics Data System (Good, 1992; Kurtz et al., 2000), the Science eXplorer[1] (SciX) (Bartlett et al., 2025) has aimed to help astronomers with bibliographic tools for both discovery and impact measurement. For example, users are not only able to filter by papers in the Hubble Space Telescope (HST) bibliography, a list of papers manually curated by the Space Telescope Science Institute (2025), but also able to see cited/citing paper for the bibliography, which authors or institutions contribute the most, activity over time, and many more advanced second-order operators (Henneken and Kurtz, 2019). While SciX already offers best practices for building and maintaining bibliographies (Observatory Bibliographers Collaboration et al., 2024), and some have automated part of of the process (Grothkopf and Treumann, 2003), it typically remains labor intensive.

Typical Curation Process & Instructions While different groups use different approaches and criteria to the problem of bibliography creation and maintenance, the steps involved typically consist of the following:

1. Use a set of full-text queries to the ADS bibli-

[1]scixplorer.org

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 142–153
November 10, 2026 ©2026 Association for Computational Linguistics

ographic database in order to find all possible relevant papers. This first step aims to identify articles that contain mention of the telescope/instrument of interest so that they can be further analyzed. For instance, the set of query terms used to find papers related to the Chandra X-Ray telescope may be "Chandra," "CXC," "CXO," "AXAF," etc.

2. Analyze the text containing mentions of the telescope/instrument and its variations in order to disambiguate the use of the terms of interest. For the Chandra example, this includes teasing apart the different entities associated with "Chandra," which may correspond to a person, a ground-based telescope, or a space-based telescope.

3. Identify whether the paper in question shows evidence of the use of datasets generated by the telescope or hosted by the archive of interest. The mention of data use may be explicit (e.g. the listing of dataset identifiers), or implied in the text (e.g. mention of analysis and results without identification of the actual dataset). Whenever dataset ids are used, they should be extracted and identified.

4. In some cases, additional classification of the dataset may be collected, such as the instrument used in the observations. This information is also correlated with the kind of data that was used (e.g. image vs. spectra vs. catalog) and its characteristics. In the case of Chandra, there are 7 different instruments that can be used for the data collection (ACIS, HRC, HETG, LETG, HRMA, PCAD, EPHIN), and their use, if explicitly mentioned in the paper, should be reported.

5. For some bibliographies, additional information is collected, such as the relevance of the paper to the scientific use of the data archive. For example, for the Chandra bibliography, the following categories are defined:

   (a) Direct use of Chandra data
   (b) Refers to published results
   (c) Predicts Chandra results
   (d) Paper on Chandra software, operations, and/or instrumentation
   (e) General reference to Chandra

Goals With modern Large Language Models (LLM) capable of ingesting ever larger quantities of text, for ever more sophisticated tasks (Minaee et al., 2024), we at SciX decided to create a dataset to help the community build tools to facilitate the creation and curation of bibliographies. This dataset is TRACS: Telescope Reference & Astronomy Categorization dataset and Shared task, a collection of texts from open access astronomy papers, categorized into three space telescope bibliographies (Chandra X-ray Observatory, Hubble Space Telescope, James Webb Space Telescope), as well as how the papers use the data from the telescope.

Contributions

- a bibliographic taxonomy based on discussion with established bibliographers
- an open dataset of space telescope bibliographies, adapted to our taxonomy from human-curated bibliographies
- a baseline analysis, evaluating off-the-shelf LLMs on the task of automating bibliographic curation

TRACS is available publicly on HuggingFace[2] and was used for the shared task challenge at the 3rd WASP @ IJCNLP-AACL 2025[3]. The scoring evaluation was run on the Kaggle platform[4] .

2 Related Work

Astronomy bibliographies and telescope curation. Telescope bibliographies have been maintained for decades as a means of measuring the scientific return of major observational facilities and of connecting published results back to the data that enabled them. Early efforts at the Space Telescope Science Institute and the Chandra X-ray Center established conventions for what counts as "use" of a facility, how to disambiguate instrument names, and how to track data products through the literature (Space Telescope Science Institute, 2025; Chandra X-ray Center, 2025). The Observatory Bibliographers Collaboration has more recently worked toward harmonizing these practices across observatories, producing shared guidelines for bibliography construction and classification (Observatory Bibliographers Collaboration et al., 2024).

[2] huggingface.co/datasets/adsabs/TRACS
[3] ui.adsabs.harvard.edu/WIESP/2025
[4] kaggle.com/competitions/tracs-wasp-2025

telescope	title/author/year	Sc/In/Me/NT Labels	abstract / body (truncated)
CHANDRA	Chandra X-Ray Observatory Observation of the High-Redshift Cluster MS 1054-0321, (Jeltema et al., 2001), 2014	true, false, false, false	Using **Chandra**, we make a more accurate temperature determination; we examine substructure in the X-ray distribution, and estimate mass/velocity dispersion of MS 1054 to assess cosmological constraints...
HST	Supernova 1996cr: SN 1987A's Wild Cousin?, (Bauer et al., 2008), 2008	true, false, false, false	Note that increasing the aperture radius in the **HST** F656N band to 1.2 yields a magnitude of 17.5 (i.e., an increase of 25% over the pointlike magnitude from SN 1996cr alone), which we attribute to the flux of the underlying H ii region...
JWST	Warm Jupiters in TESS Full-frame Images: A Catalog and Observed Eccentricity Distribution for Year 1, (Dong et al., 2021), 2021	false, false, true, false	The confirmation of these targets will help to select ideal candidates for Warm Jupiter atmospheric characterization for future missions (e.g., **JWST**). Follow-up observations on candidates with missing information listed in Table 5 are also important....
NONE	Tidal adaptive softening and artificial fragmentation in cosmological simulations, (Mostoghiu Paun et al., 2025), 2025	false, false, false, false	Traditional N-body methods introduce localized perturbations in the gravitational forces governing their evolution. These perturbations lead to an artificial fragmentation in the filamentary network of the large-scale structure...

Table 1: **Core fields view of sample records.** The full dataset contains additional fields (see §3.2); here we show the core subset: *telescope*, *author/title/year*, a short *excerpt*, and *annotation flags*. Labels are booleans in the order **Sc**ience, **In**strument, **Me**ntion of telescope, and **N**ot-**T**elescope. Each excerpt is chosen to illustrate how a telescope is referenced in context and is lightly normalized and truncated for fit.

Complementary infrastructure on the discovery side, most notably the SciX (Bartlett et al., 2025) and earlier generations of the same system (Kurtz et al., 2000), has made it possible to query the literature at the scale required for these bibliographies, but the curation itself remains predominantly manual. Prior work on (partially) automating curation has focused on full-text keyword retrieval and rule-based filtering (Grothkopf and Treumann, 2003), which scales poorly to ambiguous mentions and to the fine-grained distinctions (e.g., *science* vs. *mention*) that bibliographers actually make. TRACS differs from these efforts in that it exposes the curatorial decision itself, not just the retrieval step, as a supervised learning problem over public text.

Dataset and software mention detection in scientific text. TRACS is closely related to a growing body of work on identifying mentions of datasets, instruments, and software in scientific publications. The Rich Context / Coleridge Initiative competitions framed dataset mention detection as a named-entity and linking problem over social-science articles (Pan et al., 2023), and SciREX extended this idea to document-level extraction of datasets, methods, metrics, and tasks in machine-learning papers (Jain et al., 2020). Software mention detection has been pursued in similar terms, with corpora such as SoMeSci and the Softcite dataset providing span-level annotations of software references and their usage type (Schindler et al., 2021; Du et al., 2021). A common thread across these resources is the distinction between *mentioning* an artifact and *using* it, essentially the same distinction that motivates our `science`, `instrumentation`, `mention`, and `not_telescope` labels. TRACS adapts this framing to astronomical facilities, where the "artifact" is a telescope (and, implicitly, its data archive) rather than a dataset file or software package, and where disambiguation must contend with shared acronyms, personal names, and grant references.

Scientific NLP in astronomy. Within astronomy specifically, NLP resources have historically lagged behind those in biomedicine and computer science, though this has begun to change (Grezes et al., 2022, 2023). Domain-adapted language models such as astroBERT (Grezes et al., 2021), have established that generic encoders underperform on astronomical text and that domain-specific annotated data is a bottleneck. Most existing astronomy NLP corpora target entity recognition over celes-

tial objects, instruments, or measurements, rather than the document-level, curation-oriented labels we consider here. To our knowledge, TRACS is the first public dataset that pairs full-text astronomy articles with bibliographer-grade labels indicating both *which* facility a paper concerns and *how* the facility's data is used, making it complementary to existing span-level resources and directly usable for evaluating curation assistants.

3 Dataset Description

3.1 Data Collection and Creation

The TRACS dataset consists of papers associated with a telescope and four categories likely to be of interest to bibliographers. We have drawn the categories from a simplification of those discussed by the Observatory Bibliographers Collaboration (2024). These are `science`, `instrumentation`, `mention`, `not_telescope`. Broadly, `science` papers use data from the designated telescope to obtain new results; `instrumentation` papers describe the technical aspects of the telescope; `mention` papers do reference the designated telescope but do not produce new scientific results; and `not_telescope` are papers that include a reference that might otherwise be confused with the designated telescope, i.e. false positives. Full details are available in 3.2 below.

Bibliographic data for the Chandra X-ray Observatory (`CHANDRA`) was provided by the Chandra X-ray Center (Chandra X-ray Center, 2025), while data for the Hubble and James Webb Space Telescopes (`HST`, `JWST`) was provided by the Space Telescope Science Institute (Space Telescope Science Institute, 2025). These curated, human-verified bibliographies include more information than the scope of this dataset (ex: sub-instrument data use, links to grants) and had to be pre-processed into the categories of interest. Furthermore, the papers in this data set do not represent the full corpus of any of these human-curated bibliographies and are not an adequate substitute for them for scientific or administrative purposes. In addition, a small set of papers unrelated to any of these three nor any other space telescope was provided, labeled as `None` telescope in the dataset. This set allows users of the TRACS dataset to easily verify that their models correctly predict that a paper does not relate to a space telescope.

To validate the quality of the dataset, we performed our own manual annotations. The authors, ranging from expert astronomers to non-experts, labeled 10 papers each, giving us a Cohen's kappa score of 0.72 when comparing to the references provided by the `CHANDRA`, `HST`, and `JWST` experts.

3.2 Technical Details

TRACS entries are astronomy papers with the following features:

- `bibcode`: unique string that identifies the entry in the SciX database
- `telescope`: the telescope referenced
- `science`, `instrumentation`, `mention`, `not_telescope`: boolean labels
- `author`, `year`: metadata for the entry
- `title`, `abstract`, `body`, `acknowledgments`, `grants`: the relevant textual information for the entry.

On Kaggle, an additional `Id` column is present for automatic scoring purposes.

`science` New science papers use data from the designated telescope to obtain new results. The authors may be using new observations, using archival observations, or reanalyzing previous results. However, papers that merely refer to previous results for comparison or suggest what might be possible with future observations are Mentions, rather than Science papers. Science papers may use observations directly or indirectly, such as through a published source catalog. Indirect use must be substantive. Papers that overlay new data over images from the designated telescope without discussing the underlying image are Mentions, rather than Science papers. Papers that use catalog data, such as positions or measurements, without further discussion are Mentions, rather than Science papers. Papers that reference a grant associated with the designated telescope but provide no evidence of using data from it are Not Telescope papers, rather than Science papers or Mentions.

`instrumentation` Instrumentation papers describe the technical aspects of the telescope, its calibration activities, its data processing pipeline, or its archival procedures. These papers can discuss hardware, software, or methodologies. A paper that includes new science facilitated by use of the hardware, software, or methodology described in the paper may be both a Science and an Instrumentation paper. A paper that describes a novel

technique or software to achieve its scientific conclusions may be a Science and an Instrumentation paper. A paper that uses calibration, alignment, or engineering data to produce new results may be a Science and an Instrumentation paper.

`mention` Mentions are papers that do reference the designated telescope but do not produce new scientific results (Science) or contribute to understanding it (Instrumentation). If a paper meets the criteria for a Science paper or an Instrumentation paper anywhere, then the paper is a Science paper, even if it also contains statements that would otherwise be considered a Mention. Papers that discuss the designated telescope as part of their introductory overview of the issue, of the history of a problem, or their survey of current relevant research are Mentions. Papers that discuss the designated telescope and its scientific contributions as part of an in-depth review of a research topic are Mentions. Papers that merely refer to previous results for comparison or suggest what might be possible with future observations are Mentions, rather than Science papers. Papers that overlay new data over images from the designated telescope without discussing the underlying image are Mentions, rather than Science papers. Papers that use catalog data, such as positions or measurements without further discussion are Mentions, rather than Science papers. Papers that use a secondary catalog that incorporates data from a catalog produced directly by the designated telescope are Mentions, even if that paper acknowledges the telescope. Papers that reference a grant associated with the designated telescope but provide no evidence of using data from it are Not Telescope papers, rather than Science papers or Mentions.

`not_telescope` Not Telescope papers are papers that include a reference that might otherwise be confused with one or more designations used for the telescope of interest. An telescope may share part of their name with a historical figure for which several things are named. An telescope may share an acronym with an unrelated program. Papers that reference a grant associated with the designated telescope but provide no evidence of using data from it are Not Telescope papers, rather than Science papers or Mentions. If a paper meets the criteria for a Science paper, an Instrumentation paper, or Mention anywhere, then the paper belongs to that category, even if it also contains references to other items that share names in common with the designated telescope or instrument.

	CHANDRA	HST	JWST	None
`train`	31275	37118	11698	294
`test`	3475	4125	1300	294

Table 2: Distribution of dataset entries.

3.3 Data Segmentation for Baseline Task

The TRACS dataset comprises of 89579 entries in total. Table 2 gives the distribution of entries across the three space telescopes and the training and testing dataset splits, as well as across papers that do not feature any space telescopes.

4 Baseline Evaluation Tasks

An automated assistant able to emulate the supervised curation activities listed in 5 above would provide a valuable contribution to the human effort involved. LLMs have shown flexibility in interpreting and classifying scientific articles (Beltagy et al., 2019), which is the basis for this curation activity. They have also been successfully used for information extraction tasks (Grezes and Bartlett), which would help identify the specific datasets mentioned in the papers. This baseline task aims at improving the state of the art technologies to support these curation efforts.

4.1 Task Definition

The TRACS baseline task is composed of two subtasks: *Telescope Classification* and *Usage Classification*, each evaluating a distinct dimension of model understanding over scientific publications.

4.1.1 Telescope Classification

Given the textual fields `title`, `abstract`, `body`, `acknowledgments`, and `grants`, participants were required to predict which telescope was referenced or used in each paper. Valid predictions are limited to `CHANDRA`, `HST`, `JWST`, or `None`. This subtask focuses on assessing the model's ability to correctly identify telescope mentions and usage contexts within natural language.

4.1.2 Usage Classification

The second subtask evaluates how each paper utilizes telescope data. As defined in Section 3.2, each entry includes four boolean labels: `science`, `instrumentation`, `mention`, and `not_telescope`. Each system must output a struc-

tured CSV prediction containing one telescope label and four usage flags for every paper.
In the official Kaggle competition, participants submitted predictions as a single CSV file named `sample_submission.csv` with the following column headers:

```
Id,telescope,science, instrumentation,
mention, not_telescope
```

The `Id` uniquely identifies each paper and is used to align predictions with gold labels during scoring.

4.2 Evaluation Metrics

Each submission is evaluated by matching predictions to reference labels via the `Id` field. Because the test set is partitioned by gold telescope label into four reference splits (`CHANDRA`, `HST`, `JWST`, `NONE`), we report metrics *per split*: within each split, only one of the four telescope classes is populated, so per-split scores measure how well a model recognizes that specific telescope (or correctly abstains, in the `NONE` split). We use F1 scores adapted from the Scikit-Learn library (2011): for telescope classification, the in-class F1 against the gold label; for usage classification, the macro-average over the four boolean usage flags. We also report classification accuracy on the telescope subtask. A model's overall ranking is the average of its per-split scores. A valid example submission is shown below. The first block corresponds to the *Telescope Classification* task, and the second block lists the binary labels for the *Usage Classification* task:

```
Id,telescope
2012A&A...537A..18M,CHANDRA

sci,inst,men,not_tel
True,False,False,False
```

4.2.1 Reproducibility

All models were run with deterministic (greedy) decoding, with an identical instruction prompt and a fixed field-concatenation order (`title`, `abstract`, `body`, `acknowledgments`, `grants`). Per-field input was truncated to fixed token or character budgets calibrated to each model's context window. The full system prompt, per-model context budgets, exact model identifiers, and decoding parameters are reported in Appendix C. All code is publicly available [5] ; All inference scripts, prompt templates, and evaluation notebooks are publicly available [6] .

[5] github.com/eseefried/WASP-dataset-paper26
[6] huggingface.co/datasets/adsabs/TRACS

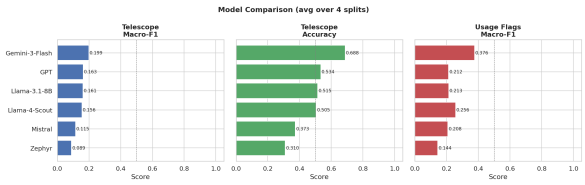


Figure 1: Model comparison averaged over the four reference splits. Gemini-3-Flash leads on every metric; the open 7–20B baselines cluster well below it.

4.3 Baseline Experiments

To establish initial performance benchmarks for the TRACS baseline task, we evaluated six large language models (LLMs) spanning a range of sizes, vendors, and training regimes: *Gemini-3-Flash* (Google, 2025), *GPT-OSS-20B* (OpenAI et al., 2025), *LLaMA-3.1-8B-Instruct* (Grattafiori et al., 2024), *LLaMA-4-Scout* (et.al. , Redacted by arXiv), *Mistral-7B-Instruct* (Jiang et al., 2023), and *Zephyr-7B-Beta* (Tunstall et al., 2023). Each model was run *out of the box*, that is, without any task-specific fine-tuning, using the same instruction set, prompt template, and token limit across all test splits to ensure comparability. Table 3 summarizes the key characteristics of each baseline.

4.3.1 Results and Analysis

Table 4 reports per-split telescope F1 and usage macro-F1 for each model, ordered by average telescope F1. Figure 1 summarizes the same data as average scores across the three primary metrics.

A strong frontier model substantially exceeds open baselines. Gemini-3-Flash dominates on every split and every metric. It reaches 82.4% F1 on `CHANDRA`, 79.0% on `HST`, and 57.1% on `JWST` for telescope recognition, and 64.4% / 46.4% / 39.6% macro-F1 on the corresponding usage subtasks. The next-best models: GPT-OSS-20B and LLaMA-3.1-8B, trail by roughly 15–20 absolute points on telescope F1 and by an even larger margin on usage classification. This indicates that the task is not intrinsically unsolvable in a zero-shot regime, but that it requires the broader world knowledge, instruction-following, and reading comprehension that currently distinguish frontier models from mid-scale open ones.

Telescope difficulty follows corpus prevalence. Across all six models, `CHANDRA` is the easiest telescope to recognize and `JWST` the hardest, with `HST` in between. This ordering matches the relative volume of `CHANDRA`, `HST`, and `JWST` content in the open

Model	Architecture Summary	Motivation	Params
Gemini-3-Flash	Proprietary mixture-of-experts decoder, accessed via API, optimized for fast multimodal reasoning.	Represents a current frontier closed model, providing an upper-bound reference for zero-shot performance.	N/A
GPT-OSS-20B	Decoder-only transformer trained on diverse web and technical text, representing a general-purpose open-source GPT design.	Closest open analog to proprietary GPT-series models; strong general baseline.	20B
LLaMA-3.1-8B-Instruct	Transformer decoder with rotary embeddings and optimized tokenization.	Balances compactness with state-of-the-art reasoning and factuality for 8B-scale models.	8B
LLaMA-4-Scout	Mixture-of-experts decoder from the LLaMA-4 family with long-context support.	Tests whether a newer-generation MoE architecture improves over its dense 8B predecessor.	17B active
Mistral-7B-Instruct	Decoder-only dense transformer using grouped-query attention (GQA) and sliding-window context.	Efficient context handling and strong instruction tuning despite small parameter size.	7B
Zephyr-7B-Beta	Transformer decoder fine-tuned via reinforcement learning from human feedback (RLHF).	Represents the Hugging Face community's open instruction-tuned family emphasizing dialogue coherence.	7B

Table 3: Baseline models evaluated on the TRACS dataset. Each model was run "out-of-the-box" with identical prompts and token limits.

	Telescope F1 (%)				Usage Macro-F1 (%)			
Model	**CH.**	**HST**	**JWST**	**NONE**	**CH.**	**HST**	**JWST**	**NONE**
Gemini-3-Flash	**82.4**	**79.0**	**57.1**	99.8	**64.4**	**46.4**	**39.6**	0.0
GPT-OSS-20B	66.9	59.1	35.4	**100.0**	34.1	27.6	23.1	0.0
LLaMA-3.1-8B	68.6	56.2	37.6	95.5	33.4	27.2	24.4	0.0
LLaMA-4-Scout	65.8	48.1	35.0	**100.0**	44.3	25.6	32.6	0.0
Mistral-7B	36.5	24.2	24.3	99.7	31.2	30.6	21.1	0.0
Zephyr-7B	27.2	7.1	8.8	**100.0**	28.9	20.1	8.6	0.0

Table 4: **Per-split baseline performance on TRACS.** Telescope F1 is computed within each gold telescope split. Usage Macro-F1 averages the four usage labels within each split. Models are ordered by average telescope F1.

scientific literature predating each model's training cutoff: Chandra has been operating since 1999 and HST since 1990, while JWST began producing science only in mid-2022. Notably, JWST recognition is still the weakest class even for Gemini-3-Flash, suggesting that recency rather than ambiguity is the limiting factor.

The `NONE` split is nearly saturated. All six models achieve $\geq 95\%$ accuracy on the `NONE` reference split (Table 4, last column), and four of them score 100%. This is consistent with the intuition that abstaining—predicting that a paper involves none of the three target telescopes—is much easier than identifying which specific telescope is involved. However, this strong `NONE` performance comes at a cost: weaker models default to `NONE` indiscriminately. The confusion matrices in Appendix D show that Mistral and Zephyr misclassify 71–94% of true telescope papers as `NONE`, while Gemini-3-Flash misclassifies only 22–29%.

Usage classification is qualitatively harder. For every model, usage macro-F1 is well below the corresponding telescope F1 (Figure 2, Appendix). Within the usage labels, `science` is consistently the strongest signal (reaching 83.1% F1 for Gemini-3-Flash on HST), while `not_telescope` is the hardest—two of the six models score 0.0% on `not_telescope` across every split. This mirrors the distinction observed in earlier dataset-mention work between identifying that an artifact is referenced versus characterizing how it is used, and reinforces the value of *negative* training examples for future work.

Model agreement decreases on hard examples. On the JWST split, pairwise agreement between models ranges from 0.34 (Gemini-3-Flash vs. Zephyr) to 0.86 (Mistral vs. Zephyr), with high agreement clustering among models that perform similarly poorly. A per-row difficulty analysis (Appendix Fig. 4) shows that roughly 60% of JWST papers are mislabeled by every model in the cohort,

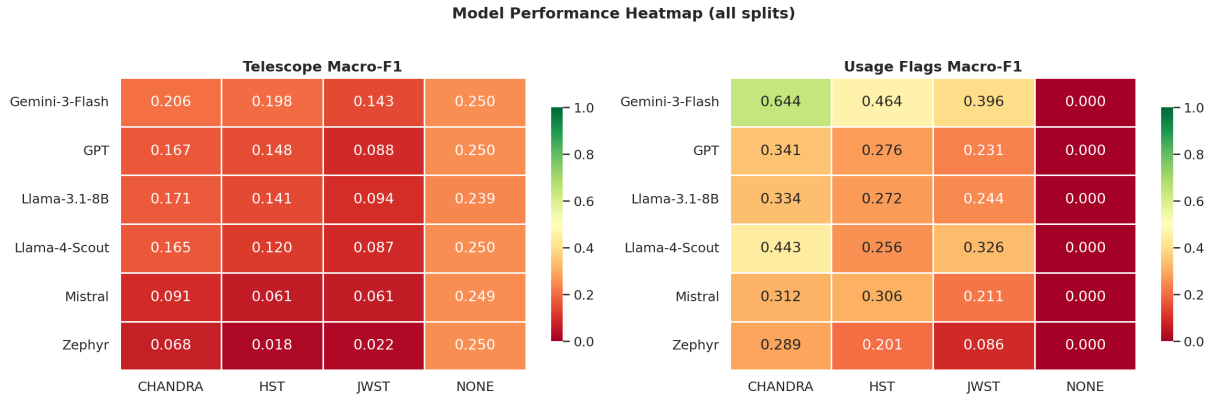


Figure 2: Per-split per-metric heatmap. Left: telescope macro-F1 (computed on the four-class label space, hence diluted by structural zeros from out-of-split classes). Right: usage macro-F1. Gemini-3-Flash is the clear top row across both panels; the NONE column of the usage panel is zero by construction since NONE papers carry no positive usage flags.

indicating that a meaningful fraction of TRACS examples require capabilities that the entire current generation of open models lacks.

Implications for future work. These results carry three practical implications. First, fine-tuning on TRACS is likely to yield substantial gains: even the best zero-shot model leaves $\geq$ 17 points of headroom on CHANDRA telescope F1 and roughly 35 points on the harder usage subtasks. Second, addressing the NONE-bias of smaller models requires either balanced sampling or explicit negative supervision, since several models effectively collapse to a single-class predictor. Third, the hardest residual examples—those that no current model handles—are the most valuable to surface to human curators, suggesting an active-learning workflow in which models triage easy cases and route ambiguous ones to bibliographers.

5 Conclusion and Future Directions

In this paper, we present three contributions to automated bibliographic curation for astronomy. First, we introduce a bibliographic taxonomy developed in collaboration with established bibliographers, grounded in real-world curatorial practices. Second, we provide an open dataset of space telescope bibliographies, manually adapted to align with our taxonomy. Third, we conduct a thorough baseline analysis evaluating the performance of off-the-shelf large language models on a bibliographic curation task. The baseline experiments on TRACS, with the best off-the-shelf LLMs achieving 38% accuracy and 11.5% F1-score on the bibliographic classification task, show that creating bibliographies for space telescopes is not a trivial task to solve, and requires dedicated tools. By releasing the TRACS dataset and taxonomy, we aim to enable further research in this specialized but critical area of scholarly infrastructure. As astronomy archives continue to grow, tools that augment curator expertise will become increasingly essential for maintaining comprehensive and accurate bibliographic records. In the future, we plan to continue to expand TRACS with as many human-curated bibliographies as possible, including ground telescopes. We have already started coordinating with curators at the European Southern Observatory[7] to add the Very Large Telescope[8] to the dataset. In addition to more data, we also would like to refine the evaluation tools. In particular, we would like to use unsupervised evaluation metrics to measure how good models are at recognizing telescope bibliographies from unseen telescopes, evaluating models on:

- Can models generalize and be used to create bibliographies from a new telescope given just a list of names and synonyms for that telescope?
- Can models cluster and detect telescopes in unlabeled astronomy data?
- Can these models be deployed and used by current curators alongside or replacing existing tools?

[7] https://www.eso.org/
[8] https://www.eso.org/public/teles-instr/paranal-observatory/vlt/

References

Jennifer Bartlett, Mugdha Polimera, Kelly Lockhart, Alberto Accomazzi, Michael Kurtz, and Science Explorer Team. 2025. ADS and SciX: Pioneering the Next Generation of Interdisciplinary Research Discovery. In *American Astronomical Society Meeting Abstracts #245*, volume 245 of *American Astronomical Society Meeting Abstracts*, page 442.04.

Jennifer Bartlett, Mugdha Polimera, Kelly Lockhart, Alberto Accomazzi, Michael Kurtz, and Science Explorer Team. 2025. Ads and scix: Pioneering the next generation of interdisciplinary research discovery.

F. E. Bauer, V. V. Dwarkadas, W. N. Brandt, S. Immler, S. Smartt, N. Bartel, and M. F. Bietenholz. 2008. Supernova 1996cr: SN 1987A's Wild Cousin? , 688(2):1210–1234.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv e-prints*, arXiv:1903.10676.

Chandra X-ray Center. 2025. Chandra bibliographic statistics. cxc.harvard.edu/cda/bibstats/bibstats.html. [Online; accessed 14-October-2025].

Jiayin Dong, Chelsea X. Huang, Rebekah I. Dawson, Daniel Foreman-Mackey, Karen A. Collins, Samuel N. Quinn, Jack J. Lissauer, Thomas Beatty, Billy Quarles, Lizhou Sha, Avi Shporer, Zhao Guo, Stephen R. Kane, Lyu Abe, Khalid Barkaoui, Zouhair Benkhaldoun, Rafael Brahm, François Bouchy, Theron W. Carmichael, and 41 others. 2021. Warm Jupiters in TESS Full-frame Images: A Catalog and Observed Eccentricity Distribution for Year 1. , 255(1):6.

Caifan Du, Johanna Cohoon, Patrice Lopez, and James Howison. 2021. Softcite dataset: A dataset of software mentions in biomedical and economic research publications. *Journal of the Association for Information Science and Technology*, 72(7):870–884.

Aaron Adcock et.al. (Redacted by arXiv). 2026. The Llama 4 Herd: Architecture, Training, Evaluation, and Deployment Notes. *arXiv e-prints*, arXiv:2601.11659.

J. C. Good. 1992. Overview of the Astrophysics Data System (ADS). In *Astronomical Data Analysis Software and Systems I*, volume 25 of *Astronomical Society of the Pacific Conference Series*, page 35.

Google. 2025. Gemini 3 flash. `https://deepmind.google/technologies/gemini/flash`. Computer software / Large language model.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 181 others. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, arXiv:2407.21783.

Felix Grezes, Thomas Allen, Tirthankar Ghosal, and Sergi Blanco-Cuaresma. 2023. Function of Citation in Astrophysics Literature (FOCAL): Findings of the Shared Task. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, page 143.

Felix Grezes and Jennifer Bartlett. KAILAS: Keyword Labeler At SciX. huggingface.co/adsabs/KAILAS.

Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S Grant, Donna M Thompson, Roman Chyla, Stephen McDonald, and 1 others. 2021. Building astrobert, a language model for astronomy & astrophysics. *arXiv preprint arXiv:2112.00590*.

Felix Grezes, Sergi Blanco-Cuaresma, Thomas Allen, and Tirthankar Ghosal. 2022. Overview of the First Shared Task on Detecting Entities in the Astrophysics Literature (DEAL). In *First Workshop on Information Extraction from Scientific Publications*, pages 1–7.

Uta Grothkopf and Angelika Treumann. 2003. Towards an Automated Retrieval of Publications based on Telescope Observations. In *Library and Information Services in Astronomy IV (LISA IV)*, page 193.

Edwin A. Henneken and Michael J. Kurtz. 2019. Usage Bibliometrics as a Tool to Measure Research Activity. In *Usage Bibliometrics as a Tool to Measure Research Activity. In: Glänzel W*, pages 819–834. Springer International Publishing, Cham.

Sarthak Jain, Madeleine Van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. Scirex: A challenge dataset for document-level information extraction. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7506–7516.

Tesla E. Jeltema, Claude R. Canizares, Mark W. Bautz, Michael R. Malm, Megan Donahue, and Gordon P. Garmire. 2001. Chandra X-Ray Observatory Observation of the High-Redshift Cluster MS 1054-0321. , 562(1):124–132.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv e-prints*, arXiv:2310.06825.

Michael J Kurtz, Guenther Eichhorn, Alberto Accomazzi, Carolyn S Grant, Stephen S Murray, and Joyce M Watson. 2000. The nasa astrophysics data system: Overview. *Astronomy and astrophysics supplement series*, 143(1):41–59.

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

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large Language Models: A Survey. *arXiv e-prints*, arXiv:2402.06196.

Robert A. Mostoghiu Paun, Darren Croton, Chris Power, Alexander Knebe, Adam J. Ussing, and Alan R. Duffy. 2025. Tidal adaptive softening and artificial fragmentation in cosmological simulations. , 542(2):735–746.

Observatory Bibliographers Collaboration, Raffaele D'Abrusco, Monique Gomez, Uta Grothkopf, Sharon Hunt, Ruth Kneale, Mika Konuma, Jenny Novacescu, Luisa Rebull, Elena Scire, and et al. 2024. Assessing your Observatory's Impact: Best Practices in Establishing and Maintaining Observatory Bibliographies. *The Open Journal of Astrophysics*, 7:85.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv e-prints*, arXiv:2508.10925.

Huitong Pan, Qi Zhang, Eduard Dragut, Cornelia Caragea, and Longin Jan Latecki. 2023. Dmdd: A large-scale dataset for dataset mentions detection. *Transactions of the Association for Computational Linguistics*, 11:1132–1146.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

David Schindler, Felix Bensmann, Stefan Dietze, and Frank Krüger. 2021. Somesci-a 5 star open data gold standard knowledge graph of software mentions in scientific articles. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4574–4583.

Space Telescope Science Institute. 2025. Stsci library and institutional archive. stsci.edu/scientific-community/stsci-library-and-institutional-archive. [Online; accessed 14-October-2025].

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct Distillation of LM Alignment. *arXiv e-prints*, arXiv:2310.16944.

Appendix

A Ethics Statement

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

B Instructions for Human Annotators

`For each row, select which telescope it relates to (CHANDRA, HST, JWST or None) and how it's used (science, instrumentation, mention, or not_telescope) (if telescope=None, leave blank. not_telescope means it has the string in it but it's about something else: ex a paper citing a person named Chandra).`

C Full Reproducibility Details

This appendix provides the complete prompt template, input handling, and decoding configuration used for the baseline experiments in §4.2.1.

C.1 System Prompt

All models received an identical instruction set. For the API model (Gemini-3-Flash) the instructions were passed as a system message; for the local open-weights models (GPT-OSS, Mistral, Zephyr, LLaMA-3.1-8B, LLaMA-4-Scout) they were prepended to the user message. The verbatim instruction text follows:

```
You are an expert research assistant tasked
with analyzing scientific papers from the
TRACS dataset. Your job is to read the
provided fields and produce structured
predictions.

Tasks:
1. Identify the Telescope:
```

```
   - Output exactly one telescope from:
     CHANDRA, JWST, HST, or None.
   - If no telescope is identifiable,
     output None.
2. Classify the Paper's Usage:
   - science -> True if the paper uses
     telescope data to produce new scientific
     results, else False.
   - instrumentation -> True if the paper
     describes technical or engineering
     aspects of the telescope, else False.
   - mention -> True if the paper references
     the telescope but does not generate new
     results, else False.
   - not_telescope -> True if the reference
     is unrelated or a false positive, else
     False.
   Multiple labels can be True.
3. Output Format:
   Return exactly one CSV row and nothing
   else. Do not include the header. The row
   must have exactly these six fields:
   Id,telescope,science,instrumentation,
   mention,not_telescope
   Example row:
   2012A&A...537A..18M_CHANDRA,JWST,True,
   False,False,False
```

C.2 User Message Template

Each paper was rendered into a user message using a fixed field- concatenation order. Fields are inserted verbatim from the dataset entry, subject to the per-model truncation budgets in Appendix C.3.

```
Id={bibcode}
Title: {title}
Abstract: {abstract}
Body: {body}
Acknowledgments: {acknowledgments}
Grants: {grants}
```

C.3 Input Truncation Budgets

Paper bodies routinely exceed model context windows, so each field is truncated to a fixed budget before concatenation. Local open-weights models are budgeted in tokens (using each model's own tokenizer); Gemini-3-Flash uses a character-based budget. Truncation is applied field by field from the end, preserving leading content.

C.4 Decoding Configuration

All baseline models were run with deterministic (greedy) decoding to remove sampling variance from the comparison. For local models we set `do_sample=False`; for Gemini-3-Flash we set `temperature=0` and `thinking_budget=0` to suppress chain-of-thought tokens from the output budget. Nucleus sampling (`top_p`) was not used by any model.

Field	Budget	Formula
GPT-OSS, Mistral, Zephyr, LLaMA-3.1-8B (ctx=8,192 tokens)		
Title	682	$\lfloor ctx/12 \rfloor$
Abstract	2,048	$\lfloor ctx/4 \rfloor$
Body	2,048	$\lfloor ctx/4 \rfloor$
Acknowledgments	512	$\lfloor ctx/16 \rfloor$
Grants	512	$\lfloor ctx/16 \rfloor$
LLaMA-4-Scout (ctx=4,096 tokens)		
Title	256	$\min(256,\ \max(64, ctx/16))$
Abstract	1,024	$\min(1024,\ \max(128, ctx/4))$
Acknowledgments	341	$\min(512,\ \max(64, ctx/12))$
Grants	341	$\min(512,\ \max(64, ctx/12))$
Body	1,622	$ctx - above - 512$ (remainder)
Gemini-3-Flash (ctx=24,000 characters)		
Title	2,000	fixed
Abstract	6,000	fixed
Acknowledgments	3,000	fixed
Grants	3,000	fixed
Body	remainder	see caption

Table 5: Per-field input budgets for every baseline model. Local models share a uniform fractional allocation of an 8,192-token context; LLaMA-4-Scout's tighter context allocates a residual to Body; Gemini-3-Flash uses fixed character budgets, with the Body field receiving $\max(1{,}000,\ 24{,}000 - used - 2{,}000)$ characters.

Model	Temp.	Max out	Sampling
Gemini-3-Flash	0.0	1,024	thinking disabled
GPT-OSS-20B	—	10,000	Greedy
LLaMA-3.1-8B	—	256	Greedy
LLaMA-4-Scout	—	64	Greedy
Mistral-7B	0.0	256	Greedy
Zephyr-7B	0.0	256	Greedy

Table 6: Decoding parameters used for each baseline model. All local models used greedy decoding (`do_sample=False`). GPT-OSS-20B was given a larger output budget to accommodate its reasoning trace preceding the final CSV row.

D Inter-Model Agreement and Per-Paper Difficulty

The aggregate metrics in §4.3.1 report *how often* each model is correct, but not *which papers* each model gets right. To probe this, we examine two complementary views on the JWST split, the hardest of the three telescope subsets.

Pairwise model agreement. Figure 3 shows the fraction of JWST papers on which each pair of models produces the same telescope label, regardless of whether that label is correct. Two patterns stand out. First, agreement is high *among the weaker models*: Mistral and Zephyr agree on 86% of papers, and the four open mid-scale baselines (Llama-3.1-8B, Llama-4-Scout, Mistral, Zephyr) form a visibly high-agreement cluster. Inspection of the

Model	Model ID
Gemini-3-Flash	`gemini-3-flash-preview`
GPT-OSS-20B	`openai/gpt-oss-20b`
LLaMA-3.1-8B	`Llama-3.1-8B-Instruct`
LLaMA-4-Scout	`Llama-4-Scout-17B-16E-Instruct`
Mistral-7B	`OpenHermes-2.5-Mistral-7B`
Zephyr-7B	`HuggingFaceH4/zephyr-7b-beta`

Table 7: Exact model identifiers used for the baseline experiments, matching the configurations in Table 6.

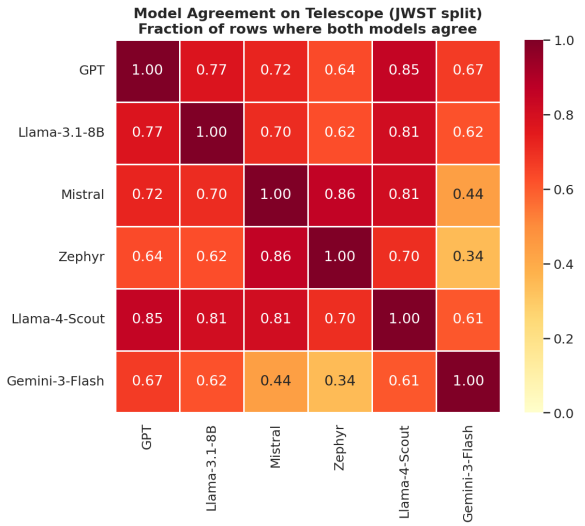


Figure 3: Pairwise model agreement on the JWST split, measured as the fraction of rows on which both models produce the same telescope prediction. Agreement clusters among the weaker models, which tend to default to NONE together; Gemini-3-Flash, the strongest model, shows the lowest agreement with every other model.

underlying predictions shows that this shared behavior is dominated by *co-abstention*: when in doubt, these models converge on NONE. Second, Gemini-3-Flash—the strongest model—shows the *lowest* agreement with every other model in the cohort (0.34 with Zephyr, 0.44 with Mistral). This is consistent with Gemini-3-Flash recovering telescope labels that the other models miss, rather than making independent errors.

Per-paper difficulty. Figure 4 reframes the same predictions from the paper's perspective. For each JWST paper, we count the fraction of the six baseline models that produce the correct telescope label, and plot the distribution. The distribution is sharply bimodal: a large left peak of roughly 750 papers (about 60% of the split) for which *no model* predicts correctly, a small right tail of papers that most or all models handle, and a sparse middle. The left peak is the practically important observation: a substantial fraction of TRACS examples lie outside the current capability frontier of the evaluated open models, and any future system that closes this gap would be providing genuinely new signal rather than re-discovering what easier methods already find. These are the examples that most warrant inclusion in a human-in-the-loop curation workflow.

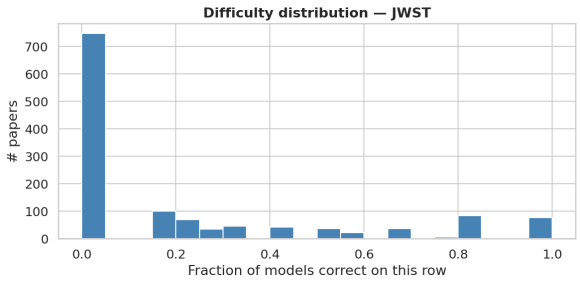


Figure 4: Per-paper difficulty distribution on the JWST split. Each bar counts the number of papers for which a given fraction of the six baseline models predicted the correct telescope. Roughly 60% of JWST papers sit in the leftmost bin, indicating that no model in the current cohort correctly identifies them.