


amc: The Automated Mission Classifier for Telescope Bibliographies

John F. Wu* Joshua E. G. Peek Sophie J. Miller
Jenny Novacescu Achu J. Usha Christopher A. Wilkinson
Space Telescope Science Institute
3700 San Martin Dr
Baltimore, MD 21218 USA
*jfwu@stsci.edu

Abstract

Telescope bibliographies record the pulse of astronomy research by capturing publication statistics and citation metrics for telescope facilities. Robust and scalable bibliographies ensure that we can measure the scientific impact of our facilities and archives. However, the growing rate of publications threatens to outpace our ability to manually label astronomical literature. We therefore present the Automated Mission Classifier (amc), a tool that uses large language models (LLMs) to identify and categorize telescope references by processing large quantities of paper text. A modified version of amc performs well on the TRACS Kaggle challenge, achieving a macro F_1 score of 0.84 on the held-out test set. amc is valuable for other telescopes beyond TRACS; we developed the initial software for identifying papers that featured scientific results by NASA missions. Additionally, we investigate how amc can also be used to interrogate historical datasets and surface potential label errors. Our work demonstrates that LLM-based applications offer powerful and scalable assistance for library sciences. 

1 Introduction

Telescope bibliographies provide one way to measure the scientific productivity of our astronomical facilities. Bibliometrics can quantify how often telescopes are discussed in scientific publications, e.g., through passing mentions or via detailed scientific analyses that originate from data taken by each telescope. Although these quantitative analyses are vital for assessing the impact of our scientific investments, they hinge on complete, homogeneous bibliographies, which can be expensive and onerous to manually curate. Librarians, archive scientists, and bibliographers maintain telescope bibliographies by consistently tracking publications, extracting metadata, and labeling the scientific *intent* of each telescope reference for all papers (see, e.g., Lagerstrom, 2015; Observatory Bibliographers Collab-

oration et al., 2024). Complete observatory bibliographies enable us to investigate publication rates, and citation statistics, links between publications and observing proposals, data product usage metrics, and archival science impact (e.g., Apai et al., 2010, for HST).

There is more scientific literature than ever before (notwithstanding gender-disparate impacts from the recent pandemic, Böhm and Liu, 2023). Some of this increase accompanies a general rise of publication rates throughout academia (Hanson et al., 2024). Additionally, very recent growth in publication rates may stem from the advent of large language models (LLMs), which can lower the barrier to writing papers (e.g., Astarita et al., 2024). These trends suggest that we need a sustainable solution for producing telescope bibliographies amid the deluge of astronomy papers.

LLMs can also be useful for compiling telescope bibliographies at scale: artificial intelligence (AI) systems are highly scalable, and are now adept at processing large amounts of text inputs. Modern LLMs can complete many tasks *without any optimization*, instead relying solely on emergent capabilities like in-context learning (e.g., Brown et al., 2020). With frontier AI labs now deploying LLMs as a service, we can easily leverage simple API (Application Programming Interface) calls and design software around cutting-edge LLMs.

Before deploying an automated bibliography system, we must first ensure that its performance is *robust*. To this end, we present and evaluate the Automated Mission Classifier (amc), an LLM-powered, bibliometric tool for identifying telescopes or NASA missions in the literature. We adapt amc for a specific shared task, TRACS (Section 2); in the Appendix, we note that similar systems are already in operations for JWST (Appendix B) and can be used for archival science with other telescopes (Appendix C). In Section 3, we describe the software’s system design,

and we present results in Section 4. In Section 5, we discuss how observatory bibliographers can leverage AI to compute bibliometrics at scale, assess (historical) data quality, and upgrade the LLM systems. We provide publicly available code on Github: <https://github.com/jwuphysics/automated-mission-classifier>.

“To LLMs! The cause of, and solution to, all of bibliographers’ problems.”¹

2 The TRACS Shared Task

The Telescope Reference and Astronomy Categorization Shared task (TRACS) is a data challenge organized as part of the 2025 Workshop for Artificial Intelligence for Scientific Publications (WASP; Grezes et al. 2025) at IJCNLP-AAACL.² The task consists of classifying astronomy papers into at least one of four categories: science, instrumentation, mention, or not_telescope.

In this data challenge, papers are decomposed into several fields (including the title, abstract, and “body” full text) and, based on keyword filtering, labeled with a candidate telescope name (CHANDRA, HST, JWST, or NONE). The objective is to predict the boolean labels for all paper categories for each of the provided bibcode + telescope combinations. However, it is important to note that the candidate telescope name may be mislabeled, and that certain paper categories impose constraints on the others (i.e., a single paper + telescope can have True labels for both science and instrumentation, but cannot for both science and not_telescope).

Training and test data sets, in CSV format with 80,385 and 9,194 entries respectively, are provided for the shared task. To participate in the challenge, entrants must submit 9,194 test-set predictions via Kaggle³ and have their predictions evaluated. The test outputs are scored according to the average between the macro F_1 score of the telescope labels and the macro F_1 score of paper labels; each class is weighted equally. Note that NONE is a valid telescope class and not_telescope is a valid paper class. In the subsections below, we note some details that we considered important for our submission.

¹Quote adapted from *The Simpsons* (Swartzwelder and Anderson, 1997).

²For details on The International Joint Conference on Natural Language Processing & Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AAACL), see <https://2025.aaclnet.org/>.

³<https://www.kaggle.com/competitions/tracs-wasp-2025>

2.1 Input Data

The full list of columns in the train data set include: (0) ID, (1) bibcode, (2) telescope, (3) author, (4) year, (5) title, (6) abstract, (7) body, (8) acknowledgments, and (9) grants, (10) science, (11) instrumentation, (12) mention, and (13) not_telescope. The test dataset does not include column (2) or columns (10) through (13). However, a preliminary telescope label is implicitly named in column (0), as the ID is simply the concatenation of {bibcode}_{telescope}.

Some rows in the datasets are missing: 3% of the test data set is missing an abstract, 19% does not have full-body text, and > 90% does not have text under the grants column. Incomplete data are likely due to a combination of parsing errors (e.g., correctly parsing out grants/acknowledgments) and publisher restrictions. Issues with publisher agreements tend to impact certain journals or publication venues (i.e., demarcated by their “bibstem” entries); in many of these cases, the body text is completely absent. Nonetheless, classifications can sometimes still be made on the basis of just the title and abstract (but see Appendix C).

Some of the input data may not be helpful. For example, the list of authors is unlikely to yield useful indicators of the paper classification, and may even produce false positives, as “Webb” or “Chandra” can show up as (sub-word) names of authors. Likewise, “Hubble” can often show up in the acknowledgments, e.g. due to funding acknowledgments from the NASA Hubble Fellowship Program. Thus, it is imperative to design a language modeling system that can flexibly understand the context surrounding telescope detections.

2.2 Paper Types

Establishing a common definition for paper types is a nontrivial task. When tasking human bibliographers to classify papers, e.g., identify science papers, disagreements often arise about the precise definition of a science paper.

In order to implement a useful LLM system for automated classification, it is necessary to unambiguously define the labels. *Observatory Bibliographers Collaboration et al. (2024)* issue the following guidance on science papers:

“To qualify as a science paper, it must be apparent that data or data product(s) from the observatory were used and that the

data or data product(s) formed the basis for reaching a new scientific conclusion.”

The authors recognize that these definitions must be continually updated.

Indeed, the taxonomy should serve the telescope or mission. Existing schemes may not be sufficient to characterize all of the edge cases, and new categories may arise. As a concrete example, STScI established a *data-influenced* category in 2019 for papers that indirectly rely on data or products, but do not directly analyze data or use data products. In the TRACS challenge taxonomy, data-influenced papers would generally be labeled under the mention category.

As part of the shared task, the TRACS website provided a narrative format description of the different paper types (Grezes et al., 2025). We used an LLM (claude-sonnet-4.1) to process this text in order to create a user prompt that includes definitions and examples of each paper type (which is manually updated and described in more detail in Section 3). The full prompts can be found in the Github repository, and we have copied the paper type definitions here (note that we remove the markdown text formatting for human readability):

- science: Paper directly uses {{telescope}} data (new or published) to obtain new scientific results in this paper.
- instrumentation: Paper describes new instrument science or engineering.
- mention: Paper references the telescope but does not produce new scientific results.
- not_telescope: Paper includes references that are false positives – names that look like the telescope but refer to something completely different.

3 The Automated Mission Classifier (amc)

Figure 1 shows a high-level overview of the amc system. The system classifies a single paper and a single telescope at a time.

First, amc performs a keyword search to filter all mentions of telescope-related keywords, and we include surrounding context (± 3 sentences). This step effectively converts the body into a list of telescope-specific text snippets (Section 3.1). Text snippets are then ranked by their relevance to the core question of “is this a {telescope} science paper?”, and we only keep the most relevant snippets

(Section 3.2). These top-ranked snippets are subsequently passed to an LLM, which is prompted to classify the paper types and provide quotes and supporting reasoning for its predictions (Section 3.3). The specialized code used for TRACS is forked from amc and can be found at https://github.com/jwuphysics/tracs_was2025.

Finally, we note that our LLM system design is strongly influenced by a prior task: classifying whether arXiv paper preprints contained JWST science. In Appendix B, we describe how these earlier motivations shaped (and biased) the design of the amc. Additional discussion of the limitations of amc are discussed in Section 4.4.

3.1 Keyword Filtering on Full Text

We concatenate the title, abstract, and body as a single text input. We extract only the most relevant portions of the text by searching for keywords. First, we divide text into sentences by using the Punkt sentence tokenizer (Kiss and Strunk, 2006) in the NLTK package (Bird and Loper, 2004). We then use a simple Python case-insensitive string search to identify sentences with keywords for the relevant telescope. We expand snippets to include the $n = 3$ prior and following sentences (i.e., such that each snippet contains $2n + 1$ sentences). If no keywords are found, then we automatically classify the paper as not_telescope.

We note that our keywords prioritize high recall at the expense of low precision; in other words, we value keyword completeness to make sure that no important keywords are missed. However, this means that false positives are expected. For example, our simple string matching over “COS” (an instrument on the Hubble Space Telescope) will also trigger matches on the words “cosmic” or “cosine.” Therefore, it is essential that we guard against false positives by ranking text snippets according to their relevance.

3.2 Reranking Excerpts

Rerankers are typically LLMs that determine the relevance of some text snippet for answering a specific question. In information retrieval systems or retrieval-augmented generation (RAG), a first-stage algorithm usually produces an initial ranking or filtering over relevant documents/snippets (e.g. via semantic similarity in an embedding space). Rerankers provide a second-stage ranking between the query and a smaller set of snippets; recent works have demonstrated them valuable for LLM

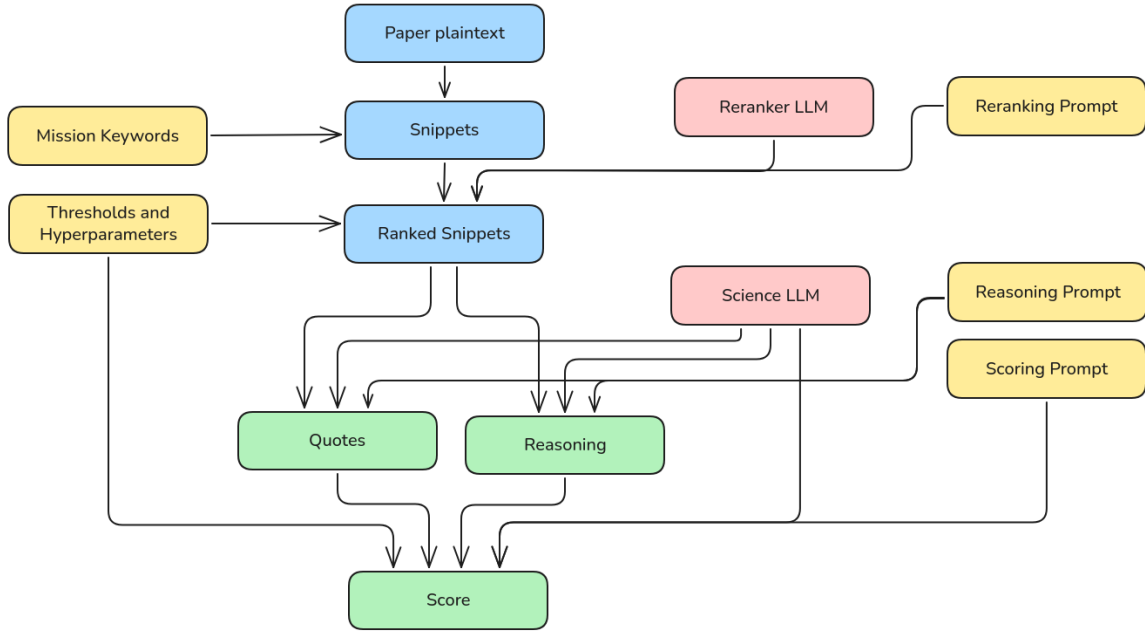


Figure 1: Schematic showing the system design for amc. Note that the version of amc adapted for TRACS does not separate the LLM generation of quotes, reasoning, and predictions; rather they are all output together. See the text in Section 3 and Appendix B for a full discussion of differences.

systems in astronomy and science more broadly (Iyer et al., 2024; Chen et al., 2025; Xu et al., 2025).

We implement a custom reranker solution⁴ that achieves similar performance to leading commercial products (e.g., Cohere Reranker v3.5; based on a few informal assessments). We use a lightweight, general-purpose, non-thinking model (gpt-4.1-nano) with a restricted vocabulary (“Yes” and “No”) that outputs logits (log-probabilities) between 0 and 1. Using a short reranker prompt, we task this model to identify whether each individual snippet discusses the telescope in a way that may be used to classify the paper type. One of the main goals of this step is to remove accidental and unrelated keyword matches.

Once every snippet has a reranker score, we can sort them and/or filter out irrelevant snippets. We keep up to $k = 15$ top-ranked snippets in order to reduce the amount of text that is sent to the next LLM call.

3.3 LLM Classifications

We combine the filtered text snippets together along with their reranker scores. The scores can serve as

⁴We were unaware (until the time of writing) that this reranking approach had been proposed in the literature before (see e.g. Liang et al., 2023).

another reference for whether snippets are useful for determining the paper type.

We use gpt-5-mini with minimal reasoning effort to make the final classification as a structured output. The LLM prompt contains the top-ranked snippets and their scores, and it defines the different paper types and provides some examples. In addition to predicting boolean classes for the science, instrumentation, mention, and not_telescope paper types, the LLM is also prompted to supply the most relevant quotes and justify its reasoning. All structured outputs and their data types are constrained via a pydantic model schema (e.g., boolean predictions, a list of strings for the quotes, and a single text string for the reasoning).

In the amc package, the quotes and justification are provided first, followed by a separate LLM call to predict the final score on these lines of thinking (see, e.g., Figure 1). However, because TRACS requires multiple classifications, we simplify the system so that all predictions and quotes/reasoning are output at the same time. The original amc also supports floating point values between 0 and 1 for scoring science paper types, which allows another hyperparameter to control the threshold for scoring science papers. For TRACS, we simplified the system by using boolean values for each prediction.

4 Results

We briefly present some limited results on the TRACS test set. Our best score in terms of F_1 is 0.84 on the held-out test set, enough for a third-place rank according to the Kaggle leaderboard. In Appendix A, we show the amc JSON-formatted outputs, including paper type examples for science (Listing 1), instrumentation (Listing 2), mention (Listing 3), and not_telescope (Listing 4). Based on a cursory review, these outputs seem accurate, the quotes do not suffer from hallucinations (although the risk is still present), and the provided reasoning largely appears to be faithful to its classification.

4.1 Evaluating amc

In order to understand our system’s strengths and weaknesses, we select $N = 100$ random entries from the training set, comprising 25 rows per telescope. This small, non-representative evaluation set enables us to investigate why our LLM system tended to make incorrect predictions. This random set is also able to surface potential issues with the dataset (see Section 4.3).

In Figure 2, we show confusion matrices displaying amc predictions on the limited validation set, for all telescopes except NONE. Each column shows a paper type (denoted “True”) against all other paper types (denoted “False”). We note that some combinations of missions and paper types tend to succeed (e.g., CHANDRA/science) or fail more frequently (e.g., CHANDRA/mention). These confusion matrices are based on the same version of amc as the final TRACS submission. However, we caution against overinterpreting results on this relatively small evaluation set.

4.2 Performance on TRACS

Our first submission to TRACS achieved a macro F_1 score of 0.80. At the time, the system included a few suboptimal settings, e.g., slightly misspecified prompts, or a non-zero reranker threshold which caused weak mention classes to occasionally be mislabeled (since the threshold might cause all text snippets to be filtered out, rendering a default verdict of not_telescope).

After removing the reranker threshold and updating the prompts, we saw a modest increase in macro F_1 score to 0.84. We examined two of our higher-scoring sets of predictions, and used an LLM as a judge (gpt-5-mini) to resolve discrepancies be-

tween them and to issue final predictions; the performance remained at $F_1 = 0.84$.

The final LLM system took less than 24 hours in wall-clock time to run, and incurred roughly \$10 in OpenAI costs. About 22% of the cost is for reranking snippets with gpt-4.1-nano, 37% is for processing top-ranked inputs with gpt-5-mini, and 41% is for generating outputs with gpt-5-mini. Batch processing could lower some costs, but would necessitate an asynchronous pipeline, where we first perform all reranker calls, followed by all LLM classifications.

4.3 Missing Data and Label Errors

The dataset likely contains errors or uncertain classifications due to the imperfect nature of manually annotating bibliographic data, and the somewhat subjective nature of label distinctions. However, it is not possible to capture this uncertainty in the discrete classes. We also cannot measure the error rate directly, as there is no *golden sample* against which we can compare. A golden sample would consist of papers that have been independently classified by multiple reviewers, where cases of disagreement are subject to deliberation and re-review until consensus is reached. Therefore, the error rate or uncertainty is unknown.

Through repeated evaluation, we can surface potential errors in the TRACS dataset. While testing our LLM system on a small ($N = 100$) subsample from the training data, we inspected all cases where the LLM prediction disagreed with the target label. Some of these appeared to be genuine error or ambiguity in the ground truth dataset, and we display them in Table 1.

We find that one paper is labeled as both science and mention, which (we assume) should not be possible. This classification may have resulted from human annotation error, or perhaps an accidental combination of a HST mention (as the paper is about the Hubble Deep Field) and CHANDRA science. Simultaneously conflicting labels like this can be easily filtered out by using boolean logic and some set rules. We find another paper that mentions the “Next Generation Space Telescope,” the original name for JWST. Arguably, this paper should be considered a JWST mention, but is instead labeled as not_telescope.

Three papers are missing their body text; they are only described by their titles, abstracts, and other metadata. For each of these three entries, we verify that (within the TRACS data) there is no

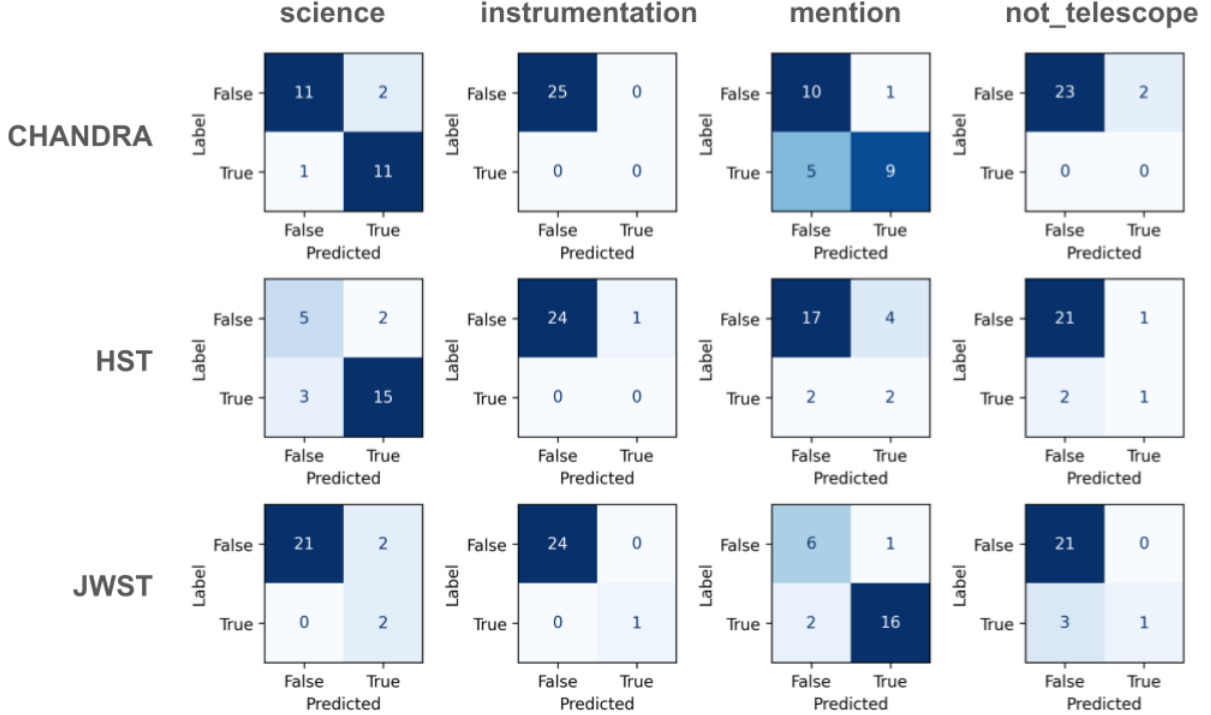


Figure 2: Binary confusion matrices, over a randomized subsample of 25 TRACS training set examples for each telescope (*rows*), shown for each paper type (*columns*) as a one-vs-all classification. Each confusion matrix shows true negatives (*top left*), false positives (*top right*), false negatives (*bottom left*), and true positives (*bottom right*).

Table 1: Potential dataset issues found in a random selection of 100 labeled examples.

Bibcode	Notes
2001AJ...122..598D	Labeled as both CHANDRA/science and CHANDRA/mention.
2001ApJ...550..104Y	Missing body text, making it impossible to correctly classify as HST/science.
2002IJMPA...17..3446T	Mentions “Next Generation Space Telescope” (the placeholder name for JWST), but the label is JWST/not_telescope rather than JWST/mention.
2004RMxAC...20..215S	Missing body text, making it impossible to correctly classify as CHANDRA/science.
2004fxra.book...89D	Missing body text, making it impossible to correctly classify as CHANDRA/science.

mention or science/instrumentation of the candidate telescope presented. Missing full body text is often a symptom of complex publisher licensing agreements, and it may not always be possible to procure the full data. In any event, such entries do not contain sufficient data for making accurate predictions.

4.4 Limitations of amc

As noted above, the amc system is designed to be general. Although we have specialized the code for the TRACS task, there are additional adjustments that could lead to improved performance. For example, the multiclass predictions would benefit from dedicated prompts for each paper type. The current system effectively uses the same prompts

for each telescope, which might also limit its performance.

We also note that amc is at the mercy of our keyword filtering. If we miss any telescope keywords, then it is possible to filter out relevant snippets, which could jeopardize the prediction task performance. Frequent keywords could be empirically learned using traditional NLP techniques like term frequency (TF; [Spark Jones 1972](#)) normalized by its document frequency (i.e., TF-IDF; [Salton et al. 1975](#)). The reranker step could potentially be replaced by a simple first-pass classifier using TF-IDF or another data-driven approach.

5 Discussion

LLMs are becoming pervasive throughout astronomy. Quantitative benchmarks (Carrit Delgado Pinheiro et al., 2025; Joseph et al., 2025; Ting et al., 2025) and human-centered studies (Fouesneau et al., 2024; Wu et al., 2024; Hyk et al., 2025) deliver complementary evaluations for how to successfully deploy LLMs for real-world benefit in astronomy. There is also rapid adoption of LLMs for navigating through and interacting with the astronomy literature (Ciucă and Ting, 2023; Iyer et al., 2024), which is particularly salient for WASP/TRACS.

As researchers are in the midst of a fundamental shift of how they interact with literature, we discuss a future vision of how the astronomical community may leverage LLMs to augment or automate bibliographies (Section 5.1), how AI systems can assist in evaluating or improving our ground truth datasets (Section 5.2), and how the amc software we presented could be improved further in future work (Section 5.3).

5.1 Scalable, AI-Supported Bibliographies

We have shown that compiling telescope bibliographies can be assisted by or partially automated with LLMs. LLM developments are built on traditional NLP techniques, which have already been vital for astronomical literature review (Iyer et al., 2024) and detecting usage of telescopes/facilities (e.g., using TF-IDF, Amado Olivo et al., 2025). While LLMs can be more expensive to put into production relative to simple NLP techniques or specialized fine-tuned models (e.g., SciBERT, Beltagy et al., 2019), LLMs that have been pre-trained on trillions of tokens of general text are also capable of in-context learning via zero- or few-shot demonstrations (Radford et al., 2019; Brown et al., 2020). Modern LLMs also have longer context windows, enabling them to ingest multiple text snippets (or even entire documents at a time). This feature is particularly valuable if the telescope classification depends on nuanced text snippets buried within the body (i.e., often the case for *archival* data sets, and rare for *flagship* NASA missions; see Appendix C).

AI systems can still be extremely useful even if manual vetting of bibliographies is necessary. We have designed amc to have high recall, so it can confidently remove from consideration papers that have no chance of being mention paper types. Accurate labels ($F_1 > 0.8$) can dramatically save

human time and mental energy.

5.2 Errors and Ground Truths

When creating LLM-augmented bibliographies at massive scale, it is imperative to understand how the LLM is susceptible to errors, and/or if those errors originate from the LLM or from the dataset. For TRACS, our analysis of a small subsample in Section 4.3 resulted in direct performance gains; we exposed some issues with our system, as well as errors in the dataset.

We emphasize the value in compiling a golden sample with consensus reviews, even if this dataset is much smaller compared to the archival set of (single-pass) human classifications. In our prior work (see Appendices B and C), we have relied on a golden sample with about $N \sim 100$ examples to serve as a benchmark for improving the LLM system (Shaw et al., *in prep*). Crucially, it also serves as a measure of *human performance*, which is often incorrectly assumed to be perfect. By setting human error rates as the error “floor,” we can quantify a goal for LLMs to achieve.

AI augmentation can also facilitate a better understanding of our datasets. For example, LLMs can easily comb through a large number of negative classes from *historical* datasets, and surface candidate missing papers or other errors (e.g., Section 4.3). An LLM can be vital for efficiently constructing such a golden sample dataset.

5.3 Future Improvements

Our solution for the TRACS task can likely benefit from additional optimization. In particular, other LLMs can help iteratively optimize the prompts used to guide the (TRACS-specific) amc code, by using meta-optimizers (see, e.g., Opsahl-Ong et al., 2024; Agrawal et al., 2025) in a prompt compilation framework like DSPy (Khattab et al., 2023). Given the large TRACS training data set, meta-optimization could be costly, and may be precariously sensitive to the training label quality. However, meta-optimization could also produce (as a byproduct) empirical definitions of paper types like science or instrumentation, which could be valuable for comparing against explicit definitions that bibliographers have historically adopted.

Another option is to use AI agents: LLMs that can call tools in a loop in order to accomplish a task.⁵ Even though an AI agent might access the

⁵For one definition of an AI “agent” that we like, see <https://simonwillison.net/2025/Sep/18/agents/>

same tools that we have described in Section 3, e.g., keyword search, reranking, filtering, or summarization, the LLM’s *agency* means that it can decide when and how to use such tool calls. The LLM agent can also maintain a working memory, allowing it to determine whether it has enough information to make a classification; for instance, if it finds immediate evidence that the paper presents scientific results, then the agent can stop the analysis and classify the paper as science.

Finally, we may wish to deploy *smaller*, specialized models for this task because they can be run locally and perhaps at lower costs. For example, our keyword filtering and reranking steps are somewhat reminiscent of “late-interaction” retrieval mechanisms (e.g., ColBERT, [Khattab and Zaharia, 2020](#)), and it may be advantageous to substitute those steps with more lightweight model like ColBERT. We might simplify further by substituting this initial stage with classical NLP algorithms like TF-IDF. Models with specialized tokenizers for scientific literature like SciBERT ([Beltagy et al., 2019](#)) may also prove to be beneficial for parsing the astronomical literature.

6 Summary

We have presented amc, an LLM-based system that can automatically categorize real astronomical papers into specific labels. Using a specialized instance of amc, we demonstrate strong performance ($F_1 = 0.84$) and secure third place on the TRACS shared task ([Grezes et al., 2025](#)). Our tool is also valuable for evaluating labeled data quality, as it provides reasoning and supporting quotes to justify its predicted labels. Given the growing volume of papers, as well as the rising capabilities of LLMs, we believe that AI tools represent scalable solutions for accomplishing or assisting with this task.

In the future, however, LLMs may completely obviate the need for predefined “classifications” that comprise current paper types; instead, we may be able to *directly* ask LLMs questions like: “How many papers present ground-based follow-up observations for targets initially discovered with HST?” or “How did the fraction of Chandra *archival* science papers change between 2010 through 2025?” We envision that, by exploiting the capabilities of AI systems, library scientists can study a broader range of bibliographic questions than ever before.

Ethical Disclosure

All of this text was written solely by the authors. The document was partially reviewed by LLMs, primarily gpt-5 and Gemini 2.5 Pro, in order to surface issues in clarity and prose. Some of the code in the associated repository is generated by LLMs, primarily via Claude Code. Having validated the software and results, the authors take full responsibility and ownership over the results presented here.

Acknowledgments

We acknowledge helpful conversations with STScI staff, including Brian Cherinka, Dick Shaw, and Jinmi Yoon. Some of the ideas about combining the LLM system with TF-IDF and SciBERT are based on exchanges with Madhusudhana Naidu after the end of the TRACS competition.

References

- Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziemis, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2025. [Gepa: Reflective prompt evolution can outperform reinforcement learning](#). *Preprint*, arXiv:2507.19457.
- Vicente Amado Olivo, Wolfgang Kerzendorf, Brian Cherinka, Joshua V. Shields, Annie Didier, and Katharina von der Wense. 2025. [Identifying Telescope Usage in Astrophysics Publications: A Machine Learning Framework for Institutional Research Management at Observatories](#). *The Astronomical Journal*, 169(1):42.
- Dániel Apai, Jill Lagerstrom, Iain Neill Reid, Karen L. Levay, Elizabeth Fraser, Antonella Nota, and Edwin Henneken. 2010. [Lessons from a High-Impact Observatory: The Hubble Space Telescope’s Science Productivity between 1998 and 2008](#). *Publications of the Astronomical Society of the Pacific*, 122(893):808.
- Simone Astarita, Sandor Kruk, Jan Reerink, and Pablo Gómez. 2024. [Delving into the Utilisation of ChatGPT in Scientific Publications in Astronomy](#). In *Proceedings of SPAICE2024: The First Joint European Space Agency / IAA Conference on AI in and for Space*, pages 241–246.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. [SciBERT: A pretrained language model for scientific text](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–

- 3620, Hong Kong, China. Association for Computational Linguistics.
- Steven Bird and Edward Loper. 2004. [NLTK: The natural language toolkit](#). In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.
- Vanessa Böhm and Jia Liu. 2023. [Impact of the COVID-19 pandemic on publishing in astronomy in the initial two years](#). *Nature Astronomy*, 7:105–112.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- Lucas Carrit Delgado Pinheiro, Ziru Chen, Bruno Caixeta Piazza, Ness Shroff, Yingbin Liang, Yuan-Sen Ting, and Huan Sun. 2025. [Large Language Models Achieve Gold Medal Performance at the International Olympiad on Astronomy & Astrophysics \(IOAA\)](#). *arXiv e-prints*, arXiv:2510.05016.
- Haotian Chen, Qingqing Long, Meng Xiao, Xiao Luo, Wei Ju, Chengrui Wang, Xuezhi Wang, Yuanchun Zhou, and Hengshu Zhu. 2025. [Scirerankbench: Benchmarking rerankers towards scientific retrieval-augmented generated llms](#). *Preprint*, arXiv:2508.08742.
- Ioana Ciucă and Yuan-Sen Ting. 2023. [Galactic ChitChat: Using Large Language Models to Converse with Astronomy Literature](#). *Research Notes of the American Astronomical Society*, 7(9):193.
- Morgan Fouesneau, Ivelina G. Momcheva, Urmila Chadayammuri, Mariia Demianenko, Antoine Dumont, Raphael E. Hviding, K. Angelique Kahle, Nadiia Pulatova, Bhavesh Rajpoot, Marten B. Scheuck, Rhys Seeburger, Dmitry Semenov, and Jaime I. Villaseñor. 2024. [What is the Role of Large Language Models in the Evolution of Astronomy Research?](#) *arXiv e-prints*, arXiv:2409.20252.
- Felix Grezes, Jennifer Lynn Bartlett, Kelly Lockhart, Alberto Accomazzi, Ethan Seefried, and Tirthankar Ghosal. 2025. Overview of TRACS: the telescope reference and astronomy categorization dataset & shared task. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.
- Mark A Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. 2024. The strain on scientific publishing. *Quantitative Science Studies*, 5(4):823.
- E. A. Henneken and A. Accomazzi. 2012. [Linking to Data: Effect on Citation Rates in Astronomy](#). In *Astronomical Data Analysis Software and Systems XXI*, volume 461 of *Astronomical Society of the Pacific Conference Series*, page 763.
- Alina Hyk, Kiera McCormick, Mian Zhong, Ioana Ciucă, Sanjib Sharma, John F Wu, J. E. G. Peek, Kartheik G. Iyer, Ziang Xiao, and Anjalie Field. 2025. [From queries to criteria: Understanding how astronomers evaluate LLMs](#). In *Second Conference on Language Modeling*.
- Kartheik G. Iyer, Mikael Yunus, Charles O’Neill, Christine Ye, Alina Hyk, Kiera McCormick, Ioana Ciucă, John F. Wu, Alberto Accomazzi, Simone Astarita, Rishabh Chakrabarty, Jesse Cranney, Anjalie Field, Tirthankar Ghosal, Michele Ginolfi, Marc Huertas-Company, Maja Jabłońska, Sandor Kruk, Huiling Liu, and 11 others. 2024. [pathfinder: A Semantic Framework for Literature Review and Knowledge Discovery in Astronomy](#). *The Astrophysical Journal Supplements*, 275(2):38.
- Sebastian Antony Joseph, Syed Murtaza Husain, Stella S. R. Offner, Stéphanie Juneau, Paul Torrey, Adam S. Bolton, Juan P. Farias, Niall Gaffney, Greg Durrett, and Junyi Jessy Li. 2025. [AstroVis-Bench: A Code Benchmark for Scientific Computing and Visualization in Astronomy](#). *arXiv e-prints*, arXiv:2505.20538.
- Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2023. [Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp](#). *Preprint*, arXiv:2212.14024.
- Omar Khattab and Matei Zaharia. 2020. [Colbert: Efficient and effective passage search via contextualized late interaction over bert](#). *Preprint*, arXiv:2004.12832.
- Tibor Kiss and Jan Strunk. 2006. [Unsupervised multi-lingual sentence boundary detection](#). *Computational Linguistics*, 32(4):485–525.
- J. Lagerstrom. 2015. Best Practices for Creating an Observatory or Telescope Bibliography from the IAU Commission 5 Working Group on Libraries. In *Open Science at the Frontiers of Librarianship*, volume 492 of *Astronomical Society of the Pacific Conference Series*, page 99.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. [Holistic evaluation of language models](#). *Transactions on Machine Learning Research*. Featured Certification, Expert Certification, Outstanding Certification.
- Observatory Bibliographers Collaboration, Raffaele D’Abrusco, Monique Gomez, Uta Grothkopf, Sharon Hunt, Ruth Kneale, Mika Konuma, Jenny Novacescu,

- Luisa Rebull, Elena Scire, and et al. 2024. [Assessing your Observatory’s Impact: Best Practices in Establishing and Maintaining Observatory Bibliographies](#). *The Open Journal of Astrophysics*, 7:85.
- Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. [Optimizing instructions and demonstrations for multi-stage language model programs](#). *Preprint*, arXiv:2406.11695.
- Joshua Peek, Vandana Desai, Richard L. White, Raffaele D’Abrusco, Joseph M. Mazzeella, Carolyn Grant, Jenny Novacescu, Elena Scire, and Sherry Winkelman. 2019. [Robust Archives Maximize Scientific Accessibility](#). In *Bulletin of the American Astronomical Society*, volume 51, page 105.
- Alberto Pepe, Alyssa Goodman, August Muench, Merce Crosas, and Christopher Erdmann. 2014. [How do astronomers share data? reliability and persistence of datasets linked in aas publications and a qualitative study of data practices among us astronomers](#). *PLOS ONE*, 9(8):1–11.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). *OpenAI Blog*.
- G. Salton, A. Wong, and C. S. Yang. 1975. [A vector space model for automatic indexing](#). *Commun. ACM*, 18(11):613–620.
- Karen Spark Jones. 1972. [A statistical interpretation of term specificity and its application in retrieval](#). *Journal of Documentation*, 28(1):11–21.
- John Swartzwelder and Bob Anderson. 1997. Homer vs. the Eighteenth Amendment. Television series episode, spoken by Homer Simpson: “To alcohol! The cause of, and solution to, all of life’s problems.”.
- Y. S. Ting, T. D. Nguyen, T. Ghosal, R. Pan, H. Arora, Z. Sun, T. de Haan, N. Ramachandra, A. Wells, S. Madireddy, and A. Accomazzi. 2025. [AstroMLab 1: Who wins astronomy jeopardy!?](#) *Astronomy and Computing*, 51:100893.
- John F. Wu, Alina Hyk, Kiera McCormick, Christine Ye, Simone Astarita, Elina Baral, Jo Ciuca, Jesse Cranney, Anjalie Field, Kartheik Iyer, Philipp Koehn, Jenn Kotler, Sandor Kruk, Michelle Ntampaka, Charles O’Neill, Joshua E. G. Peek, Sanjib Sharma, and Mikaeel Yunus. 2024. [Designing an Evaluation Framework for Large Language Models in Astronomy Research](#). *arXiv e-prints*, arXiv:2405.20389.
- Xueqing Xu, Boris Bolliet, Adrian Dimitrov, Andrew Laverick, Francisco Villaescusa-Navarro, Licong Xu, and Íñigo Zubeldia. 2025. [Evaluating Retrieval-Augmented Generation Agents for Autonomous Scientific Discovery in Astrophysics](#). *arXiv e-prints*, arXiv:2507.07155.

A Examples of amc Structured Outputs for the TRACS Test Set

In Listings 1, 2, 3, and 4, we show several representative examples of amc outputs for the TRACS test dataset.

```
"2024ApJ...977..173C_JWST": {
  "id": "2024ApJ...977..173C_JWST",
  "bibcode": "2024ApJ...977..173C",
  "telescope": "JWST",
  "classification": {
    "telescope": "JWST",
    "science": true,
    "instrumentation": false,
    "mention": false,
    "not_telescope": false,
    "quotes": [
      "we present JWST MIRI observations of the hydrocarbon-rich source, DoAr 33, a 1.1 M star.",
      "We detect the emission of H 2 O, CO 2 , OH, and C 2 H 2 , including its isotopologue 13 C
      ↳ 12 CH 2 , C 4 H 2 , and tentatively CH 4 and HC 3 N, for all of which we retrieve co
      lumn densities, temperatures, and emitting areas, as described in Section 2 .",
      "We detect the presence of H 2 O, CO 2 , OH, C 2 H 2 , HCN, C 4 H 2 , and tentatively CH 4
      ↳ and HC 3 N in the JWST MIRI-MRS spectrum of the solar-mass star DoAr 33.",
      "JWST/MIRI Detection of a Carbon-rich Chemistry in the Disk of a Solar Nebula Analog",
      "Figure 1. JWST MIRI-MRS spectrum of DoAr 33."
    ],
    "reasoning": "The excerpts clearly describe original JWST MIRI-MRS observations of the target
    ↳ DoAr 33 and report new detections and quantitative retrievals (column densities, te
    mperatures, emitting areas) derived in this paper. These are new scientific results based on JWST
    ↳ data, so 'science' is True. The text discusses data reduction using a custom JDISCS p
    ipeline and calibration references but does not present new instrument design, calibration method
    ↳ development as the primary focus or a technical/instrumentation paper, so 'instrument
    ation' is False. Because the paper actively uses JWST data to produce new results, it is not
    ↳ merely a mention; therefore 'mention' is False. There is no indication that references to
    JWST are false positives, so 'not_telescope' is False."
  }
}
```

Listing 1: amc output for an science paper type.

B The JWST Preprints Automation

The design of amc was initially devised for a specific task: automatically checking whether new arXiv preprints feature JWST science. While identifying JWST science preprints is useful on its own, it is also necessary for another goal: determining whether JWST science publications are accompanied by digital object identifiers (DOIs), which offers downstream benefits to individual scientists (Henneken and Accomazzi, 2012) and the broader science community (Pepe et al., 2014). Therefore, we built an earlier version of amc for automatically classifying whether a JWST preprint is science or not.⁶ We achieved $F_1 \approx 0.95$ on real-world tests, using a held-out golden sample dataset with $N = 114$ papers, and this system is currently being used in production at STScI.

There are two major differences between the JWST preprints automation and the TRACS task: (1) the former downloads PDF files from arXiv and converts them into a single plain-text body, while the latter provides more cleanly separated metadata and body text (when available); and (2) the former task is only concerned with identifying science papers, whereas the latter solicits binary classifications for science, instrumentation, mention, and not_telescope.

There are several important implications due to distinction #1. For example, we do not have a programmatic method for separating abstracts or other titles from the body, and so we must always proceed with the full body text. Another consequence is that the references section is included in the

⁶<https://github.com/spacetelescope/jwst-preprint-automator>

```

"1996SPIE.2808..650F_CHANDRA": {
  "id": "1996SPIE.2808..650F_CHANDRA",
  "bibcode": "1996SPIE.2808..650F",
  "telescope": "CHANDRA",
  "classification": {
    "telescope": "CHANDRA",
    "science": false,
    "instrumentation": true,
    "mention": false,
    "not_telescope": false,
    "quotes": [
      "Modeling the diffraction efficiencies of the AXAF high-energy transmission gratings: II",
      "In order to characterize the instrumentation on AXAF, each of the science instrument teams
        ↳ carries out sub-assembly calibrations.",
      "For the high energy transmission grating (HETG) group, this means individual measurements
        ↳ of the diffraction efficiencies of each of the 336 grating elements that goes into t
        he completed HETG assembly."
    ],
    "reasoning": "The excerpt discusses calibration and modeling of the AXAF high-energy
      ↳ transmission gratings (HETG), measurements, data reduction and instrumentation
      ↳ characterization. AXAF is the original name for the Chandra X-ray Observatory, and the
      ↳ content is clearly about instrument performance and calibration rather than presenting
      ↳ new astrophysical results. Therefore instrumentation = True. It does not present new
      ↳ scientific/astronomical analyses using telescope observations, so science = False.
      ↳ Because the paper actually discusses the telescope/instrument technical work, it should
      ↳ not be classified as a mere mention or as not_telescope."
    ]
  }
}

```

Listing 2: amc output for an instrumentation paper type.

plain-text body extract from arXiv preprints. Thus, the body text contains references to titles of *other* papers, which can sometimes mimic sentences that appear to support a JWST science classification.

Because we focus only on classifying whether a JWST paper is science (distinction #2), we break down the LLM output into two stages (see Figure 1). First, we write out a specialized prompt with in-context examples of low and high science scores, and prompt the LLM to output reasoning and supporting quotes. Then, given its provided justification and quotes, we force the LLM to output a science score as a floating point number between 0 and 1.

C Other Mission Bibliographies

We note that NASA flagship missions tend to be mentioned in the paper abstract, or even in the title. Thus, for TRACS and for JWST preprints automation, it is often possible to accurately classify papers using just the abstract, a tiny fraction of the available text per entry.

However, we designed amc based on our experience performing bibliometrics for other missions, including TESS, Pan-STARRS, and GALEX, which all enable *archival* science well after the telescope ceases operations (e.g., Peek et al., 2019). However, these telescopes—and, in general, archival science usage of telescopes—are less likely to appear in the title or abstract. Therefore, parsing the much longer body text is imperative for correctly predicting those telescopes’ paper types. We design amc so that it can be used just as easily to compute bibliometrics for non-flagship telescopes and/or archival science.


```

"2024A&A...691A.352W_JWST": {
  "id": "2024A&A...691A.352W_JWST",
  "bibcode": "2024A&A...691A.352W",
  "telescope": "JWST",
  "classification": {
    "telescope": "JWST",
    "science": false,
    "instrumentation": false,
    "mention": true,
    "not_telescope": false,
    "quotes": [
      "an unbiased mid-infrared survey observing H-dominated metal-enriched white dwarfs making
      ↪ use of the much improved sensitivity of JWST, compared to previous facilities, would
      ↪ likely reveal the true population of discs.",
      "infrared spectroscopy can identify the mineralogy of discs, with silicates having been
      ↪ detected using Spitzer and JWST ( Jura et al. 2007 ; Swan et al. 2024 )."
    ],
    "reasoning": "The excerpt discusses JWST in the context of future or potential surveys and
    ↪ improved sensitivity compared to previous facilities, and cites JWST detections (Swan et
    ↪ al. 2024) as literature. There is no indication that this paper presents new JWST
    ↪ observational data or original data analysis; it instead suggests that an unbiased survey
    ↪ with JWST would be useful and refers to prior JWST detections. This fits the 'mention'
    ↪ category (discussion of future/planned observations and citing others). It does not
    ↪ describe instrumentation or calibration, and the references to JWST clearly refer to the
    ↪ telescope, so not_telescope is false."
  }
}

```

Listing 3: amc output for a mention paper type.

```

"2008ApJ...673..657M_HST": {
  "id": "2008ApJ...673..657M_HST",
  "bibcode": "2008ApJ...673..657M",
  "telescope": "HST",
  "classification": {
    "telescope": "HST",
    "science": false,
    "instrumentation": false,
    "mention": false,
    "not_telescope": true,
    "quotes": [
      "Hubble diagram showing the magnitude deviation relative to an empty universe, for the three
      ↪ models considered.",
      "Fig. 4. – Hubble diagram showing the magnitude deviation relative to an empty universe, for
      ↪ the three models considered. The dotted lines separate the Tonry et al. ( 2003 ) data (
      ↪ left ) from the mock data ( right )."
    ],
    "reasoning": "The excerpts discuss a 'Hubble diagram' and cosmological supernova analyses;
    ↪ there is no mention of the Hubble Space Telescope or HST observations, instruments,
    ↪ calibrations, or use of HST data. The term 'Hubble' here refers to the Hubble diagram
    ↪ (relation to Hubble expansion), which is a false positive for the telescope name.
    ↪ Therefore this is not about the HST telescope (not_telescope = True). Because it does not
    ↪ discuss any telescope data usage or instrumentation, science and instrumentation are
    ↪ False. It also is not merely a mention of the telescope, so mention = False. The provided
    ↪ exact substrings showing 'Hubble diagram' are included as supporting quotes."
  }
}

```

Listing 4: amc output for a not_telescope paper type.