AstroConcepts: A Large-Scale Multi-Label Classification Corpus for Astrophysics

Atilla Kaan Alkan[1], **Felix Grezes**[1], **Sergi Blanco-Cuaresma**[1,2], **Jennifer Lynn Bartlett**[1], **Daniel Chivvis**[1], **Anna Kelbert**[1], **Kelly Lockhart**[1], **Alberto Accomazzi**[1]

[1]Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA
[2]Faculty of Psychology, UniDistance Suisse, Brig, Switzerland
{atilla.alkan, felix.grezes, sblancocuaresma, jennifer.bartlett, daniel.chivvis, anna.kelbert, kelly.lockhart, alberto.accomazzi}@cfa.harvard.edu

Abstract

Scientific multi-label text classification suffers from extreme class imbalance, where specialized terminology exhibits severe power-law distributions that challenge standard classification approaches. Existing scientific corpora lack comprehensive controlled vocabularies, focusing instead on broad categories and limiting systematic study of extreme imbalance. We introduce AstroConcepts, a corpus of English abstracts from 21,702 published astrophysics papers, labeled with 2,367 concepts from the Unified Astronomy Thesaurus. The corpus exhibits severe label imbalance, with 76% of concepts having fewer than 50 training examples. By releasing this resource, we enable systematic study of extreme class imbalance in scientific domains and establish strong baselines across traditional, neural, and vocabulary-constrained LLM methods. Our evaluation reveals three key patterns that provide new insights into scientific text classification. First, vocabulary-constrained LLMs achieve competitive performance relative to domain-adapted models in astrophysics classification, suggesting a potential for parameter-efficient approaches. Second, domain adaptation yields relatively larger improvements for rare, specialized terminology, although absolute performance remains limited across all methods. Third, we propose frequency-stratified evaluation to reveal performance patterns that are hidden by aggregate scores, thereby making robustness assessment central to scientific multi-label evaluation. These results offer actionable insights for scientific NLP and establish benchmarks for research on extreme imbalance.

Keywords: Multi-label Text Classification, Scientific Document Classification, Extreme Label Imbalance

1. Introduction

Scientific multi-label text classification poses significant challenges due to extreme class imbalance, with specialized terminology exhibiting severe power-law distributions that challenge standard classification methods (Liu et al., 2023). While imbalanced distributions are common in scientific domains, existing corpora (Giles et al., 1998; McCallum et al., 2000; Yang et al., 2018) typically provide limited coverage of controlled vocabulary, focus on broad disciplinary categories, or operate at scales that hinder systematic methodological investigation of extreme imbalance scenarios. This limitation is particularly problematic for comprehensive scientific text classification, where controlled vocabularies organize thousands of specialized concepts with naturally occurring power-law distributions. Existing datasets that provide controlled-vocabulary coverage either operate at prohibitive computational scales (Toney and Dunham, 2022) or focus on limited subsets of domain terminology, thereby preventing systematic investigation of how different approaches address the fundamental challenge of learning from severely imbalanced specialized terminologies.

The astrophysics literature exemplifies these challenges while providing an ideal testbed for systematic investigation. The Unified Astronomy Thesaurus (UAT) (Accomazzi et al., 2014) organizes 2,367 concepts across 11 hierarchical levels, creating a comprehensive controlled vocabulary that exhibits natural power-law distributions characteristic of scientific domains.

To provide the community with essential resources for investigating extreme multi-label classification in scientific domains, we introduce AstroConcepts, a corpus of 21,702 astrophysics papers labeled with the complete UAT vocabulary (2,367 concepts). This resource enables systematic investigation of three fundamental research questions:

- **RQ1** How do traditional methods, supervised neural models, and vocabulary-constrained LLMs compare for handling extreme label imbalance in scientific classification?
- **RQ2** Does domain-specific pretraining provide uniform benefits across frequency bins, or do improvements concentrate in specific regions of the label distribution?
- **RQ3** Can vocabulary-constrained LLMs achieve competitive performance with

Proceedings of the 3rd Int. Workshop on Natural Scientific Language Processing (NSLP 2026) @ LREC 2026, pages 1–12
12 May 2026. © ELRA Language Resources Association (ELRA), 2026

domain-adapted models in astrophysics classification?

Our main contributions are:

1. AstroConcepts[1] provides the community with the first tractable-scale corpus, enabling systematic investigation of extreme multi-label classification with comprehensive controlled-vocabulary coverage.
2. Systematic comparison across traditional, neural, and vocabulary-constrained LLM approaches reveals competitive performance for parameter-efficient methods, opening new research directions for scientific NLP.
3. Introduction of frequency-stratified evaluation framework with robustness metrics that reveal performance patterns invisible in aggregate scores, providing essential tools for extreme multi-label assessment.
4. Demonstration that domain adaptation improvements concentrate on rare specialized terminology, with implications for training strategies in scientific applications.
5. Comprehensive baseline establishment that provides essential benchmarks for scientific multi-label classification while revealing fundamental properties of extreme imbalance in specialized domains.

The remainder of this paper is structured as follows. Section 2 reviews related work in scientific multi-label classification and positions AstroConcepts within the existing landscape of scientific corpora. Section 3 describes the corpus construction methodology, annotation procedures, and key characteristics of the resulting dataset. Section 4 details our experimental setup, including baseline methods, evaluation metrics, and a frequency-stratified analysis framework. Section 5 presents results across all methods, revealing key patterns in overall performance and frequency-specific behaviors. Section 6 discusses the broader implications of our findings for scientific NLP, methodological contributions, and limitations of the current work. Section 7 concludes with a summary of key insights and directions for future research.

2. Scientific Multi-label Classification Benchmarks

Multi-label text classification spans diverse domains and scales. Early work established evaluation protocols with news categorization (Lewis, 1987), while legal document classification introduced a hierarchical structure through EUR-Lex (Loza Mencía and Fürnkranz, 2010), though annotations include complete paths rather than the flat terminal concepts typical in practice. Extreme multi-label benchmarks (Amazon-670K, Wiki-500K) scale to hundreds of thousands of labels but prioritize computational efficiency over domain-specific controlled vocabularies.

Scientific classification began with computer science papers: Giles et al. (1998) used six broad categories, McCallum et al. (2000) expanded to 70 categories with a 3-level hierarchy, and later work incorporated controlled vocabularies (Santos and Rodrigues, 2009; Kowsari et al., 2017; Yang et al., 2018), although these remained limited in scope. Recent efforts leverage Microsoft Academic Graph: Cohan et al. (2020) created embeddings across 19 fields, Sadat and Caragea (2022) scaled to 186K papers with 6-level hierarchy, and Toney and Dunham (2022) used 180–220M papers, though computational demands limit systematic experimentation. While MAG provides broad coverage, it spans multiple disciplines rather than offering deep domain specialization.

Table 1 puts AstroConcepts in context of existing resources. General benchmarks lack controlled vocabularies, whereas scientific corpora either use ad hoc categories, span multiple disciplines without deep specialization, or reach a prohibitive scale. AstroConcepts uniquely combines tractable scale (21K), deep hierarchy (11 levels), and domain-specific controlled vocabulary (UAT), enabling comprehensive evaluation of hierarchy-aware methods and few-shot learning in a realistic scientific classification scenario.

3. The AstroConcepts Corpus

AstroConcepts exhibits characteristics that make scientific multi-label classification challenging: a hierarchical controlled vocabulary (UAT) with flat expert annotations, severe label imbalance, and domain-specific terminology that requires specialized language understanding. This section describes the data collection process, the UAT taxonomy structure, the annotation methodology, and the comprehensive corpus statistics.

3.1. Source Data

We collected 21,702 abstracts from published English-language papers indexed by the NASA-funded Science Explorer (SciX; Bartlett et al. (2025)), an expansion of the Astrophysics Data System (ADS; Accomazzi et al. (2015)) to cover all NASA science disciplines. Building on the ADS legacy, SciX is the primary bibliographic database

[1] https://huggingface.co/datasets/adsabs/SciX_UAT_keywords

Corpus	Docs	Labels	Hierarchy	Supervision	Vocabulary	Domain
General Domain						
Loza Mencía and Fürnkranz (2010)	19K	7.2K	2-level	Hierarchical	EuroVoc	Legal
Lewis (1987)	21K	90	None	Flat	Ad-hoc	News
RCV1	800K	103	4-level	Flat	Ad-hoc	News
Scientific Domain						
Giles et al. (1998)	3K	6	None	Flat	Ad-hoc	CS
Santos and Rodrigues (2009)	15K	92	2-level	Mixed	ACM	CS
AstroConcepts	**21K**	**2.3K**	**11-level**	**Flat**	**UAT**	**Astrophys**
Cohan et al. (2020)	25K	19	1-level	Flat	MAG	Multi-sci
Kowsari et al. (2017)	47K	134	2-level	Hierarchical	WoS	CS+Med
McCallum et al. (2000)	53K	70	3-level	Hierarchical	Ad-hoc	CS
Yang et al. (2018)	55.8K	54	2-level	Flat	Ad-hoc	CS
Sadat and Caragea (2022)	186K	1.2K	6-level	Mixed	MAG	Multi-sci
Toney and Dunham (2022)	180M	313	2-level	Flat	MAG	Multi-sci

Table 1: Multi-label classification benchmarks sorted by corpus size. AstroConcepts provides tractable scale (21K documents) with deep hierarchical structure (11 levels) and domain-specific controlled vocabulary (UAT), enabling comprehensive experimentation on astrophysics literature classification.

for astronomy and astrophysics. It indexes papers from approximately 8,000 refereed journals, including the *Astrophysical Journal* (*ApJ*). Starting in 2018, journal editors began requiring or encouraging UAT concept assignment during submission to promote standardized concept usage in astronomy, resulting in a growing corpus of controlled-vocabulary annotations.

We selected papers meeting the following criteria: (1) published in journals where authors assign UAT concepts during submission, ensuring controlled standard concept annotation, (2) at least one UAT concept assigned, and (3) English-language abstracts. The temporal scope covers 2018-2023, reflecting the data available during corpus construction. Future work could extend coverage to more recent publications. This process yielded 21,702 documents.

3.2. The Unified Astronomy Thesaurus

The UAT (Accomazzi et al., 2014) is a community-maintained controlled vocabulary for astronomical literature, owned and openly licensed by the *American Astronomical Society*. It follows the Simple Knowledge Organization System (SKOS; Miles and Bechhofer (2009)) standard and incorporates terms from earlier astronomy thesauri. Astronomy librarians and domain experts have contributed to its development and maintenance. Released in 2017 and subsequently adopted by major journals and SciX, the UAT provides standardized terminology for indexing and retrieval of astronomy content.

The UAT version 5.1.0[2] used in this work contains 2,367 astronomical concepts organized in an eleven-level hierarchical structure forming a directed acyclic graph (DAG). Concepts range from broad top-level categories such as cosmology or observational astronomy, to highly specific concepts such as the Kreutz group. Each concept includes a unique identifier, a canonical designation, alternative names (synonyms, acronyms), and an optional textual definition (scope note), and explicit relationships to parents, children, and related concepts.

The UAT follows a polyhierarchical structure in which concepts can have multiple parents, creating a DAG topology in which nodes (concepts) can be reached through multiple paths without forming cycles. For example, *Stellar atmosphere* appears under both *Stars* and *Spectroscopy*, as studies of stellar atmospheres involve both stellar physics and spectroscopic methods. This DAG structure reflects the multidisciplinary nature of astronomical research, where concepts naturally belong to multiple semantic categories simultaneously.

3.3. Concept Assignment Process

Labels in AstroConcepts come from the standard publishing process where authors assign UAT concepts to their manuscripts during submission. Authors typically select 4 concepts per paper (see Table 2) using the journal's submission system. They choose specific concepts without marking hierarchical paths. For example, selecting *Exoplanet atmospheres* does not require selecting its broader categories, such as *Exoplanet astronomy*. This creates an interesting challenge: systems must predict specific concepts from a hierarchical vocabulary using only flat annotations. The approach has clear advantages: we collect high-quality labels from domain experts who know their work best, and the standardized UAT vocabulary ensures consistency.

[2] https://github.com/astrothesaurus/UAT/tree/v.5.0.0

However, authors naturally focus on what they consider most important rather than providing comprehensive coverage. This means some relevant concepts might be missing, creating a realistic yet challenging evaluation scenario that mirrors real-world conditions in which systems operate with an incomplete set of annotations.

3.4. Overall Statistics

Table 2 presents overall corpus statistics. AstroConcepts contains 21,702 abstracts with 93,547 total label assignments, averaging 4.31 labels per abstract. The assigned label space comprises 1,864 unique UAT concepts (92% of the UAT vocabulary), indicating that the selected literature abstracts span the conceptual space defined by UAT.

Statistic	Value
Documents	
Total abstracts	21,702
Labels	
Total assignments	93,547
Assigned unique labels	1,864
Per abstract (mean/median)	4.31 / 4
Per abstract (range)	1–12
Text	
Length in words (mean/median)	211.2 / 223
Length (range)	18–462
Vocabulary size	163,326

Table 2: Overall statistics for AstroConcepts.

Abstracts average 211 words (median: 223), typical for scientific abstracts and compatible with standard transformer context windows (512 tokens). The distribution is approximately normal with slight positive skew toward longer abstracts; 94% fit within 512 tokens. We retain abstracts up to 462 words to capture the full range of scientific writing styles without truncation artifacts.

The distribution of labels per abstract (mean: 4.31, median: 4, range: 1–12) reflects the multifaceted nature of astrophysics research. Most abstracts (70%) receive 2–5 labels, with single-label papers typically representing narrowly focused studies and papers with 6+ labels (15%) covering interdisciplinary or methodologically diverse work. This moderate label density exceeds general multi-label benchmarks such as Reuters-21578 (Lewis, 1987), which averages 1.2 labels per document (Huang et al., 2021), underscoring the conceptual complexity inherent in scientific literature.

3.5. Concept Frequency Distribution

A critical characteristic of AstroConcepts is severe label imbalance, typical of real-world multi-label scenarios. Figure 1 shows the label frequency distribution on a log-log scale, revealing a power-law pattern. The most frequent label (*Galaxy evolution*) appears in 1,106 abstracts (5.1%), while the median label appears in only 12 abstracts. We fit a power law $f(r) \propto r^{-\alpha}$ where r is rank and $f(r)$ is frequency, obtaining $\alpha = 1.50$ with $R^2 = 0.825$.

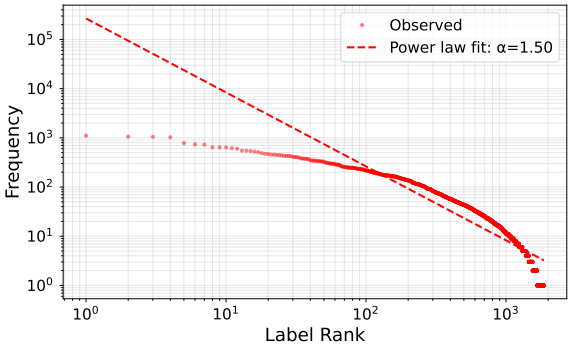

Figure 1: Label frequency distribution (log-log scale) and fitted power-law function with exponent $\alpha = 1.50$ ($R^2 = 0.825$). The long tail contains 76% of labels with fewer than 50 occurrences each, creating a severe class imbalance characteristic of scientific multi-label classification.

The exponent $\alpha = 1.50$ indicates a moderately steep long-tail distribution. This suggests that AstroConcepts presents a substantial but not extreme class imbalance compared to other scientific classification scenarios. The moderate slope means mid-frequency (torso) labels retain more training examples than in steeper distributions; by rank 100, labels still average 58 examples, whereas in datasets with $\alpha = 2.0$, rank-100 labels would have only 25 examples. Nevertheless, a severe imbalance remains: 76% of labels occur fewer than 50 times.

We partition labels into three frequency bins (see Table 3). Head labels (frequency > 500): 17 labels (0.9% of vocabulary) covering 12,288 assignments (13.1% of total). These represent core concepts studied extensively across astrophysics subfields. Torso labels ($50 \leq$ frequency ≤ 500): 429 labels (23.0%) covering 62,808 assignments (67.1%). These represent moderately common research topics with sufficient training examples for standard supervised learning. Tail labels (frequency < 50): 1,418 labels (76.1%) covering 18,451 assignments (19.7%). These represent specialized phenomena, emerging research areas, and niche methodologies with limited training examples, creating the zero-shot challenge we investigate in Section 4.

This distribution creates the multi-level challenge characteristic of extreme multi-label classification: head labels are easily learned from abundant examples (hundreds per label), torso labels require careful modeling to generalize from moderate data

Bin	# Labels	%	# Assign.	%
Head (>500)	17	0.9	12,288	13.1
Torso (50–500)	429	23.0	62,808	67.1
Tail (<50)	1,418	76.1	18,451	19.7
Total	**1,864**	**100.0**	**93,547**	**100.0**

Table 3: Label frequency distribution across bins. The long tail contains 76% of labels but only 20% of assignments, creating a severe class imbalance characteristic of scientific multi-label classification.

(50–500 examples), and tail labels present a few-shot scenario (fewer than 50 examples, median: 12) where standard supervised methods struggle.

The most frequent labels span major astrophysics subfields: extragalactic astronomy (*Galaxy evolution*, *Active galactic nuclei*), stellar physics (*Star formation*, *Neutron stars*), planetary science (*Exoplanets*, *Exoplanet atmospheres*), and observational methods (*Spectroscopy*, *Astronomy data analysis*). No single subfield dominates the top-20 labels, confirming that AstroConcepts captures the breadth of modern astrophysics research rather than concentrating on narrow phenomena. This diversity is important for multi-label classification research: the corpus contains both well-represented core concepts and sparse specialized topics, enabling evaluation across the full spectrum of label frequencies.

3.6. Concept Specificity Patterns

Table 4 reveals that authors prefer moderately specific concepts (peaking at level 4). Both very general concepts (levels 1-2) and highly specialized terminology (levels 6+) are underrepresented relative to mid-level concepts. This concentration around moderate specificity creates additional challenges for classification systems beyond frequency imbalance. Models must handle vocabularies in which training examples cluster around mid-level concepts, with limited examples at both ends of the taxonomic spectrum, requiring systems to predict across varying specificity levels with highly imbalanced training signals.

3.7. Concept Co-occurrence Patterns

We analyzed how concepts co-occur in our corpus. We computed pointwise mutual information (PMI) for concept pairs, where $\text{PMI}(\ell_i, \ell_j) = \log \frac{P(\ell_i,\ell_j)}{P(\ell_i)P(\ell_j)}$ measures whether two concepts appear together more frequently than expected by chance. Positive PMI indicates stronger-than-expected co-occurrence, while negative PMI suggests mutual exclusivity. Focusing on pairs with positive PMI and sufficient co-occurrence ($\geq$ 10 abstracts), we found that authors rarely select hierarchically related concepts together. Concepts from unrelated subtrees show 2.3× higher average PMI than parent-child pairs, indicating that authors typically choose concepts spanning multiple taxonomic branches rather than selecting both general and specific terms from the same hierarchy. This cross-branch annotation behavior represents a challenge for classification systems, which must learn to predict conceptually diverse label combinations spanning the entire taxonomic structure.

Depth	Count	%
1	106	4.5
2	106	4.5
3	405	17.1
4	**680**	**28.7**
5	554	23.4
6	341	14.4
7	110	4.6
8	36	1.5
9	13	0.5
10	11	0.5
11	5	0.2

Table 4: Distribution of UAT concept depths in AstroConcepts showing annotation preference patterns across taxonomic levels.

4. Experiments

4.1. Experimental Setup

Task Formulation We formulate astrophysics concept classification as a multi-label task where, given an abstract x concatenated with its title, the goal is to predict a subset $Y \subseteq \mathcal{L}$ of relevant UAT concepts from the complete label space $\mathcal{L}$ of 2,367 labels. We use title+abstract concatenation as input text to provide models with maximum available semantic information for concept prediction.

Data Partitioning We split the corpus into train (18,677 abstracts, 85%) and test (3,025 abstracts, 15%) sets using label-aware stratification. Labels appearing $\geq$15 times are stratified to achieve approximately 85/15 distribution per label, while labels with $<$15 occurrences are placed entirely in the training set to maximize training signal.

Evaluation Metrics Following standard practice in multi-label classification (Zhang and Zhou, 2014; Tsoumakas and Katakis, 2010), we evaluate using Macro-F1, which averages per-label F1 scores and is essential for imbalanced settings (Wu et al., 2020). For ranking evaluation, we use Precision/Recall at k ($P@k$, $R@k$) with $k \in \{1, 3, 5\}$, chosen to align with the average number of assigned concepts per paper (4.31, see Table 2).

4.2. Baseline Approaches

We establish baselines across multiple paradigms to understand effective modeling approaches for astrophysics concept classification, organized by increasing complexity: non-parametric methods, supervised neural models, and vocabulary-constrained LLMs.

4.2.1. Non-Parametric Methods

Rule-based Matching We implement lexical matching based on the assumption that if a UAT concept is explicitly mentioned in the text, it should be assigned as a label. The method searches for exact string matches of UAT concept names within the title and abstract. For each concept, we check for its canonical designation and optionally include synonyms and abbreviations from the UAT taxonomy (e.g., searching for both *active galactic nuclei* and *AGN*). We evaluate two variants: Rule-based$_{\text{w/o var}}$ uses only canonical names, while Rule-based$_{\text{w/ var}}$ includes all alternative forms provided in the UAT.

k-Nearest Neighbors This approach assumes abstracts with high contextual similarity should share similar concepts. We encode titles and abstracts using three embedding models: astroBERT (Grezes et al., 2024) (adapted specifically for astrophysics texts), INDUS (Bhattacharjee et al., 2024) (a scientific language model covering astrophysics, earth science, and general physics), and `Qwen3-Embedding-8B`[3] (chosen for strong sentence similarity performance). For each abstract from the test set, we retrieve k nearest training neighbors using cosine similarity, then predict the most frequent labels among them. We perform a grid search over $k \in \{5, 10, 20, 50\}$ and all embedding models, with detailed results provided in the appendix. Table 5 reports only the best-performing configuration (embedding model and k value) for readability.

4.2.2. Supervised Neural Models

To investigate the effects of domain adaptation, we fine-tune three transformer models representing different levels of domain specialization: BERT (Devlin et al., 2019) (general-purpose), SciBERT (Beltagy et al., 2019) (scientific domains), and astroBERT (Grezes et al., 2024). We add classification heads on $[CLS]$ representations and fine-tune end-to-end with max_length=512, batch_size=8, and AdamW optimizer. Due to computational resource constraints, we were unable to include Qwen models in the fine-tuning experiments. We perform a grid search over learning rates $\{2e-5, 3e-5, 5e-5\}$ and epochs $\{1, 2, 3, 5, 8, 10\}$ to find optimal hyperparameter configurations. Detailed results for each configuration are provided in the appendix. Table 5 reports only the best-performing configuration (lr=2e-5, epochs=8) for readability.

4.2.3. Vocabulary-Constrained LLMs

LLMs demonstrate strong capabilities on multi-label text classification tasks (Zhou et al., 2024; Tabatabaei et al., 2025) but face challenges with large label spaces. Prompting an LLM directly to select from all 2,367 UAT concepts is infeasible due to context length limitations and label complexity. Preliminary experiments on a small subset yielded very poor results when using the complete label list, as the model hallucinated concepts or became overwhelmed by the extensive vocabulary. We therefore implement a two-stage approach: (1) use our best supervised model (astroBERT; see Table 5) to generate top-50 candidate labels for each abstract, (2) prompt `DeepSeek-V3-reasoner` (DeepSeek-AI et al., 2024) via its API[4] to select relevant concepts from these candidates. We chose the top-50 threshold based on analysis showing that astroBERT's top-50 predictions covered approximately 82% of ground-truth labels (see Figure 3 in the appendix 9), providing good coverage while maintaining manageable prompt length. We evaluate `DeepSeek-V3-reasoner` using the prompt shown in Figure 2 of the appendix 9. This approach constrains the model to valid UAT terminology while leveraging its semantic understanding to select the most relevant concepts from the candidate set.

5. Results and Analysis

This section presents and analyzes results across all methods, revealing key insights about domain adaptation, frequency effects, and the comparative strengths of different learning paradigms for extreme multi-label scientific classification.

5.1. Overall Performance and Domain Impact

Table 5 presents comprehensive results across all methods, revealing fundamental insights about learning paradigms for extreme multi-label scientific classification.

Our evaluation reveals a progression of insights across methodological paradigms. Simple rule-based matching achieves reasonable precision (0.229) but faces fundamental limitations: string matching of concept mentions may not reflect the core research focus. Common terms like "*photon*"

[3] https://huggingface.co/Qwen/Qwen3-Embedding-8B

[4] https://api-docs.deepseek.com/

Method	Overall			Ranking Precision			Ranking Recall		
	Precision	Recall	F_1	P@1	P@3	P@5	R@1	R@3	R@5
Non-parametric									
Rule-based$_{\text{w/o var}}$	0.2290	0.2130	0.1950	‡	‡	‡	‡	‡	‡
Rule-based$_{\text{w var}}$	0.2170	0.2730	0.2140	‡	‡	‡	‡	‡	‡
k-NN	0.1165	**0.7509**	0.2017	0.6125	0.4607	0.3698	0.1398	0.3155	0.4221
Supervised neural models									
BERT	0.1784	0.2273	0.1880	0.2975	0.2187	0.1784	0.0810	0.1730	0.2273
SciBERT	0.2020	0.2563	0.2127	0.3213	0.2409	0.2020	0.0872	0.1864	0.2563
astroBERT	**0.3068**	0.3905	0.3243	0.4909	0.3733	0.3068	0.1360	0.2933	0.3905
Zero-shot prompting									
Deepseek	0.2891	0.6322	**0.3770**	**0.6502**	**0.4930**	**0.4017**	**0.1815**	**0.3837**	**0.5050**

Table 5: Overall performance on AstroConcepts test set. Best results shown in **bold**. ‡Not applicable for rule-based methods.

appear across diverse papers, while implicit language and incomplete variant coverage constrain effectiveness.

Moving to similarity-based approaches, k-NN achieves exceptional recall (0.751) but poor precision (0.117). Crucially, domain-adapted astroBERT embeddings outperform general models like Qwen, better capturing subtle concept distinctions essential for astrophysics. However, performance degrades beyond $k = 10$ neighbors as noise from irrelevant papers accumulates.

The most striking finding emerges with vocabulary-constrained DeepSeek, which outperforms even the best domain-adapted model (astroBERT) by 16% in F_1 score (0.377 vs 0.324). This demonstrates that domain expertise can be effectively incorporated through vocabulary constraints rather than solely through model parameters. Meanwhile, supervised domain adaptation shows clear value: astroBERT substantially outperforms SciBERT (0.324 vs 0.213) with improvements concentrated in precision and confident prediction rather than comprehensive recall.

Together, these results reveal that effective scientific text classification benefits from hybrid approaches combining general language understanding with structured domain knowledge, opening promising directions for parameter-efficient scientific NLP.

5.2. The Long-Tail Challenge

The extreme label distribution in AstroConcepts (78% of concepts have < 50 examples) enables the systematic analysis of long-tail performance patterns. Table 6 presents frequency-stratified results across Head (> 500 examples), Torso (50–500), and Tail (< 50) concepts, revealing insights that aggregate metrics cannot capture.

Three patterns emerge, providing useful insights into handling extreme imbalance. First, domain adaptation provides asymmetric benefits: astroBERT shows modest Head improvements over SciBERT (0.193 $\rightarrow$ 0.216) but larger relative gains for Tail concepts (0.023 $\rightarrow$ 0.081), though absolute performance remains limited across traditional approaches. Second, all methods exhibit a consistent "torso peak" where mid-frequency concepts achieve optimal performance. This pattern suggests fundamental properties of scientific vocabulary learning, in which models balance sufficient training signal with complexity issues at frequency extremes. Most significantly, the vocabulary-constrained approach achieves superior tail performance (F_1: 0.198 vs 0.081 for astroBERT), demonstrating that domain expertise can be effectively incorporated through structured constraints rather than parameter fine-tuning alone. This hybrid approach, combining general language understanding with domain-specific vocabulary guidance, proves particularly effective for rare concepts. Finally, we introduce frequency robustness (Δ = Head F_1 - Tail F_1) as a critical evaluation dimension. The constrained approach achieves 3× better robustness than SciBERT ($\Delta = 0.045$ vs 0.170), indicating that architectural choices and constraint mechanisms matter more for handling frequency imbalance than domain specialization alone.

6. Discussion

Our systematic evaluation reveals fundamental insights into extreme multi-label classification in scientific domains while establishing new evaluation paradigms that advance understanding beyond existing benchmarks.

Addressing the Research Questions Our findings provide clear answers to the three research questions posed. Regarding RQ1 (handling ex-

Method	F_1			Head		Torso		Tail		
	Head	Torso	Tail	P@3	R@3	P@3	R@3	P@3	R@3	Δ^*
BERT	0.180	0.167	0.021	0.084	0.205	0.168	0.183	0.012	0.024	0.159
SciBERT	0.193	0.197	0.023	0.084	0.211	0.199	0.212	0.015	0.026	0.170
astroBERT	0.216	0.312	0.081	0.092	0.230	0.311	0.333	0.046	0.088	0.135
DeepSeek	**0.243**	**0.376**	**0.198**	**0.107**	**0.266**	**0.417**	**0.443**	**0.135**	**0.267**	**0.045**

Table 6: performance across frequency bins. Head/Torso/Tail thresholds: $> 500/50$–$500/< 50$ training examples. $^*\Delta$ = Head F_1 - Tail F_1 (lower is better). Best results in **bold**.

treme imbalance), vocabulary-constrained LLMs demonstrate superior robustness across frequency bins, achieving 3× better head-tail balance than traditional supervised approaches. For RQ2 (domain adaptation benefits), we find asymmetric improvements that focus on rare, specialized terminology rather than on frequent concepts, challenging assumptions about uniform domain adaptation effects. RQ3 (competitive LLM performance) is answered affirmatively: the hybrid approach combining astroBERT candidate generation with LLM selection achieves competitive results while requiring substantially fewer computational resources than full fine-tuning.

Methodological Impact We establish frequency-stratified evaluation as essential for extreme multi-label assessment and introduce the head-tail gap (Δ) as a robustness metric that reveals performance patterns invisible in aggregate scores. The systematic comparison across paradigms demonstrates that different approaches excel in complementary areas: rule-based methods provide interpretability but limited coverage, k-NN offers high recall with domain-adapted embeddings, supervised models achieve confident predictions, and constrained LLMs balance precision-recall trade-offs effectively.

Broader Implications The torso peak phenomenon, in which all methods perform best on mid-frequency concepts, suggests fundamental limits to learning from extreme imbalance that transcend architectural choices. This finding has immediate implications for resource allocation in scientific NLP: optimization efforts should target the frequency regions where improvement is most feasible, rather than aiming for uniform performance gains across all concepts.

Relationship to Complementary Resources A related effort from Ting et al. (2025) extracted 9,999 concepts from 408,590 astrophysics papers using LLM-based pipelines and clustering over full-text content. The two resources address different but complementary needs. AstroMLab 5 produces a semantically rich, emergent vocabulary optimized for discovery, knowledge graph construction, and temporal analysis at scale. AstroConcepts, by contrast, provides controlled-vocabulary annotations grounded in the UAT, enabling reproducible evaluation of classification methods under extreme label imbalance, a use case for which a fixed label space, train/test split, and expert-assigned labels are essential. Looking forward, the two resources open natural avenues for joint investigation: AstroMLab 5 concepts could serve as additional candidate labels or weak supervision signals for tail concepts in AstroConcepts, while UAT-grounded annotations could provide an extrinsic evaluation signal for the quality of LLM-extracted concept vocabularies.

Limitations and Future Directions Our findings are domain-specific and require validation across other scientific fields before broader claims about scientific NLP can be established. The persistent tail performance challenge (best F_1: 0.198) indicates fundamental limitations in current approaches, suggesting opportunities for architectural innovations tailored to extreme imbalance scenarios. Future work should explore integrating structured domain knowledge with few-shot learning approaches. An important validation step is to audit a stratified sample of the corpus through expert review and inter-annotator agreement analysis, which would quantify annotation noise and its downstream effect on recall-based metrics, particularly for tail concepts where incomplete author-assigned labels may inflate the apparent difficulty. Extending this methodology to other scientific domains is essential to establish whether the torso-peak phenomenon, asymmetric domain adaptation benefits, and the effectiveness of vocabulary-constrained approaches generalize beyond astrophysics. Finally, more extensive experiments with the LLM filtering stage are needed: evaluating DeepSeek over candidate sets generated by BERT and SciBERT, in addition to astroBERT, would isolate the contribution of the candidate generator's quality from the LLM's own re-ranking ability, providing a cleaner assessment of where the performance gains originate.

7. Conclusion

AstroConcepts provides the NLP community with essential resources for investigating extreme multi-label classification in scientific domains, in particular for astrophysics. Through systematic evaluation across traditional, neural, and vocabulary-constrained approaches, we demonstrate three key insights that advance understanding of scientific text classification. The effectiveness of hybrid vocabulary-constrained approaches, in which astroBERT generates candidate labels and DeepSeek selects from this constrained set, demonstrates that domain expertise can be incorporated through structured vocabulary guidance rather than through extensive LLM fine-tuning. This approach achieves competitive performance (F_1: 0.377) while requiring only inference costs for the domain model and API calls for the LLM, opening promising directions for cost-effective scientific NLP that combines specialized knowledge extraction with general language understanding capabilities. Domain adaptation benefits concentrate asymmetrically on rare, specialized terminology, suggesting that specialized models primarily handle concepts beyond general model capabilities rather than improving performance uniformly. Our frequency-stratified evaluation framework, combined with robustness metrics, provides useful methods for assessing extreme multi-label systems in which aggregate scores can mask critical performance patterns. These contributions address a critical methodological gap by enabling systematic evaluation of extreme imbalance while providing actionable insights for scientific NLP practitioners. The corpus, baselines, and evaluation framework lay the foundations for future research on specialized domain classification, while our findings on vocabulary-constrained approaches indicate promising directions for resource-efficient scientific text processing systems. To facilitate future research, we make the AstroConcepts corpus publicly available. The persistent challenges in tail performance underscore opportunities for novel approaches that integrate structured knowledge with text-based classification. Future work should prioritize annotation validation through expert audits and inter-annotator agreement studies, systematic cross-domain replication, and controlled ablations of the LLM re-ranking stage across candidate generators of varying quality. As the scientific literature continues to expand and specialized terminology increases, effective handling of extreme imbalance becomes essential for scientific NLP applications.

8. Ethical Considerations

All abstracts are from published scientific papers that are publicly accessible. We include only bibliographic metadata (bibcode, title, abstract, publication year, journal) and assigned UAT concepts.

9. Bibliographical References

A. Accomazzi, N. Gray, C. Erdmann, C. Biemesderfer, K. Frey, and J. Soles. 2014. The Unified Astronomy Thesaurus. In *Astronomical Data Analysis Software and Systems XXIII*, volume 485 of *Astronomical Society of the Pacific Conference Series*, page 461.

Alberto Accomazzi, Michael J. Kurtz, Edwin A. Henneken, Roman Chyla, James Luker, Carolyn S. Grant, Donna M. Thompson, Alexandra Holachek, Rahul Dave, and Stephen S. Murray. 2015. Ads: The next generation search platform.

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.

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.

Bishwaranjan Bhattacharjee, Aashka Trivedi, Masayasu Muraoka, Muthukumaran Ramasubramanian, Takuma Udagawa, Iksha Gurung, Nishan Pantha, Rong Zhang, Bharath Dandala, Rahul Ramachandran, Manil Maskey, Kaylin Bugbee, Mike Little, Elizabeth Fancher, Irina Gerasimov, Armin Mehrabian, Lauren Sanders, Sylvain Costes, Sergi Blanco-Cuaresma, Kelly Lockhart, Thomas Allen, Felix Grezes, Megan Ansdell, Alberto Accomazzi, Yousef El-Kurdi, Davis Wertheimer, Birgit Pfitzmann, Cesar Berrospi Ramis, Michele Dolfi, Rafael Teixeira de Lima, Panagiotis Vagenas, S. Karthik Mukkavilli, Peter Staar, Sanaz Vahidinia, Ryan McGranaghan, and Tsendgar Lee. 2024. INDUS: Effective and Efficient Language Models for Scientific Applications. *arXiv e-prints*, page arXiv:2405.10725.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. DeepSeek-V3 Technical Report. *arXiv e-prints*, page arXiv:2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. 1998. Citeseer: an automatic citation indexing system. In *Proceedings of the Third ACM Conference on Digital Libraries*, DL '98, page 89–98, New York, NY, USA. Association for Computing Machinery.

F. Grezes, S. Blanco-Cuaresma, A. Accomazzi, M. J. Kurtz, G. Shapurian, E. Henneken, C. S. Grant, D. M. Thompson, R. Chyla, S. McDonald, T. W. Hostetler, M. R. Templeton, K. E. Lockhart, N. Martinovic, S. Chen, C. Tanner, and P. Protopapas. 2024. Building astroBERT, a Language Model for Astronomy & Astrophysics. In *Astronomical Data Analysis Software and Systems XXXI*, volume 535 of *Astronomical Society of the Pacific Conference Series*, page 119.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8153–8161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, K. Meimandi, Matthew S. Gerber, and Laura E. Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 364–371.

David Lewis. 1987. Reuters-21578 Text Categorization Collection. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C52G6M.

Rundong Liu, Wenhan Liang, Weijun Luo, Yuxiang Song, He Zhang, Ruohua Xu, Yunfeng Li, and Ming Liu. 2023. Recent advances in hierarchical multi-label text classification: A survey.

Eneldo Loza Mencía and Johannes Fürnkranz. 2010. *Efficient Multilabel Classification Algorithms for Large-Scale Problems in the Legal Domain*, pages 192–215. Springer Berlin Heidelberg, Berlin, Heidelberg.

Andrew McCallum, Kamal Nigam, Jason D. M. Rennie, and Kristie Seymore. 2000. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3:127–163.

Alistair Miles and Sean Bechhofer. 2009. SKOS Simple Knowledge Organization System Reference. W3c recommendation, W3C.

Mobashir Sadat and Cornelia Caragea. 2022. Hierarchical multi-label classification of scientific documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8923–8937, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

António Paulo Santos and Fátima Rodrigues. 2009. Multi-label hierarchical text classification using the acm taxonomy.

Seyed Amin Tabatabaei, Sarah Fancher, Michael Parsons, and Arian Askari. 2025. Can large language models serve as effective classifiers for hierarchical multi-label classification of scientific documents at industrial scale? In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 163–174, Abu Dhabi, UAE. Association for Computational Linguistics.

Yuan-Sen Ting, Alberto Accomazzi, Tirthankar Ghosal, Tuan Dung Nguyen, Rui Pan, Zechang Sun, and Tijmen de Haan. 2025. AstroMLab 5: Structured summaries and concept extraction for 400,000 astrophysics papers. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 170–185, Mumbai, India and virtual. Association for Computational Linguistics.

Autumn Toney and James Dunham. 2022. Multi-label classification of scientific research documents across domains and languages. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 105–114, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Grigorios Tsoumakas and Ioannis Katakis. 2010. Mining multi-label data. *Data Mining and Knowledge Discovery Handbook*.

Ximing Wu, Hao Chen, and Qinmin Zhang. 2020. Revisiting macro-f1 score for imbalanced multi-label classification. *arXiv preprint*.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE TKDE*.

Chuang Zhou, Junnan Dong, Xiao Huang, Zirui Liu, Kaixiong Zhou, and Zhaozhuo Xu. 2024. QUEST: Efficient extreme multi-label text classification with large language models on commodity hardware. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3929–3940, Miami, Florida, USA. Association for Computational Linguistics.

A.1. Complete Prompt Template (Section 4.2.3)

Figure 2 shows the designed prompt as part of our experiments.

A.2. astroBERT label coverage (Section 4.2.3)

Figure 3 shows that astroBERT's top-50 predicted labels covers approximately 82% of the ground-truth labels.

```
You are an expert astrophysicist and scientific topic classifier.
Your task is to choose between 1 to 10 labels from the candidate list that
accurately describe the main scientific themes of the following research
paper.
A label should be selected only if it is clearly relevant to the paper's
content.

--
Abstract:
{abstract}

--
Candidate topics suggested by the model:
{topk_labels}

Return your answer in valid JSON as follows:
{
  "selected_labels":  ["label1", "label2", ...]
}
Do not include explanations or text outside the JSON.
```

Figure 2: Prompt template used for vocabulary-constrained LLM classification.

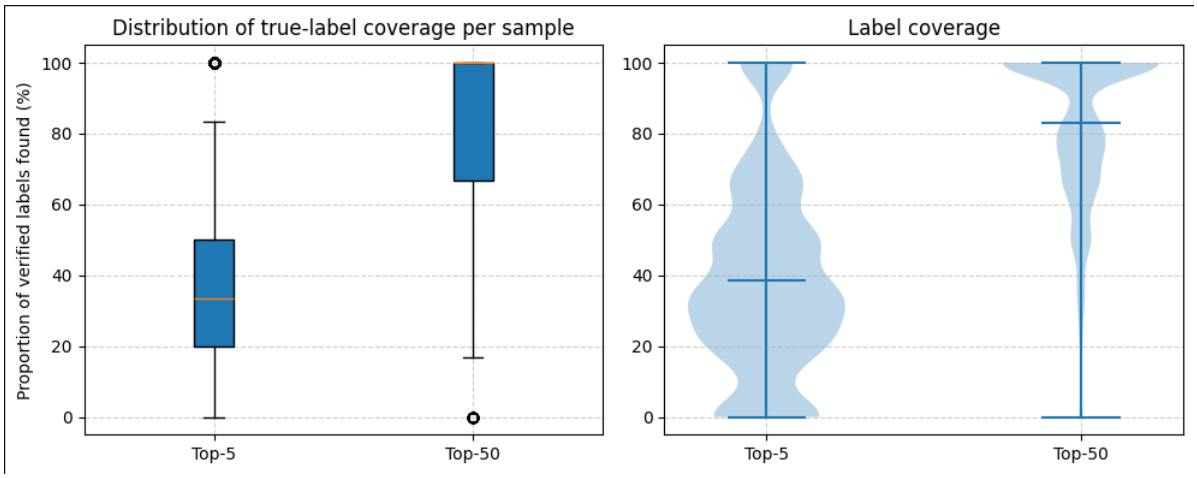

Figure 3: Fine-tuned astroBERT Label Coverage