Systematic Evaluation of Machine Learning and Transformer-Based Methods for Scientific Telescope Literature Classification

Huynh Trung Kiet[1*] **Dao Sy Duy Minh**[1*] **Tran Chi Nguyen**[1*] **Nguyen Lam Phu Quy**[1]
Pham Phu Hoa[1] Nguyen Dinh Ha Duong[1] Dinh Dien[1†] Nguyen Hong Buu Long[1†]
[1]University of Science, VNU-HCM
{23122039, 23122041, 23122044, 23122048, 23122030, 23122002}@student.hcmus.edu.vn
{ddien, nhblong}@fit.hcmus.edu.vn
[*]Equal contribution [†]Corresponding authors

Abstract

Recent space missions such as Hubble, Chandra, and JWST have produced a rapidly growing body of scientific literature. Maintaining telescope bibliographies is essential for mission assessment and research traceability, yet current curation processes rely heavily on manual annotation and do not scale. To facilitate progress in this direction, the TRACS @ WASP 2025 shared task provides a benchmark for automatic telescope bibliographic classification based on scientific publications. In this work, we conduct a comparative study of modeling strategies for this task. We first explore traditional machine learning methods such as multinomial Naive Bayes with TF–IDF and CountVectorizer representations. We then evaluate transformer-based multi-label classification using BERT-based scientific language models. Finally, we investigate a task-wise classification approach, where we decompose the problem into separate prediction tasks and train a dedicated model for each. In addition, we experiment with a limited-resource LLM-based approach, showing that even without full fine-tuning and using only a partial subset of the training data, LLMs exhibit promising potential for telescope classification. Our best system achieves a macro F1 of 0.72 with BERT-based models on the test evaluation, substantially outperforming the official openai-gpt-oss-20b baseline (0.31 macro F1).

1 Introduction

Bibliographic curation plays a central role in scientific knowledge management, enabling mission impact assessment, citation tracking, and research traceability. In astronomy, maintaining telescope bibliographies is essential to quantify the scientific output of major space missions such as Hubble, Chandra, and JWST. However, current bibliographic systems depend predominantly on manual effort, making large-scale curation impractical.

The TRACS @ WASP 2025 (Grezes et al., 2025) shared task formalizes this problem by releasing a benchmark dataset derived from the SAO/NASA Astrophysics Data System (ADS) and defining a unified evaluation framework for telescope bibliography classification. The task jointly addresses telescope detection and scientific intent categorization, reflecting real-world curation needs in astrophysical research.

Automating telescope bibliography classification is challenging due to ambiguous telescope mentions, heterogeneous writing styles across scientific disciplines, and the long-context nature of research articles. Moreover, each publication may involve multiple telescopes simultaneously, leading to a multi-label classification problem under severe label imbalance, where some telescopes (e.g., Chandra) dominate the dataset while others appear rarely. In addition, the dataset contains many hard negative cases, as papers that merely mention telescope names vastly outnumber those that reflect genuine telescope usage, making model learning even more difficult.

In this work, we conduct a systematic study of modeling strategies for telescope bibliographic classification. First, we establish classical machine learning baselines using multinomial Naive Bayes with TF–IDF and CountVectorizer representations, serving as lightweight yet competitive text classification models. Second, we investigate transformer-based multi-label classification using domain-adapted BERT variants such as SciBERT and AstroBERT, which were pre-trained or fine-tuned on large-scale scientific corpora. These models employ a sigmoid output layer with binary cross-entropy loss to support multi-label learning. Third, we explore a task-wise classification strategy by training separate models for each prediction task to reduce cross-label interference. To mitigate severe class imbalance, we incorporate focal loss(Lin et al., 2017) during fine-tuning to better emphasize

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 136–145

minority labels. Finally, we extend our study with a limited-resource LLM-based approach, where open-weight large language models (LLMs) are evaluated under partial-data and zero-shot settings, demonstrating competitive performance even without full fine-tuning.

Our contributions are as follows:

- We conduct a systematic comparison of modeling strategies for telescope bibliographic classification, covering classical machine learning, transformer-based methods, and LLM-based approaches.
- We show that domain-adapted BERT variants (e.g., SciBERT, AstroBERT) significantly outperform traditional TF–IDF baselines.
- We propose a task-wise classification pipeline with focal loss to mitigate label imbalance.
- We demonstrate that limited-resource LLM inference yields competitive performance even without full fine-tuning.

2 Related work

2.1 Text Representation Methods

Traditional text representation methods have been fundamental to NLP tasks. **TF-IDF (Term Frequency-Inverse Document Frequency)** weights terms based on their frequency in a document relative to their frequency across the corpus, effectively identifying discriminative terms while downweighting common words. **Count Vectorization** represents documents as bags-of-words with raw term frequencies, providing a simple yet effective baseline for many classification tasks. While these methods have been widely used in document classification and information retrieval, they lack semantic understanding and cannot capture contextual word meanings.

2.2 Pre-trained Language Models

The introduction of **BERT (Bidirectional Encoder Representations from Transformers)** (Devlin et al., 2019) revolutionized NLP by pre-training deep bidirectional transformers on large text corpora using masked language modeling and next sentence prediction objectives. BERT's contextualized word representations enable transfer learning across diverse downstream tasks through fine-tuning, achieving state-of-the-art performance on various benchmarks including GLUE(Wang et al., 2019) and SQuAD(Rajpurkar et al., 2016).

Building on BERT's success, **DistilBERT** (Sanh et al., 2019) applies knowledge distillation to create a smaller, faster variant that retains 97% of BERT's language understanding while reducing model size by 40% and inference time by 60%. Through distillation training, DistilBERT learns to mimic BERT's behavior using a student-teacher framework, making it suitable for resource-constrained environments and real-time applications without significant performance degradation.

2.3 Domain-Specific Language Models

Recognizing that general-purpose models may not capture domain-specific terminology and discourse patterns, researchers have developed specialized variants. **SciBERT** (Beltagy et al., 2019) is pre-trained on 1.14M scientific papers from the Semantic Scholar corpus, using a scientific vocabulary and achieving significant improvements on biomedical and computer science tasks.

SPECTER (Scientific Paper Embeddings using Citation-informed TransformERs) (Cohan et al., 2020) takes a different approach by leveraging citation graphs during pre-training. It learns document-level representations by training on triplets of papers where citing papers should have embeddings similar to cited papers, effectively encoding scientific relatedness. However, SPECTER relies on discrete citation relations, which enforce a hard cut-off to similarity and ignore that papers can be very similar despite lacking direct citations.

SciNCL (Scientific Neighborhood Contrastive Learning) (Ostendorff et al., 2022) addresses this limitation by using controlled nearest neighbor sampling over citation graph embeddings for contrastive learning. Instead of discrete citations, SciNCL learns continuous similarity by sampling hard-to-learn negatives and positives while avoiding collisions between samples through margin control. Initialized from SciBERT and trained with neighborhood contrastive objectives, SciNCL outperforms previous methods on the SciDocs (Cohan et al., 2020) benchmark and demonstrates sample-efficient training capabilities.

AstroBERT (Grèzes et al., 2021) further specializes BERT for astronomy by pre-training on astronomical literature from the Astrophysics Data System (ADS). It demonstrates superior performance on astronomy-specific tasks including named entity recognition of celestial objects, classification of as-

tronomical papers, and extraction of observational metadata. AstroBERT's domain adaptation makes it particularly relevant for our telescope bibliography curation task.

These document-level embedding models are particularly relevant to telescope bibliography curation because they capture semantic relationships between scientific papers beyond simple keyword matching. The task requires understanding nuanced distinctions between papers that use telescope data for new scientific results versus those that merely mention the telescope in passing. Citation-aware models like SPECTER and SciNCL can identify papers with similar research contexts, while domain-specific models like AstroBERT understand astronomy terminology and discourse patterns essential for disambiguating telescope references (e.g., distinguishing "Chandra" as a space telescope from other entities with the same name). Furthermore, these models' ability to generate document-level representations enables effective transfer learning for our multi-label classification objectives.

2.4 Fine-tuning Strategies for Transformer Models

While pre-trained language models have shown remarkable capabilities, their effective fine-tuning requires careful consideration of training configurations. (Mosbach et al., 2021) investigate the instability of BERT fine-tuning, revealing that performance can vary significantly across different random seeds, particularly on small datasets. They demonstrate that this instability stems from catastrophic forgetting and vanishing gradients in early layers during fine-tuning.

To address these issues, they propose several techniques:

- **Debiased training**: Using bias correction in the Adam optimizer to stabilize early training steps
- **Re-initialization**: Selectively re-initializing top layers to prevent over-fitting to pre-training tasks
- **Learning rate schedules**: Employing smaller learning rates ($2e-5$ to $5e-5$) with linear warmup and decay
- **Multiple runs**: Averaging predictions across multiple training runs with different seeds to reduce variance

These findings have significant implications for our work, as the telescope bibliography curation task involves multi-label classification on scientific texts where training stability is crucial for reliable performance. We adopt these best practices in our BERT-based approaches, including careful hyperparameter tuning, multiple seed experiments, and appropriate learning rate scheduling.

2.5 Large Language Models

Recent advances in LLMs have pushed the boundaries of language understanding. The **Qwen2.5**(Yang et al., 2024) series represents efficient multilingual language models with strong performance across diverse tasks. **Qwen2.5-1.5B**(Yang et al., 2024) and **Qwen2.5-3B** (Yang et al., 2024) offer different trade-offs between model capacity and computational efficiency. Despite their smaller size compared to models like GPT-3(Brown et al., 2020) or GPT-(OpenAI et al., 2024), these models demonstrate competitive performance on reasoning, question answering, and classification tasks. Their compact architecture makes them suitable for resource-constrained environments while maintaining strong generalization capabilities.

3 Problem definition

3.1 TRACS Dataset

We conduct our experiments on the TRACS @ WASP 2025 dataset (Grezes et al., 2025), which consists of scientific papers from the SciX bibliographic database annotated with telescope associations and usage categories. Each entry includes textual content from five fields: title, abstract, body, acknowledgments, and grants, along with four boolean labels (science, instrumentation, mention, not_telescope) indicating the paper's relationship to the referenced telescope. The multi-label classification task requires models to simultaneously identify the telescope and categorize how the paper uses or references it. Following the competition setup, we use the provided train.csv and test.csv splits. We perform minimal preprocessing steps to maintain the original text structure:

- Text cleaning: Remove HTML tags, special characters, and reference markers.

- We concatenate all text fields into a single input sequence. For transformer-based models, the input is truncated to a maximum sequence length (512 tokens for BERT-based

models and 1024 tokens for LLM-based architectures).

- No sequence truncation is applied as the model handles variable-length sequences automatically.

3.2 Task Formulation

Given a scientific publication p with associated metadata and textual content, we define the telescope bibliography curation task as a multi-label classification problem combined with telescope identification.

Let $\mathcal{D} = \{(p_i, t_i, \mathbf{y}_i)\}_{i=1}^{N}$ denote our dataset of N scientific papers, where:

- p_i represents the i-th paper consisting of five textual components: $p_i = \{p_i^{\text{title}}, p_i^{\text{abstract}}, p_i^{\text{body}}, p_i^{\text{ack}}, p_i^{\text{grants}}\}$
- $t_i \in \mathcal{T}$ denotes the associated telescope, where $\mathcal{T}$ is the set of all telescopes in our taxonomy
- $\mathbf{y}_i = [y_i^{\text{sci}}, y_i^{\text{inst}}, y_i^{\text{men}}, y_i^{\text{not}}] \in \{0,1\}^4$ represents the multi-label annotation vector

3.3 Label Definitions

The four binary labels characterize the relationship between the paper and the referenced telescope. For each paper p:

- $y^{\text{sci}} = 1$ if p uses telescope data for new scientific results, 0 otherwise
- $y^{\text{inst}} = 1$ if p describes technical or instrumental aspects, 0 otherwise
- $y^{\text{men}} = 1$ if p mentions telescope without producing new results, 0 otherwise
- $y^{\text{not}} = 1$ if p contains false positive reference, 0 otherwise

3.4 Objective

Our goal is to predict two components for each paper p:

1. The telescope identifier: $\hat{t} \in \mathcal{T}$
2. The multi-label vector: $\hat{\mathbf{y}} \in \{0,1\}^4$

This can be achieved through various modeling approaches, including joint multi-task learning, pipeline architectures, or ensemble methods.

4 Methodology

4.1 Classical Machine Learning Approaches

We establish baseline models using classical machine learning methods with two text representation strategies: TF-IDF vectorization and count-based vectorization, combined with Multinomial Naive Bayes classifiers.

4.1.1 Text Representation

Given a paper p with concatenated text from all fields, we construct feature vectors using:

TF-IDF Vectorization: For each term w in paper p, the TF-IDF weight is computed as:

$$\text{TF-IDF}(w, p) = \text{TF}(w, p) \times \log \frac{N}{\text{DF}(w)}$$

where $\text{TF}(w, p)$ is the term frequency of word w in paper p, N is the total number of documents, and $\text{DF}(w)$ is the document frequency of word w.

Count Vectorization: We represent each paper as a vector of raw term frequencies:

$$\mathbf{v}_p = [\text{TF}(w_1, p), \text{TF}(w_2, p), ..., \text{TF}(w_{|V|}, p)]$$

where $|V|$ is the vocabulary size.

4.1.2 Classification Strategy

We employ Multinomial Naive Bayes classifiers with different strategies for telescope identification and label prediction:

Telescope Identification: For the multi-class telescope classification problem, we use a One-vs-Rest (OvR) approach. For each telescope $t \in \mathcal{T}$, we train a binary classifier:

$$P(t|\mathbf{v}_p) = \frac{P(\mathbf{v}_p|t) \cdot P(t)}{\sum_{t' \in \mathcal{T}} P(\mathbf{v}_p|t') \cdot P(t')}$$

The predicted telescope is:

$$\hat{t} = \arg\max_{t \in \mathcal{T}} P(t|\mathbf{v}_p)$$

Binary Classification: For each of the four binary labels $l \in \{\text{sci}, \text{inst}, \text{men}, \text{not}\}$, we train independent binary Multinomial Naive Bayes classifiers:

$$P(y^l = 1|\mathbf{v}_p) = \frac{P(\mathbf{v}_p|y^l = 1) \cdot P(y^l = 1)}{P(\mathbf{v}_p)}$$

Each label is predicted independently, allowing multiple labels to be assigned to a single paper when appropriate.

4.2 BERT-based Approaches

We apply transformer models with the following processing pipeline:

4.2.1 Tokenization

Input text is tokenized using the tokenizer corresponding to each pre-trained model. Due to the context length limitation of transformer models, the model **automatically truncates sequences to the first 512 tokens**, which is the maximum sequence length for BERT-based models. This typically includes the entire title and most of the abstract, which contain the most important information of the paper.

Given input text x, the tokenization process produces a sequence of token IDs:

$$\mathbf{t} = \text{Tokenize}(x) = [t_1, t_2, \ldots, t_n]$$

where $n \leq 512$. These tokens are then converted to embeddings and processed through the transformer encoder to obtain contextualized representations:

$$\mathbf{H} = \text{Transformer}(\mathbf{t}) = [\mathbf{h}_{\text{[CLS]}}, \mathbf{h}_1, \ldots, \mathbf{h}_n]$$

where $\mathbf{h}_{\text{[CLS]}} \in \mathbb{R}^d$ is the representation of the [CLS] token used for classification.

4.2.2 Classification Heads

We train two separate models, each with its own specialized classification head:

Multi-label Classification Model: A fully-connected layer with sigmoid activation is attached to the transformer encoder to predict 4 labels simultaneously (each label independently):

$$\mathbf{p}_{\text{multi}} = \sigma(\mathbf{W}_{\text{multi}}\mathbf{h}_{\text{[CLS]}} + \mathbf{b}_{\text{multi}})$$

where $\mathbf{W}_{\text{multi}} \in \mathbb{R}^{4\times d}$, $\mathbf{b}_{\text{multi}} \in \mathbb{R}^4$, and σ is the sigmoid function applied element-wise.

Telescope Identification Model: A separate model with a fully-connected layer and softmax activation is used to classify telescope types:

$$\mathbf{p}_{\text{telescope}} = \text{softmax}(\mathbf{W}_{\text{telescope}}\mathbf{h}_{\text{[CLS]}} + \mathbf{b}_{\text{telescope}})$$

where $\mathbf{W}_{\text{telescope}} \in \mathbb{R}^{K\times d}$, $\mathbf{b}_{\text{telescope}} \in \mathbb{R}^K$, and K is the number of telescope types.

Both models share the same transformer encoder architecture but are trained independently with their respective loss functions.

4.2.3 Training Objective

We train models independently or jointly for different classification tasks, using task-specific loss functions optimized for their respective objectives.

Binary Classification. For the four binary labels, we employ binary cross-entropy loss:

$$\mathcal{L}_{\text{multi-label}} = -\frac{1}{4}\sum_{i=1}^{4}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

Not-Telescope Classification. Due to significant class imbalance in the `not_telescope` category, we also employ focal loss when training a independent dedicated binary classifier.

$$\mathcal{L}_{\text{not-tel}} = -\Big[y \cdot \alpha(1-p)^\gamma \log(p) + (1-y)\cdot(1-\alpha)p^\gamma \log(1-p)\Big]$$

Telescope Identification. For multi-class telescope classification over K telescope types, we use categorical cross-entropy:

$$\mathcal{L}_{\text{telescope}} = -\sum_{k=1}^{K} y_k \log(p_k)$$

where $y_k \in \{0, 1\}$ is the one-hot encoded label and p_k is the predicted probability for telescope class k.

Each model is trained independently with its respective loss function, using the same base transformer architecture but optimized separately for its specific classification task. This modular approach allows task-specific optimization strategies and hyperparameter tuning.

4.2.4 Inference

At inference time, the model takes the first 512 tokens of a paper as input and forwards through the encoder. The encoded representation is then passed through two separate classification heads: one predicts the telescope type, and the other predicts the 4 classification labels (multi-label classification).

4.3 LLM-based Approach

We leverage large language models through parameter-efficient fine-tuning using QLoRA (Quantized Low-Rank Adaptation)(Dettmers et al., 2023), which enables training on consumer hardware by quantizing the base model to 4-bit precision while training low-rank adapter matrices.

Model Architecture. We fine-tune Qwen-1B and Qwen-3B models by freezing the quantized base parameters $\mathbf{W}$ and learning low-rank decompositions $\mathbf{AB}$ with rank r. The adapted weight matrix becomes: $\mathbf{W}' = \mathbf{W}_{\text{4-bit}} + \alpha \cdot \mathbf{AB}$

Task Formulation. We formulate classification as structured generation where the model outputs

JSON with telescope identification and binary labels. Each input consists of concatenated paper fields with a detailed system prompt encoding:

- Task objectives and label definitions
- Classification rules (e.g., mutual exclusivity of `not_telescope`)
- Output constraints (strict JSON schema)

System Prompt. Our prompt explicitly defines each category:

- `science`: Uses telescope data for new results
- `instrumentation`: Describes technical/engineering aspects
- `mention`: References telescope without new contributions
- `not_telescope`: Contains false positive references

The model is trained to generate valid JSON responses that are parsed during inference to extract predictions. This approach allows the LLM to reason about complex classification rules while producing structured outputs suitable for evaluation.

5 Experiments

5.1 Baselines

The TRACS organizers provide two official baseline models for comparison. Table 1 presents their performance on the test set.

Model	Macro F1
Random	0.24
openai-gpt-oss-20b	0.31

Table 1: Baseline performance on TRACS test set.

5.2 Experimental Setup

We split the training data into training and validation sets with an 8:2 ratio for model development and hyperparameter tuning. We train our models using adamw_torch optimizer with a learning rate of 2e-5, batch size of 16, and maximum sequence length of 512 tokens. For the multi-task models, training continues for 3 epochs with early stopping based on validation performance. For per-class binary classifiers, we train for 1-2 epochs to prevent overfitting, as single-task models tend to converge faster and are more prone to overfitting. All experiments are conducted on NVIDIA A100 GPUs via Google Colab. The primary evaluation metric is macro F1-score computed across both telescope identification and the four classification labels, ensuring balanced performance across all categories.

5.3 Main Results

5.3.1 Per-Class Specialized Models

To further improve classification performance, we train separate binary classifiers for each of the four classification categories (science, instrumentation, mention, not_telescope) and the telescope identification task. Table 2 shows the performance of our best model (SciBERT) when trained independently for each class.

Classification Task	F1 Score
Multi-label Classification	
science	0.78
instrumentation	0.76
mention	0.73
not_telescope	0.61
Macro F1 (Classification)	**0.72**

Table 2: Per-class F1 scores using separate SciBERT classifiers trained independently for each task. Macro F1 is computed as the average across all four classification categories.

5.3.2 Instruction-tuned LLM Evaluation

Training Configuration Table 3 presents the hyperparameters used for QLoRA fine-tuning. We employ 4-bit quantization to reduce memory footprint while maintaining model performance. The effective batch size of 8 is achieved through gradient accumulation, allowing training on consumer-grade hardware.

Hyperparameter	Value
Learning Rate	1×10^{-4}
Batch Size (per device)	1
Gradient Accumulation	8
Effective Batch Size	8
Max Epochs	3
Max Sequence Length	1024
Quantization	4-bit

Table 3: QLoRA fine-tuning hyperparameters for Qwen models.

Prompt Design We construct a structured system prompt that includes:

- **Role definition:** Positioning the model as an expert assistant for telescope paper classification

- **Category definitions:** Explicit descriptions of *science*, *instrumentation*, *mention*, and *not_telescope*
- **Classification rules:** Constraints such as mutual exclusivity of *not_telescope* and multi-label capability for other categories
- **Edge cases:** Guidelines for handling ambiguous references, name collisions, and grant-only mentions
- **Output format:** Strict JSON schema enforcement to ensure parseable predictions

Results The complete prompt template is provided in Appendix A. This prompt is prepended to each paper's content during both training and inference phases.

Method	Parameters	Macro F1
Qwen-1B + QLoRA	1B	0.58
Qwen-3B + QLoRA	3B	0.61

Table 4: Performance comparison on the multi-label classification task, trained for **a single epoch**. Macro F1 is averaged across all four categories (science, instrumentation, mention, not_telescope).

5.3.3 Joint Task Performance

We assess all models on the unified task encompassing both telescope identification and publication classification. The overall leaderboard score is defined as the arithmetic mean of the F1 score for telescope identification and the macro-averaged F1 across the four classification categories, formulated as:

$$\text{Final Score} = \frac{\text{Telescope F1} + \text{Classification Macro F1}}{2}.$$

Table 5 summarizes the complete performance comparison across all evaluated methods.

5.3.4 Ablation Study

To examine the effect of focal loss, we fine-tuned task-specific models with and without focal loss on imbalanced tasks. Although focal loss slightly improved per-task stability, these models still performed worse than the joint multi-task model trained without focal loss, indicating that task interaction contributes more to generalization than loss reweighting alone.

6 Conclusion

In this work, we presented a systematic study of modeling strategies for automatic telescope bibliographic classification in the TRACS @ WASP 2025 shared task. We evaluated a diverse range of approaches, from classical machine learning methods to transformer-based architectures and limited-resource LLM-based inference.

Our experiments demonstrate that domain-adapted BERT variants significantly outperform traditional ML, with SciBERT achieving the best performance of 0.73 macro F1 on the leaderboard evaluation—more than doubling the official baseline score of 0.31. We show that pre-training on scientific corpora provides substantial benefits for this task, as evidenced by the strong performance of SciBERT and AstroBERT compared to general-domain models.

While ensemble methods did not yield improvements in our experiments, we attribute this primarily to the multi-label, multi-class complexity of the task and our computational constraints. Our task-wise classification approach with focal loss showed promise in addressing class imbalance, though further investigation with larger models and more extensive hyperparameter tuning could yield additional gains.

Importantly, our limited-resource LLM experiments suggest that instruction-tuned models can achieve competitive performance even without full fine-tuning and with only partial training data. This opens promising directions for low-resource scenarios in scientific bibliography curation.

Future work should explore more sophisticated long-document processing strategies to better leverage complete paper content, investigate advanced techniques for handling severe class imbalance in multi-label settings, and examine larger-scale LLM fine-tuning with expanded computational resources. Additionally, incorporating metadata such as author affiliations, publication venues, and citation networks may further improve classification accuracy.

Limitations

This study faces several important constraints:

Computational Resource Constraints: As students, we faced significant GPU and computational limitations. This restricted our ability to experiment with larger models (e.g., full fine-tuning of

Method	Multi-label Classification			Telescope Identification		Leaderboard
	Macro F1	Micro F1	Samples F1	Accuracy	Macro F1	Score
Traditional ML						
TF-IDF	0.52	0.58	-	0.64	0.51	0.51
CountVectorizer	0.54	0.59	-	0.67	0.56	0.56
Transformer Models						
DistilBERT (66M)	0.71	0.73	0.72	0.81	0.68	0.69
SciNCL (110M)	0.71	0.73	0.72	0.80	0.68	0.68
AstroBERT (110M)	0.72	0.73	0.74	0.80	0.69	0.68
SPECTER (110M)	0.70	0.72	0.72	0.81	0.68	0.69
SciBERT (110M)	**0.77**	**0.79**	**0.78**	**0.81**	**0.73**	**0.72**
Ensemble Methods						
Soft Voting	0.70	0.72	0.71	0.73	0.65	0.67
Weighted Voting	0.71	0.73	0.72	0.74	0.66	0.68
Hard Voting	0.69	0.71	0.70	0.72	0.64	0.66

Table 5: Comparison of traditional ML, transformer-based models, and ensemble methods on joint telescope identification and paper classification tasks. The leaderboard score is computed as the average of Telescope Macro F1 and Classification Macro F1. Ensemble methods combine SciBERT, DistilBERT, and AstroBERT using different voting strategies but show slight performance degradation compared to the best single model (SciBERT). With SciBERT, our system achieves a **Top-2** ranking on the leaderboard.

models beyond 3B parameters) and limited the hyperparameter search space we could explore.

Ensemble Methods Underperformance: Despite theoretical advantages, our ensemble approaches did not yield substantial improvements. This is likely due to the multi-label, multi-class nature of the task where predictions must simultaneously classify both the telescope type and four binary labels (science, instrumentation, mention, not_telescope). The complexity of combining predictions across these dimensions without introducing conflicting classifications proved challenging within our resource constraints.

Class Imbalance: The dataset exhibits significant class imbalance across both telescope types and label categories. Certain telescope-label combinations are severely underrepresented, making it difficult for models to learn robust patterns for minority classes and potentially biasing predictions toward more frequent categories.

Long Document Processing: Scientific papers often contain extensive text spanning abstracts, full body text, and acknowledgments. Processing these long sequences requires either truncation (risking information loss) or sophisticated chunking strategies. Our computational constraints limited our ability to fully leverage the complete textual context, particularly for papers exceeding typical transformer input limits (512 tokens).

Acknowledgments

We thank the organizers of the TRACS @ WASP 2025 shared task for providing this valuable learning opportunity and the benchmark dataset. We are grateful to the SAO/NASA Astrophysics Data System (ADS) for maintaining the telescope bibliography infrastructure that made this work possible.

Through this shared task, we gained significant insights into challenges of long-document processing, resource-constrained LLM deployment, and domain-specific NLP in astronomy. We acknowledge the developers of the Hugging Face Transformers library and the creators of SciBERT, AstroBERT, SPECTER, and Qwen models for making their pre-trained models publicly available.

Despite computational limitations as students, this experience deepened our understanding of practical deployment considerations for scientific text classification systems and the unique characteristics of astronomical literature.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *Preprint*, arXiv:1903.10676.

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.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

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.

Félix Grèzes, Sergi Blanco-Cuaresma, and .... 2021. Building astrobert, a language model for astronomy astrophysics. *arXiv preprint arXiv:2112.00590*. Preprint; also presented at ADASS 2021/ADASS 2022.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *Preprint*, arXiv:2006.04884.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11670–11688. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Preprint*, arXiv:1804.07461.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

A System Prompt for TRACS Classification

Below is the complete system prompt used to fine-tune Qwen models and guide their inference on the TRACS shared task.

You are an expert assistant for the TRACS (Telescope Reference and Astronomy Categorization Shared Task) at WASP @ IJCNLP-AACL 2025. Your task is to classify scientific papers according to telescope usage categories defined by the shared task guidelines.

Given the paper content, identify:

1. telescope: The specific telescope referenced in the paper.

2. science: True if the paper uses telescope data to produce new scientific results.

3. instrumentation: True if the paper describes technical aspects of the telescope (hardware/software/calibration/data pipeline).

4. mention: True if the paper references the telescope but does not produce new results nor address technical aspects.

5. not_telescope: True if the paper contains misleading telescope-like references or false positives unrelated to an actual telescope.

Classification Rules:

- A paper can be classified into multiple categories except 'not_telescope', which is mutually exclusive.

- If a paper qualifies for 'science', it must be labeled science=True even if it also mentions the telescope.

- If a paper discusses telescope engineering or data processing, label instrumentation=True.

- Papers that only cite a telescope historically, in background, or for comparison → mention=True.

- If the telescope name is used ambiguously (e.g. name collision with a person, project, or acronym) → not_telescope=True.

- Referencing telescope-funded grants alone without data use → not_telescope=True.

Output Format:

Respond strictly in valid JSON only:

```
{
  "telescope": "<string>",
  "science": <true/false>,
  "instrumentation": <true/false>,
  "mention": <true/false>,
  "not_telescope": <true/false>
}
```