ORBIT: Cost-Effective Dataset Curation for Large Language Model Domain Adaptation with an Astronomy Case Study

Anonymous ACL submission

Abstract

Recent advances in language modeling demonstrate the need for high-quality domain-specific training data, especially for tasks that require specialized knowledge. General-purpose models, while versatile, often lack the depth needed for expert-level tasks because of limited domain-specific information. Domain adaptation training can enhance these models, but it demands substantial, high-quality data. To address this, we propose ORBIT, a cost-efficient methodology for curating massive, high-quality domain-specific datasets from noisy web sources, tailored for training specialist large language models. Using astronomy as a primary case study, we refined the 1.3T-token FineWeb-Edu dataset into a high-quality, 10B-token subset focused on astronomy. Fine-tuning LLAMA-3-8B on a 1B-token astronomy subset improved performance on the MMLU astronomy benchmark from 69% to 76% and achieved top results on AstroBench, an astronomy-specific benchmark. Moreover, our model (Orbit-LLaMA) outperformed LLAMA-3-8B-BASE, with GPT-4o evaluations preferring it in 73% of cases across 1000 astronomy-specific questions. Additionally, we validated ORBIT's generalizability by applying it to law and medicine, achieving a significant improvement of data quality compared to an unfiltered baseline. We open-source the ORBIT methodology, including the curated datasets, the codebase, and the resulting model.

1 Introduction

The rapid advancement of large language models (LLMs) has transformed natural language processing (NLP) and artificial intelligence (AI), with general-purpose models like GPT-4 and LLaMA demonstrating versatility across tasks such as knowledge retrieval, open-domain question answering, and linguistic applications. However, these models often struggle in specialized domains, such as astronomy, where deep, nuanced understanding and up-to-date factual accuracy are crucial (Singhal et al., 2023). This performance gap arises because general-purpose LLMs must balance performance across a wide range of tasks, diluting domain-specific knowledge (Li et al., 2024; Yang et al., 2024b).

To address this limitation, domain-specialized LLMs can allocate their capacity toward mastering specific domains, offering greater depth and accuracy. However, building these models is challenging due to the need for high-quality, domain-specific datasets. Conventional approaches, such as using academic sources like arXiv papers, tend to focus on highly technical content, neglecting the breadth and diversity needed for effective model generalization. Alternatively, web-sourced datasets offer greater diversity but are often noisy, containing irrelevant or low-quality content. Traditional filtering methods, such as keyword-based or rule-based approaches, frequently fail to balance coverage and quality, potentially excluding relevant data while admitting suboptimal material.

In this work, we propose **ORBIT**, a novel, scalable data curation framework for creating high-quality, domain-specific datasets. ORBIT combines embedding-based similarity matching with a BERT-based regression model to filter large-scale web datasets efficiently. By focusing on both semantic relevance and educational value, this methodology ensures that the curated datasets are both diverse and tailored to specific domains. Using astronomy as the primary case study, we curated a 10-billion-token dataset derived from FineWeb-Edu (Penedo et al., 2024), incorporating a broader range of content compared to prior approaches like AstroLLaMA (Nguyen et al.), which rely solely on arXiv abstracts. The inclusion of web-sourced educational content alongside academic texts enables ORBIT to balance depth and diversity, capturing a more comprehensive understanding of domain-specific knowledge.

To demonstrate the generalizability of ORBIT, we also applied it to law and medicine, achieving significant quality improvements in these domains. GPT-4o evaluations rated the curated datasets at an average educational value of 3.05 and 2.9 on a scale of 0-5 per document, respectively, compared to an unfiltered baseline of approximately 0.4. These results highlight ORBIT's ability to extract domain-relevant, high-quality data across diverse fields.

Fine-tuning a LLAMA-3-8B model on a randomly sampled 1B-token astronomy subset of the ORBIT-curated dataset results in substantial improvements on astronomy-specific tasks. Our model (Orbit-LLaMA) achieves a 7-point accuracy gain over the base LLaMA-3-8B model (from 69.08% to 76.3%) on the MMLU astronomy benchmark and outperforms AstroLLaMA (66.45%) by a significant margin. Furthermore, ORBIT-trained models surpass state-of-the-art performance on various astronomy baselines, receiving higher ratings from both GPT-4o evaluations and domain experts in the vast majority of cases. These results underscore the value of ORBIT's methodology in producing specialized datasets that enhance both the depth and breadth of domain-specific knowledge in LLMs.

The key contributions of this paper are:

- We introduce **ORBIT**, a generalizable, scalable framework for filtering noisy web data into high-quality, domain-specific datasets, addressing challenges of scalability, noise, and coverage balance.
- We demonstrate ORBIT's generalizability by applying it to multiple domains, including astronomy, law, and medicine, achieving significant quality improvements in each field with minimal computational overhead.
- We present a **specialized astronomy dataset** curated using ORBIT, comprising 10 billion tokens that combine academic rigor with web-scale diversity, advancing prior work limited to arXiv-based sources.
- We train a **state-of-the-art astronomy-specific language model** (which we call Orbit), fine-tuned on a subset of the ORBIT-curated dataset, achieving significant performance gains on astronomy-related benchmarks and surpassing existing models, including AstroLLaMA, in expert evaluations.

By presenting ORBIT and its application to astronomy, as well as its successful extension to law and medicine, we provide a generalizable framework for developing targeted, domain-specific AI tools. This methodology has the potential to accelerate scientific research, education, and practical applications across a wide range of specialized fields.

2 Related Work

2.1 Data Curation for Language Models

Recent research has demonstrated the paramount role of high-quality data in the development of large language models. For instance, the technical reports of models like LLama-3 (Grattafiori et al.) and Qwen-2 (Yang et al., 2024a) emphasize extensive data curation methodologies for general-purpose language models. These efforts have led to significant performance gains, even when model architectures and parameter sizes remain largely unchanged (e.g., the transition from LLama-2 to LLama-3).

Several efforts have focused on automated data curation techniques. Chen et al. (2023) proposed a method to automatically filter and clean web-crawled data to build high-quality training corpora, while Gururangan et al. (2020) developed a data selection method for identifying domain-relevant examples within large datasets. Furthermore, Kreutzer et al. (2022) demonstrated that smaller, carefully curated datasets often outperform larger but noisier datasets.

However, these methods often face limitations when applied to highly specialized domains. Many automated filtering techniques rely on general quality metrics or term whitelisting, which can inadvertently include irrelevant or low-quality content while excluding high-quality data that does not fit predefined patterns. For instance, filtering by specific terms or phrases, such as LaTeX commands, may be effective in domains like mathematics but fails in more diverse fields like astronomy where specialized exact terms do not exist or are more varied. Additionally, many datasets rely on scraped web data, which presents risks related to copyright issues, noise, and incomplete data extraction from APIs, further limiting the potential for domain-specific curation.

2.2 Domain-Specific Language Models

Advances in natural language processing have led to the rise of domain-specific language models that are fine-tuned on specialized corpora. These models are designed to perform well within particular domains, outperforming general-purpose models on domain-specific tasks (Beltagy et al., 2019). However, each of these approaches has notable limitations.

For example, Azerbayev et al. (2024) introduced LLEMMA, an open-source language model for mathematics that achieves state-of-the-art results on the MATH benchmark. LLEMMA filters data based on whether it contains LaTeX syntax, a technique well-suited to mathematics but restrictive when applied to other fields, such as astronomy or biology, where such syntactic markers do not exist. This method risks excluding valuable content that lacks LaTeX or including low-quality data simply because it contains LaTeX markup.

Similarly, Singhal et al. (2023) developed Med-PaLM 2, a medical domain model that achieved 85.4% accuracy on US Medical Licensing Examination (USMLE) questions. However, its approach to fine-tuning is relatively limited, relying primarily on instruction fine-tuning without deep post-training adjustments specific to medical literature, limiting its adaptability for more niche medical tasks.

Other domain-specific models face similar limitations in data sourcing. Yang et al. (2023) introduced FinGPT, which demonstrates strong performance on financial tasks, but it heavily relies on domain-specific data sources like SEC filings and NYSE transaction reports. These data sources are highly specific to the financial domain and do not generalize well to other fields, limiting the flexibility of such models.

Nguyen et al. introduced AstroLLaMA, a 7-billion-parameter model fine-tuned on the abstracts of 300,000 astronomy papers from arXiv. Furthermore, Ting et al. (2024) builds upon this work with larger and more modern models. While these works show strong performance in generating scientifically relevant text completions, limiting the dataset to only arXiv papers (and in this case, only to certain sections such as the Abstract and Introduction) restricts the breadth and depth of the information available for fine-tuning. The homogeneous distribution of similarly formatted research abstracts leads to a lack of data diversity that reduces the model's capacity to generalize across broader applications within the domain.

These models highlight the importance of high-quality, domain-specific datasets for effective model performance but also demonstrate the challenges in collecting and curating sufficiently diverse and representative datasets.

Algorithm 1 Domain-Specific Dataset Curation Pipeline

Input: Corpus of documents, astronomy-related terms, similarity threshold τ, educational value threshold η
Output: Filtered astronomy-specific dataset
Initialize astronomy vector $\mathbf{A}$ by averaging embeddings of astronomy-related terms
Stage 1: Embedding-Based Threshold Filtering
for each document D in the corpus **do**
  Compute document vector $\mathbf{B}$ by aggregating embeddings of tokens in D
  Calculate similarity: $\text{Similarity}(D) = \frac{\mathbf{A} \cdot \mathbf{B}}{|A| * |B|}$
  if $\text{Similarity}(D) > \tau$ **then**
    Retain document D
  end if
end for
Stage 2: BERT-Based Regressor Evaluation
for each retained document D **do**
  Compute educational value score $EV(D)$ using BERT-based regressor
  if $EV(D) > \eta$ **then**
    Retain document D
  end if
end for
Return filtered dataset

3 Dataset Curation Methodology

3.1 Choice of Corpus

For this study, we selected the FineWeb-Edu dataset (Penedo et al., 2024) as our primary corpus. FineWeb-Edu is a specialized subset of FineWeb, which is a large-scale, high-quality dataset derived from CommonCrawl web data, specifically designed for pretraining large language models. The FineWeb-Edu dataset uses the Open Data Commons License Attribution family. FineWeb-Edu focuses on "educational content" based on prompt engineering strategies and contains approximately 1.3 trillion tokens, curated by filtering out content

with lower educational value. This subset allowed us to begin with a high-quality dataset that is more focused and manageable for the specific tasks required in astronomy. Figure 1 illustrates the comprehensive filtering pipeline from FineWeb-Edu to ORBIT (our method), highlighting the quality and size at each step with examples of what has been eliminated. Our full filtering method is shown programmatically in Algorithm 1.

3.2 Methodology for Domain-Specific Dataset Curation

Our research presents a novel approach to curating a high-quality, domain-specific dataset for astronomical language models. This methodology combines advanced natural language processing techniques with rigorous quality assurance measures to produce a dataset that balances complex reasoning tasks with factual content in the field of astronomy. Our approach is designed for cost-effectiveness, using a combination of broad initial filtering and more thorough assessments at later stages to optimize the dataset's quality and relevance.

3.2.1 Stage 1: Initial Domain-Specific Filtering

We developed a lexicon of 101 single-word astronomy-related terms, encompassing concepts from astrophysics, cosmology, and space exploration. To efficiently process large volumes of text, we implemented an embedding-based matching technique utilizing GloVe word embeddings (Pennington et al., 2014). A representative astronomy aggregated embedding vector $\mathbf{A}$ was computed by averaging the embeddings of all terms. For each document in FineWeb-Edu, we calculated a document vector and computed its cosine similarity with $\mathbf{A}$. Documents exceeding a similarity threshold of $\tau = 0.2$ were retained for further analysis. This threshold was empirically determined to balance dataset size and quality, resulting in approximately 10 billion tokens of high-quality, astronomy-relevant content. After this stage, approximately 20B tokens of the corpus remained.

3.2.2 Stage 2: Educational Value Assessment

After the initial filtering, we applied a more thorough evaluation to refine the dataset further, focusing on its educational merit. Without this second phase, we would be left with a number of low-qualtiy documents, as shown in Figure 3. Furthremore, if only Stage 2 was applied, the computational cost would increase significantly. For example, if Stage 1 keeps $\frac{1}{100}$ of the total data, the number of NVIDIA A100 GPU hours needed for stage 2 would decrease by 100x. See Table 1 for more information.

We developed a BERT-based regressor model (Devlin et al., 2019), using Huggingface's HUGGINGFACEFW/FINEWEB-EDU-CLASSIFIER model, trained to evaluate the educational value of astronomy-related text on a scale of 0 to 5. We kept any value above or equal to our threshold $\eta = 3$

The training dataset for this model was meticulously curated through a multi-step process:

1. Random sampling of 50,000 documents from the embedding-filtered corpus to ensure topic diversity.

2. Automated evaluation of each sampled document using GPT-4o model (OpenAI et al., 2024), which was prompted to assess the educational value on a 6-point scale (0-5).

3. Collection of both quantitative scores and qualitative justifications for each evaluation, used for prompt engineering.

The language model was instructed to consider factors such as depth of astronomical content, clarity of explanations, relevance to a general audience, and the presence of advanced concepts. Our prompt, inspired by Yuan et al. (2024) (see Appendix), emphasized educational value specific to the domain of astronomy. See Figure 2 for a visual of our Stage 2 pipeline.

3.2.3 Cross-Domain Validation: Law and Medicine

To assess the generalizability of ORBIT, we extended the dataset curation pipeline to two additional domains: law and medicine. Using the same methodology applied to astronomy, we developed domain-specific lexicons for these fields. For law, the lexicon included terms such as "litigation," "precedent," and "contract," while for medicine, it featured terms like "pathology," "oncology," and "metastasis." The complete lists of terms for each domain are provided in the Appendix.

Stage 1 filtering, based on embedding-based similarity, was adapted to these domains by computing aggregated embedding vectors from their respective lexicons. For each document in FineWeb-Edu,

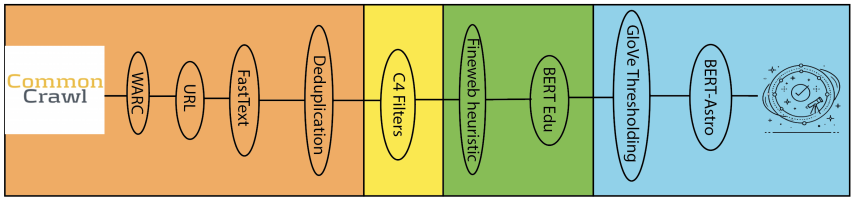

Figure 1: Comprehensive Filtering Pipeline from FineWeb-Edu to ORBIT. The pipeline emphasizes the quality and size of the dataset. The orange includes common filtering methods formalized in Wenzek et al. (2020). The yellow summarizes large-scale semantic filters from Raffel et al. (2023). The green includes the additional semantic filters and the BERT-based classifier used to filter for educational relevance in FineWeb-Edu. The blue outlines our contributions: GloVe-based embedding thresholding and a BERT classifier for educational relevance specific to astronomy. See subsections 3.2 and 3.2.2 for details on our contributions.

Table 1: Comparison of Processing Time and Cost for Dataset Filtering. Stage 1 filtering retains 1% of documents (and thus tokens), drastically reducing the effective dataset size for Stage 2. Stage 2 alone processes the full dataset. The combined approach significantly lowers the time and cost of Stage 2. Pricing estimates are based on current market rates and hardware usage. Furthermore, both stages are fully parallelizable, meaning additional hardware can cause linear decrease in time for an approximately constant price.

Scenario	Processing Unit	Total Time	Total Cost	Quality
Stage 1 Only	Intel Core i9 (16 cores)	177 hours	$44	Medium
Stage 2 Only	A100 PCIe GPU (1 unit)	12,000 hours	$16,200	Highest
Stage 1 + Stage 2	Intel Core i9 + A100 PCIe GPU	297 hours	$206	Highest

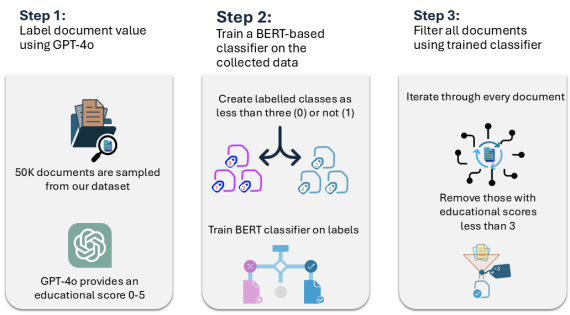

Figure 2: Full Stage 2 pipeline visualized.

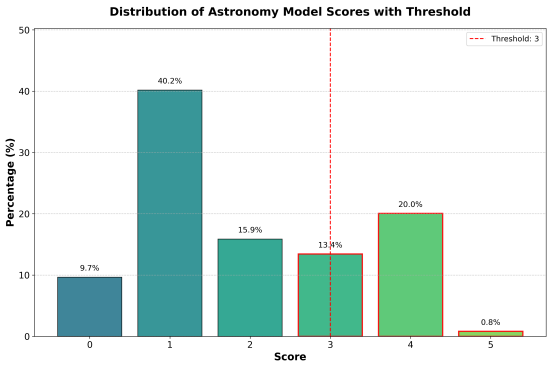

Figure 3: Distribution of educational value scores (ranging from 0 to 5) assigned by the BERT-based regressor model to a sample of 1000 astronomy-related documents. This visualization demonstrates the validity of the classifier by showing alignment with expected distributions based on held-out test sets and expert evaluations.

the cosine similarity between its embedding vector and the domain-specific aggregated vector was calculated. Documents exceeding the similarity threshold of 0.2 were retained for further analysis.

4 Experiments

To validate the effectiveness of the ORBIT methodology, we conducted a series of experiments focusing on the quality of the curated dataset, the impact of fine-tuning on model performance, and the influence of different thresholding values within the pipeline. These experiments aim to assess how ORBIT's two-stage filtering approach improves dataset relevance and educational value while balancing dataset size and computational cost. Additionally, we evaluate the performance of models fine-tuned on ORBIT-curated datasets with varying similarity and educational value thresholds, examining their impact on downstream tasks. The results provide insights into the trade-offs between dataset size, quality, and curation efficiency, while demonstrating the effectiveness of ORBIT for training astronomy-specialized language models. Below, we outline the experimental setup, datasets, and evaluation metrics used to address these questions.

4.1 Experimental Setup

For our experiments, we utilized the Delta GPU cluster at the National Center for Supercomputing Applications, equipped with 8 NVIDIA A100 GPUs, each with 40GB of memory. The model, named Orbit-LLaMA, was derived from Meta's LLaMA-3-8B, an 8-billion-parameter language model optimized for large-scale training. We used the Punkt tokenizer from NLTK for sentence segmentation during preprocessing. LLaMA-3 operates under the LLaMA 3 Community License Agreement. See the Appendix for more training details.

4.2 Effect of Thresholding, Embedding Methods, and Keyword Search

To explore the effectiveness of various filtering strategies, we tested the impact of:

1. Different threshold values in embedding similarity filters.

2. Multiple embedding methods, including fastText, 100-dimensional, and 300-dimensional embeddings.

3. Keyword filtering approaches compared to unfiltered datasets.

This analysis assessed how these methods balance dataset quality and coverage. The performance of each filtering strategy was measured based on average scores obtained from downstream tasks, as shown in Figure 4. Error bars indicate the standard error of the mean (SEM), highlighting variability. The results underscore how keyword filters and embedding-based thresholds can improve dataset curation by focusing on the most relevant content.

The results demonstrate that:

- Higher threshold values generally reduce dataset size while maintaining or improving average scores.

- Embedding methods showed slightly varying efficacy.

- Keyword filtering, while simpler, achieved competitive performance by focusing on domain-specific terminology.

- No filtering resulted in the largest datasets but the lowest scores.

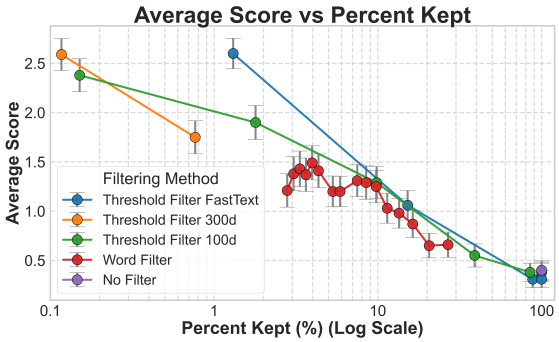

Figure 4: Average Score vs Percent Kept, comparing different filtering methods: embedding thresholds (fastText, 100d, 300d), keyword filtering, and no filtering. The x-axis is log-scaled for clarity.

4.3 Cross-Domain Validation: Law and Medicine

To evaluate ORBIT's generalizability, we applied the dataset curation pipeline to two additional domains: law and medicine. Stage 1 filtering was adapted to these domains by constructing domain-specific lexicons, following the methodology described in Section 3.2. For law, the lexicon included terms such as "litigation," "precedent," and "contract," while for medicine, it featured terms like "pathology," "oncology," and "metastasis" (see Appendix for full term lists).

Embedding-based similarity filtering retained approximately 1.0% of the initial corpus for law and 1.0% for medicine, similar to the retention rate observed for astronomy. The average educational value scores, evaluated using GPT-4o, showed significant improvements over the unfiltered baseline (0.3), with 2.9 for medicine and 3.05 for law.

These scores align closely with the results obtained for astronomy, indicating that Stage 1 filtering alone is sufficient to extract high-quality, domain-specific content across diverse fields.

4.4 Benchmarks and Baselines

We evaluated Orbit-LLaMa using multiple datasets, including the astronomy section of the MMLU benchmark (Hendrycks et al., 2021) and two versions of AstroBench. Baseline models were AstroLLaMA-3-8B (Pan et al., 2024), the prior state-of-the-art in astronomy language modeling, and Meta-LLaMA-3-8B, a general-purpose model.

A total of three datasets were used for quantitative analysis:

1. **Hugging Face AstroBench Subcategories**:

Organized into subcategories:

- **Basic Knowledge (BK)**: Tests core astronomy concepts.
- **Scientific Calculation (SC)**: Involves solving astrophysical numerical problems.
- **Knowledge Application (KA)**: Assesses applying knowledge to novel scenarios.

Each subcategory is scored separately for detailed performance analysis.

2. **Official AstroBench Benchmark**: A comprehensive dataset of 4,425 multiple-choice questions from 885 *Annual Review of Astronomy and Astrophysics* articles (1963–2023). It provides an aggregated performance score, covering diverse topics such as quasars, cosmological simulations, and the circumgalactic medium.

3. **MMLU Benchmark**: The astronomy section of MMLU evaluates factual knowledge and reasoning across topics like stellar formation and cosmology, testing scientific depth in language models.

Evaluation Methodology. We used multiple-choice perplexity prediction to select answers and conducted qualitative pairwise comparisons rated by expert astronomers and GPT-4 for accuracy, clarity, and reasoning.

4.4.1 Qualitative Analysis

Qualitative evaluations compared responses from Orbit-LLaMA, AstroLLaMA, and Meta-LLaMA using 24 test questions developed by Astronomy Ph.D. students and faculty. Responses were ranked for accuracy (or, for active areas of research, likelihood), clarity, and reasoning using two methods:

1. **Preference Ratings**: Four graduate students selected the best response for each question. Majority consensus was reached for 83% of questions, with Orbit-LLaMA preferred for 66% of total responses (Table 2).

2. **Detailed Feedback**: Reviewers noted:

   - **Meta-LLaMA**: Responses often repeated content and lacked focus.
   - **Orbit-LLaMA**: Delivered clear and concise answers resembling student-created work.

Model	Selected Output(%)
Meta-LLaMA	22.1%
Orbit-LLaMA	66.3%
AstroLLaMA	11.6%

Table 2: The total number of times each model's response was selected from total votes cast (N = 95).

   - **AstroLLaMA**: Long, research-style responses with structural and coherence issues.

4.5 Experiment Results

Orbit-LLaMa outperformed baselines on all metrics. On the MMLU astronomy section, Orbit-LLaMa scored 76 compared to 69 (Meta-LLaMA) and 66.45 (AstroLLaMA). On AstroBench subcategories, Orbit-LLaMa excelled in Basic Knowledge (45.53%), Scientific Calculation (30.28%), and Knowledge Application (45.53%). On the official AstroBench, Orbit-LLaMa scored 69.7, surpassing AstroLLaMA (66.4) and Meta-LLaMA (61.5).

Table 3 summarizes the results, showing Orbit's superior performance in both specific tasks and overall benchmarks.

Pairwise comparisons confirmed Orbit-LLaMA's superiority, with win rates over 92% against baselines (Table 4). Expert feedback highlighted its accuracy, clarity, and reasoning improvements. See the appendix for detailed examples.

5 Discussion

The results demonstrate the utility of the ORBIT methodology in addressing key challenges in domain-specific dataset curation and fine-tuning. By using a two-stage filtering process, ORBIT balances relevance and quality while remaining computationally efficient. Stage 1's embedding-based similarity filtering significantly reduces the dataset size, while Stage 2's educational value assessment ensures the retained data is highly relevant and informative. This layered approach enables the creation of datasets that are both comprehensive and focused, as evidenced by its application to astronomy, law, and medicine.

Fine-tuning Orbit-LLaMA on the ORBIT-curated dataset led to notable improvements across multiple benchmarks, including MMLU astronomy and AstroBench. The gains in both quantitative metrics and qualitative evaluations highlight the

Table 3: Performance Comparison of Models on MMLU and AstroBench.

Model	AS	CC	CP	CPH	HSC	HSP	KA	SC	BK	AstroBench
AstroLLaMA	66.45	47.00	38.24	55.74	53.20	41.06	39.84	29.48	63.75	66.4
Meta-LLaMA	69.08	44.00	37.25	54.04	52.22	41.72	41.46	25.90	65.50	61.5
Orbit-LLaMA	**76.30**	**52.00**	**47.10**	**56.20**	**53.70**	**43.10**	**45.53**	**30.28**	**69.96**	**69.7**

Table 4: Win Rates and Tie Percentages Between Models.

Models Compared	Meta-LLaMA	Orbit-LLaMA	AstroLLaMA	Tie
Meta-LLaMA vs Orbit-LLaMA	25.4	**73.0**	-	1.6
Meta-LLaMA vs AstroLLaMA	**84.3**	-	10.5	5.22
Orbit-LLaMA vs AstroLLaMA	-	**93.0**	5.0	2.0

impact of curating diverse and high-quality domain-specific data. The inclusion of a mix of academic and educational content allowed the model to excel in tasks requiring both factual knowledge and nuanced reasoning, demonstrating the value of combining depth with breadth in training corpora.

The success of ORBIT in multiple domains also suggests its scalability and adaptability. However, differences in domain-specific challenges, such as interdisciplinary overlaps or evolving knowledge in fields like medicine, highlight the need for further refinement. Future work could focus on automating lexicon creation and optimizing threshold selection to streamline application to new domains.

Overall, the experiments validate the potential of domain-adapted LLMs when supported by robust curation pipelines like ORBIT. This approach addresses limitations in general-purpose models for specialized tasks, emphasizing the importance of targeted datasets for achieving state-of-the-art performance in specific fields.

6 Conclusion

This paper presents a novel approach to creating high-quality, domain-specific datasets for training language models, with a focus on the field of astronomy. Our methodology, combining embedding-based matching and BERT-based regression for data filtering and selection, has demonstrated significant potential for enhancing the performance of language models in specialized scientific domains. Furthermore, we validated the scalability and generalizability of this approach by extending it to the domains of law and medicine, achieving similar improvements in dataset quality.

The key findings of our study include:

1. The effectiveness of our data curation methodology in creating balanced, high-quality datasets that support both complex reasoning and factual knowledge across multiple domains, including astronomy, law, and medicine.

2. Significant improvements in model performance on astronomy-related tasks, even with relatively small-scale training data, highlighting the potential for efficient resource utilization.

3. The adaptability of our methodology to diverse scientific and professional fields, demonstrating that domain-specific models can outperform general-purpose models in specialized tasks.

In conclusion, our work represents a significant step toward more efficient and effective AI tools for specialized scientific and professional domains. As this field continues to evolve, we anticipate that domain-specific language models will play an increasingly important role in supporting research, education, and decision-making across a wide range of disciplines. Moreover, we believe that ongoing collaboration between AI researchers and domain experts will be essential to unlocking the full potential of these models in addressing complex, real-world challenges.

Acknowledgments

This research was supported in part by the Delta compute cluster at the National Center for Supercomputing Applications (NCSA). We thank the NCSA for providing computational resources and support that significantly contributed to the success of this work.

7 Limitations

While the ORBIT methodology and the resulting Orbit model show significant promise, it is essential to acknowledge several limitations that may impact their applicability and effectiveness. These limitations are categorized into technical and social aspects to provide a comprehensive understanding of the challenges involved.

7.1 Technical Limitations

The primary technical limitations of the ORBIT methodology and the Orbit model are as follows:

- **Domain-Specific Generalizability.** Although ORBIT has proven effective in the field of astronomy, its applicability to other domains remains untested. Domains with less structured data or those that are highly interdisciplinary may require additional adaptations to the filtering and evaluation processes. Defining domain-specific terms and educational value criteria in such fields could pose unique challenges that the current methodology does not address.

- **Dependence on Embedding Models.** The embedding-based filtering approach relies heavily on the quality and coverage of pretrained word embeddings, such as fastText. These embeddings may not fully capture the nuances of highly specialized or emerging astronomical terminology, potentially leading to the exclusion of relevant content or the inclusion of less pertinent material. Enhancing embedding models to better represent domain-specific language could mitigate this limitation.

- **Computational and Resource Constraints.** Despite the efficiency gains from using frameworks like DeepSpeed and FlashAttention v2, the fine-tuning process for large models like Orbit demands substantial computational resources. This requirement may limit accessibility for smaller research teams or institutions with limited budgets. Additionally, scaling the methodology to accommodate larger datasets or models with higher parameter counts may encounter practical barriers related to memory and processing power.

- **Evaluation Scope.** The current evaluations are primarily focused on astronomy-specific tasks and benchmarks such as MMLU and AstroBench. This narrow scope may limit the generalizability of the findings, as broader benchmarks that include interdisciplinary or collaborative tasks have not been assessed. Expanding the evaluation to encompass a wider range of benchmarks would provide a more comprehensive assessment of the model's utility.

- **Dynamic Nature of Scientific Knowledge.** Astronomy is a rapidly evolving field, and the curated dataset represents a specific temporal snapshot. As new discoveries and theories emerge, the model's relevance and accuracy may decline without ongoing updates. Developing methods for efficiently integrating new knowledge into existing models is necessary to maintain their effectiveness over time.

Addressing these technical limitations will require future work to explore the adaptability of the ORBIT methodology across domains, enhance embedding models for better domain-specific representation, and develop scalable solutions to manage computational demands.

We acknowledge the assistance of ChatGPT for paraphrasing and shortening text in this document. All content generated with AI was carefully reviewed and validated by the authors.

8 Ethical Considerations

The development of domain-specific language models like Orbit raises several ethical considerations that warrant careful examination:

- **Transparency and Open Sourcing.** Open-sourcing the methodology, dataset, and codebase promotes transparency and ensures that other researchers can replicate and validate our findings. However, this accessibility also increases the risk of misuse. For example, malicious actors could adapt the approach to create highly specialized LLMs for unethical purposes, such as generating misleading or pseudoscientific content within specialized domains.

- **Mitigation of Misuse.** To mitigate risks of misuse, safeguards such as dataset provenance disclosure, ethical use guidelines, and community oversight should be implemented. Openly documenting the sources and filtering criteria

ensures clarity about the data used, while ethical use guidelines can provide clear boundaries for the responsible use of the dataset and methodology. Encouraging the research community to establish and enforce standards for domain-specific LLMs can help prevent misuse.

- **Bias and Representation.** While we have curated a dataset with a focus on educational value and scientific rigor, the model could inadvertently propagate biases present in the source data. Historical datasets may reflect outdated or unbalanced perspectives, such as overrepresenting contributions from certain geographic regions or underrepresenting emerging subfields within astronomy. These biases can perpetuate systemic inequities if not carefully addressed.

- **Bias Mitigation Strategies.** Post-hoc audits can analyze representation across subfields, geographic regions, and demographics of authorship. Iterative refinement, through periodic dataset updates and expanding coverage of underrepresented areas, can further reduce bias. Engaging a diverse group of domain experts to guide future dataset expansions ensures inclusive curation processes.

- **Representation and Inclusivity.** The curated dataset may inadvertently exclude contributions from underrepresented groups or regions, thereby limiting the model's inclusivity. Ensuring diverse representation in the data sources is crucial for developing models that reflect a wide range of perspectives and knowledge bases. Failure to address these disparities can perpetuate existing inequities within the scientific community.

- **Transparency and Accountability.** While documenting dataset provenance and filtering criteria promotes transparency, ensuring accountability in the development and deployment of domain-specific models requires ongoing efforts. Establishing clear ethical guidelines and engaging in community oversight are essential steps toward responsible AI development.

FineWeb-Edu, our baseline dataset, explicitly addresses the removal of personally identifying and offensive content, as well as trying to address the mentioned issues above. By proactively addressing these ethical considerations, we aim to promote responsible development and deployment of domain-specific language models that support equitable and transparent scientific advancement.

References

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, and et al. 2024. Llemma: An Open Language Model For Mathematics. *arXiv preprint*. ArXiv:2310.10631.

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.

Daoyuan Chen, Yilun Huang, Zhijian Ma, and et al. 2023. Data-Juicer: A One-Stop Data Processing System for Large Language Models. *arXiv preprint*. ArXiv:2309.02033.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and et al. 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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and et al. The Llama 3 Herd of Models. *arXiv preprint*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, and et al. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, and et al. 2021. Measuring Massive Multitask Language Understanding. *arXiv preprint*. ArXiv:2009.03300.

Jared Kaplan, Sam McCandlish, Tom Henighan, and et al. 2020. Scaling Laws for Neural Language Models. *arXiv preprint*. ArXiv:2001.08361.

Julia Kreutzer, Isaac Caswell, Lisa Wang, and et al. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Hongyu Li, Liang Ding, Meng Fang, and et al. 2024. Revisiting Catastrophic Forgetting in Large Language Model Tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308, Miami, Florida, USA. Association for Computational Linguistics.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciucă, and et al. AstroLLaMA: Towards Specialized Foundation Models in Astronomy. *arXiv preprint*.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, et al. 2024. GPT-4o System Card. *arXiv preprint*. ArXiv:2410.21276.

Rui Pan, Tuan Dung Nguyen, Hardik Arora, and et al. 2024. AstroMLab 2: AstroLLaMA-2-70B Model and Benchmarking Specialised LLMs for Astronomy. *arXiv preprint*. ArXiv:2409.19750.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, and et al. 2024. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *arXiv preprint*. ArXiv:2406.17557.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, and et al. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint*. ArXiv:1910.10683.

Karan Singhal, Shekoofeh Azizi, Tao Tu, and et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Yuan-Sen Ting, Tuan Dung Nguyen, Tirthankar Ghosal, and et al. 2024. AstroMLab 1: Who Wins Astronomy Jeopardy!? *arXiv preprint*. ArXiv:2407.11194.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, and et al. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

An Yang, Baosong Yang, Binyuan Hui, and et al. 2024a. Qwen2 Technical Report. *arXiv preprint*. ArXiv:2407.10671.

Haoran Yang, Yumeng Zhang, Jiaqi Xu, and et al. 2024b. Unveiling the Generalization Power of Fine-Tuned Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 884–899, Mexico City, Mexico. Association for Computational Linguistics.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-Source Financial Large Language Models. *arXiv preprint*. ArXiv:2306.06031.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, and et al. 2024. Self-Rewarding Language Models. *arXiv preprint*. ArXiv:2401.10020.

A Appendix

A.1 Evaluation Prompt for Educational Value of Astronomy Texts

The following prompt is utilized to assess the educational value of astronomy-related texts. This scoring system assigns a score from 0 to 5 based on the depth, clarity, and relevance of the content. The prompt guides evaluators in determining the quality of information to ensure only high-value educational material is selected for domain-specific training.

```
prompt = f"""Please evaluate the
    educational value of the following
    astronomy-related text from a web
    document. Use this 6-point scoring
    system:

0 points: No astronomy content at all.
1 point: Minimal astronomy information,
    or astronomy mixed with non-
    astronomical content.
2 points: Covers basic astronomical
    concepts but lacks depth or
    comprehensive explanation.
3 points: Clear explanation of concepts
    with relevant examples, educational
    for a general audience.
4 points: In-depth knowledge, covers
    advanced concepts or recent
    discoveries, well-structured and
    engaging.
5 points: Exceptionally high educational
     value, expert-level insights,
    connects multiple concepts,
    addresses misconceptions, inspires
    further learning.

Provide a brief justification (up to 100
     words) and conclude with the score
    in the format "Score: X".

Here's the text to evaluate:

{text}"""
```

A.2 Domain-Relevant Astronomy Key Terms

The following list comprises astronomy-related terms used to construct the domain-specific "astronomy vector" within the embedding-based filtering process. This selection encompasses key concepts in astrophysics, observational astronomy,

and cosmology, ensuring comprehensive coverage of critical terms relevant to domain-specific filtering.

A.3 Training Details

The training of the Orbit-LLaMA model was conducted using the `DeepSpeed` framework, leveraging `Zero-2` optimization for efficient memory management and scaling. `FlashAttention v2` was employed to enhance the efficiency of the self-attention mechanism, improving both memory usage and computational speed.

Training Configuration:

- **Epochs**: 1
- **Block Size**: 512 tokens
- **Effective Batch Size**: 8
- **Learning Rate**: 2×10^{-5}
- **Learning Rate Schedule**: Linear warmup over 500 steps followed by cosine decay
- **Optimizer**: AdamW with parameters $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay of 0.01
- **Gradient Clipping**: 1.0
- **Precision**: Mixed precision training enabled with `bf16` to reduce memory usage and accelerate training

Optimization Techniques:

- **DeepSpeed Zero-2 Optimization**: Reduced memory footprint by partitioning optimizer states, gradients, and parameters across GPUs, enabling effective training of large models.
- **FlashAttention v2**: Minimized memory usage during self-attention computations, allowing for faster training without compromising accuracy.

A.4 Qualitative Evaluation Methodology

A.4.1 Test Questions and Development Process

A set of 24 test questions was developed by three Ph.D.-track astronomy graduate students and a faculty member from an anonymized university. These questions were designed to evaluate the models' capabilities across a broad range of topics, including:

- **Basic Definitions and Conceptual Knowledge**: For example, defining astronomical terms.
- **Problem-Solving in Complex or Ambiguous Scenarios**: For instance, addressing under-explored areas of astronomy.
- **Support for Research-Oriented Tasks**: Such as code generation for data analysis or simulations.

Each question was carefully reviewed to ensure it was appropriate for benchmarking a wide range of tasks and model competencies.

A.4.2 Evaluation Framework

The responses from Orbit, AstroLLaMA, and Meta-LLaMA were evaluated using the following criteria:

- **Accuracy of Content**: How well the response aligned with established scientific understanding.
- **Clarity of Explanation**: The readability and coherence of the generated response.
- **Reasoning**: The logical structure and handling of complex, ambiguous topics.

The evaluation process consisted of two phases:

1. **Preference Ratings**: Four graduate students were asked to select their preferred response for each question, considering the aforementioned criteria. This approach allowed reviewers to account for cases where correctness was unclear or when responses were partially correct.
2. **Detailed Feedback**: Two astronomy graduate students with Teaching Assistant grading experience analyzed the responses in-depth, identifying strengths, weaknesses, and idiosyncrasies of each model.

A.4.3 Model-Specific Observations

Detailed grader comments highlighted the following characteristics:

- **Meta LLaMA-3**: Generated verbose and repetitive answers, occasionally diverging from the original prompt. It was prone to creating irrelevant subtopics and exhibited issues with coherence.

- **Orbit LLaMA**: Produced concise, relevant responses that were easier to follow. It occasionally adopted a teaching-style format, resembling lecture notes or student-generated answers.

- **AstroLLaMA**: Generated longer, more research-like responses. While it included citations, these were not always accurate or relevant. Responses tended to drift off-topic and lacked logical progression.

These qualitative observations, combined with the preference ratings, helped identify Orbit LLaMA as the most effective overall. Full responses from all models for each test question, along with grader comments, are provided in the supplementary material.

A.5 Model Generations Comparison

The following section presents a comparison of responses from three models—Meta LLaMA-3, Orbit LLaMA, and AstroLLaMA—to a sample question. Each response is evaluated based on accuracy, clarity, and reasoning.

A.5.1 Sample Question

> What challenges do spacecraft face when crossing the heliopause, and how do these challenges impact the design of interstellar missions?

A.5.2 Responses and Evaluations

- **Meta LLaMA-3 Response:**

  > Crossing the heliopause poses several challenges for spacecraft. Firstly, the heliopause region is a highly dynamic environment, with intense solar wind and interstellar magnetic fields. These can cause significant disturbances in the spacecraft's trajectory and navigation, making it difficult to accurately predict the spacecraft's position and velocity. Additionally, the solar wind can cause ionization and plasma flow, which can affect the spacecraft's communication and power systems. The interstellar magnetic field can also induce currents in the spacecraft's structure, which can lead to heating and potential damage. To address these challenges, interstellar missions typically use advanced navigation and control systems, as well as sophisticated shielding and power systems. The shielding is used to protect the spacecraft from ionization and plasma flow, while the power systems are used to generate and distribute electricity to the spacecraft's systems and instruments. Overall, crossing the heliopause is a complex task that requires careful planning and design to ensure the safety and success of the spacecraft.

  Evaluation: *Score: 3.* Meta LLaMA-3 provides a reasonable overview of the challenges spacecraft face when crossing the heliopause, mentioning solar wind disturbances, ionization, and the need for advanced systems to mitigate these challenges. However, the response could have provided more specificity about the unique nature of the heliopause and its implications for spacecraft design. There was an emphasis on general navigational and shielding challenges that, while relevant, did not delve deeply into the unique aspects of the heliopause.

- **Orbit LLaMA Response:**

  > The heliopause, the boundary between the solar wind and interstellar space, presents unique challenges for spacecraft crossing it. One of the primary challenges is navigation. Without the solar wind to provide a reference point, spacecraft must rely on other methods such as star-tracking or pulsar navigation to determine their position and trajectory. This requires complex navigation systems that can operate in the absence of the familiar solar wind. Another challenge is the radiation environment. The heliopause is surrounded by a region of intense radiation, including cosmic rays and solar wind particles. This radiation can damage spacecraft electronics and pose a threat to the health of the crew. To mitigate these risks, spacecraft must

be equipped with advanced radiation shielding and backup systems. Additionally, the distance and communication lag between Earth and the spacecraft make real-time control of the mission impossible. This requires spacecraft to be fully autonomous, capable of making complex decisions and adjustments on their own. Overall, these challenges highlight the need for advanced technology and innovative solutions to enable successful interstellar missions.

Evaluation: *Score: 4.* Orbit LLaMA provides a concise and relevant description of the challenges at the heliopause, such as navigation difficulties due to the lack of solar wind and the increased radiation environment. It clearly explains the need for radiation shielding, autonomy, and advanced navigation systems, directly correlating these challenges with the design implications for interstellar missions. This response was clear, accurate, and focused on the question.

- **AstroLLaMA Response:**

  Crossing the heliopause presents several challenges for interstellar missions, including the interaction of the spacecraft with the solar wind and its magnetic field, as well as the presence of interstellar material such as interstellar dust and the interstellar magnetic field. The heliopause acts as a barrier, preventing the spacecraft from directly interacting with the interstellar medium. However, a boundary layer where interactions between the solar wind and the interstellar medium occur is expected to form at the heliopause. The spacecraft may encounter several layers of plasma and magnetic fields across the boundary layer, which can affect its trajectory and instruments. The spacecraft may also be exposed to interstellar dust, which can be a potential risk to the spacecraft and instruments. The spacecraft designers must consider these challenges and design the spacecraft accordingly, including protective measures and additional instruments to study the interstellar medium.

  Evaluation: *Score: 3.* AstroLLaMA offers an extensive explanation of the heliopause and discusses various physical phenomena associated with it. However, while it provides a broad scientific context, it is overly detailed in parts not directly pertinent to the question, which could overwhelm or confuse readers seeking a direct answer. The response lacks specific mention of direct spacecraft design implications, which are crucial for answering the question on interstellar mission design.

B Supplementary Figures and Tables

B.1 Residual Components Distribution

The distribution of residual components for the domain-specific embeddings ($m = 100$) is visualized below. The residuals exhibit a normal distribution centered near zero, validating that noise diminishes with an increasing number of domain-relevant terms. This result supports the robustness of our astronomy vector in representing domain relevance while minimizing noise.

B.2 Sample Qualitative Evaluation

B.3 Instructions to Reviewers and Annotator Details

To evaluate the quality of the model outputs, we recruited four graduate students in astronomy who volunteered to participate in the evaluation process. The primary goal was to compare responses generated by three models—Orbit LLaMA, AstroLLaMA, and Meta LLaMA-3—on a set of astronomy-related questions, focusing on accuracy, clarity, and reasoning.

B.3.1 Instructions to Annotators

Annotators were provided with a detailed set of instructions that outlined the evaluation criteria and process. They were asked to:

1. **Read and Assess**: Carefully review the responses generated by the three models for each test question.

2. **Evaluate Against Criteria**:

- **Accuracy**: Determine if the content of the response is factually correct and relevant to the question.
- **Clarity**: Assess whether the response is well-structured, easy to read, and free of ambiguity.
- **Reasoning**: Evaluate the logical structure and whether the response adequately addresses complex or ambiguous aspects of the question.

3. **Rank the Responses**: Rank the responses from the three models in order of preference for each question, with justifications for their rankings.

4. **Provide Feedback**: Offer detailed comments on strengths and weaknesses for each response, particularly highlighting areas where models excelled or failed.

B.3.2 Recruitment and Voluntary Participation

All annotators were graduate students specializing in astronomy. They volunteered for the task, and their contributions were limited to evaluating a manageable set of 24 test questions, ensuring the workload was minimal. No monetary or material compensation was provided as they were willing participants.

B.3.3 Consent

All participants were informed about the purpose of the evaluation and the intended use of their feedback for research and publication. Consent was obtained prior to their participation, ensuring ethical compliance.

B.3.4 Anonymity

To protect the privacy of the annotators, we do not report individual demographic or geographic characteristics. This decision aligns with their preference for anonymity while contributing to the research.

This structured process ensured high-quality, expert feedback for evaluating model performance while adhering to ethical standards and respecting the privacy of participants.

B.3.5 Sample Question

> What challenges do spacecraft face when crossing the heliopause, and how do these challenges impact the design of interstellar missions?

B.3.6 Responses and Evaluations

- **Meta LLaMA-3 Response:**

> Crossing the heliopause poses several challenges for spacecraft. Firstly, the heliopause region is a highly dynamic environment, with intense solar wind and interstellar magnetic fields. These can cause significant disturbances in the spacecraft's trajectory and navigation, making it difficult to accurately predict the spacecraft's position and velocity. Additionally, the solar wind can cause ionization and plasma flow, which can affect the spacecraft's communication and power systems. The interstellar magnetic field can also induce currents in the spacecraft's structure, which can lead to heating and potential damage. To address these challenges, interstellar missions typically use advanced navigation and control systems, as well as sophisticated shielding and power systems. The shielding is used to protect the spacecraft from ionization and plasma flow, while the power systems are used to generate and distribute electricity to the spacecraft's systems and instruments. Overall, crossing the heliopause is a complex task that requires careful planning and design to ensure the safety and success of the spacecraft.

Evaluation: *Score: 3.* Meta LLaMA-3 provides a reasonable overview of the challenges spacecraft face when crossing the heliopause, mentioning solar wind disturbances, ionization, and the need for advanced systems to mitigate these challenges. However, the response could have provided more specificity about the unique nature of the heliopause and its implications for spacecraft design. There was an emphasis on general navigational and shielding challenges that, while relevant, did not delve deeply into the unique aspects of the heliopause.

- **Orbit LLaMA Response:**

The heliopause, the boundary between the solar wind and interstellar space, presents unique challenges for spacecraft crossing it. One of the primary challenges is navigation. Without the solar wind to provide a reference point, spacecraft must rely on other methods such as star-tracking or pulsar navigation to determine their position and trajectory. This requires complex navigation systems that can operate in the absence of the familiar solar wind. Another challenge is the radiation environment. The heliopause is surrounded by a region of intense radiation, including cosmic rays and solar wind particles. This radiation can damage spacecraft electronics and pose a threat to the health of the crew. To mitigate these risks, spacecraft must be equipped with advanced radiation shielding and backup systems. Additionally, the distance and communication lag between Earth and the spacecraft make real-time control of the mission impossible. This requires spacecraft to be fully autonomous, capable of making complex decisions and adjustments on their own. Overall, these challenges highlight the need for advanced technology and innovative solutions to enable successful interstellar missions.

Evaluation: *Score: 4.* Orbit LLaMA provides a concise and relevant description of the challenges at the heliopause, such as navigation difficulties due to the lack of solar wind and the increased radiation environment. It clearly explains the need for radiation shielding, autonomy, and advanced navigation systems, directly correlating these challenges with the design implications for interstellar missions. This response was clear, accurate, and focused on the question.

- **AstroLLaMA Response:**

  Crossing the heliopause presents several challenges for interstellar missions, including the interaction of the spacecraft with the solar wind and its magnetic field, as well as the presence of interstellar material such as interstellar dust and the interstellar magnetic field. The heliopause acts as a barrier, preventing the spacecraft from directly interacting with the interstellar medium. However, a boundary layer where interactions between the solar wind and the interstellar medium occur is expected to form at the heliopause. The spacecraft may encounter several layers of plasma and magnetic fields across the boundary layer, which can affect its trajectory and instruments. The spacecraft may also be exposed to interstellar dust, which can be a potential risk to the spacecraft and instruments. The spacecraft designers must consider these challenges and design the spacecraft accordingly, including protective measures and additional instruments to study the interstellar medium.

  Evaluation: *Score: 3.* AstroLLaMA offers an extensive explanation of the heliopause and discusses various physical phenomena associated with it. However, while it provides a broad scientific context, it is overly detailed in parts not directly pertinent to the question, which could overwhelm or confuse readers seeking a direct answer. The response lacks specific mention of direct spacecraft design implications, which are crucial for answering the question on interstellar mission design.

C Additional Mathematical Foundations

C.1 Efficient Domain-Specific Dataset Curation

The core objective of language models is to estimate the probability distribution over sequences of words by predicting each token based on preceding ones. This is achieved through training on large datasets, where the model minimizes the negative log-likelihood (cross-entropy loss) across the corpus. Model performance tends to improve predictably with the number of parameters, as greater capacity enables capturing more complex

patterns—up to a limit governed by dataset quality and complexity (Kaplan et al., 2020).

Domain-specific models, such as the astronomy-focused variant presented here, face unique challenges in obtaining sufficient, high-quality data, as general-purpose datasets often include noise or irrelevant content. A refined dataset requires filtering methods that prioritize domain relevance without extensive computational costs.

To address this, we developed a method that leverages cosine similarity between token embeddings and a representative aggregated word embedding derived from a predefined list of astronomy-related terms. This approach enables efficient filtering by identifying documents based on their semantic similarity to the target domain.

C.1.1 Decomposition of Embeddings

We assume that each astronomy-related term's embedding can be decomposed into two components:

$$\mathbf{e}_{t_i} = \mathbf{a} + \mathbf{r}_i, \tag{1}$$

where:

- $\mathbf{e}_{t_i} \in R^d$ is the normalized embedding vector of the i-th astronomy-related term.
- $\mathbf{a} \in R^d$ is the domain-specific astronomy component common to all astronomy-related terms.
- $\mathbf{r}_i \in R^d$ is the random noise component unique to each term, with $E[\mathbf{r}_i] = \mathbf{0}$.

The astronomy aggregated embedding vector $\mathbf{A}$ is defined as the average of the embeddings of all astronomy-related terms:

$$\mathbf{A} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{e}_{t_i} = \mathbf{a} + \frac{1}{m}\sum_{i=1}^{m} \mathbf{r}_i. \tag{2}$$

By the Law of Large Numbers, as the number of astronomy-related terms m increases, the average of the random components converges to zero:

$$\lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} \mathbf{r}_i = \mathbf{0}. \tag{3}$$

Therefore, for sufficiently large m, the astronomy aggregated vector $\mathbf{A}$ approximates the domain-specific component $\mathbf{a}$:

$$\mathbf{A} \approx \mathbf{a}. \tag{4}$$

The mean vector $\mathbf{A}$ also serves as the mathematical minimum point for minimizing the sum of squared Euclidean distances between $\mathbf{A}$ and each individual astronomy-related embedding $\mathbf{e}_{t_i}$. Formally, $\mathbf{A}$ minimizes the following objective:

$$\mathbf{A} = \arg\min_{\mathbf{x}\in R^d} \sum_{i=1}^{m} \|\mathbf{e}_{t_i} - \mathbf{x}\|^2. \tag{5}$$

This property ensures that $\mathbf{A}$ is the most representative point in the embedding space for the set of astronomy-related terms.

C.1.2 Error Analysis

The error introduced by the random components $\mathbf{r}_i$ can be quantified by analyzing the difference between the astronomy aggregated vector $\mathbf{A}$ and the true domain-specific component $\mathbf{a}$:

$$\mathbf{E} = \mathbf{A} - \mathbf{a} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{r}_i.$$

We aim to analyze the expected and actual error rates to ensure that our theoretical results are sound. Specifically, we verify that the random vectors $\mathbf{r}_i \in R^d$ are independently and identically distributed (i.i.d.) with mean zero.

To validate the properties of the residual components $\mathbf{r}_i$, we conducted experiments using the GloVe word embeddings (Pennington et al., 2014). We assessed whether the residual components for a significant number of astronomy-related terms have an expected value around zero and how the error $\|\mathbf{E}\|$ behaves as a function of m.

We selected a list of 101 astronomy-related terms (see Section A.2) and extracted their corresponding embeddings from the pre-trained GloVe model. We computed the astronomy aggregated vector $\mathbf{A}$ as the average of m randomly selected term embeddings and calculated the error vector $\mathbf{E} = \mathbf{A} - \mathbf{a}$, where $\mathbf{a}$ is the true average embedding computed using all available astronomy-related terms.

C.1.3 Computational Efficiency

To filter a large corpus efficiently, we employ a combination of embedding-based similarity filtering and BERT-based classification. The pipeline's goal is to reduce the dataset to the most relevant documents before applying more computationally expensive processes.

Given a document D with s tokens $\{w_1, w_2, \ldots, w_s\}$, each token has a corresponding normalized embedding $\mathbf{e}_{w_j} \in R^d$.

The document vector $\mathbf{B}$ is the average of these embeddings:

$$\mathbf{B} = \frac{\sum_{j=1}^{s} \mathbf{e}_{w_j}}{m}.$$

The relevance to the astronomy domain is assessed using cosine similarity between $\mathbf{B}$ and the astronomy vector $\mathbf{A}$:

$$\text{Similarity}(D) = \frac{\mathbf{A} \cdot \mathbf{B}}{|A| * |B|}.$$

A document is retained if this similarity exceeds a threshold τ.

C.1.4 Formalized Pipeline

1. **Embedding Lookup**: For each token w_j in document D, retrieve its embedding $\mathbf{e}_{w_j}$ from a hashmap. **Runtime**: $O(1)$
2. **Document Vector Computation**: Calculate $\mathbf{B} = \sum_{j=1}^{s} \mathbf{e}_{w_j}$. **Runtime**: $O(s \cdot d)$
3. **Similarity Calculation**: Compute cosine similarity between $\mathbf{A}$ and $\mathbf{B}$. **Runtime**: $O(d)$
4. **Thresholding**: Retain the document if the similarity exceeds τ. **Runtime**: $O(1)$

Total Complexity per Document: $O(s \cdot d)$

Given N documents, each with s tokens on average, the overall complexity for the filtering step is:

$$O(N \cdot s \cdot d)$$

Optimizations Implemented:

- **Precomputation of Normalized $\mathbf{A}$**: Eliminates repeated division during similarity computation.
- **Vectorized Operations**: Speeds up vector calculations using optimized libraries.
- **Parallel Processing**: Distributes the workload across multiple cores.

D Mathematical Foundations

D.1 Efficient Domain-Specific Dataset Curation

The core objective of language models is to estimate the probability distribution over sequences of words by predicting each token based on preceding ones. This is achieved through training on large datasets, where the model minimizes the negative log-likelihood (cross-entropy loss) across the corpus. Model performance tends to improve predictably with the number of parameters, as greater capacity enables capturing more complex patterns—up to a limit governed by dataset quality and complexity (Kaplan et al., 2020).

Domain-specific models, such as the astronomy-focused variant presented here, face unique challenges in obtaining sufficient, high-quality data, as general-purpose datasets often include noise or irrelevant content. A refined dataset requires filtering methods that prioritize domain relevance without extensive computational costs.

To address this, we developed a method that leverages cosine similarity between token embeddings and a representative aggregated word embedding derived from a predefined list of astronomy-related terms. This approach enables efficient filtering by identifying documents based on their semantic similarity to the target domain.

D.1.1 Decomposition of Embeddings

We assume that each astronomy-related term's embedding can be decomposed into two components:

$$\mathbf{e}_{t_i} = \mathbf{a} + \mathbf{r}_i, \tag{6}$$

where:

- $\mathbf{e}_{t_i} \in R^d$ is the normalized embedding vector of the i-th astronomy-related term.
- $\mathbf{a} \in R^d$ is the domain-specific astronomy component common to all astronomy-related terms.
- $\mathbf{r}_i \in R^d$ is the random noise component unique to each term, with $E[\mathbf{r}_i] = \mathbf{0}$.

The astronomy aggregated embedding vector $\mathbf{A}$ is defined as the average of the embeddings of all astronomy-related terms:

$$\mathbf{A} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{e}_{t_i} = \mathbf{a} + \frac{1}{m} \sum_{i=1}^{m} \mathbf{r}_i. \tag{7}$$

By the Law of Large Numbers, as the number of astronomy-related terms m increases, the average of the random components converges to zero:

$$\lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \mathbf{r}_i = \mathbf{0}. \tag{8}$$

Therefore, for sufficiently large m, the astronomy aggregated vector $\mathbf{A}$ approximates the domain-specific component $\mathbf{a}$:

$$\mathbf{A} \approx \mathbf{a}. \tag{9}$$

The mean vector $\mathbf{A}$ also serves as the mathematical minimum point for minimizing the sum of squared Euclidean distances between $\mathbf{A}$ and each individual astronomy-related embedding $\mathbf{e}_{t_i}$. Formally, $\mathbf{A}$ minimizes the following objective:

$$\mathbf{A} = \arg\min_{\mathbf{x} \in R^d} \sum_{i=1}^{m} \|\mathbf{e}_{t_i} - \mathbf{x}\|^2. \tag{10}$$

This property ensures that $\mathbf{A}$ is the most representative point in the embedding space for the set of astronomy-related terms.

D.1.2 Error Analysis

The error introduced by the random components $\mathbf{r}_i$ can be quantified by analyzing the difference between the astronomy aggregated vector $\mathbf{A}$ and the true domain-specific component $\mathbf{a}$:

$$\mathbf{E} = \mathbf{A} - \mathbf{a} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{r}_i.$$

We aim to analyze the expected and actual error rates to ensure that our theoretical results are sound. Specifically, we verify that the random vectors $\mathbf{r}_i \in R^d$ are independently and identically distributed (i.i.d.) with mean zero.

To validate the properties of the residual components $\mathbf{r}_i$, we conducted experiments using the GloVe word embeddings (Pennington et al., 2014). We assessed whether the residual components for a significant number of astronomy-related terms have an expected value around zero and how the error $\|\mathbf{E}\|$ behaves as a function of m.

We selected a list of 101 astronomy-related terms (see Section A.2) and extracted their corresponding embeddings from the pre-trained GloVe model. We computed the astronomy aggregated vector $\mathbf{A}$ as the average of m randomly selected term embeddings and calculated the error vector $\mathbf{E} = \mathbf{A} - \mathbf{a}$, where $\mathbf{a}$ is the true average embedding computed using all available astronomy-related terms.

D.1.3 Computational Efficiency

To filter a large corpus efficiently, we employ a combination of embedding-based similarity filtering and BERT-based classification. The pipeline's goal is to reduce the dataset to the most relevant documents before applying more computationally expensive processes.

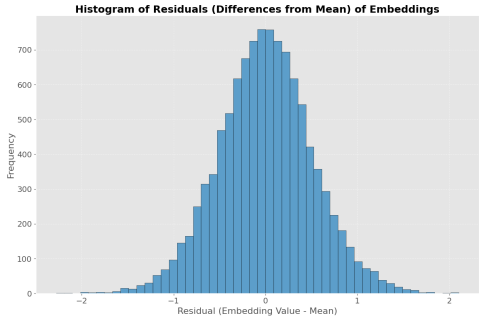

Figure 5: Distribution of residual components for the domain-specific embeddings ($m = 100$). The residuals exhibit a normal distribution centered near zero, validating that noise diminishes with an increasing number of domain-relevant terms. This result supports the robustness of our astronomy vector in representing domain relevance while minimizing noise.

Given a document D with s tokens $\{w_1, w_2, \ldots, w_s\}$, each token has a corresponding normalized embedding $\mathbf{e}_{w_j} \in R^d$. The document vector $\mathbf{B}$ is the average of these embeddings:

$$\mathbf{B} = \frac{\sum_{j=1}^{s} \mathbf{e}_{w_j}}{m}.$$

The relevance to the astronomy domain is assessed using cosine similarity between $\mathbf{B}$ and the astronomy vector $\mathbf{A}$:

$$\text{Similarity}(D) = \frac{\mathbf{A} \cdot \mathbf{B}}{|A| \cdot |B|}.$$

A document is retained if this similarity exceeds a threshold τ.

D.1.4 Formalized Pipeline

1. **Embedding Lookup**: For each token w_j in document D, retrieve its embedding $\mathbf{e}_{w_j}$ from a hashmap. **Runtime**: $O(1)$

2. **Document Vector Computation**: Calculate $\mathbf{B} = \sum_{j=1}^{s} \mathbf{e}_{w_j}$. **Runtime**: $O(s \cdot d)$

3. **Similarity Calculation**: Compute cosine similarity between $\mathbf{A}$ and $\mathbf{B}$. **Runtime**: $O(d)$

4. **Thresholding**: Retain the document if the similarity exceeds τ. **Runtime**: $O(1)$

Total Complexity per Document: $O(s \cdot d)$

Given N documents, each with s tokens on average, the overall complexity for the filtering step is:

$$O(N \cdot s \cdot d)$$

Optimizations Implemented:

- **Precomputation of Normalized A**: Eliminates repeated division during similarity computation.
- **Vectorized Operations**: Speeds up vector calculations using optimized libraries.
- **Parallel Processing**: Distributes the workload across multiple cores.

Terms	**Terms**	**Terms**
Albedo	Aphelion	Apogee
Asteroid	Astronomy	Aurora
Axion	Azimuth	Barycenter
Baryon	Blackbody	Bolide
Brilliance	Cepheid	Comet
Constellation	Corona	Cosmic
Cosmology	DESC	Dyne
Eclipse	Ecliptic	Emission
Erg	Exoplanet	Extinction
Fluence	Frequency	Galaxy
Geocentric	Gibbous	Gravity
Heliocentric	Interferometry	Isotropic
JWST	kpc	Light-Year
LSST	Luminosity	Magnetar
Magnetosphere	Metallicity	Meteor
Meteorite	Microlensing	Moon
Morphology	Multiverse	Nebula
Neutrino	Noctilucent	Nova
Nucleosynthesis	Orbit	Parallax
Parsec	Perihelion	Phase
Photometry	Photosphere	Planck
Planetesimal	Pulsar	Quasar
Quiescence	Recombination	Reddening
Redshift	Reionization	Satellite
Seyfert	Simulation	Singularity
Spectroscopy	SPT	Sublimation
Sunspot	Supercomputer	Supermassive
Supernova	Telescope	Transit
Universe	Voids	Wavelength
Waxing	Wormhole	X-ray
Zenith	Zodiac	Optical
Infrared	Ultraviolet	Microwave
Proton	Neutron	Electron
Flux	Intensity	Companion
Outflow	QSO	Pulse
Progenitor		

Table 5: Domain-Relevant Astronomy Key Terms

Terms	Terms	Terms
Anatomy	Pathology	Physiology
Oncology	Cardiology	Neurology
Radiology	Pharmacology	Surgery
Pediatrics	Dermatology	Gastroenterology
Endocrinology	Hematology	Immunology
Nephrology	Pulmonology	Psychiatry
Rheumatology	Urology	Obstetrics
Gynecology	Orthopedics	Ophthalmology
Otolaryngology	Infectious	Microbiology
Epidemiology	Toxicology	Genetics
Biochemistry	Histology	Embryology
Virology	Bacteriology	Parasitology
Cytology	Prognosis	Diagnosis
Treatment	Therapy	Vaccination
Antibiotic	Antiviral	Pathogen
Tumor	Cancer	Leukemia
Diabetes	Hypertension	Cardiomyopathy
Stroke	Sepsis	Inflammation
Autoimmune	Fibrosis	Circulation
Respiration	Homeostasis	Anesthesia
Trauma	Fracture	Hemorrhage
Venous	Arterial	Renal
Hepatic	Liver	Kidney
Lung	Heart	Brain
Spinal	Nerve	Bone
Muscle	Skin	Blood
Plasma	Lymph	Hormone
Enzyme	Protein	Gene
DNA	RNA	Chromosome
Cell	Tissue	Organ
Organism	Metabolism	Nutrition
Obesity	Malnutrition	Infection
Immunity	Allergy	Vaccine
Mutation	Carcinogen	Biopsy
MRI	CT	X-ray
Ultrasound	PET	Radiotherapy
Chemotherapy	Surgical	Endoscopy
Laparoscopy	Thermography	Pharmacokinetics
Pharmacodynamics	Clinical	Hospital
Ambulance	ICU	Ward
Therapist	Psychologist	Psychiatrist
Physician	Surgeon	Nurse
Paramedic	Dentist	Optometrist
Audiologist	Dietitian	Nutritionist
Emergency	CPR	Defibrillator