AstroMLab 3: Achieving GPT-4o Level Performance in Astronomy with a Specialized 8B-Parameter Large Language Model

Tijmen de Haan[1], Yuan-Sen Ting (丁源森)[2,3], Tirthankar Ghosal[4], Tuan Dung Nguyen[5], Alberto Accomazzi[6], Azton Wells[7], Nesar Ramachandra[7], Rui Pan (潘瑞)[8], Zechang Sun (孙泽昌)[9]
[1]*International Center for Quantum-field Measurement Systems for Studies of the Universe and Particles (QUP-WPI), High Energy Accelerator Research Organization (KEK), Tsukuba, Ibaraki, Japan*
[2]*Department of Astronomy, The Ohio State University, Columbus, OH, USA*
[3]*Center for Cosmology and AstroParticle Physics (CCAPP), The Ohio State University, Columbus, OH, USA*
[4]*National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA*
[5]*Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA*
[6]*Center for Astrophysics, Harvard & Smithsonian, Cambridge, MA, USA*
[7]*Computational Science Division, Argonne National Laboratory, Lemont, IL, USA*
[8]*Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong and*
[9]*Department of Astronomy, Tsinghua University, Beijing, People's Republic of China*

AstroSage-Llama-3.1-8B is a domain-specialized natural-language AI assistant tailored for research in astronomy, astrophysics, and cosmology. Trained on the complete collection of astronomy-related arXiv papers from 2007-2024 along with millions of synthetically-generated question-answer pairs and other astronomical literature, AstroSage-Llama-3.1-8B demonstrates remarkable proficiency on a wide range of questions. AstroSage-Llama-3.1-8B scores 80.9% on the AstroMLab-1 benchmark, greatly outperforming all models—proprietary and open-weight—in the 8-billion parameter class, and performing on par with GPT-4o. This achievement demonstrates the potential of domain specialization in AI, suggesting that focused training can yield capabilities exceeding those of much larger, general-purpose models. AstroSage-Llama-3.1-8B is freely available, enabling widespread access to advanced AI capabilities for astronomical education and research.

Keywords: AI assistant, large-language model, continued pretraining, supervised fine-tuning

I. INTRODUCTION

Large-language model (LLM) assistants are rapidly gaining traction across all sectors of knowledge work worldwide. In astronomy, these models are used for providing factual information, as programming assistants, for brainstorming ideas, and for providing explanations tailored to the level of understanding or preferred style of the user. LLMs exhibit a remarkable robustness, often delivering useful outputs even when the input is malformed, lacks context, or contains inaccuracies.

Despite their potential, the development of specialized LLMs has been limited due to their recent emergence and the substantial resources required for training. Previous studies [1–4] have shown that models narrowly tailored to a specific domain can perform on par with, or even exceed, much larger general-purpose models. This suggests that a large, highly domain-specific model could achieve state-of-the-art performance.

In astronomy, however, high-performing specialized language models have not yet been achieved. While models like AstroLLaMA [5, 6] have gained attention, they lack comprehensive benchmarking of their astronomical knowledge recall capabilities. Recent studies [7] have shown that many of these models, due to limited specialized training data and fine-tuning for instruction-following, suffer from either catastrophic forgetting or an inability to follow precise question-answering instructions, often performing worse than their baseline models (in this case, the Llama models).

Building on the previous efforts of cosmosage [8] and AstroLLaMA, we have developed AstroSage-Llama-3.1-8B, a natural language assistant specialized in astronomy, astrophysics, cosmology, and astronomical instrumentation. For the remainder of this paper, we will refer to these subdomains collectively as "astronomy". Through the use of a substantially more extensive and well-curated training dataset, we demonstrate for the first time that our specialized language model significantly outperforms baseline models in downstream tasks, particularly in astronomical knowledge recall.

In the long term, we envision an agentic research assistant capable of autonomously conducting literature reviews, identifying relevant hypotheses, carrying out data analysis, and even formulating new research questions. The development of such scientific agents (LLMs capable of solving scientific problems end-to-end) is already a rapidly growing field in astronomy. Recent studies have shown promising results in automating research tasks, such as analyzing James Webb Space Telescope data through multi-agent collaboration and self-play reinforcement learning [9]. However, these studies have been largely constrained by the substantial API costs associated with proprietary models.

Realizing this level of agency will require extensive experimentation and careful optimization. Given the substantial compute costs and data requirements inherent in large-scale model training, keeping the model size manageable while maintaining high performance is crucial. Our approach, demonstrated through astronomy knowledge recall, shows that specialized models can achieve state-of-the-art performance in specific domains. This not only makes the development of advanced research assistants more feasible but also ensures their accessibility to a wider range of institutions and researchers, potentially transforming the landscape of astronomical research and education.

II. CONTINUED PRETRAINING

For AstroSage-Llama-3.1-8B, we selected Meta's Llama-3.1 8-billion parameter model [10] as our foundation model. This base model was chosen for both its strong general-purpose capabilities and its availability under the permissive Llama 3.1 Community License. Furthermore, among models in the 8-billion parameter class, it demonstrated superior performance in astronomical knowledge recall compared to both general-purpose models [11] and specialized astronomical LLMs [7], making it an ideal baseline.

To begin the development process, we first focused on curating, obtaining, and cleaning a continued pretraining (CPT) dataset.

A. Dataset Preparation

The scaling laws of Hoffmann *et al.* [2] show that model capability increases predictably with increased training data volume and computational resources. More recently, it has been found [12–15] that the power-law index of these scaling laws depends on data quality. Therefore, our general approach to assembling a corpus was to focus on maintaining a high quality threshold, maximizing data volume at that quality level.

We employed a multi-faceted approach to create a comprehensive, high-quality, high-variety CPT corpus. The primary components of our dataset include:

- Approximately 250,000 arXiv preprints from 2007-2024 with primary or cross-listed categories in astro-ph (astrophysics) or gr-qc (general relativity and quantum cosmology). We deliberately excluded the Annual Review of Astronomy and Astrophysics papers used in Ting *et al.* [11, AstroMLab-1] to generate the benchmark questions, ensuring our evaluation would test the model's ability to generalize knowledge rather than recall specific source materials.

- Relevant articles from a depth-2 search through Wikipedia's astronomy and astrophysics categories.

- A selection of textbooks that are available as PDFs or ebooks online.

We processed the data into markdown format. For the vast majority of the dataset, the rendered PDF files were converted to markdown using Nougat OCR [16]. For the remaining sources, the data was either already in markdown format, or was left as plain text.

B. Pretraining run

The pretraining was conducted on the ORNL OLCF Frontier 1.6 exaflop supercomputer, leveraging its substantial computational resources. We used 184 nodes simultaneously, each of which is equipped with four AMD MI250X, which in turn have 2 Graphics Compute Dies (GCDs) each for a total of 8 GCDs per node, or a total of 1472 GCDs.

Further statistics and our choices of pretraining hyperparameters are summarized in Table I.

Hyperparameter	**Value**
Sequence length	8,192 tokens
Micro batch size	3
Epochs	2
Learning rate	1.5e-4
Base model	Llama 3.1 8B
Optimizer	AdamW
Adam beta2	0.95
Adam epsilon	1e-5
Learning rate schedule	Constant with quadratic warmup
Max gradient norm	3.0
Weight decay	0.001
Warmup steps	40
Precision	BF16
FSDP	Full shard, auto wrap
Resource	**Value**
Plaintext filesize	19.9 GB
Token count	3.3 billion
Nodes	184
GCDs (effective # of GPUs)	1,472
Training wall time	10 hours
Total time spent	11.5 hours
Effective GPU-hours spent	14,872 hours
VRAM Usage	96% of 64 GB/GCD

TABLE I. Summary of pretraining hyperparameters and resource usage.

The Llama-3.1 tokenizer, which uses a variant of tiktoken with UTF-flexible encoding, was sufficient for our purposes, so we did not introduce any astronomy-specific tokens to the vocabulary. We considered incorporating arXiv identifiers (in the format arXiv:YYMM.numbervV) for papers in the training set but ultimately decided against it due to the substantial increase in vocabulary size this would entail. Future work may explore expanding the vocabulary to include numerical representations relevant to astronomy, such as common quantities and units.

During training, we tracked the loss function and step sizes. Minimal hyperparameter tuning was performed, relying on the hyperparameters from de Haan [8] for all parameters other than the learning rate. The learning rate was extrapolated from smaller runs, but remained problematic, as Frontier requires high levels of parallelization with short wall times. This caused our initial runs to suffer from either insufficient learning due to a low learning rate, or catastrophic exploding gradients due to an excessively high learning rate. The final run described in this work used a tuned learning schedule with a learning rate as high as possible but still allowing convergence. In future efforts, we plan to further optimize the efficiency of the training procedure along the lines of the work presented in Dash *et al.* [17], incorporating tensor, pipeline, and data paralellism through libraries such as Megatron-Deepspeed. We also aim to request a longer walltime with lower parallelization factor in order to be able to reduce loss and improve downstream performance.

C. Dataset Cleaning

We followed the cleaning procedures from de Haan [8], including a perplexity-based cleaning approach. This method first splits the corpus into individual paragraphs and calculates their respective perplexity scores. Perplexity measures how well a language model can predict a given text sequence. Lower perplexity indicates text that follows expected patterns of natural language, while very high perplexity values often signal anomalous or corrupted content. Outliers with such high perplexity frequently stem from OCR errors, malformed text, or non-prose content such as tables.

Based on the distribution of perplexity scores (Figure 1), we established a threshold that excluded the top 2% of paragraphs with the highest perplexity scores. We then reconstructed each document using only the paragraphs below this threshold. This cleaning procedure removed approximately 2% of the total data volume. Figure 1 illustrates this process by showing the distribution of perplexity scores and our chosen threshold.

III. SUPERVISED FINE-TUNING

To improve the model's ability to follow instructions and answer questions effectively, we performed supervised fine-tuning (SFT). In this process, the model was trained to predict appropriate responses to given prompts, learning from a collection of high-quality example conversations. Below, we describe our approach to generating and curating the SFT dataset.

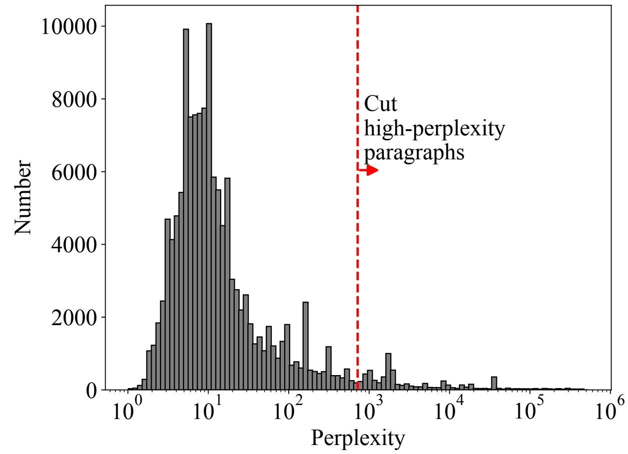

FIG. 1. Histogram of paragraph-level perplexity used for dataset cleaning. The perplexity is calculated as $\exp(-\langle \ln P \rangle)$, where $\langle \ln P \rangle$ is the average log probability per token in each paragraph. The red dashed line indicates the manually chosen threshold used to filter out high-perplexity paragraphs, which comprised approximately 2% of the total data volume. Paragraphs with perplexity above this threshold were removed from the dataset. Note that the highest observed perplexity values (around 10^{12}) extend beyond the right edge of the plot.

A. SFT Dataset

Pan *et al.* [7] identified a critical limitation in the AstroLLaMA series of specialized astronomical LLMs: their inability to outperform even their own starting base models, partly due to inadequate SFT. While these models showed marginal improvements in basic next-token prediction tasks, they performed worse than their baseline models on instructional Q&A tasks, even for straightforward astronomical knowledge recall. This shortcoming fundamentally undermines the purpose of specialized training. Therefore, in our study, we paid particular attention to the SFT process, generating training datasets orders of magnitude larger than previously available in astronomy.

The largest component of our SFT dataset consists of question-and-answer (Q&A) pairs. Using the method from de Haan [8], we generated over 11 million synthetic Q&A pairs from papers in our CPT dataset. These Q&A pairs were then evaluated using an LLM based on four criteria:

1. Correctness: The factual accuracy of the answer in relation to the question, ensuring that each Q&A pair adheres to current scientific understanding and accurately reflects the information presented in the

source material.

2. Stand-alone: The ability of each Q&A pair to be understood in isolation, without needing additional context beyond the content provided. This ensures that an expert in astronomy, astrophysics, or cosmology could answer the question based solely on the information within the pair.

3. Pertinence: The relevance and importance of the question to researchers or students in astronomy, astrophysics, or cosmology. We want the questions to be ones that a professional in the field might ask or find valuable for deeper understanding.

4. Overall Quality

The LLM was presented with several hand-written examples of ratings in the form Question–Answer–Score–Explanation, followed by the Question–Answer pair to be judged. The model was then asked to complete the score values. A small number of resulting scores were verified and confirmed to be sufficiently accurate. Only Q&A pairs with perfect scores in all four categories were kept, resulting in 8.8 million high-quality Q&A pairs.

We also included a filtered version of Infinity-Instruct-7M [18], keeping only entries with at least 70% alphanumeric characters, as well as filtering out entries with certain keywords. The inclusion of this dataset was to ensure that AstroSage-Llama-3.1-8B would gain instruction-following abilities such as multi-turn conversation.

Additionally, we generated synthetic summaries for all of the papers in the CPT dataset. The SFT dataset for these summaries consists of a user prompt asking to summarize a certain paper, with the assistant completion being a small preamble followed by the summary. The user prompt was created through a series of random choices about the way the question is asked, a small preamble, and the way in which the paper is referenced, yielding a high variety of question styles.

Furthermore, we generated a metadata-based dataset, where a again a series of custom rules and random selections result in diverse questions about titles, dates of publication, arXiv IDs, and first author names from the papers in the CPT dataset. This was included in an effort to memorize the paper metadata so that users can reference papers in their conversations with AstroSage-Llama-3.1-8B. However, as we will discuss, our training procedure was relatively shallow and insufficient memorization of this information took place, as spot checks show that the final model can not reliably answer metadata-based questions correctly.

These datasets were combined with five further datasets which were assembled by hand from various sources on the web. The combined dataset comprised approximately 2 billion tokens.

B. SFT procedure

The SFT process was also conducted on the Frontier supercomputer with a configuration summarized in Table II.

Hyperparameter	**Value**
Epochs	6
Learning rate	1e-4
Base model	CPT model
Learning rate schedule	Cosine with quadratic warmup
Weight decay	0.0
Resource	**Value**
Plaintext filesize	9.8 GB
Token count	2.0 billion
Training wall time	9.5 hours
Total time spent	11.5 hours
Effective GPU-hours spent	13,738 hours

TABLE II. Summary of supervised fine-tuning hyperparameters and resource usage. Parameters that are not stated here were kept the same as in Table I.

Throughout the fine-tuning process, we again monitored loss and step sizes, which are shown in Figure 2 (CPT curves look similar). The learning was—like in §II B—limited by the maximum wall-time allowed by the Frontier HPC system, which for 184 nodes or more is a maximum of 12 hours. With the learning rate as high as it could comfortably be set to avoid exploding gradients, this limitation on walltime strongly limited the number of steps that could be taken and therefore the final loss that could be obtained.

IV. MODEL MERGING

Model merging, also known as parameter averaging, has emerged as a powerful technique for combining capabilities of multiple expert models into a single language model [19, 20]. While our CPT+SFT procedure significantly improved the model's astronomical knowledge recall in few-shot prompts, we observed that performance in conversational Q&A scenarios such as multi-turn conversations and instructions regarding the output style still fell slightly short of optimal. This challenge likely stems from the fact that the "instruct" version of Llama-3.1-8b provided by Meta underwent substantially more extensive supervised fine-tuning than what we could achieve as an academic group. We found that merging our specialized model with Meta's instruct model significantly improved these conversational capabilities.

To create the final version of AstroSage-Llama-3.1-8B, we employed MERGEKIT [21], using the DARE-TIES method to combine our SFT-trained model described in §III with Meta-Llama-3.1-8B-Instruct [10]. The merge

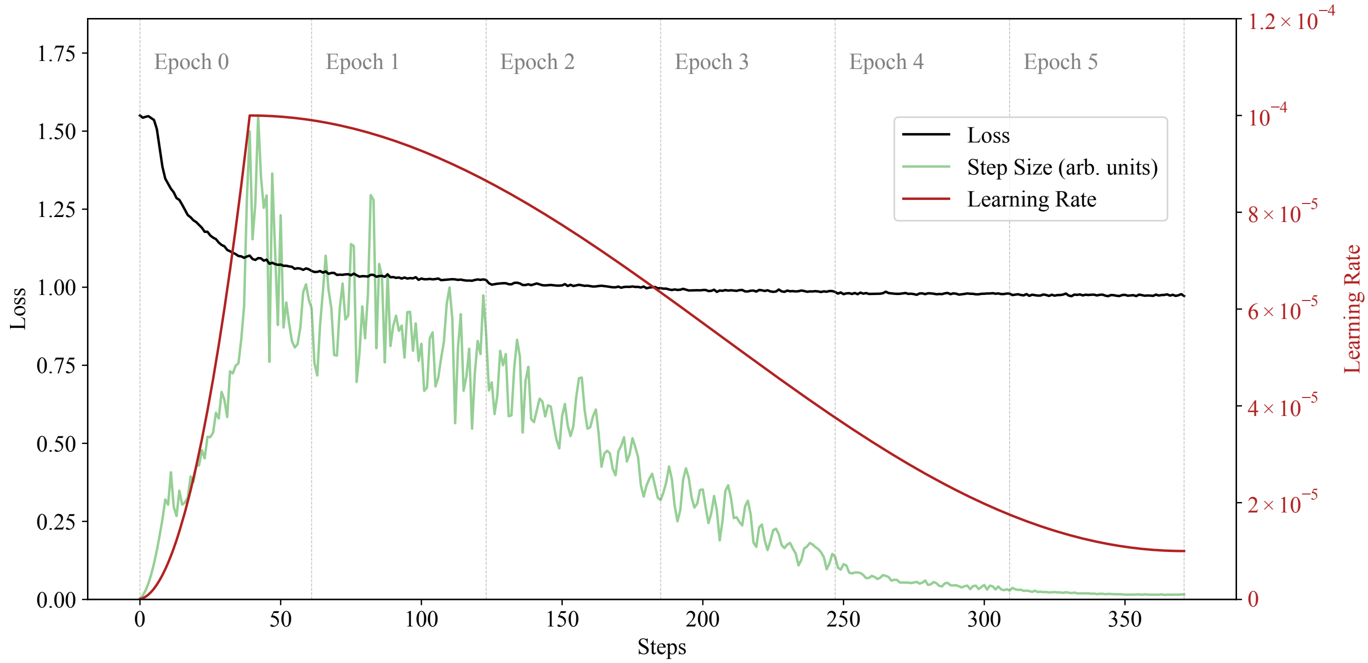

FIG. 2. Supervised fine-tuning loss curve. The learning rate schedule, shown in brick red, follows a quadratic warmup followed by a cosine schedule that ends at 10% of the peak learning rate. The peak learning rate was chosen to prevent exploding gradients. The loss curve in black shows no significant signs of overfitting, as evidenced by minimal discontinuities at epoch boundaries. The green curve represents the L2 norm of the parameter update, shown in arbitrary units.

was performed at full density, bf16 precision, and with the weight parameters set to 0.75 and 0.25 for AstroSage-Llama-3.1-8B-SFT and Meta-Llama-3.1-8B-Instruct, respectively.

The resulting merged model exhibits enhanced instruction-following capabilities and improved performance on the AstroMLab-1 multiple-choice question benchmark in both few-shot and structured output scenarios. To determine whether these improvements stemmed from enhanced instruction-following rather than additional astronomical knowledge, we conducted a control experiment. We fine-tuned a separate version of the CPT+SFT model on unrelated multiple-choice questions using identical output formatting. This control model achieved near-identical scores on the AstroMLab-1 benchmark in the structured output scenario without any merging, suggesting that the process of merging in a small fraction of the Meta-Llama-3.1-8B-Instruct model weights transferred general question-answering capabilities rather than domain-specific knowledge.

To recap, we began with Meta-Llama-3.1-8B as our base model, then performed CPT on a large corpus of astronomy literature to instill domain knowledge. This was followed by SFT using carefully curated instruction-response pairs to improve task performance and instruction following. Finally, we merged the resulting model with Meta-Llama-3.1-8B-Instruct to enhance general instruction-following capabilities while preserving the astronomical expertise, resulting in our final model which we are releasing as AstroSage-Llama-3.1-8B.

V. EVALUATION

To evaluate AstroSage-Llama-3.1-8B's performance, we employed the multiple-choice question benchmark from the first paper in this series [11, AstroMLab 1]. This benchmark consists of diverse astronomy-related questions generated from selected Annual Review of Astronomy and Astrophysics (ARAA) papers and remains, to our knowledge, the only comprehensive astronomy-specific benchmarking effort available. We refer interested readers to the original paper for detailed benchmark specifications.

Importantly, we deliberately excluded the ARAA papers from AstroSage-Llama-3.1-8B's training dataset. This strategic exclusion enables us to evaluate the model's broader understanding of astronomical concepts rather than its ability to recall specific information from the source materials. This approach helps ensure that the benchmark scores reflect AstroSage-Llama-3.1-8B's genuine comprehension of astronomy rather than mere memorization of the content used to create the questions.

Our choice to primarily evaluate AstroSage-Llama-3.1-8B with a knowledge-based benchmark was motivated by two key factors. First, this benchmark represents the only extensively tested and human-vetted dataset available for astronomical knowledge assessment. Second, while astronomical knowledge recall represents just one aspect of LLM capabilities, it serves as a critical foundation for more advanced applications such as scientific agents. The primary goal is to demonstrate that proper

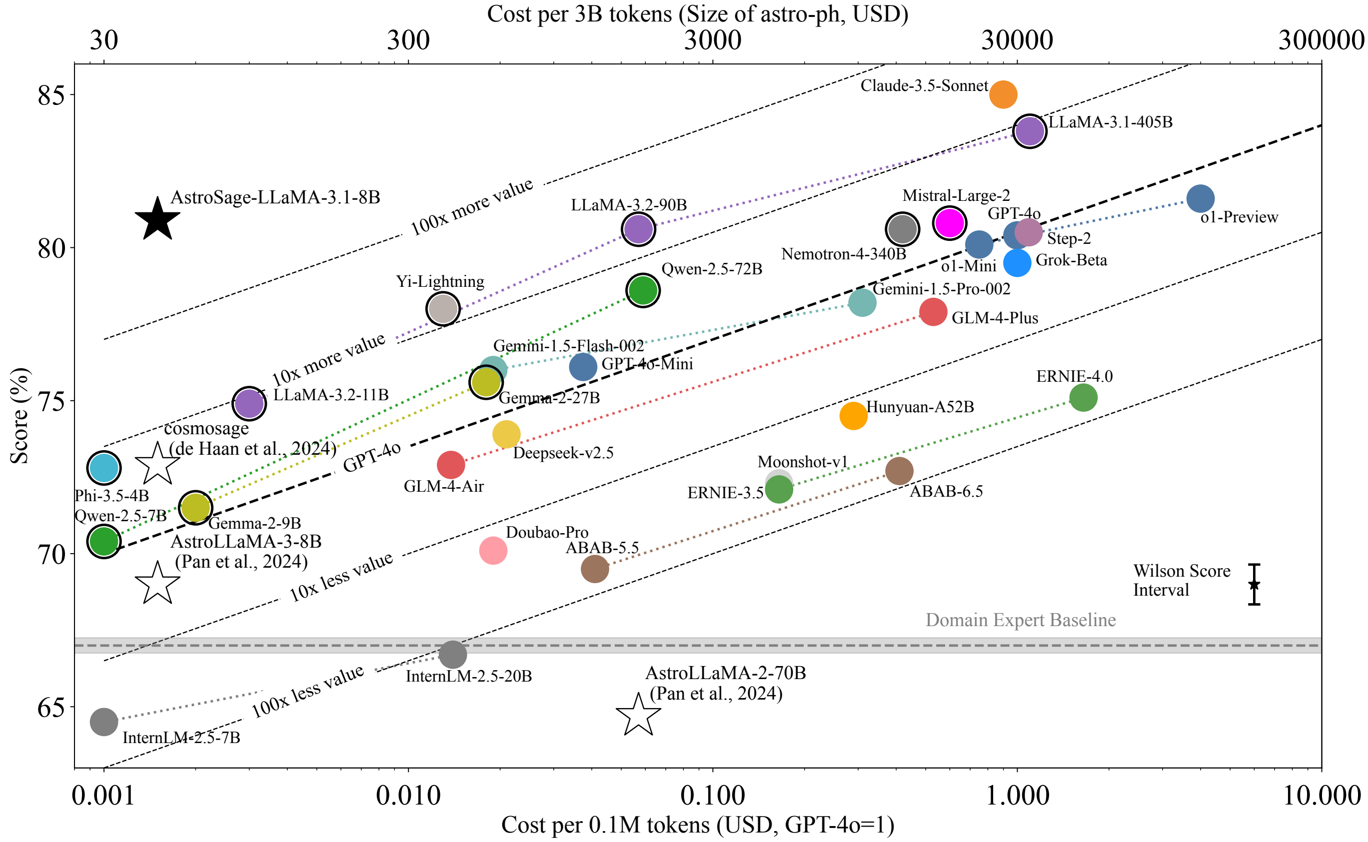

FIG. 3. Performance comparison on the AstroMLab 1 benchmark, which contains 4,425 high-quality, human-verified multiple-choice questions across astronomy, astrophysics, cosmology, and instrumentation. The gray shaded region indicates the performance range of human domain experts. We present updated results as of November 2024, incorporating both cutting-edge proprietary and open-weight models. AstroSage-Llama-3.1-8B outperforms all other models in the 8-billion parameter class and achieves performance comparable to OpenAI's latest models, including GPT-4o, while Claude-3.5-Sonnet maintains the highest performance overall. The diagonal dashed lines represent cost-efficiency trade-offs as determined in AstroMLab 1 (see text for details). The Wilson Score interval shows the typical uncertainty in the score due to the finite number of questions. Star symbols indicate all published specialized LLMs for astronomy to our knowledge. Previously, these specialized models often failed to outperform their baseline models in astronomical recall due to various training limitations. AstroSage-Llama-3.1-8B represents a significant advancement in specialized astronomical LLMs, demonstrating that extensive data curation, massive continued pre-training and supervised fine-tuning, and model merging techniques can substantially improve performance on specific astronomical tasks. This result highlights the effectiveness of domain specialization even in relatively smaller models.

fine-tuning of a relatively small model can significantly improve performance on a specific task—an achievement not previously demonstrated in astronomy.

The performance score is calculated as the fraction of correctly answered multiple-choice questions in the benchmarking dataset. The resulting scores are shown in Figure 3, where round symbols represent scores for cutting-edge proprietary and open-weight models. The open-weight models are also marked with an outer circle. The x-axis displays the cost per 10^5 tokens, a metric chosen based on practical applications: in the first (and to our knowledge, only) implementation of astronomical agents [9], analyzing a celestial source's spectral energy distribution from James Webb Space Telescope data requires approximately 10^5 tokens. The top x-axis shows costs scaled to 3B (3×10^9) tokens, roughly equivalent to the entire astro-ph section of the arXiv. For proprietary models, we use current token costs (averaging input and output costs where they differ), while open-weight model costs are estimated based on typical pricing of commercial API platforms.

Specialized astronomical LLMs are denoted by star symbols, except for the first AstroLLaMA model [5], whose score falls below the plot's lower limit. The bottom right panel shows the typical uncertainty (calculated using the Wilson score interval), demonstrating that our dataset of 4,425 multiple-choice questions provides sufficiently small sampling noise to establish robust performance differences. We have updated all scores using the latest model versions following the methodology from Ting *et al.* [11, AstroMLab 1].

The diagonal dashed lines represent a universal cost-efficiency trade-off observed across major model series (e.g. Llama, GPT, GLM) that simultaneously released

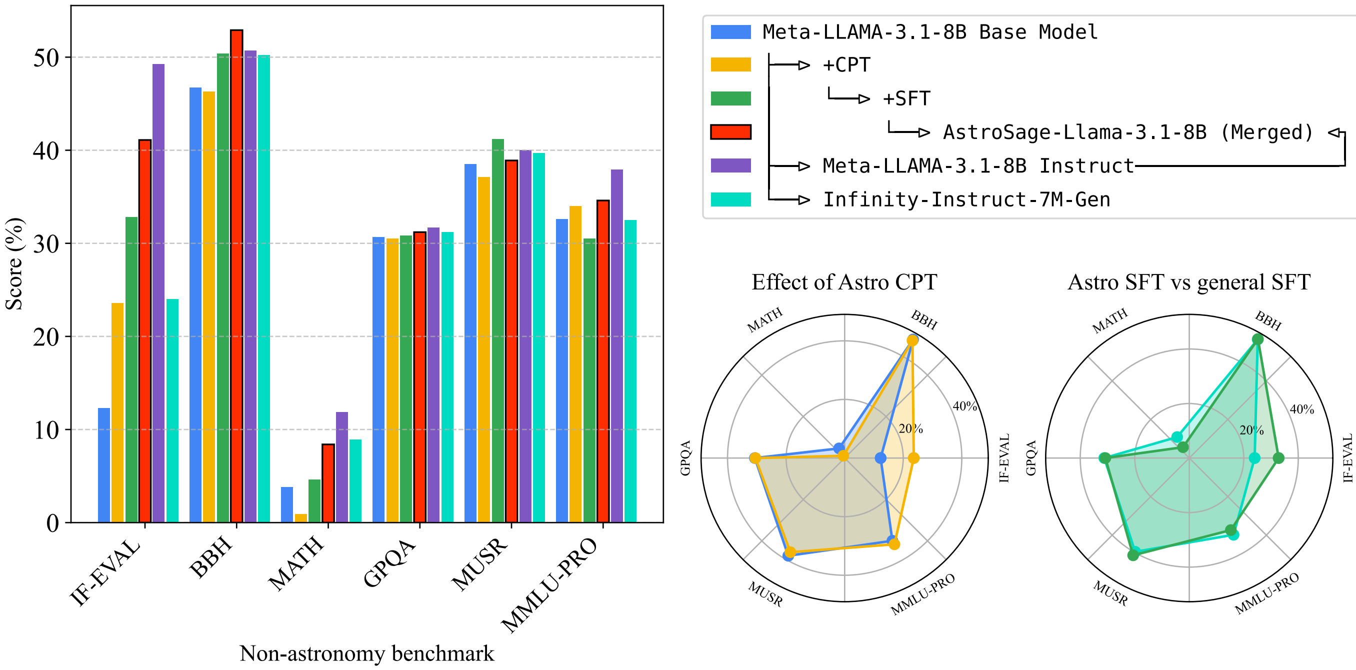

FIG. 4. Performance comparison across general language model benchmarks. The left panel shows a bar chart comparing model performances on standard benchmarks. The center radar chart visualizes the effect of continued pretraining. The right radar chart compares our model post-SFT against BAAI/Infinity-Instruct-7M-Gen-LLaMA3_1-8B, which was trained on the same base model and SFT data but without our astronomy-specific data. Despite optimization for astronomy tasks, the merged model we are releasing as AstroSage-Llama-3.1-8B maintains strong general capabilities in reasoning, mathematics, and coding, demonstrating that domain specialization did not come at the cost of other abilities.

models at multiple sizes. We consistently observe a 3.5-point improvement in performance for every 10-fold increase in cost across model families. Each dashed line represents this equivalent trade-off, offset by 3.5 percentage points (equivalent to a 10-fold gain in cost-effectiveness). Despite similar performance on general benchmarks, cutting-edge models can differ by up to 1000-fold in cost-effectiveness on astronomical tasks, highlighting the importance of specialized astronomical benchmarks for evaluating performance on niche technical domains.

To establish a human performance baseline, two domain experts from our team independently completed a random subset of benchmark questions under controlled conditions. Each expert was allowed approximately 30 seconds per question and prohibited from consulting external references, including web searches or language model assistance. Both experts achieved remarkably consistent scores of approximately 68%, which we designate as the “human domain expert baseline.” The fact that most evaluated LLMs significantly surpassed this baseline demonstrates both the benchmark’s comprehensive scope and difficulty, while highlighting the remarkable capabilities of current LLMs in capturing and applying complex astronomical knowledge.

As previously noted in [7, AstroMLab 2], existing specialized astronomical LLMs (shown as open stars in Figure 3) fail to outperform baseline models of comparable parameter size. In many cases, suboptimal specialization techniques actually led to performance degradation. In contrast, AstroSage-Llama-3.1-8B, despite its modest size of 8 billion parameters, achieved an accuracy of 80.9% on this benchmark—comparable to OpenAI’s latest flagship models (GPT-4o: 80.4%) and the best 90B-parameter open-weight Meta-Llama models (80.6%). This performance is particularly notable because AstroSage-Llama-3.1-8B achieves these results at approximately one-thousandth the inference cost of proprietary models and one-hundredth the cost of open-weight models. Furthermore, it demonstrates an 8-point improvement over its baseline model, Meta-Llama-3.1-8B (72.9%). To our knowledge, this represents the first demonstration of a specialized astronomical LLM achieving objectively verified improvements through model fine-tuning.

To ensure our domain specialization didn’t compromise general capabilities, we evaluated AstroSage-Llama-3.1-8B across a comprehensive suite of standard language model benchmarks. These include IF-EVAL (instruction following), BBH (binary hypothesis testing), MATH (mathematical reasoning), GPQA (graduate-level science questions), MUSR (real-world decision-making scenarios), and MMLU-PRO (an expanded version of MMLU with more challenging reasoning questions). As shown in Figure 4, our CPT+SFT model (green, initialized from the Llama-3.1 base model) initially performed below the

Llama-3.1 instruct model (purple) on five out of the six non-astronomy benchmarks. This was expected, given that Meta's proprietary SFT dataset for their instruct model likely far exceeds what's feasible for an academic research group to reproduce. The merging procedure, pulling in only 25% of its weight from Meta-Llama-3.1-8B-Instruct, allowed us to recover much of this performance deficit.

Crucially, this performance recovery through model merging did not compromise AstroSage-Llama-3.1-8B's astronomical expertise—it maintained its 8-point improvement (representing more than 100-fold increase in cost-effectiveness) on astronomical Q&A tasks while largely preserving capabilities across most general benchmarks. The only notable performance decrease occurred in IF-EVAL, which tests instruction following. This limited decline is unsurprising, as instruction following remains one of the more brittle capabilities in language models and likely heavily depends on the proprietary training data used in Meta's instruct model. In fact, when compared AstroSage-Llama-3.1-8B to BAAI/Infinity-Instruct-7M-Gen-LLaMA3 1-8B, the latter shows an even more severe performance deficit, highlighting how our refined training strategy and expanded SFT dataset represent crucial improvements. Ultimately, our model merging approach successfully preserved most general capabilities without sacrificing the gained astronomical expertise. This balance is essential, as it enables AstroSage-Llama-3.1-8B to engage in natural conversations and assist with broader tasks while excelling in astronomy-specific applications.

VI. AVAILABILITY

To promote reproducibility and advance the field of domain-specific AI assistants, we are making AstroSage-Llama-3.1-8B freely available under the highly permissive Llama 3.1 Community License. The full model weights can be accessed and downloaded from our project repository on Hugging Face: `https://huggingface.co/AstroMLab/AstroSage-8B` in either PyTorch or safetensors format.

The code used to prepare the datasets and perform the training will be made available upon reasonable request.

By making AstroSage-Llama-3.1-8B widely available, we aim to foster collaboration and innovation in the astronomy community. We encourage researchers to build upon our work and contribute to the ongoing development of specialized AI assistants for scientific domains.

VII. DISCUSSION AND FUTURE WORK

This work demonstrates the potential of specialized language models in astronomy through a systematic approach to model development and evaluation. While previous efforts like laid important groundwork in domain-specific modeling, the field has faced persistent challenges in achieving performance gains over baseline models, especially in instruction-following tasks. Our multi-stage training process—combining continued pretraining, extensive supervised fine-tuning, and strategic model merging—addresses these challenges, achieving a notable improvement over the baseline model.

These results demonstrate that powerful AI assistants can be developed with relatively small language models when sufficiently specialized. Despite its modest size of 8 billion parameters, AstroSage-Llama-3.1-8B achieves performance comparable to latest flagship models at a fraction of the cost—approximately one-thousandth of proprietary models and one-hundredth of open-weight models. This remarkably favorable performance-to-parameter ratio suggests even greater potential for improvement through scaling. Given access to the necessary computational resources, we plan to apply our successful CPT/SFT procedure to a 70B-class model to pursue state-of-the-art astronomy-specific performance.

Beyond the performance achievements, our work establishes a more systematic approach to model evaluation in astronomy. Through tailored astronomy-specific benchmarking in, we provide a more rigorous and transparent assessment than previously available. However, significant challenges remain in comprehensive model evaluation. The field currently lacks standardized, astronomy-specific benchmarks capable of assessing understanding across the full spectrum of astronomical tasks, particularly in exact problem-solving capabilities like those tested in ScienceAgentBench [22]. This limitation restricts our ability to validate comparisons in more direct scientific agent contexts.

The constraints of an 8B-parameter model also become apparent in certain scenarios. While AstroSage-Llama-3.1-8B demonstrates impressive performance in subjective testing, the AstroMLab-1 benchmark, and general benchmarks, it encounters natural limitations in memory capacity and reasoning depth. Particularly challenging are questions requiring complex multi-step reasoning or sophisticated calculations, where larger general-purpose models still maintain an advantage.

To address these limitations, our future work will pursue several complementary directions. While scaling up model size remains a primary goal, we will also focus on developing more specialized benchmarking tools and exploring retrieval-augmented generation for improved knowledge access. Additional initiatives include creating multilingual astronomy assistants, implementing mechanisms for real-time knowledge updates, and providing public inference capabilities.

The broader implications of this work extend well beyond its immediate achievements. AstroSage-Llama-3.1-8B serves as a compelling proof of concept for highly specialized, smaller-scale language models in astronomy. Our approach of extensive data curation, continued pretraining, and careful supervised fine-tuning demonstrates how domain-specific expertise can be enhanced while preserving general capabilities. As the field progresses toward agentic research assistants capable of autonomous literature review, data analysis, and hypothesis generation, the need for affordable, highly competent domain-specific models will only grow. While challenges remain, AstroSage-Llama-3.1-8B charts a promising course for developing the next generation of specialized scientific AI assistants, potentially transforming how we approach astronomical research and education.

ACKNOWLEDGEMENTS

This research used resources of the Oak Ridge Leadership Computing Facility (OLCF), which is a DOE Office of Science User Facility at the Oak Ridge National Laboratory supported by the U.S. Department of Energy under Contract No. DE-AC05-00OR22725 and support from Microsoft's Accelerating Foundation Models Research (AFMR) program. TdH was supported by World Premier International Research Center Initiative (WPI), MEXT, Japan. YST is supported by the National Science Foundation under Grant No. 2406729.

[1] X.-Y. Fu, M. T. R. Laskar, E. Khasanova, C. Chen and S. B. TN, "Tiny Titans: Can Smaller Large Language Models Punch Above Their Weight in the Real World for Meeting Summarization?" (2024), arXiv:2402.00841 [cs].

[2] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark et al., "Training Compute-Optimal Large Language Models," (2022), arXiv:2203.15556 [cs].

[3] T. Schick and H. Schütze, "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners," (2021), arXiv:2009.07118 [cs].

[4] I. Turc, M.-W. Chang, K. Lee and K. Toutanova, "Well-Read Students Learn Better: On the Importance of Pre-training Compact Models," (2019), arXiv:1908.08962 [cs].

[5] T. D. Nguyen, Y.-S. Ting, I. Ciucă, C. O'Neill, Z.-C. Sun, M. Jabłońska, S. Kruk, E. Perkowski, J. Miller, J. Li et al., "AstroLLaMA: Towards Specialized Foundation Models in Astronomy," https://arxiv.org/abs/2309.06126v1 (2023).

[6] E. Perkowski, R. Pan, T. D. Nguyen, Y.-S. Ting, S. Kruk, T. Zhang, C. O'Neill, M. Jablonska, Z. Sun, M. J. Smith et al., "AstroLLaMA-Chat: Scaling AstroLLaMA with Conversational and Diverse Datasets," https://arxiv.org/abs/2401.01916v2 (2024).

[7] R. Pan, T. D. Nguyen, H. Arora, A. Accomazzi, T. Ghosal and Y.-S. Ting, "AstroMLab 2: AstroLLaMA-2-70B Model and Benchmarking Specialised LLMs for Astronomy," (2024), arXiv:2409.19750.

[8] T. de Haan, "Cosmosage: A Natural-Language Assistant for Cosmologists," https://arxiv.org/abs/2407.04420v1 (2024).

[9] Z. Sun, Y.-S. Ting, Y. Liang, N. Duan, S. Huang and Z. Cai, arXiv e-prints , arXiv:2409.14807 (2024), arXiv:2409.14807 [astro-ph.IM].

[10] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan et al., "The Llama 3 Herd of Models," (2024), arXiv:2407.21783.

[11] Y.-S. Ting, T. D. Nguyen, T. Ghosal, R. Pan, H. Arora, Z. Sun, T. de Haan, N. Ramachandra, A. Wells, S. Madireddy and A. Accomazzi, "AstroMLab 1: Who Wins Astronomy Jeopardy!?" https://arxiv.org/abs/2407.11194v1 (2024).

[12] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young et al., "Scaling Language Models: Methods, Analysis & Insights from Training Gopher," https://arxiv.org/abs/2112.11446v2 (2021).

[13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language Models are Few-Shot Learners," https://arxiv.org/abs/2005.14165v4 (2020).

[14] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei and J. Launay, "The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only," https://arxiv.org/abs/2306.01116v1 (2023).

[15] R. Li, Y. Wei, M. Zhang, N. Yu, H. Hu and H. Peng, "ScalingFilter: Assessing Data Quality through Inverse Utilization of Scaling Laws," https://arxiv.org/abs/2408.08310v1 (2024).

[16] L. Blecher, G. Cucurull, T. Scialom and R. Stojnic, "Nougat: Neural Optical Understanding for Academic Documents," (2023), arXiv:2308.13418 [cs].

[17] S. Dash, I. Lyngaas, J. Yin, X. Wang, R. Egele, G. Cong, F. Wang and P. Balaprakash, "Optimizing Distributed Training on Frontier for Large Language Models," https://arxiv.org/abs/2312.12705v2 (2023).

[18] Beijing Academy of Artificial Intelligence (BAAI), "BAAI/Infinity-Instruct · Datasets at Hugging Face," https://huggingface.co/datasets/BAAI/Infinity-Instruct (2024).

[19] P. Yadav, T. Vu, J. Lai, A. Chronopoulou, M. Faruqui, M. Bansal and T. Munkhdalai, "What Matters for Model Merging at Scale?" (2024), arXiv:2410.03617.

[20] A. Dassanaike-Perera, S. Waiwitlikhit and K. Gilbai, *Cuts and Stitches: Does Model Merging Produce Better Multitask Learners?*, Stanford CS224N Default Project (Stanford, 2023).

[21] C. Goddard, S. Siriwardhana, M. Ehghaghi, L. Meyers, V. Karpukhin, B. Benedict, M. McQuade and J. Solawetz, "Arcee's MergeKit: A Toolkit for Merging Large Language Models," (2024), arXiv:2403.13257.

[22] Z. Chen, S. Chen, Y. Ning, Q. Zhang, B. Wang, B. Yu, Y. Li, Z. Liao, C. Wei, Z. Lu et al., arXiv e-prints , arXiv:2410.05080 (2024), arXiv:2410.05080 [cs.CL].