AstroSage: Leading Performance in Astronomy Q&A with a 70B-Parameter Domain-Specialized Model

Tijmen de Haan [1 2] **Yuan-Sen Ting** [3 4] **Tirthankar Ghosal** [5] **Tuan Dung Nguyen** [6] **Alberto Accomazzi** [7] **Emily Herron** [5] **Vanessa Lama** [5] **Rui Pan** [8] **Azton Wells** [9] **Nesar Ramachandra** [9]

Abstract

General-purpose large language models (LLMs), despite their broad capabilities, often struggle with specialized domain knowledge, a limitation particularly pronounced in fields such as astronomy. This study introduces AstroSage-Llama-3.1-70B. Developed from the Meta-Llama-3.1-70B foundation, AstroSage underwent extensive continued pre-training on a vast corpus of astronomical literature, followed by supervised fine-tuning and model merging. Beyond its 70-billion parameter scale, this model improves on our previous 8-billion parameter version with refined datasets, optimized hyperparameters, and reasoning capabilities. Evaluated on the Astrobench, AstroSage achieves 86.2% accuracy, surpassing all tested models including o3, Claude-3.7-Sonnet, GPT-4.1, and Deepseek-R1. This work demonstrates that domain specialization, when applied to large-scale models, can enable specialized systems to outperform even the most advanced commercial alternatives within their domain while achieving approximately 100x improvement in cost-efficiency.

1. Introduction

Astronomy and its related fields demand sophisticated tools that can process vast amounts of specialized knowledge. Large Language Models (LLMs) have emerged as promising assistants for this domain, offering capabilities as research collaborators, educational resources, and knowledge repositories (Perkowski et al., 2024). Domain-specialized models demonstrate particular cost-effectiveness in such contexts, as their parameters can be optimized for specific knowledge domains rather than distributed across the breadth of general internet content (Turc et al., 2019).

AstroSage-Llama-3.1-8B (de Haan et al., 2025), established that a relatively modest 8-billion parameter LLM, when extensively trained on astronomical content, could match or exceed the performance of much larger general-purpose models on astronomical knowledge tasks. This finding highlighted the potential of domain specialization for creating efficient, high-performing AI assistants (Schick and Schütze, 2021).

In this study, we introduce AstroSage-Llama-3.1-70B, a 70-billion parameter language model that represents an advancement in specialized AI for astronomy. Our central research question asks whether domain specialization merely improves efficiency or can enable specialized models to outperform even the largest commercial alternatives (Rae et al., 2021). Following Meta-Llama-3.1-8B, we applied similar domain specialization techniques to the larger Meta-Llama-3.1-70B foundation (Dubey et al., 2024). Beyond the increased parameter count, we implemented several key enhancements: expanded and refined datasets for both continued pre-training and supervised fine-tuning; optimized learning hyperparameters based on public benchmarks and our own experimentation; and an explicit reasoning capability that enables step-by-step analytical processes before generating answers, often referred to as chain-of-thought (Suzgun et al., 2022).

The core hypothesis driving this work is that a larger specialized model can elevate AI performance across astronomy, astrophysics, space science, cosmology, astroparticle physics, astronomical instrumentation, and related fields. While AstroSage-8B successfully matched larger models' performance, AstroSage-70B aims to surpass even advanced commercial alternatives. Beyond testing this hypothesis, we are making our trained model openly available to serve as a resource for researchers, educators, and students in the field.

[1]Institute of Particle and Nuclear Studies (IPNS), High Energy Accelerator Research Organization (KEK), Tsukuba, Ibaraki, Japan [2]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 [3]Department of Astronomy, The Ohio State University, Columbus, OH, USA [4]Center for Cosmology and AstroParticle Physics (CCAPP), The Ohio State University, Columbus, OH, USA [5]National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA [6]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA [7]Center for Astrophysics, Harvard & Smithsonian, Cambridge, MA, USA [8]Siebel School of Computing and Data Science, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA [9]Computational Science Division, Argonne National Laboratory, Lemont, IL, USA. Correspondence to: Tijmen de Haan <tijmen.dehaan@gmail.com>.

2. Model Architecture and Training

AstroSage-70B is derived from the Meta-Llama-3.1-70B architecture (Dubey et al., 2024). This base model was selected for consistency with AstroSage-8B, which chose Meta-Llama-3.1-8B for its state-of-the-art general capabilities and permissive licensing. The tokenizer from Meta-Llama-3.1-70B-Instruct was used without modification. Following our established methodology (Nguyen et al., 2023), the development process comprised three main stages: continued pre-training (CPT), supervised fine-tuning (SFT), and model merging (Dassanaike-Perera et al., 2023).

The objective of CPT is to imbue the base model with extensive domain-specific knowledge from astronomical literature (Blecher et al., 2023). The CPT dataset for AstroSage-70B builds upon the comprehensive corpus developed previously, which included approximately 250,000 arXiv preprints from astro-ph and gr-qc categories spanning 2007-2024, nearly 30,000 Wikipedia articles related to astronomy, and internet-available textbooks. The knowledge cutoff for the astronomical papers remains January 2024.

This dataset was enhanced through application of ftfy (Speer, 2019) for consistent Unicode text normalization and rule-based repetition removal to correct OCR failures, supplementing our perplexity-based cleaning methods (Li et al., 2024). To preserve general language understanding and mitigate catastrophic forgetting due to specialization (Pan et al., 2024), we incorporated a random selection of samples from the FineWeb dataset (Penedo et al., 2023) into each training epoch. This addition of previous pretraining tokens during CPT, sometimes known as "replay," proved crucial. Notably, the specific FineWeb samples were varied for each epoch, ensuring diverse exposure to general web text.

The CPT and SFT stages were conducted on the Oak Ridge Leadership Computing Facility (OLCF) Frontier supercomputer using AMD Instinct MI250X GPUs. Our implementation employed the GPT-NeoX framework (Andonian et al., 2022; Smith et al., 2022), which we adapted for compatibility with the Llama-3.1 architecture. Training was distributed across 2,048 Graphics Compute Dies (GCDs) using a multi-dimensional parallelism strategy: tensor parallelism 8, pipeline parallelism 8, and data parallelism 32. As GPT-NeoX does not currently support DeepSpeed ZeRO stage 2/3 with pipeline parallelism, we used ZeRO stage 1 (Rasley et al., 2020) with activation checkpointing enabled. This configuration achieved a computational throughput of approximately 50 TFLOPS/s per GCD, consistent with performance metrics reported by Dash et al. (2023).

Following CPT, the model underwent SFT to develop its instruction-following (Zhou et al., 2023) and conversational capabilities, and to instill behaviors such as chain-of-thought and self-reflection. Figure 2 illustrates the composition of the SFT dataset. Its largest component is NVIDIA's Llama-Nemotron-Post-Training-Dataset (Bercovich et al., 2025), which was used to train models that consistently demonstrate excellent performance on public benchmarks such as LMArena (Chiang et al., 2024), suggesting it is a strong dataset for eliciting reasoning and aligning with human preferences. This dataset provides reasoning components covering science, code, mathematics, and general chat, establishing a foundation for analytical thinking across different domains.

We also included the OpenHermes 2.5 dataset, which helps build general instruction-following capabilities and adherence to the system prompt. To enhance domain expertise, we incorporated custom domain-specific Q&A datasets from both our previous work and (de Haan, 2025), which together comprise approximately 30% of the training data. After combination, the dataset was deduplicated and shuffled. A loss mask was applied to train the model exclusively on assistant completions, excluding user queries and system prompts. The chat template adheres to the Llama-3.1 standard.

The model was fine-tuned on this SFT dataset for 0.6 epochs, consuming approximately 13,000 GPU-hours on the same infrastructure. Hyperparameters mirrored those of the CPT stage, with the exception of weight decay, which was removed. Figure 1 illustrates the training dynamics during both phases, showing consistent improvement without overfitting indicators.

To create the final, publicly released AstroSage-70B, we employed model merging using the mergekit library (Goddard et al., 2024). This technique allows us to combine the strengths of our specialized SFT model with the robust instruction-following capabilities of other popular fine-tuned Llama-3.1-70B variants (Yadav et al., 2024). The final mixture was created using the DARE-TIES method (Yu et al., 2024) with the AstroSage-70B-CPT model as the base. The components include 70% AstroSage-70B-SFT, 15% Llama-3.1-Nemotron-160-Instruct, 7.5% Meta-Llama-3.3-70B-Instruct, and 7.5% Meta-Llama-3.1-70B-Instruct.

3. Features and Capabilities

AstroSage-70B is designed for a wide range of applications within the astronomical domain. Potential applications include addressing factual queries, literature review

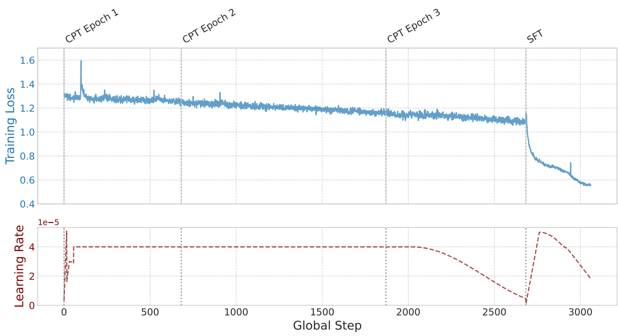

Figure 1. Training dynamics for continued pre-training (CPT) and supervised fine-tuning (SFT). The top panel shows loss trajectory across 2.5 epochs of CPT followed by 0.6 epochs of SFT. The bottom panel shows the learning rate schedule including warm-up periods, learning rate decay, and manual adjustments.

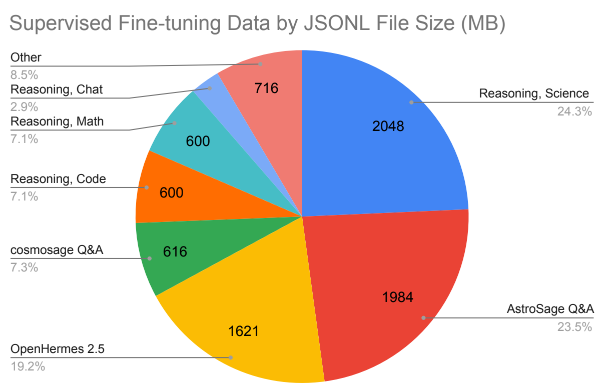

Figure 2. Composition of the AstroSage-70B SFT training dataset. The combination of reasoning-focused datasets (41.8%) with domain-specific astronomy Q&A (30.8%) reflects our strategy to develop a model combining analytical thinking with specialized knowledge.

and summarization, assisting with manuscript preparation, brainstorming and hypothesis formulation, concept learning, programming support, and serving as a component in agentic systems (Sun et al., 2024). The general capabilities of large models for such few-shot tasks were extensively demonstrated by Brown et al. (2020).

A notable feature of AstroSage-70B is its explicit reasoning capability. This aligns with recent advances in the broader LLM field, where explicit reasoning has emerged as a critical development for enhancing model performance on complex tasks (Suzgun et al., 2022; Sprague et al., 2023). The integration of reasoning mechanisms has become increasingly common in state-of-the-art models, including OpenAI's o1 through o4 series, Anthropic's Claude models with "thinking mode," DeepSeek-R1, and others. These developments demonstrate that exposing and structuring the internal reasoning process allows models to tackle complex problems more systematically, resulting in improved accuracy and reliability.

Building on these industry-wide insights, AstroSage-70B implements explicit reasoning that can be activated at inference time by setting the system prompt to "detailed thinking on" and prefilling the assistant completion with <think>. When enabled, the model generates a step-by-step reasoning process before providing the final answer. This is particularly beneficial for complex astronomical problems requiring multi-step analysis. As the reasoning tokens are enclosed within tags, they can easily be hidden from the end-user if desired.

4. Evaluation

To evaluate the performance of AstroSage-70B, we utilize the Astrobench (Ting et al., 2024). This benchmark consists of 4,425 high-quality, human-verified multiple-choice questions spanning astronomy, astrophysics, cosmology, and astronomical instrumentation. These questions are derived from Annual Review of Astronomy and Astrophysics papers that were explicitly withheld from the AstroSage training corpus. This ensures the model is evaluated on genuinely unseen material, and its performance is not merely an artifact of training on the benchmark's source texts.

On this benchmark, AstroSage-70B achieves a score of 86.2% without enabling reasoning. As illustrated in Figure 3, this performance establishes AstroSage-70B as the leading model, outperforming all other tested open-weight and proprietary models. This includes a notable improvement over AstroSage-8B, superseding also contemporary large-scale general-purpose LLMs such as o3, GPT-4.1, Claude-3.7-Sonnet, and Deepseek-R1. For context, professional astronomers score around 67% on this benchmark.

Our evaluation substantially updates the results presented in (Ting et al., 2024), which was published in July 2024. The benchmark analysis includes a cost-accuracy trade-off visualization, represented by diagonal dashed lines in Figure 3. This analysis reveals that within a model family, a tenfold increase in API cost typically corresponds to an improvement of approximately 3.5 percentage points in accuracy, a scaling trend consistent with observations in more general contexts (Hoffmann et al., 2022; Rae et al., 2021). Consequently, the distance between adjacent lines represents an order of magnitude gain in cost efficiency.

As highlighted by the vertical red arrows in Figure 3, our domain specialization approach delivers remarkable efficiency gains. Both AstroSage models jump approximately two cost-efficiency lines compared to their respective base models, representing improvement of approximately $100\times$

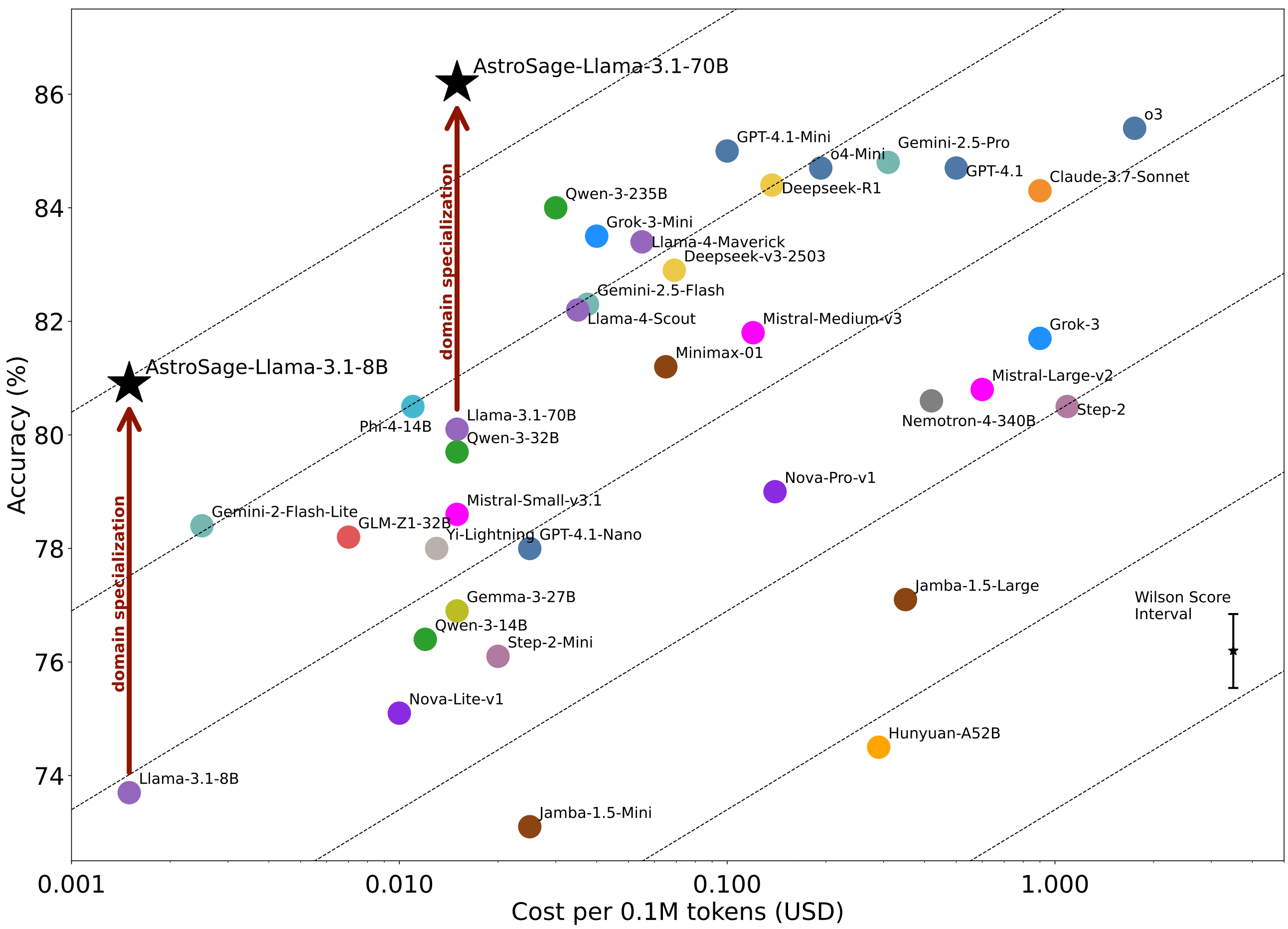

Figure 3. Performance comparison on the AstroMLab-1 benchmark across 38 LLMs as of May 2025. Models are plotted by accuracy (y-axis) versus inference cost per 0.1M tokens in USD (x-axis, logarithmic scale). AstroSage-70B achieves 86.2% accuracy, establishing state-of-the-art performance and surpassing all tested models including more expensive proprietary offerings like o3, Claude-3.7-Sonnet, and GPT-4.1. The diagonal dashed lines represent iso-efficiency contours where a tenfold increase in cost typically yields a 3.5 percentage point improvement in accuracy. Both AstroSage models (8B and 70B) jump approximately two efficiency lines above their respective base Llama models, representing a $\sim$100$\times$ improvement in cost-efficiency. The Wilson Score interval shown indicates typical uncertainty due to the finite question set.

in cost-efficiency. This demonstrates that targeted domain specialization can achieve performance levels that would otherwise require models costing two orders of magnitude more at inference time.

The original study predicted model improvements would shift performance to the next diagonal line every three to six months, a forecast that has proven accurate. An interesting observation from our updated evaluation is that while the 3.5 percentage point trade-off slope still holds for some series, models like Qwen-3 and GPT-4.1 exhibit steeper drop-offs in performance across their tiered offerings. This suggests that current distillation approaches for creating more affordable variants of powerful models may be less effective for specialized knowledge domains like astronomy (Turc et al., 2019).

In our evaluation methodology, we applied a consistent approach to models with reasoning capabilities. For models with explicit reasoning modes, we enabled this feature during testing. Interestingly, we found that enabling reasoning modes generally did not significantly improve scores for most models on this benchmark, including AstroSage-70B. This finding may be due to the Astrobench questions primarily testing fast, intuitive knowledge recall rather than complex multi-step reasoning where such modes typically demonstrate advantages.

We acknowledge this as a limitation of current astronomy benchmarks, which offer limited evaluation of problem-solving capabilities requiring deep reasoning. Other work has focused on creating benchmarks to specifically test these abilities in domains like mathematics and general problem solving (Rein et al., 2023; Suzgun et al., 2022; Hendrycks et al., 2021; Sprague et al., 2023; Wang et al., 2024). Nevertheless, since a primary goal of our specialized training is to imbue the model with comprehensive domain knowledge, the benchmark results demonstrate successful achievement of this objective.

5. Conclusion and Broader Impact

The development of AstroSage-70B advances specialized language models for astronomy. Building on the foundation established by AstroSage-8B, this 70-billion parameter model incorporates a more powerful base architecture, enriched training datasets, refined training methodologies, and explicit reasoning capabilities. Our results support the central hypothesis of this work: domain specialization, when scaled to larger models, can enable specialized systems to surpass even the most advanced general-purpose commercial models within their domain of expertise. The improvement of approximately $100\times$ in cost-efficiency highlights the practical value of domain specialization, particularly important as the field moves toward deploying AI assistants at scale (Fu et al., 2024).

An interesting trend emerged regarding the effectiveness of model distillation and scaling. The performance drop-off between flagship models and their smaller variants appears more pronounced than observed previously, particularly in the 30-70B parameter range. This trend becomes even more pronounced at smaller scales, highlighting the potential importance of specialized training for deploying cost-effective models (Pan et al., 2024).

Looking forward, two key areas warrant further investigation. First, development of more comprehensive benchmarks that specifically evaluate reasoning capabilities in astronomical contexts, including problem-solving tasks that more closely resemble real research challenges (Wang et al., 2024). Second, integration of AstroSage-70B with astronomy-specific tools and workflows, moving toward more comprehensive AI research assistants that can handle both domain knowledge and practical research tasks (Chen et al., 2024; Sun et al., 2024).

AstroSage-70B advances the integration of AI assistants into astronomical research and education. By making our specialized tools openly available, we aim to democratize access to specialized LLMs and accelerate scientific discovery (Perkowski et al., 2024).

References

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Shivanshu Purohit, Tri Songz, Wang Phil, and Samuel Weinbach. Gpt-neox: Large scale autoregressive language modeling in pytorch, April 2022.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, Gerald Shen, Ameya Sunil Mahabaleshwarkar, Bilal Kartal, Yoshi Suhara, Olivier Delalleau, Zijia Chen, Zhilin Wang, David Mosallanezhad, Adi Renduchintala, Haifeng Qian, Dima Rekesh, Fei Jia, Somshubra Majumdar, Vahid Noroozi, Wasi Uddin Ahmad, Sean Narenthiran, Aleksander Ficek, Mehrzad Samadi, Jocelyn Huang, Siddhartha Jain, Igor Gitman, Ivan Moshkov, Wei Du, Shubham Toshniwal, George Armstrong, Branislav Kisacanin, Matvei Novikov, Daria Gitman, Evelina Bakhturina, Jane Polak Scowcroft, John Kamalu, Dan Su, Kezhi Kong, Markus Kliegl, Rabeeh Karimi, Ying Lin, Sanjeev Satheesh, Jupinder Parmar, Pritam Gundecha, Brandon Norick, Joseph Jennings, Shrimai Prabhumoye, Syeda Nahida Akter, Mostofa Patwary, Abhinav Khattar, Deepak Narayanan, Roger Waleffe, Jimmy Zhang, Bor-Yiing Su, Guyue Huang, Terry Kong, Parth Chadha, Sahil Jain, Christine Harvey, Elad Segal, Jining Huang, Sergey Kashirsky, Robert McQueen, Izzy Putterman, George Lam, Arun Venkatesan, Sherry Wu, Vinh Nguyen, Manoj Kilaru, Andrew Wang, Anna Warno, Abhilash Somasamudramath, Sandip Bhaskar, Maka Dong, Nave Assaf, Shahar Mor, Omer Ullman Argov, Scot Junkin, Oleksandr Romanenko, Pedro Larroy, Monika Katariya, Marco Rovinelli, Viji Balas, Nicholas Edelman, Anahita Bhiwandiwalla, Muthu Subramaniam, Smita Ithape, Karthik Ramamoorthy, Yuting Wu, Suguna Varshini Velury, Omri Almog, Joyjit Daw, Denys Fridman, Erick Galinkin, Michael Evans, Katherine Luna, Leon Derczynski, Nikki Pope, Eileen Long, Seth Schneider, Guillermo Siman, Tomasz Grzegorzek, Pablo Ribalta, Monika Katariya, Joey Conway, Trisha Saar, Ann Guan, Krzysztof Pawelec, Shyamala Prayaga, Oleksii Kuchaiev, Boris Ginsburg, Oluwatobi Olabiyi, Kari Briski, Jonathan Cohen, Bryan Catanzaro, Jonah Alben, Yonatan Geifman, Eric Chung, and Chris Alexiuk. Llama-nemotron: Efficient reasoning models, May 2025.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural Optical Understanding for Academic Documents, August 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen

Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html`.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery, October 2024.

Wei-Lin Chiang, Li Zheng, Ying Sheng, Siyan Zhuang, Zhangyue Xing, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Chatbot arena: An open platform for evaluating llms by human preference, January 2024.

Sajal Dash, Isaac Lyngaas, Junqi Yin, Xiao Wang, Romain Egele, Guojing Cong, Feiyi Wang, and Prasanna Balaprakash. Optimizing Distributed Training on Frontier for Large Language Models, December 2023.

Akshana Dassanaike-Perera, Suppakit Waiwitlikhit, and Koren Gilbai. Cuts and Stitches: Does Model Merging Produce Better Multitask Learners? Stanford CS224N Default Project, Stanford, 2023.

Tijmen de Haan. cosmosage: A natural-language assistant for cosmology. *Astronomy and Computing*, 51:100934, April 2025. ISSN 2213-1337. doi: 10.1016/j.ascom.2025.100934. URL `https://www.sciencedirect.com/science/article/pii/S2213133725000071`.

Tijmen de Haan, Yuan-Sen Ting, Tirthankar Ghosal, Tuan Dung Nguyen, Alberto Accomazzi, Azton Wells, Nesar Ramachandra, Rui Pan, and Zechang Sun. Achieving GPT-4o level performance in astronomy with a specialized 8B-parameter large language model. *Scientific Reports*, 15(1):13751, April 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-97131-y. URL `https://www.nature.com/articles/s41598-025-97131-y`. Publisher: Nature Publishing Group.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias

Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The Llama 3 Herd of Models, August 2024.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and Shashi Bhushan TN. Tiny Titans: Can Smaller Large Language Models Punch Above Their Weight in the Real World for Meeting Summarization?, April 2024.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's MergeKit: A Toolkit for Merging Large Language Models, March 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset, November 2021. URL `http://arxiv.org/abs/2103.03874`.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models, March 2022.

Ruihang Li, Yixuan Wei, Miaosen Zhang, Nenghai Yu, Han Hu, and Houwen Peng. ScalingFilter: Assessing Data Quality through Inverse Utilization of Scaling Laws, August 2024.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciucă, Charlie O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Li, Josh Peek, Kartheik Iyer, Tomasz Różański, Pranav Khetarpal, Sharaf Zaman, David Brodrick, Sergio J. Rodríguez Méndez, Thang Bui, Alyssa Goodman, Alberto Accomazzi, Jill Naiman, Jesse Cranney, Kevin Schawinski, and UniverseTBD. AstroLLaMA: Towards Specialized Foundation Models in Astronomy, September 2023.

Rui Pan, Tuan Dung Nguyen, Hardik Arora, Alberto Accomazzi, Tirthankar Ghosal, and Yuan-Sen Ting. AstroMLab 2: AstroLLaMA-2-70B Model and Benchmarking Specialised LLMs for Astronomy. In *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 87–96, November 2024. doi: 10.1109/SCW63240.2024.00019. URL `https://ieeexplore.ieee.org/abstract/document/10820712`.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only, June 2023.

Ernest Perkowski, Rui Pan, Tuan Dung Nguyen, Yuan-Sen Ting, Sandor Kruk, Tong Zhang, Charlie O'Neill, Maja Jablonska, Zechang Sun, Michael J. Smith, Huiling Liu, Kevin Schawinski, Kartheik Iyer, Ioana Ciucă, and UniverseTBD. AstroLLaMA-Chat: Scaling AstroLLaMA with Conversational and Diverse Datasets. *Research Notes of the AAS*, 8(1):7, January 2024. ISSN 2515-5172. doi: 10.3847/2515-5172/ad1abe. URL `https://dx.doi.org/10.3847/2515-5172/ad1abe`. Publisher: The American Astronomical Society.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling Language Models: Methods, Analysis & Insights from Training Gopher, December 2021.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, Yuxiong He, Feng Yan, Elton Li, Kurt Keutzer, and Dario Amodei. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters, October 2020.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark, November 2023. URL `http://arxiv.org/abs/2311.12022`.

Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners, April 2021.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model, February 2022. URL http://arxiv.org/abs/2201.11990.

Robyn Speer. ftfy: fixes text for you, 2019. URL https://doi.org/10.5281/zenodo.3257570.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning, October 2023. URL arXiv:2310.16049v2.

Zechang Sun, Yuan-Sen Ting, Yaobo Liang, Nan Duan, Song Huang, and Zheng Cai. Interpreting Multi-band Galaxy Observations with Large Language Model-Based Agents, September 2024.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them, October 2022. URL http://arxiv.org/abs/2210.09261.

Yuan-Sen Ting, Tijmen de Haan, Tirthankar Ghosal, Tuan Dung Nguyen, Alberto Accomazzi, Azton Wells, Nesar Ramachandra, and Rui Pan. Astromlab-1: Who is the best at astronomy?, July 2024.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models, September 2019.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark, November 2024.

Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqui, Mohit Bansal, and Tsendsuren Munkhdalai. What Matters for Model Merging at Scale?, October 2024.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch, June 2024.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models, November 2023. URL http://arxiv.org/abs/2311.07911.