



AstroLLaMA: Towards Specialized Foundation Models in Astronomy

Tuan Dung Nguyen^{1,2*}, Yuan-Sen Ting^{2,3*}, Ioana Ciucă^{2*},
Charles O’Neill^{2†}, Ze-Chang Sun^{4†}, Maja Jabłońska^{2†}, Sandor Kruk^{5†},
Ernest Perkowski⁵, Jack Miller², Jason Jingshi Li⁶, Josh Peek⁷, Kartheik Iyer⁸,
Tomasz Rózański^{2,9}, Pranav Khetarpal¹⁰, Sharaf Zaman², David Brodrick²,
Sergio J. Rodríguez Méndez², Thang Bui², Alyssa Goodman¹¹, Alberto Accomazzi¹²,
Jill Naiman¹³, Jesse Cranney², Kevin Schawinski¹⁴, Roberta Răileanu¹⁵, UniverseTBD

¹University of Pennsylvania, United States ²Australian National University, Australia

³Ohio State University, United States ⁴Tsinghua University, China

⁵European Space Agency, ESAC, Spain ⁶Learning Machines, Australia

⁷Space Telescope Science Institute, United States

⁸Columbia University, United States ⁹Wrocław University, Poland

¹⁰Indian Institute of Technology Delhi, India ¹¹Harvard University, United States

¹²NASA Astrophysics Data System, Center for Astrophysics, United States

¹³University of Illinois at Urbana-Champaign ¹⁴Modulos AG, Switzerland

¹⁵University College London, United Kingdom

Abstract

Large language models often excel in many human-language tasks but tend to falter in highly specialized domains like scholarly astronomy. To bridge this gap, we introduce AstroLLaMA, a 7-billion-parameter model fine-tuned from LLaMA-2 using over 300,000 astronomy abstracts from arXiv. Optimized for traditional causal language modeling, AstroLLaMA shows marked domain adaptation by achieving a 30% lower perplexity than LLaMA-2. Compared to state-of-the-art foundation models, AstroLLaMA generates more insightful and scientifically relevant text completions and embedding extraction despite having significantly fewer parameters. AstroLLaMA serves as a highly domain-specific model with broad fine-tuning potential: Its public release aims to spur astronomy-focused research, including automatic paper summarization, conversational agent development and hypothesis generation.

1 Introduction

The advent of Large Language Models (LLMs) has sparked interdisciplinary interest thanks to a confluence of factors: accumulation of massive datasets, leaps in computational power, and breakthroughs in neural architectures. Flagship models like GPT-4 (OpenAI, 2023), PaLM (Chowdhery et al., 2022; Goo) and LLaMA (Touvron et al., 2023; Meta, 2023) have exhibited exceptional versatility in a variety of tasks from logical reasoning and comprehension to creative writing, often accomplished via

methods like prompting, fine-tuning, and human-in-the-loop reinforcement learning.

The astronomy discipline presents both a unique challenge and a fertile ground for the application of LLMs. The corpus of scholarly texts in astronomy likely constitutes but a minuscule portion of the data on which generic LLMs are trained, resulting in limitations like hallucinations in favor of more “generic” responses. Only about 2.5% of LLaMA-2’s training set, for example, likely comes from arXiv, of which less than 5% belongs to the astronomy literature. The nature of astronomical research, on the other hand, often involves cross-disciplinary insights due to universally applicable physical processes. When well-curated, LLMs could meaningfully assist with this effort, such as through hypothesis generation.

Existing scales based on in-context prompting and instruction learning, primarily involving GPT-4, have already demonstrated significant potential for generating substantive hypotheses (Ciucă and Ting, 2023; Ciucă et al., 2023). Further, the astronomy community’s “open sky” policy, which grants public access to the majority of its datasets either immediately or after a brief proprietary period (Almeida et al., 2023; Fabricius et al., 2021), pairs well with the wealth of resources available in archives like NASA’s Astrophysics Data System (Accomazzi et al., 2015; Borgman and Wofford, 2021). Such an open-access policy can facilitate deep engagement with the astronomical literature.

Despite their general capabilities, LLMs frequently lag behind specialized, smaller models in

*Lead contribution. Email: joshtn@seas.upenn.edu

†Major contribution.

domain-specific applications. This disparity stems from two primary factors: (i) the eclectic nature of the pre-training datasets, which dilutes the focus on specialized subjects in favor of general predictive performance, and (ii) the design ethos of LLMs as “foundation models” aimed at subsequent fine-tuning tailored to specific tasks. The existing landscape for LLMs in astronomy remains limited, however. To our knowledge, the only specialized model is astroBERT (Grezes et al., 2021), which has 110 million parameters, fine-tuned on nearly 400,000 ADS papers. As a non-generative model, however, astroBERT’s utility remains primarily limited to discriminative tasks.

Motivated by these gaps, we present AstroLLaMA, a state-of-the-art generative language model fine-tuned from LLaMA-2. Our model leverages a corpus of 300,000 astronomy abstracts from arXiv and boasts an architecture approximately 67 times larger than that of astroBERT. AstroLLaMA aspires to build upon astroBERT’s foundation by offering more improved performance in generating specialized information and broader fine-tuning opportunities for astronomical research. We describe our methodology in Sec. 2, provide some evaluation results in Sec. 3, and finally concluding with some remarks in Sec. 4.

2 AstroLLaMA

In this section, we discuss AstroLLaMA’s implementation, focusing on the curation of its dataset, base model architecture, and fine-tuning settings.

2.1 Dataset

We derive our dataset from the arXiv repository, available on Kaggle.^a Our curated subset focuses on papers classified under the astrophysics category (astro-ph), resulting in a collection of 326,238 articles spanning from April 1992 to July 2023. We extract these papers’ abstracts to form a corpus consisting of approximately 95 million tokens. The median length of these abstracts is 291 tokens. To enable effective model evaluation, we randomly designate 20% of this curated dataset for testing.

2.2 Base model

Our base model is LLaMA-2, a 6.7 billion-parameter model developed by Meta (Meta, 2023). Originally pre-trained on a corpus containing 2 trillion tokens, LLaMA-2 features a context window

^a<https://www.kaggle.com/Cornell-University/arxiv>

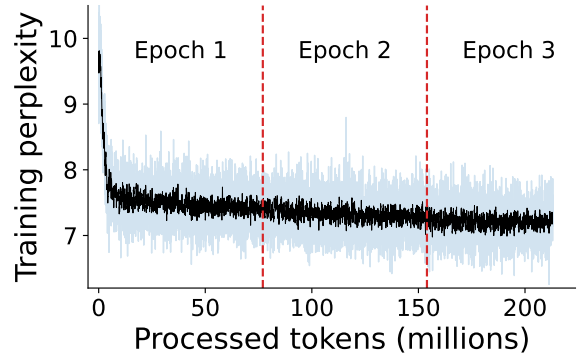


Figure 1: Learning curve of AstroLLaMA during its fine-tuning on the arXiv astrophysics dataset. The figure tracks the evolution of perplexity, a measure of the model’s next-token prediction performance. The light blue curve shows the training perplexity after each parameter update step, while the dark black curve provides a smoothed average of the same metric taken over every 10-step interval.

of 4,096 tokens. For tokenization, the model employs a bytepair encoding strategy (Sennrich et al., 2016; Kudo and Richardson, 2018), with a vocabulary of 32,000 unique tokens.

2.3 Fine-tuning settings

We rely on our curated training set, which includes 77 million tokens. The setting of the fine-tuning phase largely follows from Meta (2023). First, special [BOS] (Beginning Of Sequence) and [EOS] (End Of Sequence) tokens are prepended and appended to each training sequence. These sequences are then concatenated and divided into fixed-length chunks, each comprising 512 tokens.

We follow the causal language modeling objective employed during the model’s pre-training phase, where the next token is to be predicted using its preceding context. We use the AdamW optimizer (Loshchilov and Hutter, 2018) with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-5}$ and a batch size of 32. The learning rate follows a cosine schedule with a linear warmup to a peak value of 3×10^{-4} in the first 10% of the optimization steps and a final learning rate of 10% of its peak. Additional settings include weight decay and gradient clipping values of 0.1 and 1.0, respectively. Note that these hyperparameters are set according to LLaMA-2’s pre-training phase.

We fine-tune LLaMA over nearly three epochs, corresponding to about 230 million processed tokens, using four NVIDIA A100 GPUs each equipped with 40GB of VRAM. To achieve re-

Original abstract

The Magellanic Stream (MS) - an enormous ribbon of gas spanning 140° of the southern sky trailing the Magellanic Clouds - has been exquisitely mapped in the five decades since its discovery. However, despite concerted efforts, no stellar counterpart to the MS has been conclusively identified. This stellar stream would reveal the distance and 6D kinematics of the MS, constraining its formation and the past orbital history of the Clouds. We have been conducting a spectroscopic survey of the most distant and luminous red giant stars in the Galactic outskirts. From this dataset, we have discovered a prominent population of 13 stars matching the extreme angular momentum of the Clouds, spanning up to 100° along the MS at distances of 60–120 kpc. Furthermore, these kinematically-selected stars lie along a $[α/Fe]$ -deficient track in chemical space from $-2.5 < [Fe/H] < -0.5$, consistent with their formation in the Clouds themselves. We identify these stars as high-confidence members of the Magellanic Stellar Stream. Half of these stars are metal-rich and closely follow the gaseous MS, whereas the other half are more scattered and metal-poor. We argue that the metal-rich stream is the recently-formed tidal counterpart to the MS, and speculate that the metal-poor population was thrown out of the SMC outskirts during an earlier interaction between the Clouds. The Magellanic Stellar Stream provides a strong set of constraints - distances, 6D kinematics, and birth locations - that will guide future simulations towards unveiling the detailed history of the Clouds.

Completed by LLaMA-2

The Magellanic Stream (MS) - an enormous ribbon of gas spanning 140° of the southern sky trailing the Magellanic Clouds - has been exquisitely mapped in the five decades since its discovery. However, despite concerted efforts, no stellar counterpart to the MS has been conclusively identified. This stellar stream would reveal the distance and 6D kinematics of the MS, constraining its formation and the past orbital history of the Clouds. We find a stellar stream in the MS, with a 6D velocity of 100 km s⁻¹, and a distance of 100 kpc. The stream is 100 pc wide and 1000 pc long, and is consistent with the MS. The stream is 1000 times more massive than the Magellanic Clouds, and is likely to be a tidal stream from the Large Magellanic Cloud.

Completed by GPT-4

The Magellanic Stream (MS) - an enormous ribbon of gas spanning 140° of the southern sky trailing the Magellanic Clouds - has been exquisitely mapped in the five decades since its discovery. However, despite concerted efforts, no stellar counterpart to the MS has been conclusively identified. This stellar stream would reveal the distance and 6D kinematics of the MS, constraining its formation and the past orbital history of the Clouds. We propose a novel approach to search for the elusive stellar counterpart of the Magellanic Stream. By employing advanced photometric and spectroscopic techniques, combined with state-of-the-art data mining and machine learning algorithms, we aim to systematically analyze and cross-reference multi-wavelength survey data to identify potential candidate stars belonging to the hypothesized stellar stream. Our methodology involves leveraging the distinctive chemical signatures and proper motions associated with stars originating from the Magellanic Clouds, allowing us to differentiate them from the foreground and background contaminants. If successful, this discovery would not only provide essential insights into the origin and evolution of the Magellanic Clouds, but also enhance our understanding of the interaction between the Magellanic Stream and the Milky Way. Furthermore, the proposed approach establishes a powerful framework for investigating other unresolved galactic structures and tracing their underlying dynamical histories.

Completed by AstroLLaMA

The Magellanic Stream (MS) - an enormous ribbon of gas spanning 140° of the southern sky trailing the Magellanic Clouds - has been exquisitely mapped in the five decades since its discovery. However, despite concerted efforts, no stellar counterpart to the MS has been conclusively identified. This stellar stream would reveal the distance and 6D kinematics of the MS, constraining its formation and the past orbital history of the Clouds. We 1) present a new method to select MS candidate stars from wide field optical imaging, 2) apply this method to the MS, 3) compare the results to existing catalogs, and 4) identify the most promising MS candidate stars. We apply our method to 2547 stars in the Gaia EDR3 catalog that have Gaia proper motions consistent with the MS and use these as a pilot sample to optimize our method. We then apply our method to the entire Gaia EDR3 catalog to identify 46,000 stars that are candidates for the MS, and we use the Gaia-ESO Survey to identify 13,000 of these stars that are likely to be MS candidates. We identify 266 stars with (G-K) colours and $[Fe/H]$ that are consistent with previous estimates of the MS. These stars have proper motions and radial velocities that are consistent with the MS, and we place them on a 3D map of the MS. We find a previously unknown stellar stream, which we call the Southeastern Stream (SES), in the direction of the Clouds and MS. The SES has an M31 distance of $32.2^{+/-}0.5$ kpc and a 6D velocity of $136^{+/-}10$ km/s. The SES is about 20 times brighter in stars than the MS and about 40 times brighter in stars than the SMC. The SES is also 2.7 times more metal-rich than the MS, which suggests that it is a separate population.

Figure 2: Completion of an astronomy abstract from the arXiv database (ID: 2306.15719) using three different models: GPT-4, LLaMA-2, and AstroLLaMA. Each model is prompted with the same short text snippet, highlighted in their respective boxes, and then produces the rest of the abstract. Two authors of this paper subsequently judge the quality of each completed abstract. GPT-4 tends to produce over-generic statements, while LLaMA-2 often gives off-topic generations. AstroLLaMA demonstrates the most robust completion, offering more relevant concepts and deeper insights specific to the field of astronomy, thus significantly outperforming LLaMA-2 and GPT-4.

source efficiency, we employ 4-bit quantization of the model’s parameters and utilize LoRA, a fine-tuning technique based on low-rank matrix decomposition (Hu et al., 2021). Specifically, we set LoRA’s hyperparameters α and dropout rate to 32 and 0.05, respectively. This process is implemented using Hugging Face’s library in Python.

2.4 Fine-tuning evaluation

Fig. 1 depicts the performance of AstroLLaMA during its fine-tuning phase. Here, we present perplexity, a commonly used metric for evaluating causal language models. Perplexity is defined as the exponentiation of the training loss, with lower values indicating a better fit.

Our initial observations reveal that LLaMA-2 performs suboptimally on our dataset, with an average perplexity close to 10. By the conclusion of three epochs, AstroLLaMA achieves an average perplexity of 6.55. This represents a 32.5% reduction in perplexity compared to the base LLaMA-2 model, signifying a substantial improvement in the model’s new-token prediction accuracy. Considering LLaMA-2 as a strong pre-trained baseline for language modeling, we believe this performance improvement is substantial in this application.

3 Results

As illustrated in the previous section, AstroLLaMA outperforms its pre-trained counterpart, LLaMA-2, in terms of context-awareness during token prediction within astronomy abstracts. To delve deeper into the advantages of fine-tuning, we assess AstroLLaMA’s general abilities in two key aspects: *text generation* and *embedding space quality*. We compare its performance against multiple models, including LLaMA-2, GPT-4 and GPT-3 (ada-002) to provide a comprehensive evaluation.

3.1 Text generation

We task AstroLLaMA, LLaMA-2 and GPT-4 with completing a number of astronomy abstracts, allowing us to gauge their ability to comprehend the context and generate a meaningful continuation. Fig. 2 presents an example. In particular, we give each model the first few sentences of an abstract as a prompt and use that model to generate the rest of the abstract. For GPT-4, we utilize ChatGPT and instruct it to limit the completion to a single paragraph. AstroLLaMA and LLaMA-2 are deployed using standard sampling methods, with the temperature set to 0.3 and a maximum new tokens limit of 1,024. We find that altering the temperature setting

does not substantively improve LLaMA-2’s results.

Our observations on all generated abstracts largely echo the patterns depicted in Fig. 2. LLaMA-2 frequently deviates from the intended context after generating only a short and often off-topic continuation, resulting in inferior completions. While GPT-4 produces more coherent text, its responses are too generic to capture the nuanced understanding required in the astronomy domain. Even when explicitly prompted to focus on astronomy-related topics, GPT-4’s generated text remains largely off-target or generically applicable rather than domain-specific.

In stark contrast, AstroLLaMA exhibits remarkable context-awareness in its completions by showing a deep understanding of astronomical concepts. In Fig. 2, for example, AstroLLaMA comprehends that an effective search for stars in the Magellanic Stream involves a three-step process: initial wide-field imaging, followed by refinement using astrometric data from Gaia, and then further curation with spectroscopic data. The model also understands Gaia-ESO is surveying the southern sky and hence can observe (part of) the Magellanic Stream. It also demonstrates nuanced knowledge of the Magellanic Stream, understanding the importance of bifurcation within the stream. As a result, it appropriately completes the text by discussing the southeast stream and exploring metallicity differences to ascertain their origins.

3.2 Embedding space quality

We assess models’ ability to reflect semantic similarities among astronomy texts. We randomly choose 10,000 abstracts from our dataset and embed them using AstroLLaMA and GPT-3. Specifically, we use OpenAI’s API to invoke the text embedding function for GPT-3 (ada-002). To get text embeddings from AstroLLaMA, we pass an input through the model and extract its final hidden states, which contain embeddings for all tokens in the input. Then, we omit the [BOS] token and take the average of all other tokens’ embeddings to get the final result. For each pair of abstracts we calculate their cosine similarity (the normalized dot product) between on their vector embeddings.

The top panel of Fig. 3 presents the distribution of these pairwise similarities for the two embedding methods. We find that the embeddings by GPT-3 are overly generic with similarities clustering around relatively high values of 0.7–0.9, sug-

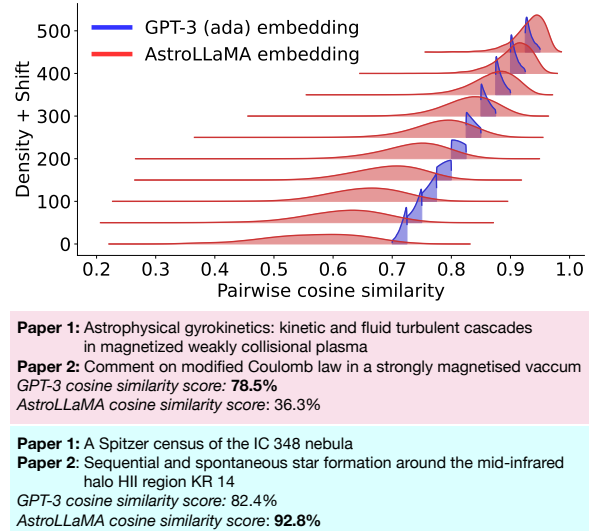


Figure 3: *Top*: Distribution of pairwise cosine similarities among 10,000 randomly selected abstracts from our corpus, divided into 10 equal bins based on similarity levels from GPT-3. *Bottom*: Two representative examples illustrating divergent cosine similarity values when comparing AstroLLaMA and GPT-3 embeddings.

gesting a lack of discriminative power (most papers are embedded very similarly). AstroLLaMA’s embeddings, on the other hand, exhibit much higher variance within each bin. This suggests that our fine-tuned model is more adept at representing the specialized semantic variance inherent to the field of astronomy, which may enable a more granular representation of astronomical content and can facilitate higher-quality document retrieval and semantic analysis.

The bottom panel of Fig. 3 provides two representative examples where AstroLLaMA and GPT-3 classifications diverge. In the first example, GPT-3 fixates on the keyword “magnetized,” resulting in an inflated similarity score despite the contents being markedly different. AstroLLaMA, on the other hand, successfully distinguishes between these disparate contexts. In the second example, AstroLLaMA accurately identifies that the study of Spitzer is closely related to star formation. GPT-3, however, fails to make this connection due to the absence of matching keywords.

4 Conclusion

In this work, we introduce AstroLLaMA, a 7-billion-parameter language model fine-tuned on a dataset encompassing over 300,000 abstracts from astronomical research papers. Compared to its base model, LLaMA-2, and even GPT-4, a cur-

rent state-of-the-art general LLM, AstroLLaMA exhibits marked improvements in generating high-quality abstracts and a competent grasp of relevant information in this specialized literature.

The efficacy of AstroLLaMA demonstrated in this paper suggests a multitude of avenues worthy of exploration for subsequent work. With well-curated instruction datasets, researchers can fine-tune our model to perform tasks such as question answering, scientific paper summarization and academic writing assistance. Combining AstroLLaMA with other information retrieval models can lead to promising systems for hypothesis generation. Finally, AstroLLaMA is a potential candidate to be incorporated into specialized multi-modal models (Liu et al., 2023), going beyond the limits of text in astronomical research.

AstroLLaMA, nevertheless, is not without limitations. During its evaluation, the most salient drawback we find is the model’s knowledge gaps in certain areas of astronomy. In Fig. 2, for example, AstroLLaMA’s estimation of potential star candidates from Gaia-ESO data is notably inaccurate. Another concern lies in the model’s tendency to generate hallucinated or fictitious numerical data, an issue most likely attributed to our simple focus on next-token prediction—a pure NLP objective—rather than explicitly steering the model toward factual accuracy. Achieving a desirable balance of “faithfulness” (respecting scientific evidence and accuracy) and “creativity” (being able to come up with interesting hypotheses) remains an open challenge in research at the intersection of generative models and other scientific disciplines.

There are a number of on-going efforts to address the limitations of AstroLLaMA as well as explore its broad capabilities in this sphere. We are in the process of enriching AstroLLaMA’s training data by including each paper’s full LaTeX sources, going beyond its abstracts and thereby increasing the token count by approximately two orders of magnitude. Although this requires a non-trivial data quality control procedure, it will almost certainly improve our model’s predictive performance substantially, making it even more adapted to this literature and less prone to hallucination. A more systematic evaluation of AstroLLaMA—including a larger set of candidate abstracts for completion, a more well-defined evaluation scheme and a larger, more diverse set of judging experts—will lead to more grounded comparison with state-

of-the-art models. Finally, the potential of AstroLLaMA to generate high-quality and creative hypotheses through novel prompting and fine-tuning techniques is being extensively studied.

AstroLLaMA stands as a compelling prototype for specialized LLMs in astronomy, showing superior context-aware capabilities compared to GPT-4 despite having much fewer parameters. Our methodology is simple and general enough for researcher to explore even more specific areas of astrophysics or even to be adapted to other areas of scientific research.

We have made AstroLLaMA’s weights, training data and code for reproducibility publicly available to researchers who are aiming to leverage LLMs for astronomy-centric applications. Along with this, we are establishing various “playgrounds” on Hugging Face to invite interested readers to explore AstroLLaMA and further refine this robust starting point for a variety of relevant downstream applications.^b

Acknowledgments

We thank the Microsoft Accelerate Foundation Models Academic Research Initiative. Access to advanced AI capabilities from Microsoft Research has greatly accelerated our work in applying language models to automate the analysis of the astronomical literature. We also thank the anonymous reviewers who gave useful insights and suggestions, especially on the potential applications of AstroLLaMA within and beyond astronomy.

Ethics Statement

We obtain the pre-trained weights for LLaMA-2 from Meta, which offers these models for download on Hugging Face. The arXiv dataset used in this paper is publicly available on Kaggle. While we have demonstrated that AstroLLaMA is capable of generating high-quality, relevant abstracts for astronomical research papers, we have noted that it has the potential to generate inaccurate data and measurements. This should serve as a caution for researchers aiming to use this model for downstream tasks, and we invite the adoption of alignment strategies in future work to ameliorate this issue.

^bAll details can be found at <https://huggingface.co/universeTBD/astrollama>.

References

- Google AI PaLM 2. <https://ai.google/discover/palm2/>.
- A. Accomazzi, M. J. Kurtz, E. A. Henneken, R. Chyla, J. Luker, C. S. Grant, D. M. Thompson, A. Holachek, R. Dave, and S. S. Murray. 2015. *ADS: The Next Generation Search Platform*. In *Open Science at the Frontiers of Librarianship*, volume 492 of *Astronomical Society of the Pacific Conference Series*, page 189.
- Andrés Almeida, Scott F. Anderson, Maria Argudo-Fernández, Carles Badenes, Kat Barger, Jorge K. Barrera-Ballesteros, Chad F. Bender, Erika Benitez, Felipe Besser, Dmitry Bizyaev, Michael R. Blanton, John Bochanski, Jo Bovy, William Nielsen Brandt, Joel R. Brownstein, Johannes Buchner, Esra Bulbul, Joseph N. Burchett, Mariana Cano Díaz, Joleen K. Carlberg, Andrew R. Casey, Vedant Chandra, Brian Cherinka, Cristina Chiappini, Abigail A. Coker, Johan Comparat, Charlie Conroy, Gabriella Contardo, Arlin Cortes, Kevin Covey, Jeffrey D. Crane, Katia Cunha, Collin Dabbieri, James W. Davidson Jr. au2, Megan C. Davis, Nathan De Lee, José Eduardo Méndez Delgado, Sebastian Demasi, Francesco Di Mille, John Donor, Peter Dow, Tom Dwelly, Mike Eracleous, Jamey Eriksen, Xiaohui Fan, Emily Farr, Sara Frederick, Logan Fries, Peter Frinchaboy, Boris T. Gaensicke, Junqiang Ge, Consuelo González Ávila, Katie Grabowski, Catherine Grier, Guillaume Guiglion, Pramod Gupta, Patrick Hall, Keith Hawkins, Christian R. Hayes, J. J. Hermes, Lorena Hernández-García, David W. Hogg, Jon A. Holtzman, Hector Javier Ibarra-Medel, Alexander Ji, Paula Jofre, Jennifer A. Johnson, Amy M. Jones, Karen Kinemuchi, Matthias Kluge, Anton Koekemoer, Juna A. Kollmeier, Marina Kounkel, Dhanesh Krishnarao, Mirko Krumpel, Ivan Lacerna, Paulo Jakson Assuncao Lago, Chervin Laporte, Ang Liu, Chao Liu, Xin Liu, Alexandre Roman Lopes, Matin Macktoobian, Viktor Malanushenko, Dan Maoz, Thomas Masseron, Karen L. Masters, Gal Matijevic, Aidan McBride, Ilija Medan, Andrea Merloni, Sean Morrison, Natalie Myers, Szabolcs Mészáros, C. Alenka Negrete, David L. Nidever, Christian Nitschelm, Audrey Oravetz, Daniel Oravetz, Kaike Pan, Yingjie Peng, Marc H. Pinsonneault, Rick Pogge, Dan Qiu, Anna Barbara de Andrade Queiroz, Solange V. Ramirez, Hans-Walter Rix, Daniela Fernández Rosso, Jessie Runnoe, Mara Salvato, Sebastian F. Sanchez, Felipe A. Santana, Andrew Saydjari, Conor Sayres, Kevin C. Schlafman, Donald P. Schneider, Axel Schwöpe, Javier Serna, Yue Shen, Jennifer Sobeck, Ying-Yi Song, Diogo Souto, Taylor Spoo, Keivan G. Stassun, Matthias Steinmetz, Ilya Straumit, Guy Stringfellow, José Sánchez-Gallego, Manuchehr Taghizadeh-Popp, Jamie Tayar, Ani Thakar, Patricia B. Tissera, Andrew Tkachenko, Hector Hernandez Toledo, Benny Trakhtenbrot, Jose G. Fernandez Trincado, Nicholas Troup, Jonathan R. Trump, Sarah Tuttle, Natalie Ullio, Jose Antonio Vazquez-Mata, Pablo Vera Alfaro, Sandro Villanova, Stefanie Wachter, Anne-Marie Weijmans, Adam Wheeler, John Wilson, Leigh Wojno, Julien Wolf, Xiang-Xiang Xue, Jason E. Ybarra, Eleonora Zari, and Gail Zasowski. 2023. *The eighteenth data release of the sloan digital sky surveys: Targeting and first spectra from sdss-v*.
- Christine L. Borgman and Morgan F. Wofford. 2021. *From Data Processes to Data Products: Knowledge Infrastructures in Astronomy*. *arXiv e-prints*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pilla, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. *PaLM: Scaling Language Modeling with Pathways*.
- Ioana Ciucă and Yuan-Sen Ting. 2023. *Galactic ChitChat: Using Large Language Models to Converse with Astronomy Literature*. *arXiv e-prints*.
- Ioana Ciucă, Yuan-Sen Ting, Sandor Kruk, and Kartheik Iyer. 2023. *Harnessing the Power of Adversarial Prompting and Large Language Models for Robust Hypothesis Generation in Astronomy*. *arXiv e-prints*.
- C. Fabricius, X. Luri, F. Arenou, C. Babusiaux, A. Helmi, T. Muraveva, C. Reylé, F. Spoto, A. Vallenari, T. Antoja, E. Balbinot, C. Barache, N. Bauchet, A. Bragaglia, D. Busonero, T. Cantat-Gaudin, J. M. Carrasco, S. Diakité, M. Fabrizio, F. Figueras, A. Garcia-Gutierrez, A. Garofalo, C. Jordi, P. Kervella, S. Khanna, N. Leclerc, E. Licata, S. Lambert, P. M. Marrese, A. Masip, P. Ramos, N. Robichon, A. C. Robin, M. Romero-Gómez, S. Rubele, and M. Weiler. 2021. *igaia/iearly data release 3*. *Astronomy & Astrophysics*, 649:A5.
- Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J. Kurtz, Golnaz Shapourian, Edwin Henneken, Carolyn S. Grant, Donna M. Thompson, Roman Chyla, Stephen McDonald, Timothy W. Hostetler, Matthew R. Templeton, Kelly E. Lockhart, Nemanja Martinovic, Shinyi Chen, Chris Tanner, and Pavlos Protopapas. 2021. *Building astroBERT, a language model for Astronomy & Astrophysics*.

- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [LoRA: Low-Rank Adaptation of Large Language Models](#). *arXiv e-prints*.
- Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#).
- Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
- Meta. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. <https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/>.
- OpenAI. 2023. [GPT-4 Technical Report](#).
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [LLaMA: Open and Efficient Foundation Language Models](#).