The Scaling Law in Stellar Light Curves

Jia-Shu Pan [*1 2 3] **Yuan-Sen Ting** [*2 4 5] **Yang Huang** [1 3] **Jie Yu** [2 6 7] **Ji-Feng Liu** [3 1]

Abstract

Analyzing time series of fluxes from stars, known as stellar light curves, can reveal valuable information about stellar properties. However, most current methods rely on extracting summary statistics, and studies using deep learning have been limited to supervised approaches. In this research, we investigate the scaling law properties that emerge when learning from astronomical time series data using self-supervised techniques. By employing the GPT-2 architecture, we show the learned representation improves as the number of parameters increases from 10^4 to 10^9, with no signs of performance plateauing. We demonstrate that a self-supervised Transformer model achieves 3-10 times the sample efficiency compared to the state-of-the-art supervised learning model when inferring the surface gravity of stars as a downstream task. Our research lays the groundwork for analyzing stellar light curves by examining them through large-scale auto-regressive generative models.

1. Introduction

Light curves, temporal brightness variations of celestial objects, can reveal much of the astrophysics of these systems and aid in the discovery of fleeting transient phenomena (Aerts et al., 2010). This realization has prompted a rapidly growing field in the last decade, with a myriad of missions such as *Kepler* (Koch et al., 2010), TESS (Ricker et al., 2014), and ZTF (Bellm et al., 2018) drastically transforming the landscape of modern-day astronomy. This trend is further accelerating, with upcoming surveys such as the Rubin Observatory (Ivezić et al., 2019) and SiTian (Liu et al., 2020) aiming to collect light curves from up to 10 billion objects from the sky within this decade.

Despite the flood of time series data in astronomy, characterizing these light curves to perform downstream label inferences or identify out-of-distribution objects remains an unresolved problem. Traditional methods often rely on crude summary statistics (Bastien et al., 2013). For example, in the field of asteroseismology, which involves characterizing light curves from stars, the established method calculates the power spectrum of the light curves and extracts the frequency of the maximum oscillation power, $\nu_{\rm max}$, which highly correlates with stellar masses and ages (Aerts et al., 2010). Therefore, asteroseismology has been the primary way to achieve golden measurements of how old stars are. However, such summary statistics do not fully harness all the information contained in the light curves and are often difficult to determine due to heterogeneous noise levels (Stello et al., 2022) and irregular cadences (Hambleton et al., 2023).

To overcome this limitation, a growing body of literature has developed machine learning methods for time-domain astronomy (Blancato et al., 2022; Pan et al., 2024). In particular, for stellar light curves, Transformer models (Vaswani et al., 2017) is receiving an increasing interest (Donoso-Oliva et al., 2023). For example, Pan et al. (2024) customized Transformer models for light curves and demonstrated that, in the case of extracting the surface gravity (or $\log g$) of stars, the ability of Transformer models to extract long-range information can yield state-of-the-art performance in a supervised learning setting.

While deep learning has shown promise in characterizing stellar light curves, most studies have been limited to discriminative supervised learning (Hon et al., 2017; Sayeed et al., 2021). Such methods often fall short in applications due to two main reasons. First, there is a lack of existing ground truth labels for the data. For example, while the

[*]Equal contribution [1]School of Astronomy and Space Science, University of Chinese Academy of Sciences, Beijing 100049, China [2]Research School of Astronomy & Astrophysics, Australian National University, Cotter Rd., Weston, ACT 2611, Australia [3]CAS Key Lab of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China [4]School of Computing, Australian National University, Acton, ACT 2601, Australia [5]Department of Astronomy, The Ohio State University, 140 West 18th Avenue, Columbus, OH 43210, USA [6]Max Planck Institute for Solar System Research, Justus-von-Liebig-Weg 3, 37077 Göttingen, Germany [7]Heidelberg Institute for Theoretical Studies (HITS) gGmbH, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany. Correspondence to: Jia-Shu Pan <jspan@smail.nju.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Rubin Observatory will collect tens of billions of time series data, the number of existing labels (obtained from other more observationally expensive means – such as monitoring eclipsing binaries) will remain on the order of thousands (Kirk et al., 2016). Second, there is a need for a generative model that can extract robust, generalizable representations suitable for other downstream tasks and out-of-distribution search. And solving these bottlenecks is critical to limit redundant resource dedications for different pipelines (Hon et al., 2017; Sayeed et al., 2021; Pan et al., 2024) and downstream tasks.

This limitation has called for a different approach that can utilize unlabeled data and can generalise for various downstream tasks. A prominent candidate would be training foundational models through self-supervised learning as such models have shown promises in different domains such as text data and images (Brown et al., 2020; Bai et al., 2023b). These models, especially Large Language Models (LLMs), exhibit performance unmatched by their smaller variants and emergent abilities such as in-context learning (Brown et al., 2020). Furthermore, in recent years, foundational models in other domains has observed the scaling law, which is the empirical observation that the next-token prediction loss of Transformer-based autoregressive generative models scales in a power law with increasing compute budgets, model size and the number of training tokens (Kaplan et al., 2020; Henighan et al., 2020; Hoffmann et al., 2022). Applying the same idea to astronomical time series data invites two questions: (1) Would the same self-supervised generative models based on Transformer exhibit the scaling law on astronomical time series data? (2) Would a better next-token prediction loss obtained by scaling up translate to better performance in downstream tasks of interest?

2. *Kepler* Mission's Stellar Light Curves

In this study, we focus on the stellar light curves from the *Kepler* mission. The *Kepler* mission was a flagship NASA mission that operated from 2009 to 2013. It yielded 4 years of light curves for approximately 200,000 stars at an equal cadence of about 29.4 minutes, containing $\sim$70,000 timestamps and corresponding observed fluxes. In order to compare with the supervised deep learning Transformer-based models (Pan et al., 2024), we follow the same *Kepler* data selection, selecting 17,201 high-quality light curves as our dataset. Some light curve examples are shown in Fig 1. The different panels show the cases with four representative stars with different surface gravity values, where a smaller $\log g$ indicates a larger stellar radius and, consequently, less oscillation in the stellar fluxes.

The light curves are selected based on the availability of the asteroseismic $\log g$ values for these stars, as provided in previous studies (Mathur et al., 2017; Yu et al., 2018; 2020). These studies focus on extracting the properties of stars with the maximum oscillation frequency $\nu_{\max}$, and the prominence of $\nu_{\max}$ makes it an effective way to select light curves that have high signal-to-noise ratios. Furthermore, the radii of the stars are also provided in Berger et al. (2020), which allow for the application of a radius-dependent high-pass filter to remove instrumental effects (García et al., 2011).

After filtering and outlier removal, we segment the light curves into non-overlapping adjacent observations in patches of length 80, with each patch representing a single sample in our training batch. We use every single timestamp as an input token (see Section 3), amounting to 0.7B tokens. We chose a relatively short context window (containing only 80 tokens) because *Kepler* observations are not contiguous and have observational time gaps between the data points. For simplicity, we opted for these short context windows to ensure that all light curves are contiguous without the distraction of time gaps. It is possible that the autoregressive generative models studied here would remain robust even with the presence of time gaps, but we chose to leave this investigation for future study, as the primary goal of this work is to demonstrate the scaling law of the autoregressive generative models. Finally, our dataset is split into a training set and a test set with a ratio of 25:1.

3. Learning Stellar Light Curves with GPT-2

In this study, we adopt the GPT-2 architecture (Radford et al., 2019) to investigate the scaling law for astronomical stellar light curves. We chose the GPT-2 because, unlike some of the later variants like Llama (Touvron et al., 2023) and Qwen (Bai et al., 2023a), GPT-2 provides a more principled and simple architecture that allows us to easily scale the models with different numbers of parameters and with minimal ambiguity. The model has demonstrated its effectiveness and transferability in text (Radford et al., 2019), image (Chen et al., 2020) and beyond in galaxy images[1].

GPT-2 is a decoder-only Transformer with learnable and straightforward positional embeddings and Pre-layer Normalization, in contrast to more recent LLM architecture such as Llama (Touvron et al., 2023) and Qwen (Bai et al., 2023a), which instead use Rotary Positional Embedding (Su et al., 2023) and RMSNorm (Zhang & Sennrich, 2019). We train all our models following the training scheme in Radford et al. (2019). We adopt the AdamW optimizer (Loshchilov & Hutter, 2019). The learning rate is increased

[1]Previously, to the best of our knowledge, scaling laws have only been established in supervised learning in astronomy (Walmsley et al., 2024). Complementary to this paper, as our work was being developed, we recognize that a similar study was independently conducted for galaxy images and released around the same time as ours (Smith et al., 2024).

Table 1. The different autoregressive generative models investigated in this study. We study the four GPT-2 models (default, medium, large, and XL) as proposed in (Radford et al., 2019). We also vary the default GPT-2 model by changing the depth (number of layers), head (number of self-attention heads), as well as width (hidden dimension size) of the Transformer architecture, to allow us to test a wider dynamic range in terms of model complexity.

Parameter (Million)	Depth	Head	Width
0.04	3	4	32
0.15	3	4	64
0.59	3	4	128
2.36	3	4	256
9.44	3	8	512
37.8	12	8	512
85 (GPT-2, default)	12	12	768
308 (medium)	24	16	1024
708 (large)	36	20	1280
1477 (XL)	48	25	1600

linearly to the peak value in 2,000 iterations and then annealed to 2×10^{-4} with a cosine schedule while ensuring the learning rate remains higher than 2×10^{-5}. We also adopt the default batch size of 12 and accumulate gradients for 40 batches, leading to an effective batch size of 480 light curve patches or 38,400 tokens. All models are trained for 200,000 iterations, summing up to 7.7B processed tokens.

For the GPT-2 model to work with time series data, some minor modifications are required. In particular, unlike discrete variable sequences such as texts, where the modeling of conditional distribution can be handled by minimizing the KL divergence between the text distribution and the output categorical distribution, continuous distribution is less straightforward to model. Instead, we carry out next-token prediction by regressing the next token with Huber loss (Huber, 1964), a rectified version of mean square error (MSE) that mitigates the influence of outlier values. Note that this is equivalent to modeling the expectation of the conditional distribution, or performing maximum likelihood estimation with the assumption that the conditional distribution is a constant-variance normal distribution.

Furthermore, the input of GPT-2 models is in the form of tokens from text data. To adapt it to light curves, we replace the token embedding layer and language modeling layer in the model with two learnable Multilayer Perceptrons (MLPs) whose dimension is listed in Table 1.

In addition to the original GPT-2, GPT-2 medium, GPT-2 large, and GPT-2 XL models proposed in Radford et al. (2019), we also train models of different sizes by adjusting the number of layers (depth), number of self-attention heads (Head), and hidden dimension size (width), as listed in Table 1. This is to ensure that we can evaluate the performance of the models with a larger dynamic range in terms of model size, ranging from 10^4 to 10^9 parameters. We note that the largest model that we train, with ~ 1.5B parameters, starts to rival some of the "lightweight" modern-day LLMs (Bai et al., 2023a).

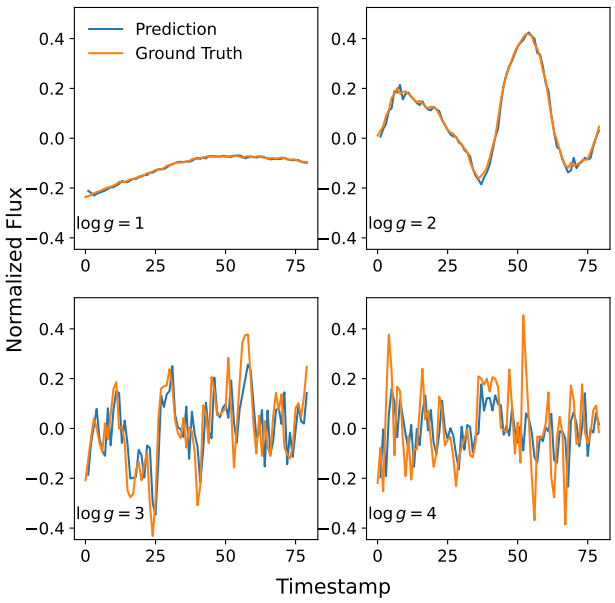

Figure 1. Autoregressive one-step predictions from our GPT-2 XL model with 1.5B parameters. The different panels show four representative light curves with varying surface gravity ($\log g$) values of stars. We perform the next-step prediction conditioning on all the previous steps. The generative model demonstrates the ability to capture the general trend of the light curves, leading to a robust representation of the light curves.

4. Results

The study of *Kepler*'s light curves is a cornerstone in modern-day stellar astrophysics. The temporal variations of the fluxes of stars are the manifestation of physical processes where light bounces within the stellar interior. As such, the light curves are tell-tale signatures of the stellar interior physics that are otherwise inaccessible to us. In this study, we train GPT-2-like models as autoregressive generative models to capture this inherent interior physics. Note that, the models are all initialized completely randomly from $\mathcal{N}(0, 0.02)$ to study the emergence of the scaling law in the generative models trained from scratch.

We will start by showcasing the qualitative results from the GPT-2 XL model with 1.5B parameters. In Figure 1, we show the one-step predictions of all fluxes, where at each step, we perform the prediction conditioning on all

the previous (true) observed fluxes. We note that, given the predominantly stochastic nature of stellar flux variations, i.e., given all N previous steps, the next "token" prediction complies to a distribution, just like in natural languages. As such, we do not expect the predictions to be exact. Specifically, since our prediction is based upon maximal likelihood estimation from a multi-mode distribution, the multi-step prediction would deviate from the truly observed fluxes although the predicted light curve could be statistically the same as the original light curves. Due to this consideration, we only examine the trend with one-step predictions.

As shown in Figure 1, like in natural languages, the autoregressive generative model does learn the "stylistic patterns" of the stellar light curves, where it shows a good understanding of the long timescale oscillations from stars with small surface gravity as opposed to short oscillations from stars with large surface gravity. More importantly, the ability to generally grasp the patterns of the light curves suggests that we could further improve the predictions if a scaling law applies to the model, and the learned latent representations might be robust for further downstream tasks such as predicting the surface gravity (and hence ages) of the stars.

4.1. The Emergence of Scaling Law for Light Curves

While the autoregressive model shows some promise, a key to understanding if astronomical time series data can benefit from a foundational model would be determining if the same scaling law that has been observed in real-life applications also translates to stellar light curves. Such a study has not been established before, and it is also not immediately clear if that would necessarily be the case.

Unlike text data, the sample size of astronomical time series adopted in this study is much smaller. With the current data described in Section 2, the high-quality *Kepler* light curves in this study only have 0.7B tokens. Even if we scale it up to the entire *Kepler* catalog, it would only lead to an order of magnitude larger in terms of training data. This is minuscule compared to modern-day LLMs, which are typically trained on trillions of tokens. But this limitation in data might be compensated by the fact that, the underlying structure of stellar light curves is also much simpler than natural languages, with only finite dynamos driving the stochastic variations of fluxes.

We investigate the scaling law between pretraining loss and compute budget with model sizes spanning $10^4 - 10^9$ parameters. Figure 2 shows the pretraining loss of the test set as a function of the compute budget in PetaFlop(PF)-days[2] for different GPT-2-like models, with individual lines of different colors representing the test loss curve for various compute budgets. We define the critical point as 3.4 epoch, where the local slope of the loss curve becomes shallower than the overall scaling law defined below. The models are trained for 11.4 epochs. The loss curves before the critical point are shown in solid colors, while the curves after the critical points are depicted with more translucent colors.

Following Henighan et al. (2020), we fit a power law to the loss at the critical points as a function[3] of the compute budget C, $\log_{10} L(C) = \alpha_C \log_{10} C + \beta_C$ with $\alpha_C = -0.00448, \beta_C = -2.386$, demonstrating the emergence of a scaling law in stellar light curve data. The scaling suggests that, for a 10 fold larger compute budget, the next-token prediction Huber loss (again, a rectified version of MSE) would improve by 1%. Already, we can predict that next-token prediction Huber loss will improve 3.8% compared to our most capable 1.5B GPT-2 XL model at a compute budget of 10^4 PF-days.

We suspect the somewhat shallow slope, i.e., smaller gain as model size increases, is due to the limited training data size, as scaling laws observed in (Kaplan et al., 2020) are also functions of the dataset size. We aim to explore the dependency on the training set size in our future work.

4.2. Scaling Law in Inferring Stellar Surface Gravity

While we have demonstrated the scaling law in the pretraining loss, in our case, the Huber loss of the next-token prediction, perhaps a more important question pertaining to most astronomical tasks is whether the latent representation acquired from such self-supervised learning also sufficiently exhibits the scaling law. The transferability and scalability to downstream tasks will ultimately determine if a foundational model for astronomical time series data is attainable.

We first investigate how representation quality varies with model depth in the GPT-2 model. For each light curve patch, we extract the embedding of the last token across all layers. The embeddings are then visualized with a UMAP (McInnes et al., 2020) projection color-coded by their $\log g$ in Figure 3. As demonstrated in the figure, the representation at deeper layers shows a higher level of abstraction, leading to more distinct and discernible patterns with the $\log g$ labels. Since the radius, and hence $\log g$, of the stars is the primary factor that determines stellar oscillation (and consequently the emergent variation in the fluxes of stars), this gradual distinct representation shows that the learned representation has a high level of generalizability. We also compare with using only the last token's embedding in specific layers, as

[2] 1 PF-day = $10^{15} \times 24 \times 3600 = 8.64 \times 10^{19}$ floating point operations $\approx$ 3.2 A100 days.

[3] The original expression in Henighan et al. (2020) is $L(C) = (\frac{C_0}{C})^{\alpha_C}$, we apply a base-10 logarithm to the expression for better readability. Throughout this study, we consistently utilize the modified equivalent expression.

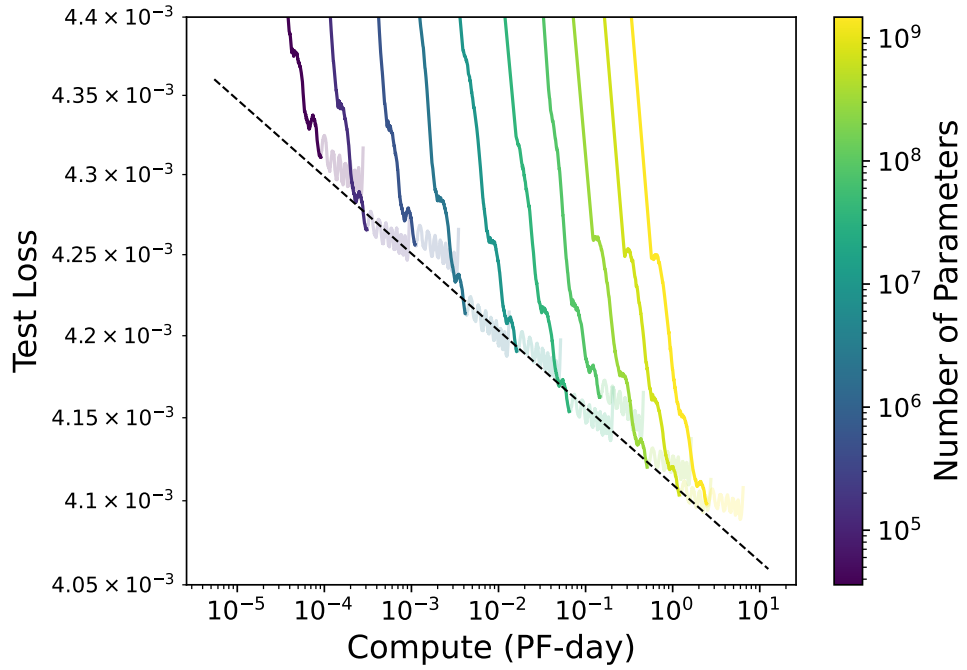

Figure 2. The emergence of scaling law for stellar light curves. We train autoregressive generative models on stellar light curves with different model sizes, ranging from 10^4 to 10^9 parameters. The different lines show the pretraining loss (Huber loss of next-token prediction) as a function of the compute budget for different model sizes. We truncate the training of the models when the slope of the tangent line connecting the loss from the last three epochs becomes shallower than the overall scaling law. The prediction loss plateaus at lower values for larger models, demonstrating that the scaling law also applies to Transformer-based autoregressive generative models when applied to stellar light curves.

shown on the left panels, versus the one where we adopt an average of all tokens at the layer, as shown on the right panels, and found the latter yields cleaner representations. Hence, in the following, we will only adopt the latter for the downstream task.

For the label inference head (mapping from the last layer representation to $\log g$), we adopt a 3-layer MLP with a hidden dimension of 64. As different models have different width, the dimensions for the representation and hence the input dimension of the MLP head also varies. The largest MLP head has 0.11M parameters. We freeze the main generative model and train using a dataset consisting of 10,000 patches for training and 20,000 patches for validation. As mentioned in Section 2, our pretraining dataset contains over 8 million patches from more than 16,000 stars. The 10,000 patches used for the downstream task are selected from this pretraining dataset, corresponding to about $10,000$ stars. This sample size is representative of the number of golden labels obtained through other means in a practical setting. To ensure no information leakage, we use a separate test set of 20,000 patches drawn from the test set, which was not used during the pretraining of the generative models.

All our tests below for the label inference are on this unseen test set. For each model, we train 10 MLP heads, and the one with the best MSE loss on the validation set mentioned above is adopted as the final model. For the optimization, we use Adam optimizer (Kingma & Ba, 2017) with a constant learning rate of 0.001 and adopt the full batch of the training data, as we find that stochastic gradient descent with minibatches does not alter the results, and the former is computationally more efficient. We set a maximum number of epochs to be 1000 and assume an early stop if the validation loss does not improve for over the last 50% of the iterations.

Figure 4 shows the precision of the $\log g$ inference as a function of the number of parameters, represented by the solid orange line. For reference, the next-token prediction loss, as previously discussed, is plotted in blue. The results demonstrate that the $\log g$ inference from the GPT-2-like model also exhibits a prominent scaling law with respect to the number of parameters in the model, and the improvement in this downstream label inference mirrors the next-token predictions. This suggests that the scaling law observed in the next-token prediction goes beyond simply memorizing

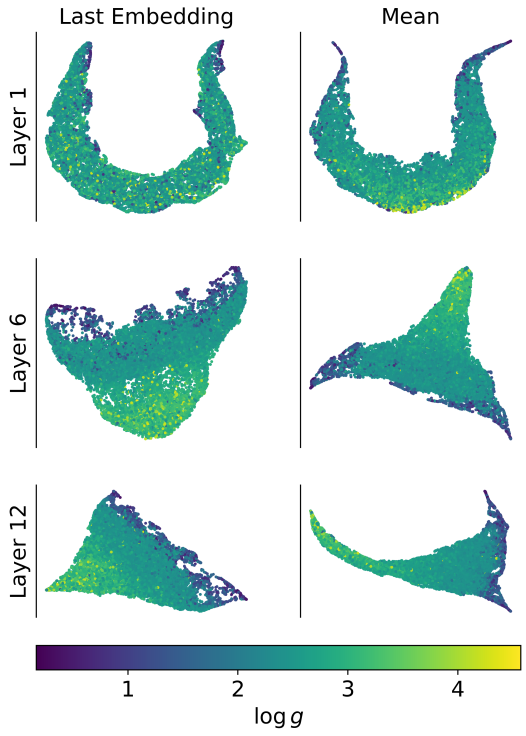

Figure 3. The latent representations extracted from the GPT-2 model. The native representation is a vector with 768 dimensions, which we subsequently visualize in 2D with the UMAP projection. Different columns show the latent embedding representation extracted at different depths in the generative model. The left panels show the case where only the last token is extracted as the embedding, while the right panels show the average of all 80 tokens at any given layer. Embeddings in the deeper layer show a higher level of abstraction. The symbols are color-coded by the surface gravity values of the stars.

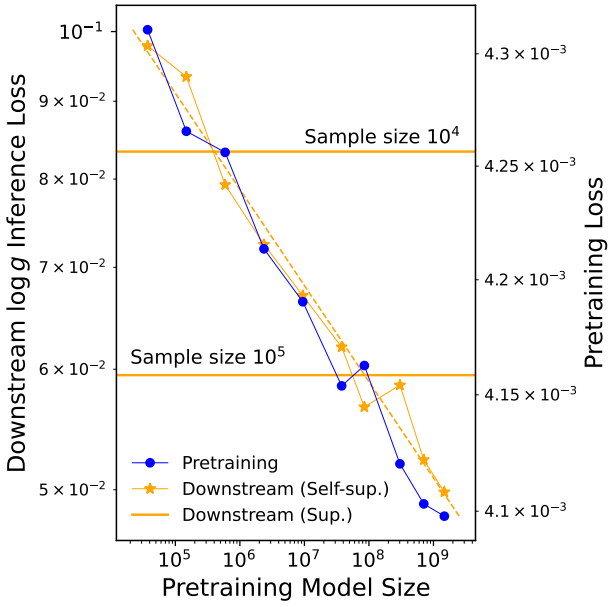

Figure 4. The downstream $\log g$ inference exhibits the scaling law. The solid orange line shows the MSE of the $\log g$ inference derived from the representation of the GPT-2 models through a final MLP head. The MSE is plotted as a function of the number of parameters. We also plot the Huber loss from the next-token prediction, as illustrated in Figure 2, as the blue solid line. The downstream $\log g$ prediction MSE closely follows the next-token prediction loss. We also compare the $\log g$ prediction with the Astroconformer models trained through supervised learning on 10,000 and 100,000 samples. We found that the generative approach in this study, at the 10^8 parameter level, rivals the performance of the state-of-the-art supervised learning model but achieves this with 10 times fewer training data.

the training set, and the model actually learns more robust abstractions from the light curves.

We compare our results from self-supervised learning with GPT-2 models to the current state-of-the-art (SOTA) supervised learning method, Astroconformer (Pan et al., 2024), represented by the solid orange horizontal lines. Astroconformer is a Transformer-based model with 0.86M parameters that learns from the direct relationship between the stellar light curves and $\log g$. They argue that introducing a strong inductive bias is necessary because the labeled training sample is limited. To address this, they combine convolution and self-attention mechanisms to simultaneously capture localized and global correlations, which they found to outperform other existing supervised learning methods primarily based on convolutional neural networks or K-nearest neighbor search (Sayeed et al., 2021). To ensure a fair comparison, the Astroconformer results presented here are trained with the same labeled training samples used for evaluating self-supervised representations. Further, to explore the sample efficiency gain by self-supervised learning, we also evaluate Astroconformer's performance based on 10 times more training samples drawn from the pretraining dataset.

The figure shows that the scaling law enables sufficiently large self-supervised autoregressive models, even with the 0.7B pretraining tokens, to surpass the performance of the current supervised learning models. Astroconformer reaches a $\log g$ prediction MSE of 0.083 and 0.059 when trained on 10,000 samples and 100,000 samples, respectively. Even with the 85M parameters GPT-2 model, the learned latent representation is robust to surpass supervised learning trained on 10 times of data, which indicates they are > 10 times more label-efficient. We fit the power law as before, and find the scaling law for the $\log g$ prediction MSE follows $\log_{10} L(N) = -0.063 \log_{10} N - 0.725$. In other words, if we extrapolate that to the typical industrial

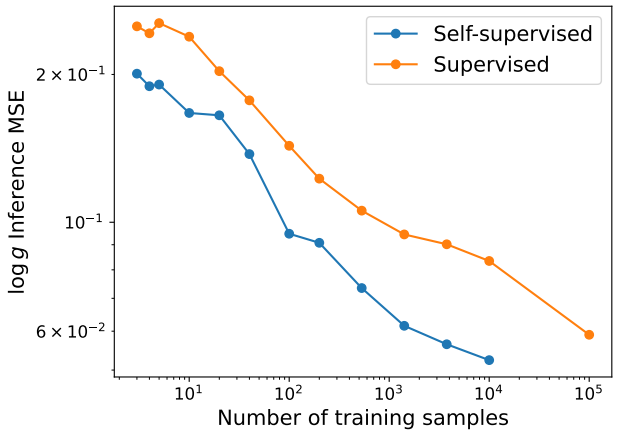

Figure 5. The impact of the labeled training dataset on the downstream $\log g$ inference task. We compare the performance of GPT-2 XL with that of Astroconformer. For any given fixed set of labeled data, the autoregressive generative approach adopted in this study leads to better precision than the supervised method. The self-supervised learning approach requires less than one third of training samples to achieve the same precision in $\log g$ inference.

scale of 10B or 100B, the $\log g$ precision MSE would be 11% and 23% lower than our most capable 1.5B GPT-2 XL model.

In the comparison above, we assume a somewhat optimistic number of labeled samples (10,000) for downstream training. However, a key advantage of training a potentially large-scale foundation model is its ability to generalize with a small number of labeled samples. In Figure 5, we explore how the downstream MSE loss varies as a function of the labeled dataset size using the GPT-2 XL model and contrast the results with Astroconformer's. As shown in Figure 5, pretraining the autoregressive generative models typically leads to the same $\log g$ inference precision but with less than one third of the number of labeled samples.

5. Discussion and Conclusion

In this study, we explore training autoregressive generative models, based on the GPT-2 architecture, on stellar light curves. Stellar light curves have been an instrumental part of modern-day astrophysics to decipher our knowledge about the stellar interior and infer fundamental parameters of the stars. However, by nature, the number of stars with ground-truth labels obtained through other observational means is highly limited. As such, a scalable deep learning approach based on self-supervision, a method that has shown promise primarily in text data, but now also in other modalities, is critical to move the field of asteroseismology forward.

We demonstrate that with stellar light curves, there is also an emergence of a scaling law, where the GPT-2-like models continue to improve the next-token prediction huber loss with a scaling law of $\log_{10} L_{\text{pretraining}}(C) = -0.00448 \log_{10} C - 2.386$. Such a scaling law also translates into a more robust and higher level of abstraction in the learned latent representation. The robust latent representation leads to a scaling law on the downstream inference MSE of the surface gravity of stars, with a scaling law of $\log_{10} L_{\text{downstream}}(N) = -0.063 \log_{10} N - 0.725$. With a labeled training set of 10,000 samples and a model size of > 85 M parameters, the autoregressive models outperform the supervised Transformer model trained from scratch on 100,000 samples, and these results continue to extrapolate to a higher precision with up to 10^9 parameters. At a fixed precision for $\log g$ inference, GPT-2 XL with 1.5B parameters generative model can achieve such precision consistently with 3-10 times fewer labeled samples compared to the mere supervised learning approach.

In this study, we chose the GPT-2 model to demonstrate more principally the scalability of autoregressive generative models for astronomical time series data. Different parts of the models can no doubt be fine-tuned for better performance, both in terms of the choice of tokenization (in our case, simple MLP tokenization) and the context window (we limited to a very short context of 80 tokens). A longer context window, or following ideas from vision Transformers (Bai et al., 2023b), the segmentation of light curves into patches and tokenizing individual patches might be a way forward. Further, here we are limited to the exploration of the scaling law at a fixed training set of 0.7B tokens; an additional ablation study on the training size will also be informative.

Nonetheless, the study demonstrates that the Transformer-based models that have shown wild success in real-world applications are also promising for astronomical time series data. The scalability of the models up to 1.5B parameters with our small training set vividly demonstrates the possibility of having a robust and generalizable foundational model for astronomical time series data with sufficient compute. Beyond asteroseismology, the same method can clearly also translate into other forms of astronomical time series data, such as understanding solar flares (Wei et al., 2006) or classifying and finding out-of-distribution transient events (Ivezić et al., 2019).

Ongoing and future surveys, TESS, ZTF, Rubin and SiTian, are estimated to collect observations at scales of 1T, 14T, 14P, and 100P, respectively. There will be a myriad of time series data, perhaps rivaling the text data on the internet and perhaps beyond. The scaling law motivates the dedication of compute resources to characterize astronomical time series data with these models. We forecast that the $\log g$ prediction MSE can be improved two folds than our

GPT-2 XL model given 10^4 PF-days, i.e., $\sim 10^5$ A100 GPU days. Looking forward, just as with much of the highly non-linear and complex data in real-life scenarios, while the still somewhat simple physical drivers of the light curves of celestial objects can be understood, to a certain extent, with analytical derivations and methods, robust inference and ultimate characterization might still require massive compute. The scalable model demonstrated here is a key step toward unlocking the full potential of astronomical time series data and paving the way for expediting discoveries in the ever-changing dark sky.

Acknowledgements

We thank Mike Smith for constructive discussion and insights. YST acknowledges financial support received from the Australian Research Council via the DECRA Fellowship, grant number DE220101520. We also recognize the invaluable contribution of public data from the *Kepler* mission, supplied by NASA's Ames Research Center. The *Kepler* mission is funded by NASA's Science Mission Directorate. Our special thanks go to the *Kepler* team for making their data openly accessible, thereby facilitating this research.

References

Aerts, C., Christensen-Dalsgaard, J., and Kurtz, D. *Asteroseismology*. Springer, Netherlands, 2010. ISBN 978-1-4020-5178-4. doi: 10.1007/978-1-4020-5803-5.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., and Zhu, T. Qwen technical report, 2023a.

Bai, Y., Geng, X., Mangalam, K., Bar, A., Yuille, A., Darrell, T., Malik, J., and Efros, A. A. Sequential modeling enables scalable learning for large vision models, 2023b.

Bastien, F. A., Stassun, K. G., Basri, G., and Pepper, J. An observational correlation between stellar brightness variations and surface gravity. *Nature*, 500(7463):427–430, aug 2013. doi: 10.1038/nature12419. URL https://doi.org/10.1038%2Fnature12419.

Bellm, E. C., Kulkarni, S. R., Graham, M. J., Dekany, R., Smith, R. M., Riddle, R., Masci, F. J., Helou, G., Prince, T. A., Adams, S. M., Barbarino, C., Barlow, T., Bauer, J., Beck, R., Belicki, J., Biswas, R., Blagorodnova, N., Bodewits, D., Bolin, B., Brinnel, V., Brooke, T., Bue, B., Bulla, M., Burruss, R., Cenko, S. B., Chang, C.-K., Connolly, A., Coughlin, M., Cromer, J., Cunningham, V., De, K., Delacroix, A., Desai, V., Duev, D. A., Eadie, G., Farnham, T. L., Feeney, M., Feindt, U., Flynn, D., Franckowiak, A., Frederick, S., Fremling, C., Gal-Yam, A., Gezari, S., Giomi, M., Goldstein, D. A., Golkhou, V. Z., Goobar, A., Groom, S., Hacopians, E., Hale, D., Henning, J., Ho, A. Y. Q., Hover, D., Howell, J., Hung, T., Huppenkothen, D., Imel, D., Ip, W.-H., Ivezić, Ž ., Jackson, E., Jones, L., Juric, M., Kasliwal, M. M., Kaspi, S., Kaye, S., Kelley, M. S. P., Kowalski, M., Kramer, E., Kupfer, T., Landry, W., Laher, R. R., Lee, C.-D., Lin, H. W., Lin, Z.-Y., Lunnan, R., Giomi, M., Mahabal, A., Mao, P., Miller, A. A., Monkewitz, S., Murphy, P., Ngeow, C.-C., Nordin, J., Nugent, P., Ofek, E., Patterson, M. T., Penprase, B., Porter, M., Rauch, L., Rebbapragada, U., Reiley, D., Rigault, M., Rodriguez, H., van Roestel, J., Rusholme, B., van Santen, J., Schulze, S., Shupe, D. L., Singer, L. P., Soumagnac, M. T., Stein, R., Surace, J., Sollerman, J., Szkody, P., Taddia, F., Terek, S., Sistine, A. V., van Velzen, S., Vestrand, W. T., Walters, R., Ward, C., Ye, Q.-Z., Yu, P.-C., Yan, L., and Zolkower, J. The zwicky transient facility: System overview, performance, and first results. *Publications of the Astronomical Society of the Pacific*, 131(995):018002, dec 2018. doi: 10.1088/1538-3873/aaecbe. URL https://doi.org/10.1088%2F1538-3873%2Faaecbe.

Berger, T. A., Huber, D., van Saders, J. L., Gaidos, E., Tayar, J., and Kraus, A. L. The gaia-kepler stellar properties catalog. i. homogeneous fundamental properties for 186,301 kepler stars, 2020.

Blancato, K., Ness, M. K., Huber, D., Lu, Y., and Angus, R. Data-driven Derivation of Stellar Properties from Photometric Time Series Data Using Convolutional Neural Networks. *The Astrophysical Journal*, 933(2):241, July 2022. doi: 10.3847/1538-4357/ac7563.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative pretraining from pixels. In *International conference on machine learning*, pp. 1691–1703. PMLR, 2020.

Donoso-Oliva, C., Becker, I., Protopapas, P., Cabrera-Vives, G., Vishnu, M., and Vardhan, H. Astromer: A transformer-based embedding for the representation of light curves. *Astronomy & Astrophysics*, 670:

A54, February 2023. ISSN 1432-0746. doi: 10.1051/0004-6361/202243928. URL http://dx.doi.org/10.1051/0004-6361/202243928.

García, R. A., Hekker, S., Stello, D., Gutiérrez-Soto, J., Handberg, R., Huber, D., Karoff, C., Uytterhoeven, K., Appourchaux, T., Chaplin, W. J., Elsworth, Y., Mathur, S., Ballot, J., Christensen-Dalsgaard, J., Gilliland, R. L., Houdek, G., Jenkins, J. M., Kjeldsen, H., McCauliff, S., Metcalfe, T., Middour, C. K., Molenda-Zakowicz, J., Monteiro, M. J. P. F. G., Smith, J. C., and Thompson, M. J. Preparation of Kepler light curves for asteroseismic analyses. *Monthly Notices of the Royal Astronomical Society: Letters*, 414(1):L6–L10, 06 2011. ISSN 1745-3925. doi: 10.1111/j.1745-3933.2011.01042.x. URL https://doi.org/10.1111/j.1745-3933.2011.01042.x.

Hambleton, K. M., Bianco, F. B., Street, R., Bell, K., Buckley, D., Graham, M., Hernitschek, N., Lund, M. B., Mason, E., Pepper, J., Prša, A., Rabus, M., Raiteri, C. M., Szabó, R., Szkody, P., Andreoni, I., Antoniucci, S., Balmaverde, B., Bellm, E., Bonito, R., Bono, G., Botticella, M. T., Brocato, E., Bricman, K. B., Cappellaro, E., Carnerero, M. I., Chornock, R., Clarke, R., Cowperthwaite, P., Cucchiara, A., D'Ammando, F., Dage, K. C., Dall'Ora, M., Davenport, J. R. A., de Martino, D., de Somma, G., Criscienzo, M. D., Stefano, R. D., Drout, M., Fabrizio, M., Fiorentino, G., Gandhi, P., Garofalo, A., Giannini, T., Gomboc, A., Greggio, L., Hartigan, P., Hundertmark, M., Johnson, E., Johnson, M., Jurkic, T., Khakpash, S., Leccia, S., Li, X., Magurno, D., Malanchev, K., Marconi, M., Margutti, R., Marinoni, S., Mauron, N., Molinaro, R., Möller, A., Moniez, M., Muraveva, T., Musella, I., Ngeow, C.-C., Pastorello, A., Petrecca, V., Piranomonte, S., Ragosta, F., Reguitti, A., Righi, C., Ripepi, V., Sandoval, L. R., Stassun, K. G., Stroh, M., Terreran, G., Trimble, V., Tsapras, Y., van Velzen, S., Venuti, L., and Vink, J. S. Rubin observatory lsst transients and variable stars roadmap. *Publications of the Astronomical Society of the Pacific*, 135(1052):105002, nov 2023. doi: 10.1088/1538-3873/acdb9a. URL https://dx.doi.org/10.1088/1538-3873/acdb9a.

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., Hallacy, C., Mann, B., Radford, A., Ramesh, A., Ryder, N., Ziegler, D. M., Schulman, J., Amodei, D., and McCandlish, S. Scaling laws for autoregressive generative modeling, 2020.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models, 2022.

Hon, M., Stello, D., and Yu, J. Deep learning classification in asteroseismology. *Monthly Notices of the Royal Astronomical Society*, 469(4):4578–4583, May 2017. ISSN 1365-2966. doi: 10.1093/mnras/stx1174. URL http://dx.doi.org/10.1093/mnras/stx1174.

Huber, P. J. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

Ivezić, Ž., Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., Alonso, D., AlSayyad, Y., Anderson, S. F., Andrew, J., Angel, J. R. P., Angeli, G. Z., Ansari, R., Antilogus, P., Araujo, C., Armstrong, R., Arndt, K. T., Astier, P., Aubourg, É., Auza, N., Axelrod, T. S., Bard, D. J., Barr, J. D., Barrau, A., Bartlett, J. G., Bauer, A. E., Bauman, B. J., Baumont, S., Bechtol, E., Bechtol, K., Becker, A. C., Becla, J., Beldica, C., Bellavia, S., Bianco, F. B., Biswas, R., Blanc, G., Blazek, J., Blandford, R. D., Bloom, J. S., Bogart, J., Bond, T. W., Booth, M. T., Borgland, A. W., Borne, K., Bosch, J. F., Boutigny, D., Brackett, C. A., Bradshaw, A., Brandt, W. N., Brown, M. E., Bullock, J. S., Burchat, P., Burke, D. L., Cagnoli, G., Calabrese, D., Callahan, S., Callen, A. L., Carlin, J. L., Carlson, E. L., Chandrasekharan, S., Charles-Emerson, G., Chesley, S., Cheu, E. C., Chiang, H.-F., Chiang, J., Chirino, C., Chow, D., Ciardi, D. R., Claver, C. F., Cohen-Tanugi, J., Cockrum, J. J., Coles, R., Connolly, A. J., Cook, K. H., Cooray, A., Covey, K. R., Cribbs, C., Cui, W., Cutri, R., Daly, P. N., Daniel, S. F., Daruich, F., Daubard, G., Daues, G., Dawson, W., Delgado, F., Dellapenna, A., de Peyster, R., de Val-Borro, M., Digel, S. W., Doherty, P., Dubois, R., Dubois-Felsmann, G. P., Durech, J., Economou, F., Eifler, T., Eracleous, M., Emmons, B. L., Fausti Neto, A., Ferguson, H., Figueroa, E., Fisher-Levine, M., Focke, W., Foss, M. D., Frank, J., Freemon, M. D., Gangler, E., Gawiser, E., Geary, J. C., Gee, P., Geha, M., Gessner, C. J. B., Gibson, R. R., Gilmore, D. K., Glanzman, T., Glick, W., Goldina, T., Goldstein, D. A., Goodenow, I., Graham, M. L., Gressler, W. J., Gris, P., Guy, L. P., Guyonnet, A., Haller, G., Harris, R., Hascall, P. A., Haupt, J., Hernandez, F., Herrmann, S., Hileman, E., Hoblitt, J., Hodgson, J. A., Hogan, C., Howard, J. D., Huang, D., Huffer, M. E., Ingraham, P., Innes, W. R., Jacoby, S. H., Jain, B., Jammes, F., Jee, M. J., Jenness, T., Jernigan, G., Jevremović, D., Johns, K., Johnson, A. S., Johnson, M. W. G., Jones, R. L., Juramy-Gilles, C., Jurić, M., Kalirai, J. S., Kallivayalil, N. J., Kalmbach, B., Kantor, J. P., Karst, P., Kasliwal, M. M., Kelly, H., Kessler, R., Kinnison, V., Kirkby, D., Knox, L., Kotov, I. V., Krabbendam, V. L., Krughoff,

K. S., Kubánek, P., Kuczewski, J., Kulkarni, S., Ku, J., Kurita, N. R., Lage, C. S., Lambert, R., Lange, T., Langton, J. B., Le Guillou, L., Levine, D., Liang, M., Lim, K.-T., Lintott, C. J., Long, K. E., Lopez, M., Lotz, P. J., Lupton, R. H., Lust, N. B., MacArthur, L. A., Mahabal, A., Mandelbaum, R., Markiewicz, T. W., Marsh, D. S., Marshall, P. J., Marshall, S., May, M., McKercher, R., McQueen, M., Meyers, J., Migliore, M., Miller, M., Mills, D. J., Miraval, C., Moeyens, J., Moolekamp, F. E., Monet, D. G., Moniez, M., Monkewitz, S., Montgomery, C., Morrison, C. B., Mueller, F., Muller, G. P., Muñoz Arancibia, F., Neill, D. R., Newbry, S. P., Nief, J.-Y., Nomerotski, A., Nordby, M., O'Connor, P., Oliver, J., Olivier, S. S., Olsen, K., O'Mullane, W., Ortiz, S., Osier, S., Owen, R. E., Pain, R., Palecek, P. E., Parejko, J. K., Parsons, J. B., Pease, N. M., Peterson, J. M., Peterson, J. R., Petravick, D. L., Libby Petrick, M. E., Petry, C. E., Pierfederici, F., Pietrowicz, S., Pike, R., Pinto, P. A., Plante, R., Plate, S., Plutchak, J. P., Price, P. A., Prouza, M., Radeka, V., Rajagopal, J., Rasmussen, A. P., Regnault, N., Reil, K. A., Reiss, D. J., Reuter, M. A., Ridgway, S. T., Riot, V. J., Ritz, S., Robinson, S., Roby, W., Roodman, A., Rosing, W., Roucelle, C., Rumore, M. R., Russo, S., Saha, A., Sassolas, B., Schalk, T. L., Schellart, P., Schindler, R. H., Schmidt, S., Schneider, D. P., Schneider, M. D., Schoening, W., Schumacher, G., Schwamb, M. E., Sebag, J., Selvy, B., Sembroski, G. H., Seppala, L. G., Serio, A., Serrano, E., Shaw, R. A., Shipsey, I., Sick, J., Silvestri, N., Slater, C. T., Smith, J. A., Smith, R. C., Sobhani, S., Soldahl, C., Storrie-Lombardi, L., Stover, E., Strauss, M. A., Street, R. A., Stubbs, C. W., Sullivan, I. S., Sweeney, D., Swinbank, J. D., Szalay, A., Takacs, P., Tether, S. A., Thaler, J. J., Thayer, J. G., Thomas, S., Thornton, A. J., Thukral, V., Tice, J., Trilling, D. E., Turri, M., Van Berg, R., Vanden Berk, D., Vetter, K., Virieux, F., Vucina, T., Wahl, W., Walkowicz, L., Walsh, B., Walter, C. W., Wang, D. L., Wang, S.-Y., Warner, M., Wiecha, O., Willman, B., Winters, S. E., Wittman, D., Wolff, S. C., Wood-Vasey, W. M., Wu, X., Xin, B., Yoachim, P., and Zhan, H. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *The Astrophysical Journal*, 873(2): 111, March 2019. doi: 10.3847/1538-4357/ab042c.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

Kirk, B., Conroy, K., Prša, A., Abdul-Masih, M., Kochoska, A., Matijevič, G., Hambleton, K., Barclay, T., Bloemen, S., Boyajian, T., Doyle, L. R., Fulton, B. J., Hoekstra, A. J., Jek, K., Kane, S. R., Kostov, V., Latham, D., Mazeh, T., Orosz, J. A., Pepper, J., Quarles, B., Ragozzine, D., Shporer, A., Southworth, J., Stassun, K., Thompson, S. E., Welsh, W. F., Agol, E., Derekas, A., Devor, J., Fischer, D., Green, G., Gropp, J., Jacobs, T., Johnston, C., LaCourse, D. M., Saetre, K., Schwengeler, H., Toczyski, J., Werner, G., Garrett, M., Gore, J., Martinez, A. O., Spitzer, I., Stevick, J., Thomadis, P. C., Vrijmoet, E. H., Yenawine, M., Batalha, N., and Borucki, W. Kepler eclipsing binary stars. vii. the catalog of eclipsing binaries found in the entire kepler data set. *The Astronomical Journal*, 151(3):68, feb 2016. doi: 10.3847/0004-6256/151/3/68. URL https://dx.doi.org/10.3847/0004-6256/151/3/68.

Koch, D. G. et al. Kepler Mission Design, Realized Photometric Performance, and Early Science. *The Astrophysical Journals Letter*, 713(2):L79–L86, April 2010. doi: 10.1088/2041-8205/713/2/L79.

Liu, J., Soria, R., Wu, X.-F., Wu, H., and Shang, Z. The sitian project, 2020.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019.

Mathur, S., Huber, D., Batalha, N. M., Ciardi, D. R., Bastien, F. A., Bieryla, A., Buchhave, L. A., Cochran, W. D., Endl, M., Esquerdo, G. A., Furlan, E., Howard, A., Howell, S. B., Isaacson, H., Latham, D. W., MacQueen, P. J., and Silva, D. R. Revised stellar properties of kepler targets for the q1-17 (DR25) transit detection run. *The Astrophysical Journal Supplement Series*, 229(2):30, mar 2017. doi: 10.3847/1538-4365/229/2/30. URL https://doi.org/10.3847/1538-4365/229/2/30.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

Pan, J.-S., Ting, Y.-S., and Yu, J. Astroconformer: The prospects of analysing stellar light curves with transformer-based deep learning models. *Monthly Notices of the Royal Astronomical Society*, 528(4):5890–5903, January 2024. ISSN 1365-2966. doi: 10.1093/mnras/stae068. URL http://dx.doi.org/10.1093/mnras/stae068.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ricker, G. R., Winn, J. N., Vanderspek, R., Latham, D. W., Bakos, G. Á., Bean, J. L., Berta-Thompson, Z. K., Brown, T. M., Buchhave, L., Butler, N. R., Butler, R. P., Chaplin, W. J., Charbonneau, D., Christensen-Dalsgaard, J., Clampin, M., Deming, D., Doty, J., De Lee, N., Dressing, C., Dunham, E. W., Endl, M., Fressin, F., Ge, J., Henning,

T., Holman, M. J., Howard, A. W., Ida, S., Jenkins, J., Jernigan, G., Johnson, J. A., Kaltenegger, L., Kawai, N., Kjeldsen, H., Laughlin, G., Levine, A. M., Lin, D., Lissauer, J. J., MacQueen, P., Marcy, G., McCullough, P. R., Morton, T. D., Narita, N., Paegert, M., Palle, E., Pepe, F., Pepper, J., Quirrenbach, A., Rinehart, S. A., Sasselov, D., Sato, B., Seager, S., Sozzetti, A., Stassun, K. G., Sullivan, P., Szentgyorgyi, A., Torres, G., Udry, S., and Villasenor, J. Transiting Exoplanet Survey Satellite (TESS). In Oschmann, Jacobus M., J., Clampin, M., Fazio, G. G., and MacEwen, H. A. (eds.), *Space Telescopes and Instrumentation 2014: Optical, Infrared, and Millimeter Wave*, volume 9143 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pp. 914320, August 2014. doi: 10.1117/12.2063489.

Sayeed, M., Huber, D., Wheeler, A., and Ness, M. K. The Swan: Data-driven Inference of Stellar Surface Gravities for Cool Stars from Photometric Light Curves. *The Astronomical Journal*, 161(4):170, April 2021. doi: 10.3847/1538-3881/abdf4c.

Smith, M. J., Roberts, R. J., Angeloudi, E., and Huertas-Company, M. Astropt: Scaling large observation models for astronomy, 2024.

Stello, D., Saunders, N., Grunblatt, S., Hon, M., Reyes, C., Huber, D., Bedding, T. R., Elsworth, Y., Garcí a, R. A., Hekker, S., Kallinger, T., Mathur, S., Mosser, B., and Pinsonneault, M. H. TESS asteroseismology of the kepler red giants. *Monthly Notices of the Royal Astronomical Society*, 512(2):1677–1686, feb 2022. doi: 10.1093/mnras/stac414. URL https://doi.org/10.1093%2Fmnras%2Fstac414.

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding, 2023.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2017.

Walmsley, M., Bowles, M., Scaife, A. M. M., Makechemu, J. S., Gordon, A. J., Ferguson, A. M. N., Mann, R. G., Pearson, J., Popp, J. J., Bovy, J., Speagle, J., Dickinson, H., Fortson, L., Géron, T., Kruk, S., Lintott, C. J., Mantha, K., Mohan, D., O'Ryan, D., and Slijepevic, I. V. Scaling laws for galaxy images, 2024.

Wei, D., Yan, T., and Fan, Y. The optical flare and afterglow light curve of grb 050904 at redshift z= 6.29. *The Astrophysical Journal*, 636(2):L69, 2006.

Yu, J., Huber, D., Bedding, T. R., Stello, D., Hon, M., Murphy, S. J., and Khanna, S. Asteroseismology of 16,000 kepler red giants: Global oscillation parameters, masses, and radii. *The Astrophysical Journal Supplement Series*, 236(2):42, Jun 2018. ISSN 1538-4365. doi: 10.3847/1538-4365/aaaf74. URL http://dx.doi.org/10.3847/1538-4365/aaaf74.

Yu, J., Bedding, T. R., Stello, D., Huber, D., Compton, D. L., Gizon, L., and Hekker, S. Asteroseismology of luminous red giants with kepler i: long-period variables with radial and non-radial modes. *Monthly Notices of the Royal Astronomical Society*, 493(1):1388–1403, jan 2020. doi: 10.1093/mnras/staa300. URL https://doi.org/10.1093%2Fmnras%2Fstaa300.

Zhang, B. and Sennrich, R. Root mean square layer normalization, 2019.