Teaching LLMs to Speak Spectroscopy

Nesar Ramachandra [1] **Yuan-Sen Ting** [2 3] **Zechang Sun** [4] **Azton Wells** [1] **Salman Habib** [1]

Abstract

Pre-trained Large Language Models (LLMs) have revolutionized text processing, yet adapting Transformer-based neural networks to non-textual scientific modalities typically requires specialized architectures and extensive computational resources. We demonstrate that LLaMA-3.1-8B can be efficiently repurposed to predict galaxy redshifts from spectroscopic data through Low-Rank Adaptation (LoRA), achieving competitive performance while preserving its linguistic capabilities. Using only 16 GPU-hours and adapting 0.04% of model parameters, our approach achieves a mean absolute error of 0.04 in redshift prediction while retaining over 85% of performance on AstroBench and 89% on general QA tasks from eval-harness. This minimal-effort adaptation—requiring only simple standard fine-tuning APIs—lowers barriers to entry for domain scientists and enables integrated agentic workflows where a single model handles both spectroscopic data for quantitative analysis and natural language for reasoning.

1. Introduction

Transformer-based models have revolutionized natural language processing through Large Language Models (LLMs) (Brown et al., 2020; Radford et al., 2019; Zhang et al., 2023), leveraging the scalable transformer architecture's self-attention mechanism (Vaswani et al., 2017). This mechanism's ability to capture long-range dependencies has enabled successful extensions beyond text to images (Dosovitskiy et al., 2020), graphs (Kipf & Welling, 2017), and spectral data (Liu et al., 2021; Fu et al., 2021).

[1]Computational Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Lemont, IL, USA [2]Department of Astronomy, The Ohio State University, Columbus, OH, USA [3]Center for Cosmology and AstroParticle Physics (CCAPP), The Ohio State University, Columbus, OH, USA [4]Department of Astronomy, Tsinghua University, Beijing, China. Correspondence to: Nesar Ramachandra <nramachandra@anl.gov>.

ML4Astro 2025, Vancouver, CA.

In astronomy, transformers have shown promise for processing time series (e.g., Pan et al., 2024b) and spectroscopic data (e.g., Leung & Bovy, 2024; Różański et al., 2025), where long-range correlations encode critical physical information. These models exhibit neural scaling laws, demonstrating potential for scaling to larger architectures (Pan et al., 2024a; Różański & Ting, 2025). However, astronomical applications typically train specialized transformers from scratch, requiring computational resources and domain expertise. These models often employ custom tokenization schemes, specialized positional encodings, and domain-specific masking strategies—each requiring careful design and validation (Różański et al., 2025).

These specialized models face several practical limitations. First, they cannot leverage the rapidly evolving LLM ecosystem, including optimized inference frameworks (Yuan et al., 2024), quantization techniques (Zhao et al., 2024), and deployment tools designed for text transformers. Second, astronomy-specific architectures often lack compatibility with fast inference systems like vLLM or TensorRT-LLM (Kwon et al., 2023), limiting their deployment at scale. Third, integrating these models into agentic workflows like (Moss, 2025) requires building custom interfaces between LLMs and domain-specific components, often increasing system complexity, maintenance burden, and high token consumption.

This raises a fundamental question: can we repurpose existing pre-trained LLMs to process entirely new scientific modalities through efficient adaptation? Such approaches have started to gain attention in other fields, including chemistry (Jablonka et al., 2024), material design (Gruver et al., 2024), and protein design (Lv et al., 2024), but have not been demonstrated in astronomy yet. Such an approach would lower the barrier to entry for astronomers while benefiting from mature LLM infrastructure. Crucially, any adaptation must preserve the model's original text processing and reasoning capabilities—the goal is augmentation, not replacement.

We demonstrate that LLaMA-3.1-8B, fine-tuned via Low-Rank Adaptation (LoRA) (Hu et al., 2021), can effectively predict galaxy redshifts from spectroscopic data while retaining its language capabilities. This parameter-efficient approach shows that generic transformer models can serve

as versatile scientific tools, processing both textual and spectroscopic modalities without requiring specialized architectures or extensive training from scratch.

2. Dataset

As a proof of concept, we chose galaxy redshift prediction—a fundamental cosmology task where accurate photometric redshifts enable large-scale structure studies (Newman & Gruen, 2022). While focusing on this high-impact application, our approach should generalize to other spectroscopic tasks where sequential patterns encode physical information. We compiled galaxy spectra from SDSS DR16, selecting galaxies (type = 3) with $0 < z < 0.50 < z < 0.5$ and dereddened $i < 18$ to ensure nearby, luminous sources (Ahumada et al., 2020). The query retrieved identifiers, coordinates, redshifts, photometric measurements, and spectroscopic details. We obtained FITS format spectra from the SDSS Science Archive Server, handling data reduction differences between SDSS and eBOSS pipelines. After converting from logarithmic to linear wavelength scales and normalizing fluxes, we obtained 10,000 galaxy samples. With equal-frequency binning, we sample 3,000 galaxies for training that uniformly span the redshift range. A validation set of 1,000 galaxies across the full redshift range is reserved for unbiased evaluation.

3. Methodology

To adapt pre-trained LLMs for spectroscopic analysis, we must address two fundamental challenges: representing continuous spectral data in a format LLMs can process, and fine-tuning the model without sacrificing its linguistic capabilities.

3.1. Tokenization

A key challenge lies in tokenization. Specialized astronomy transformers (Pan et al., 2024b; Różański et al., 2025) train custom MLP-based tokenizers from scratch, learning optimal representations for spectral features. However, this approach requires modifying the model architecture and training pipeline—precisely the barriers we aim to avoid.

Instead, we test whether standard LLM tokenizers can handle spectra with minimal adaptation. We serialize each flux value into digits using a configurable base representation with specified precision. For example, with base=10 and prec=2, the value $4.56 \rightarrow$ `"4|5|6"` where the leading space indicates positive sign and "—" separates individual digits. Complete spectra become concatenated strings with `" ,"` delimiting values. For instance, [4.56, 7.54, 11.2] becomes `"4|5|6 , 7|5|4 , 1|1|2|0 ,"`. The serialization handles signed values, removes leading zeros for variable-length representation, and includes proper separators for unambiguous parsing. Each input is prepended with `"Galaxy spectrum is rescaled and encoded to an input series:"` and target with `"Redshift: "`. With a total of 3,000 galaxies, this amounts to roughly 1.6M tokens. While suboptimal compared to learned tokenization, this approach requires zero architectural changes and tests the lower bound of what's achievable with minimal effort.

3.2. Model Selection and Fine-tuning

Having established a tokenization strategy, we selected a model balancing capability with accessibility. LLaMA-3.1-8B-Instruct represents an optimal trade-off: smaller models might lack the capacity to retain linguistic abilities while learning new modalities, while larger models exceed typical astronomy computing budgets. The 8B parameter scale provides sufficient capacity for multi-modal learning while remaining trainable on modest GPU clusters.

We employed Low-Rank Adaptation (LoRA), which decomposes weight updates into low-rank matrices $W + \Delta W = W + BA$ where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ with $r \ll \min(d, k)$ (Hu et al., 2021). This approach freezes the original weights W while training only the compact matrices A and B. LoRA has become the standard fine-tuning method across both open and proprietary models—OpenAI, Anthropic, and other providers offer LoRA-based fine-tuning APIs. While we use open-weight models for experimental flexibility, our approach translates directly to proprietary platforms where astronomers could upload their data and utilize built-in fine-tuning pipelines.

Using rank-8 adapters (3.4M parameters, $\sim$0.04% of total model parameters), two training epochs require only 16 A100 GPU hours total—well within typical astronomy computing allocations. In addition, each galaxy training point occupies less than 7% of the 8K context window of LLaMA-3.

4. Results

After fine-tuning, our model serializes galaxy spectra flux values into tokens, generating responses as `"Redshift: [value]"`. We extract numerical predictions and compare against true spectroscopic redshifts for performance metrics. Beyond spectroscopic accuracy (mean absolute error, MAE), we evaluate pre-trained capability retention using eval-harness benchmarks (Gao et al., 2023) and AstroBench, following Ting et al. (2025).

The validation set comprised 20% of galaxy spectra spanning the full redshift range. We analyze how learning rate, LoRA rank, training epochs, and dataset size affect the trade-off between modality adaptation and knowledge retention. Figure 1 illustrates the learning rate's critical role in bal-

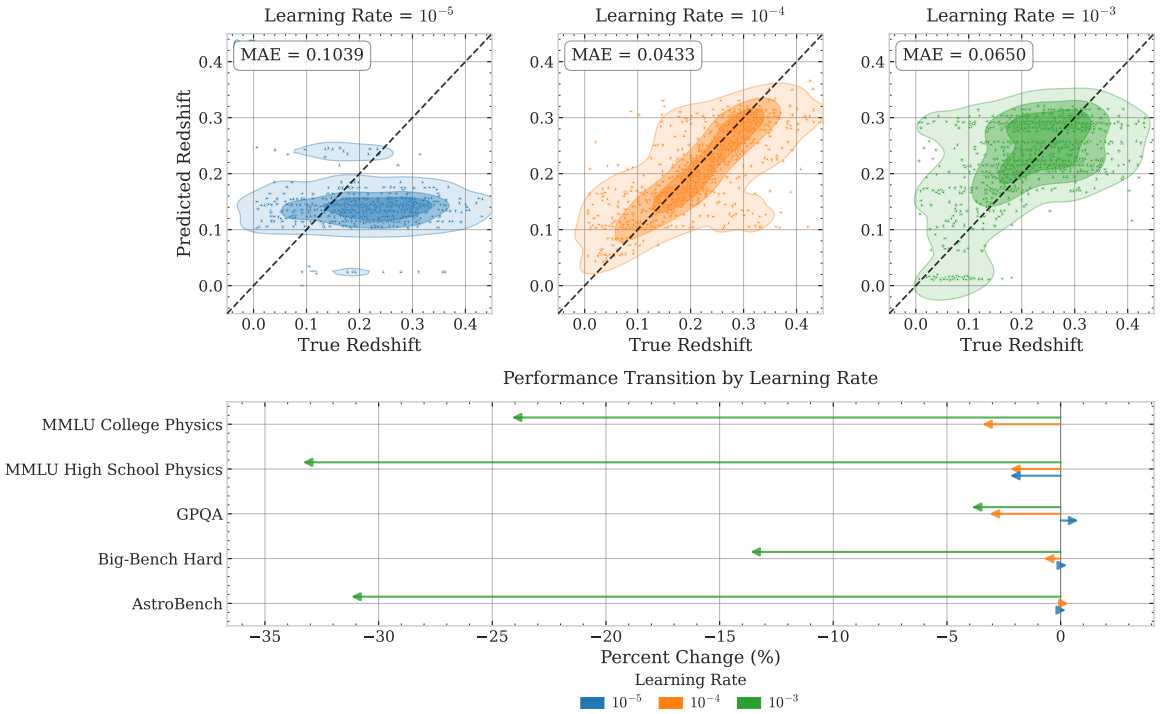

Figure 1. Trade-off between spectroscopic accuracy and language benchmark retention across learning rates. **Top:** Predicted vs. true redshifts for validation galaxies, with contours representing the full validation set of 2,000 spectra and individual points shown for clarity. Learning rate 10^{-5} (left) preserves language capabilities but yields poor redshift predictions (MAE=0.104), while 10^{-4} (middle) achieves optimal spectroscopic accuracy (MAE=0.043) with acceptable language degradation, and 10^{-3} (right) shows intermediate spectroscopic performance (MAE=0.065) with substantial language degradation. **Bottom:** Percent change in language benchmarks after fine-tuning. The optimal learning rate 10^{-4} (orange) balances accurate redshift prediction with less than 15% degradation in scientific reasoning tasks while maintaining strong general knowledge performance.

ancing spectroscopic accuracy and language preservation. At 10^{-4}, we achieve MAE=0.043 with less than 15% decline in scientific reasoning and 89.4% retention of general QA performance. The lowest rate (10^{-5}) preserves over 95% of language capabilities but yields poor spectroscopic performance (MAE=0.104), while the highest rate (10^{-3}) improves redshift prediction (MAE=0.065) but degrades reasoning tasks by less than 20%. Higher learning rates enable faster modality adaptation but disrupt the pre-trained representations more aggressively.

Beyond learning rate, our ablation studies reveal consistent patterns across other hyperparameters (Table 1). Increasing LoRA rank from 4 to 16 improves redshift accuracy (MAE decreases from 0.078 to 0.057) by allowing more parameters to adapt, but higher ranks show diminishing returns while causing greater language degradation. This aligns with LoRA theory: higher ranks enhance expressive capacity but risk overwriting original knowledge. In addition, the training duration also matters: one epoch yields incomplete adaptation (MAE=0.099), whereas three epochs only marginally improve spectroscopy (MAE=0.074) at the cost of language degradation. These trade-offs highlight the importance of careful hyperparameter tuning to align scientific performance with language retention.

Our chosen fiducial configuration—learning rate 10^{-4}, LoRA rank 8, two epochs—achieves MAE = 0.043 while retaining over 85% of original language capabilities, demonstrating an effective balance between spectroscopic accuracy and linguistic preservation. The 85.2% retention of QA performance in AstroBench indicates that the model can still understand complex scientific concepts in astronomy and answer technical questions for integrated scientific workflows. This minimal degradation enables agentic analysis where the model processes raw spectra, derives physical parameters, and then reasons about the results in natural language. Below is a condensed conversation illustrating its dual capability:

Step 1 - User provides spectrum: [Galaxy spectrum data]
Model prediction: Photometric redshift: 0.315

Step 2 - User: What type of galaxy is expected at the redshift you just predicted? Which surveys could observe it?

Table 1. Performance Comparison Across Different Fine-tuning Configurations

Learning Rate	LoRA Rank	Training size	Epochs	Redshift MAE†	Scientific QA*	General QA*
Varying Learning Rate (rank = 8, epochs = 2, samples = 3,000)						
10^{-5}	8	3,000	2	0.104	96.5%	95.1%
10^{-4}	8	3,000	2	**0.043**	**85.2%**	**89.4%**
10^{-3}	8	3,000	2	0.065	76.2%	79.8%
Varying LoRA Rank (rate = 10^{-4}, epochs = 2, samples = 3,000)						
10^{-4}	4	3,000	2	0.078	87.8%	91.2%
10^{-4}	8	3,000	2	**0.043**	85.2%	89.4%
10^{-4}	**16**	3,000	2	0.057	82.1%	86.7%
Varying Number of Epochs (rate = 10^{-4}, rank = 8, samples = 3,000)						
10^{-4}	8	3,000	1	0.099	87.9%	91.5%
10^{-4}	8	3,000	**2**	**0.043**	**85.2%**	**89.4%**
10^{-4}	8	3,000	3	0.074	83.7%	88.1%

† Lower is better. * Higher is better. Fiducial configurations in gray. Blue: best, second.

Model: At redshift z=0.315, this is likely a massive elliptical galaxy. These are among the most massive and evolved galaxies, typically formed through mergers. The redshift corresponds to a lookback time of ∼3.8 billion years, during an active epoch of galaxy evolution.

Observable by: SDSS, DES, HST, ALMA (dust/-gas), and future surveys like Euclid and LSST. The galaxy is receding at ∼84,000 km/s, with spectral features like Hα shifted from 656nm to 863nm.

For context, specialized spectroscopic redshift estimation methods can achieve lower MAE, even at sub-0.01 level (Bolton et al., 2012), suggesting our approach reaches competitive though not state-of-the-art performance. However, the value lies not in surpassing specialized methods but in demonstrating that general-purpose LLMs can achieve reasonable accuracy on scientific tasks while maintaining their broader capabilities, enabling truly integrated workflows where a single model handles the complete pipeline from raw data to scientific interpretation.

5. Conclusion and Broader Impact

This work demonstrates that pre-trained LLMs can be efficiently adapted to process non-textual scientific data through parameter-efficient fine-tuning. However, the tokenization, fine-tuning process, and the resulting performance may vary across the type of modalities. This approach contrasts sharply with training specialized models from scratch (e.g. Leung & Bovy, 2024; Różański et al., 2025), which requires orders of magnitude more computational resources while sacrificing the ability to process natural language. Our findings have several important implications for scientific computing:

Democratizing Access: The approach substantially lowers barriers to entry for domain scientists. Rather than developing specialized architectures or training models from scratch, astronomers can leverage existing LLM infrastructure, fine-tuning APIs, and established deployment pipelines.

Enabling Integrated Workflows: Our approach enables truly end-to-end scientific analysis within a single model. As demonstrated in our results, the same model that processes raw spectroscopic data to derive redshifts can immediately reason about the physical implications—discussing galaxy types, evolutionary stages, and observational strategies. This eliminates the need for complex interfaces between specialized components and enables more natural human-AI collaboration in scientific discovery.

Revealing Transferable Representations: Our results suggest that foundation models trained on text contain remarkably transferable representations applicable to diverse scientific modalities. The success of adapting a language model to spectroscopic analysis—achieving less than 15% degradation in reasoning capabilities—implies that transformer pre-training captures general computational strategies for processing sequential information that transcend specific data types. While there are successful demonstrations of unified prediction frameworks like this in other specific scientific modalities (Hu et al., 2025; Zhang et al., 2025), this area requires a deeper investigation.

As LLM infrastructure continues its rapid advancement,

parameter-efficient adaptation offers a practical path to democratize AI-driven scientific analysis. By building on existing foundation models rather than starting from scratch, the scientific community can benefit from ongoing improvements in language modeling while maintaining the flexibility to incorporate domain-specific data. This approach promises to accelerate scientific discovery by enabling models that can seamlessly navigate between quantitative analysis and conceptual reasoning—a capability increasingly essential for tackling complex agentic challenges.

Acknowledgements

Work at Argonne National Laboratory is supported by UChicago Argonne LLC, Operator of Argonne National Laboratory. Argonne, a U.S. Department of Energy Office of Science Laboratory, is operated under Contract No. DE-AC02-06CH11357. The training is carried out on Swing, a GPU system at the Laboratory Computing Resource Center (LCRC) of Argonne National Laboratory. This material is based upon work supported by Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DEAC02-06CH11357. YST is supported by the National Science Foundation under Grant No. AST-2406729. AZ and SH are supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics, High Energy Physics Center for Computational Excellence (HEP-CCE) at Argonne National Laboratory under B&R KA2401045.

References

Ahumada, R., Prieto, C. A., Almeida, A., Anders, F., Anderson, S. F., Andrews, B. H., Anguiano, B., Arcodia, R., Armengaud, E., Aubert, M., et al. The 16th data release of the sloan digital sky surveys: first release from the apogee-2 southern survey and full release of eboss spectra. *The Astrophysical Journal Supplement Series*, 249(1):3, 2020.

Bolton, A. S., Schlegel, D. J., Aubourg, É., Bailey, S., Bhardwaj, V., Brownstein, J. R., Burles, S., Chen, Y.-M., Dawson, K., Eisenstein, D. J., et al. Spectral classification and redshift measurement for the sdss-iii baryon oscillation spectroscopic survey. *The Astronomical Journal*, 144(5): 144, 2012.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition. *arXiv preprint arXiv:2010.11929*, 2020.

Fu, Z., Zhang, Y., Li, X., Wang, S., and Chen, D. Transformers for 1d spectral data: A survey. In *Proceedings of the International Conference on Learning Representations*, 2021.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Gruver, N., Sriram, A., Madotto, A., Wilson, A. G., Zitnick, C. L., and Ulissi, Z. Fine-Tuned Language Models Generate Stable Inorganic Materials as Text. *arXiv e-prints*, art. arXiv:2402.04379, February 2024. doi: 10.48550/arXiv.2402.04379.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., and Wang, P. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hu, X., Liu, G., Chen, C., Zhao, Y., Zhang, H., and Liu, X. 3dmolformer: A dual-channel framework for structure-based drug discovery. *arXiv preprint arXiv:2502.05107*, 2025.

Jablonka, K. M., Schwaller, P., and Ortega-Guerrero, A. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6:161–169, 2024. doi: 10.1038/s42256-023-00788-1.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Leung, H. W. and Bovy, J. Towards an astronomical foundation model for stars with a transformer-based model. *MNRAS*, 527(1):1494–1520, January 2024. doi: 10.1093/mnras/stad3015.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Lv, L., Lin, Z., Li, H., Liu, Y., Cui, J., Yu-Chian Chen, C., Yuan, L., and Tian, Y. ProLLaMA: A Protein Language Model for Multi-Task Protein Language Processing. *arXiv e-prints*, art. arXiv:2402.16445, February 2024. doi: 10.48550/arXiv.2402.16445.

Moss, A. The ai cosmologist i: An agentic system for automated data analysis. *arXiv preprint arXiv:2504.03424*, 2025.

Newman, J. A. and Gruen, D. Photometric redshifts for next-generation surveys. *Annual Review of Astronomy and Astrophysics*, 60(1):363–414, 2022.

Pan, J.-S., Ting, Y.-S., Huang, Y., Yu, J., and Liu, J.-F. The Scaling Law in Stellar Light Curves. *arXiv e-prints*, art. arXiv:2405.17156, May 2024a. doi: 10.48550/arXiv.2405.17156.

Pan, J.-S., Ting, Y.-S., and Yu, J. Astroconformer: The prospects of analysing stellar light curves with transformer-based deep learning models. *MNRAS*, 528(4): 5890–5903, March 2024b. doi: 10.1093/mnras/stae068.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.

Różański, T. and Ting, Y.-S. Scaling Laws for Emulation of Stellar Spectra. *arXiv e-prints*, art. arXiv:2503.18617, March 2025. doi: 10.48550/arXiv.2503.18617.

Różański, T., Ting, Y.-S., and Jabłońska, M. Transformer-Payne: Enhancing Spectral Emulation Accuracy and Data Efficiency by Capturing Long-range Correlations. *ApJ*, 980(1):66, February 2025. doi: 10.3847/1538-4357/ad9b99.

Ting, Y. S., Nguyen, T. D., Ghosal, T., Pan, R., Arora, H., Sun, Z., de Haan, T., Ramachandra, N., Wells, A., Madireddy, S., and Accomazzi, A. AstroMLab 1: Who wins astronomy jeopardy!? *Astronomy and Computing*, 51:100893, April 2025. doi: 10.1016/j.ascom.2024.100893.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Yuan, Z., Shang, Y., Zhou, Y., Dong, Z., Zhou, Z., Xue, C., Wu, B., Li, Z., Gu, Q., Lee, Y. J., et al. Llm inference unveiled: Survey and roofline model insights. *arXiv preprint arXiv:2402.16363*, 2024.

Zhang, G., Li, Y., Luo, R., Hu, P., Zhao, Z., Li, L., Liu, G., Wang, Z., Bi, R., Gao, K., et al. Unigenx: Unified generation of sequence and structure with autoregressive diffusion. *arXiv preprint arXiv:2503.06687*, 2025.

Zhang, J. et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2307.09288*, 2023.

Zhao, Y., Lin, C.-Y., Zhu, K., Ye, Z., Chen, L., Zheng, S., Ceze, L., Krishnamurthy, A., Chen, T., and Kasikci, B. Atom: Low-bit quantization for efficient and accurate llm serving. *Proceedings of Machine Learning and Systems*, 6:196–209, 2024.