Cross-Modal Adaptation of Decoder-only Models to Partial Differential Equation Data

Anonymous authors
Paper under double-blind review

Abstract

Different methods of fine-tuning Large Language Models to new modalities have been introduced in recent years, particularly for Scientific ML tasks such as time-dependent simulation tasks based on Partial Differential Equations (PDEs). Most of these approaches are based on encoder-only models, even though decoder-only models have gained popularity in NLP and ML more broadly, given their scaling capabilities. However, the impact of model architecture on these approaches has not been investigated before. In this ongoing work, we perform a series of ablation studies that compare encoder-only and decoder-only models. We find that encoder-only models perform better than decoder-only models (with a great variation between tasks). This is because of how the data is introduced into decoder-only models, which get heavily penalized for being autoregressive. We also find that, in contrast to other tasks, scaling decoder-only models does not change performance. Pending more experimentation, these results show that we need to find new ways to harness the potential of decoder-only models in the context of cross-modal adaptation.

1 Introduction

Over the years, we have seen an undeniable rise in the popularity of pre-trained Large Language Models (LLMs). These models can then be adapted to new tasks, using different approaches, like fine-tuning or in-context learning. Recent work has used fine-tuning techniques to adapt models across modalities (Lu et al., 2022; Shen et al., 2023; Ma et al., 2024; Shen et al., 2024a), achieving competitive performance across a wide range of tasks. Given the popularity and accessibility of these pre-trained models, these approaches can be of great utility for Scientific Machine Learning tasks, and are being used for tasks such as seismic monitoring (Wang et al., 2025) and time series forecasting (Liu et al., 2025).

However, the reasons behind this success are unclear, since few ablation studies have been performed, varying the originally-proposed configurations. For example, most of these approaches are based on encoder-only models, even though decoder-only models have gained popularity in the fields of NLP and ML, given their scaling capabilities.

To determine whether decoder-only models can be of use for cross-modal adaptation approaches, we introduce a series of ablation studies. Our research questions are:

- How does model architecture affect cross-modal adaptation? (§4)
- How does scaling decoder-only models affect cross-modal adaptation? (§5)

With these questions, we try to get a better understanding of the capabilities of cross-modal adaptation methods and to broaden the potential available models used to perform cross-modal fine-tuning. In our results, we see that decoder-only models do not outperform encoder-only models, even when scaled up. This is due to autoregressive attention over the input, as well as how the outputs are computed, which is by averaging the representations of the last hidden layer, rather than generating outputs. Our results point to the need for future work to propose custom methods to leverage the potential of decoder-only models.

2 Background

Large language models for science LLMs are increasingly being used for scientific tasks, including to improve text quality, coding, clinical research tasks, and more (Almarie et al., 2023). Recent work has even studied the potential of LLMs as hypothesis generators (Zhou et al., 2024).

LLMs can also be very useful for scientific tasks where data can be processed sequentially. One example of this is protein data, where a growing number of LLM methods are proposed for different tasks (Xiao et al., 2025), including for protein understanding and prediction (Xiao et al., 2024; Wu et al., 2024; Truong Jr & Bepler, 2023), protein engineering, generation, and translation (Ghafarollahi & Buehler, 2024; Zheng et al., 2023; Shen et al., 2024b).

Cross-modal adaptation In recent years, several approaches have been proposed to fine-tune large language models for different modalities unseen during pre-training. These methods include Frozen Pretrained Transformers (FPT) (Lu et al., 2022), ORCA (Shen et al., 2023), Patch Replacement (PaRe) (Cai et al., 2024), Modality kNowledge Alignment (MoNA) (Ma et al., 2024), Unified PDE Solver (UPS) (Shen et al., 2024a), etc. All these methods are based on taking advantage of the acquired knowledge of the model during pre-training to minimize the amount of fine-tuning necessary to adapt it to a new modality.

These techniques have a lot of potential to be adapted to different Scientific Machine Learning tasks, and recently, some practical applications have been presented, for example, for seismic monitoring (Wang et al., 2025) and time series forecasting (Liu et al., 2025).

3 Experimental Setup

To evaluate the effects of model architecture and scale on cross-modal adaptation with partial differential equation data, we experiment with several models, scales, and cross-modal adaptation methods as described below.

Methods We choose two popular methods for cross-modal adaptation in the literature – FPT and ORCA. FPT adapts pre-trained models to new tasks by fine-tuning only the input and output layers and the layernorm parameters. ORCA performs an embedder training step before fine-tuning using Optimal Transport Dataset Distance (OTDD) (Alvarez-Melis & Fusi, 2020) between the new task dataset and a pre-selected proxy dataset, as a loss function. All parameters are trained during the fine-tuning step. We follow ORCA's (Shen et al., 2023) implementation for both ORCA and FPT (Lu et al., 2022).

Models We select RoBERTa base (Liu et al., 2019) as our encoder-only model, following ORCA (Shen et al., 2023), and GPT-2 (Radford et al., 2019) as our decoder-only model, since it is used in Lu et al. (2022). Both of these models have a similar size (125M vs. 137M parameters). For the scaling experiments, we consider the larger versions of the GPT-2 family: GPT-2 Medium (380M), GPT-2 Large (812M), and GPT-2 XL (1.61B).

Datasets Following Shen et al. (2023), we use four different PDE datasets (Advection, Diffusion-Reaction, Diffusion-Sorption, and Navier-Stokes) from PDEBench Takamoto et al. (2022), explained in Appendix A. In addition to the target dataset, the ORCA method also requires a proxy dataset for embedder training. For RoBERTa base, we use CoNLL-2003, the original proxy dataset generated by Shen et al. (2023). Due to difficulties with replicating their approach, we used CoNLL-2000 to generate proxy datasets for the rest of the models, trying to remain as close as possible to their implementation. A detailed explanation of the proxy dataset generation can be found in Appendix B.

As in previous literature (Shen et al., 2023; Ma et al., 2024; Shen et al., 2024a; Cai et al., 2024), we evaluate the tasks using normalized Root Mean Squared Errors (nRMSE), since it is scale-independent. As the metric is error-based, lower values are better.

4 How Does Model Architecture Affect Cross-Modal Adaptation?

In this section, we experiment with two different kinds of transformer architectures, encoder-only and decoder-only models, represented by RoBERTa base and GPT-2, respectively. Prior work generally assumes that pre-training results in better cross-modal adaptation results, but we ablate for this factor as well by including randomly-initialized versions of these models. This allows us to disentangle the effects of both architecture and pre-training.

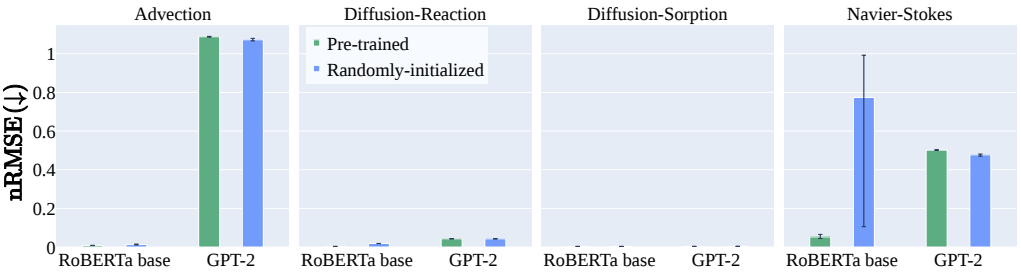

Figure 1: Comparison of model performance using ORCA (Shen et al., 2023) for cross-modal adaptation, using both pre-trained and randomly-initialized versions of RoBERTa base and GPT-2 models.

As shown in Figure 1, encoder-only models outperform decoder-only models for three of the four selected tasks (Advection, Diffusion-Reaction, and Navier-Stokes), with very different performance depending on the task. For both Advection and Navier-Stokes, we can see that GPT-2 is unable to solve the task, but RoBERTa base shows good results (only when using ORCA). On the other hand, Diffusion-Reaction shows just a small performance deterioration when using GPT-2 instead of RoBERTa base. Performance with FPT (shown in Appendix C) shows identical patterns to ORCA-based adaptation.

The remaining task, Diffusion-Sorption, shows similarly good performance for all models, indicating that the task is simple enough to be solved without pre-training. Similar to what García de Herreros et al. (2024) found with the Satellite dataset for satellite image time series analysis, this highlights the importance of selecting tasks that allow us to better evaluate cross-modal adaptation methods. We contend that applying these approaches should only be done when the pre-training in the original modality is necessary; otherwise, there is no gain from pre-training a model at all.

To better understand these results, we analyzed the way these tasks were fed into the model, as well as the way the predictions were made. By doing this, we discovered that decoder-only models are doubly penalized. First, they are penalized for being autoregressive, since each point in the sequence is treated as an individual token and GPT-2 cannot condition on the sequence bidirectionally, which is necessary for waveforms with symmetry. Secondly, the predictions are not computed generatively, but instead, the representations of the last hidden layer are just averaged.

This shows that **while encoder-only models outperform decoder-only models** for PDE tasks using cross-model adaptation methods, **the setup for this comparison is biased**.

5 How Does Scaling Decoder-Only Models Affect Cross-Modal Adaptation?

The previous results motivated us to find potential ways in which decoder-only models can overcome this penalization and improve their performance. In this section, we test scaling the selected decoder-only model to see if this has an effect.

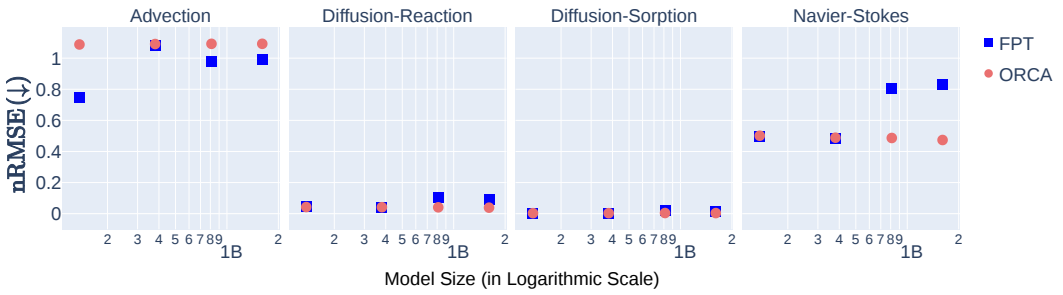

Figure 2: Performance of different sizes of models of the GPT-2 family using both ORCA (Shen et al., 2023) and FPT (Lu et al., 2022) for cross-modal adaptation.

As Figure 2 shows, scaling does not improve performance for any of the selected tasks, neither when using ORCA, nor with FPT. For Advection and Navier-Stokes, we can even see some performance deterioration when using FPT. We leave it to future work to investigate why scaling does not work despite its success in other areas (Kaplan et al., 2020; Caillaut et al., 2024; Cai et al., 2025), and put forth two hypotheses:

Hypothesis 1 It could be that being able to condition bidirectionally on the whole sequence context is important for the time prediction, as in the previous section.

Hypothesis 2 Given that predictions are computed by averaging the last hidden layer of the model, increasing the embedding dimension could make this task harder.

6 Limitations

Model selection As we only evaluate one model per architecture, we caution against drawing conclusions about the performance of, e.g., other encoder-only models beyond just RoBERTa. To this end, we intend to experiment with a wider range of models.

Proxy dataset Given our difficulties replicating the original proxy dataset from ORCA (Shen et al., 2023), more testing is required to determine the potential influence this could have on all models.

Cross-modal adaptation methods We only experiment with two popular cross-modal adaptation methods, and leave it to future work to investigate whether the same patterns hold for PARE (Cai et al., 2024) and UPS (Shen et al., 2024a).

7 Conclusion and Future Work

In this work, we perform a series of ablation studies to study the effect of model architecture and size on cross-modal adaptation approaches. Contrary to our expectations, we found that decoder-only models cannot outperform encoder-only models, even when scaled. We found that the unidirectional attention plays a key role in this performance difference, not allowing the model to get an overall understanding of the PDE data.

Based on these results, future work should introduce new ways to overcome this penalization to decoder-only models to be able to take advantage of their full capabilities and scaling potential. These results should also be validated with a wider variety of models, cross-modal adaptation methods, and datasets, to address the limitations described in the previous section.

References

Bassel Almarie, Paulo EP Teixeira, Kevin Pacheco-Barrios, Carlos Augusto Rossetti, and Felipe Fregni. Editorial - the use of large language models in science: Opportunities and challenges. *Principles and Practice of Clinical Research*, 9(1):1–4, Jul. 2023. doi: 10.21801/ppcrj.2023.91.1. URL https://journal.ppcr.org/index.php/ppcrjournal/article/view/259.

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21428–21439. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f52a7b2610fb4d3f74b4106fb80b233d-Paper.pdf.

Hongru Cai, Yongqi Li, Ruifeng Yuan, Wenjie Wang, Zhen Zhang, Wenjie Li, and Tat-Seng Chua. Exploring training and inference scaling laws in generative retrieval. *arXiv preprint arXiv:2503.18941*, 2025.

Lincan Cai, Shuang Li, Wenxuan Ma, Jingxuan Kang, Binhui Xie, Zixun Sun, and Chengwei Zhu. Enhancing cross-modal fine-tuning with gradually intermediate modality generation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5236–5257. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cai24c.html.

Gaëtan Caillaut, Mariam Nakhlé, Raheel Qader, Jingshu Liu, and Jean-Gabriel Barthélemy. Scaling laws of decoder-only models on the multilingual machine translation task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 1318–1331, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.124. URL https://aclanthology.org/2024.wmt-1.124/.

Paloma García de Herreros, Vagrant Gautam, Philipp Slusallek, Dietrich Klakow, and Marius Mosbach. What explains the success of cross-modal fine-tuning with orca? In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pp. 8–16, 2024.

Alireza Ghafarollahi and Markus J Buehler. Protagents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 3(7):1389–1409, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Peiyuan Liu, Hang Guo, Tao Dai, Naiqi Li, Jigang Bao, Xudong Ren, Yong Jiang, and Shu-Tao Xia. Calf: Aligning llms for time series forecasting via cross-modal fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18915–18923, 2025.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Frozen pretrained transformers as universal computation engines. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 7628–7636, 2022.

Wenxuan Ma, Shuang Li, Lincan Cai, and Jingxuan Kang. Learning modality knowledge alignment for cross-modality transfer. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 33777–33793. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ma24d.html.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Erik F Sang and Sabine Buchholz. Introduction to the conll-2000 shared task: Chunking. *arXiv preprint cs/0009008*, 2000.

Junhong Shen, Liam Li, Lucio M. Dery, Corey Staten, Mikhail Khodak, Graham Neubig, and Ameet Talwalkar. Cross-modal fine-tuning: Align then refine. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31030–31056. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/shen23e.html.

Junhong Shen, Tanya Marwah, and Ameet Talwalkar. Ups: Towards foundation models for pde solving via cross-modal adaptation. *arXiv preprint arXiv:2403.07187*, 2024a.

Yiqing Shen, Zan Chen, Michail Mamalakis, Yungeng Liu, Tianbin Li, Yanzhou Su, Junjun He, Pietro Liò, and Yu Guang Wang. Toursynbio: A multi-modal large model and agent framework to bridge text and protein sequences for protein engineering. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2382–2389, 2024b. doi: 10.1109/BIBM62325.2024.10822695.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. Pdebench: An extensive benchmark for scientific machine learning. *Advances in Neural Information Processing Systems*, 35:1596–1611, 2022.

Timothy Truong Jr and Tristan Bepler. Poet: A generative model of protein families as sequences-of-sequences. *Advances in Neural Information Processing Systems*, 36:77379–77415, 2023.

Xinghao Wang, Feng Liu, Rui Su, Zhihui Wang, Lei Bai, and Wanli Ouyang. Seismollm: Advancing seismic monitoring via cross-modal transfer with pre-trained large language model. *arXiv preprint arXiv:2502.19960*, 2025.

Kevin E Wu, Howard Chang, and James Zou. Proteinclip: enhancing protein language models with natural language. *bioRxiv*, pp. 2024–05, 2024.

Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. Proteingpt: Multimodal llm for protein property prediction and structure understanding. *arXiv preprint arXiv:2408.11363*, 2024.

Yijia Xiao, Wanjia Zhao, Junkai Zhang, Yiqiao Jin, Han Zhang, Zhicheng Ren, Renliang Sun, Haixin Wang, Guancheng Wan, Pan Lu, et al. Protein large language models: A comprehensive survey. *arXiv preprint arXiv:2502.17504*, 2025.

Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei Ye, and Quanquan Gu. Structure-informed language models are protein designers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42317–42338. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zheng23a.html.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis generation with large language models. In Lotem Peled-Cohen, Nitay Calderon, Shir Lissak, and Roi Reichart (eds.), *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pp. 117–139, Miami, FL, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4science-1.10. URL https://aclanthology.org/2024.nlp4science-1.10/.

A PDE Datasets Details

As we saw in Section 3, we tested the models in a collection of PDE datasets from PDEBench (Takamoto et al., 2022). We follow Shen et al. (2023) for the download, pre-processing, and loading of the data. The specifications of the selected datasets can be seen in Table 1

Dataset	Dimension	Resolution	Coefficients
Advection	1D	1024	$\beta = 0.4$
Diffusion-Reaction	1D	1024	$\nu = 0.5, \rho = 1.0$
Diffusion-Sorption	1D	1024	-
Compressible Navier-Stokes	1D	1024	$\eta = \zeta = 0.1$, rand periodic

Table 1: List of PDE dataset used as target datasets and their corresponding specifications.

B Proxy Datasets

To create the specific proxy datasets for GPT-2, GPT-2 Medium, GPT-2 Large, and GPT-2 XL, we follow the specifications given by Shen et al. (2023).

Instead of using CoNLL-2003, like stated in Shen et al. (2023), we used the CoNLL-2000 dataset (Sang & Buchholz, 2000). Following Shen et al. (2023), we selected a random sample of 2000 sequences containing less than 32 tokens. We unified the length by adding the padding token until all sequences have a length of 32 tokens. Lastly, we calculate the embeddings using the selected models.

C How Does Model Architecture Affect FPT Adaptation?

In this section, we include the results for FPT for Section 4.

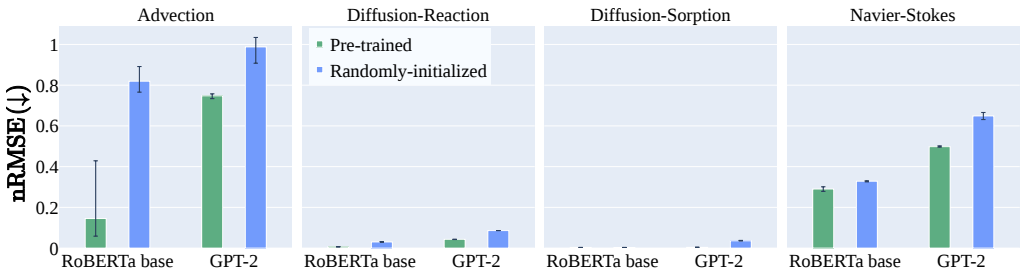

Figure 3: Comparison of model performance using FPT (Lu et al., 2022) for cross-modal adaptation, using both pre-trained and randomly-initialized version of RoBERTa base and GPT-2 models.