Decoding Partial Differential Equations: Cross-Modal Adaptation of Decoder-only Models to PDEs

Anonymous ACL submission

Abstract

Large language models have shown great success on natural language tasks in recent years, but they have also shown great promise when adapted to new modalities, e.g., for scientific machine learning tasks. Even though decoder-only models are more popular within NLP and scale exceedingly well at generating natural language, most proposed approaches for cross-modal adaptation focus on encoder-only models, raising the question of how model architecture affects these approaches. In this paper, we therefore perform a series of ablation studies to answer this question, systematically comparing encoder-only and decoder-only models on cross-modal adaptation for time-dependent simulation tasks based on partial differential equations (PDEs). We find that decoder-only models are far worse than encoder-only models, when existing approaches are applied unmodified. In contrast to several other domains, scaling decoder-only models also does not help. To harness the potential of decoder-only models in this context, we introduce two novel approaches, **Parallel Flipping** and **Sequence Doubling**, attempting to mimic bidirectionality in autoregressive models. Both our methods improve overall performance using decoder-only models for all tasks and all cross-model adaptation methods, closing the gap to encoder-only model performance. We hope that our findings broaden the spectrum of models used on cross-modal adaptation tasks to further scientific ML.

1 Introduction

Pre-trained large language models (LLMs) have seen unprecedented improvements in processing natural language in recent years. 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 even adapt models across modalities, achieving competitive performance across a wide range of tasks including detecting atrial cardiac disease from ECG recordings, and time-dependent simulation tasks based on Partial Differential Equations (Lu et al., 2022; Shen et al., 2023; Ma et al., 2024; Shen et al., 2024). These approaches can be of great utility for scientific machine learning tasks, and are currently used for tasks such as seismic monitoring (Wang et al., 2025) and time series forecasting (Liu et al., 2025).

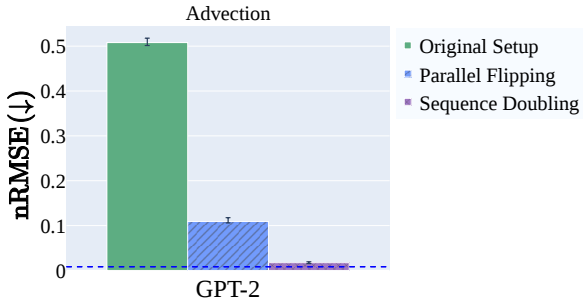

Figure 1: Cross-modal adaptation of GPT-2, a decoder-only model, with ORCA-based adaptation on the Advection dataset of time-dependent PDE simulation. Although the original setup shows high error, our proposed methods (Parallel Flipping and Sequence Doubling) close the gap to encoder-only model performance.

However, it is how unclear how general cross-modal adaptation methods are, since few ablation studies have been performed that vary the originally proposed configurations. For example, most approaches are based on encoder-only models, even though decoder-only models are by far the more popular transformer-based model for NLP tasks, as they scale impressively, mimic human-like language convincingly, and can be used for a wide variety of tasks. Through better, more general representations of natural language, today's best decoder-only models may well provide a better starting point for cross-modal adaptation.

In our work, we attempt to leverage the potential of decoder-only models to broaden the range of models available for cross-modal adaptation. First,

we directly apply two existing cross-modal adaptation methods to encoder-only and decoder-only models, and find that decoder-only models perform much worse. Next, we test whether scaling up decoder-only models helps performance, but find that it does not.

We hypothesize that the reasons for the lack of success of these approaches 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, as in natural language. Addressing these issues, we introduce two different methods to improve cross-modal adaptation with decoder-only models by simulating bidirectionality:

Parallel Flipping: In parallel, we run the original setup and the same pipeline with the data inverted, and we then combine the predictions by taking the second half of both of them.

Sequence Doubling: We concatenate every sequence in the original data with itself before introducing it to the model, and use only the second half of the last hidden layer to compute predictions.

Both methods extend the sequence context that models can access at different points, and results show that our methods outperform the original setup for all tasks and cross-modal adaptation methods, closing the gap to encoder-only models, as shown in Figure 1. Each method comes with particular tradeoffs, making this a promising direction for future work. We hope that our findings broaden the spectrum of models used for cross-modal adaptation and further the field of scientific ML. Our code is available here: **REDACTED FOR REVIEW**.

2 Related Work

2.1 Large Language Models For Science

LLMs are increasingly 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 are also beginning to be used on complex mathematical tasks, such as enhancing analytical PDE approximations (Bhatnagar et al., 2025), moving towards using LLMs to find the analytical solution to differential equations, with Surkov et al. (2024) and Zakharov et al. (2025) respectively proposing a baseline and a dataset for fine-tuning LLMs to solve differential equations.

Some work has even used LLMs to generate code (Li et al., 2025) to simulate the data modeled by certain PDEs, as proposed in Takamoto et al. (2022). In contrast, we use cross-modal adaptation of LLMs to solve these tasks, similar to Shen et al. (2023), and Shen et al. (2024). This data consists of time series predictions of continuous observations over a space domain, similar to other scientific ML data such as Satellite (Petitjean et al., 2012) and MegaFlow2D (Xu et al., 2023), which our PDE-focused work could also be relevant to.

2.2 Cross-Modal Adaptation

In recent years, a more extreme approach for large language model adaptation has been introduced, known as cross-modal adaptation. Such approaches involve adapting models to new modalities unseen by the model during pre-training. Most of the proposed methods focus on the fine-tuning stage of large language models. 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 (UPS; Ma et al., 2024, MoNA), Unified PDE Solver (UPS; Shen et al., 2024), and more. All these methods purport to take advantage of the knowledge and skills the model acquires 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 used for various scientific machine learning tasks, and recently, some practical applications have been presented, including seismic monitoring (Wang et al., 2025) and time series forecasting (Liu et al., 2025).

2.3 Architecture Differences

Modern transformer-based large language models include the original encoder-decoder architecture (Vaswani et al., 2017), encoder-only architectures such as BERT (Devlin et al., 2018), as well as decoder-only architectures such as GPT (Radford et al., 2019), the latter two of which are more popular in modern NLP. Due to differences between the architectures, including pre-training objectives and attention mechanisms, several works compare them, finding differences in phenomena such as pronoun use (Gautam et al., 2024) and various linguistic probes (Waldis et al., 2024). In cross-modal adaptation, however, there have been no systematic architectural comparisons, to the best of our knowl-

edge. Some papers, like us, also try to close the gap between encoder-only and decoder-only models in various contexts, including language embeddings (Springer et al., 2025; BehnamGhader et al., 2024) and cognitively plausible language models (Charpentier and Samuel, 2024).

3 Experimental Setup

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

3.1 Methods

We choose two popular methods for cross-modal adaptation in the literature – Frozen Pretrained Transformers (FPT) (Lu et al., 2022) and ORCA (Shen et al., 2023). In both cases, a task-specific embedder and predictor are created to account for mismatches in dimensions between the target modality data and the original model. Then, FPT adapts the pre-trained models to new tasks by fine-tuning only the input and output layers, as well as the layer normalization parameters. ORCA, on the other hand, first trains the embedder on its own minimizing the Optimal Transport Dataset Distance (OTDD) (Alvarez-Melis and Fusi, 2020) between the target task dataset and a pre-selected proxy dataset. After this, all parameters are trained on the target task dataset. We use ORCA's implementation for both ORCA and FPT, with the same hyperparameters. As our Sequence Doubling method doubles the sequence length of inputs, we reduce the batch size for some configurations.

3.2 Models

We select RoBERTa-BASE (Liu et al., 2019) and BERT as our encoder-only models, following ORCA (Shen et al., 2023), and GPT-2 (Radford et al., 2019) and Pythia (Biderman et al., 2023) as our decoder-only models, since GPT-2 is used in Lu et al. (2022) and Pythia-160M has a large range of model sizes. All of these models have similar sizes (respectively, 125M, 110M, 160M, and 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), as well as the Pythia family: Pythia-14M, Pythia-70M, Pythia-410M, Pythia-1B, and Pythia-1.4B. We did not consider larger Pythia model sizes to keep the comparison with the GPT2 family fair.

3.3 Datasets

We use four different datasets of time-dependent simulation tasks based on partial differential equations: Advection, Diffusion-Reaction, Diffusion-Sorption, and Navier-Stokes, all taken from PDEBench (Takamoto et al., 2022). We follow the configurations in Shen et al. (2023) as detailed in Appendix A.

Proxy Datasets In addition to the target dataset, the ORCA method also requires a proxy dataset for training the embedder. For RoBERTa-BASE, we use the original proxy dataset generated by Shen et al. (2023) using CoNLL-2003. We follow their approach with CoNLL-2000 to generate proxy datasets for the rest of the models. A detailed explanation of the proxy dataset generation can be found in Appendix B.

3.4 Evaluation Metric

As in previous literature (Shen et al., 2023; Ma et al., 2024; Shen et al., 2024; Cai et al., 2024; Li et al., 2025), we report normalized Root Mean Squared Errors (nRMSE) for all tasks, as it is scale-independent. As the metric is error-based, lower values are better, which we also note in all figure captions. We report averages over five runs; given high variance for some configurations, we show best (minimum) and worst (maximum) performance with error bars.

4 Decoder-Only Models Perform Much Worse than Encoder-Only Models

In this section, we experiment with two transformer architectures, encoder-only and decoder-only models, represented by RoBERTa-BASE and BERT-BASE, and GPT-2 and Pythia-160M, respectively, plugged directly into the existing cross-modal adaptation approaches, FPT and ORCA. Prior work generally assumes that pre-training results in better cross-modal adaptation performance, 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.

We start by considering the performance of randomly-initialized versions of the models, to evaluate whether pre-training on language data actually helps at all with these tasks. Using ORCA,

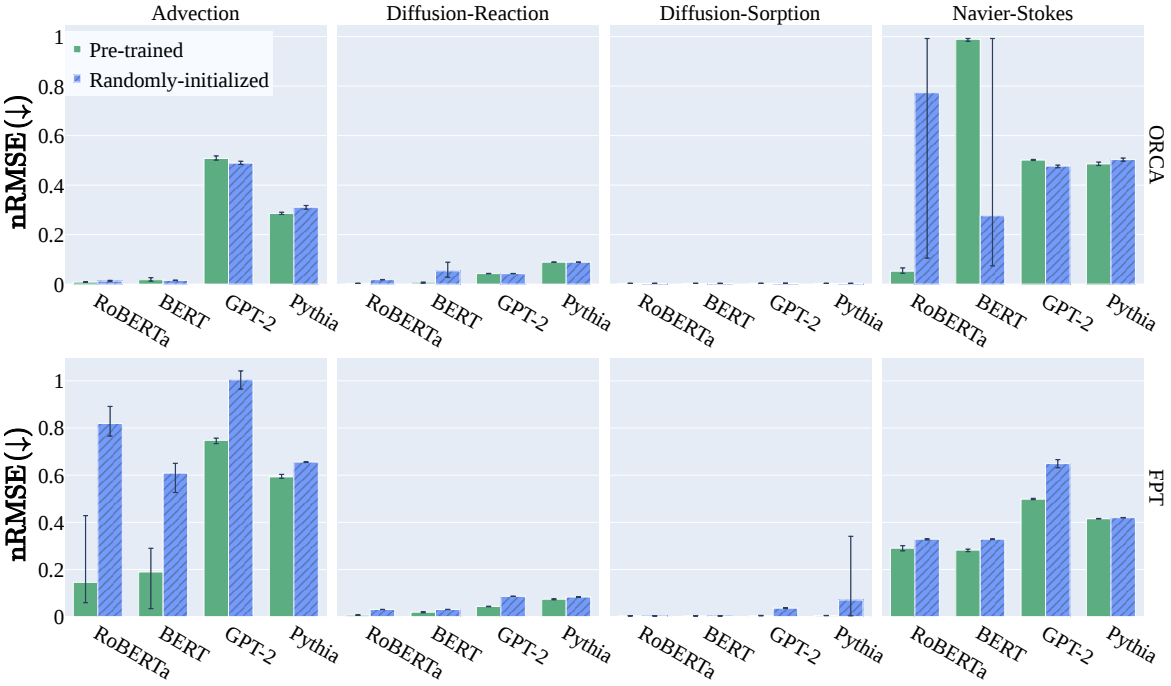

Figure 2: Comparison of model performance with ORCA- (above) and FPT-based (below) cross-modal adaptation, using both pre-trained and randomly-initialized versions of encoder-only models (RoBERTa, BERT) and decoder-only models GPT-2 and Pythia). Performance is measured using nRSME, where lower is better; the plots show average performance over 5 random seeds, and the error bars represent the best and worst runs.

as Figure 2 shows, decoder-only models do not outperform their randomly-initialized counterparts for any of the tasks, while encoder-only models do so for all of the tasks, at least with one of the tested models. On the other hand, when using FPT, both encoder-only and decoder-only models do outperform their randomly-initialized versions for most of the tasks (except Pythia for Navier Strokes). However, both sets of models still show very large error compared to ORCA-based adaptation, and in some cases (for example, Pythia for Diffusion-Reaction), the performance gain is small. 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.

When comparing the performance pre-trained models of different architectures, as Figure 2 shows, **encoder-only models outperform decoder-only models overall** for three of the four selected tasks (Advection, Diffusion-Reaction, and Navier-Stokes), with very different performance depending on the task. The remaining task, Diffusion-Sorption, shows equally good performance for all models and cross-modal adaptation methods, indicating that the task is simple enough to be solved without pre-training. Similarly to what García de Herreros et al. (2024) report 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. Broadly, we also observe that ORCA achieves better results than FPT on three of the four tasks, as previously described in Shen et al. (2023).

Lastly, it must be noted that for some of these tasks, there is large variance between runs. For ORCA, all tasks are stable except for Navier-Stokes, where we can see high variance when using encoder-only models that have been randomly initialized. When using their pre-trained counterparts, this variance reduces dramatically. On the other hand, when using FPT, all tasks except for Advection seem stable. For Advection, pre-trained encoder-only models show high variance between runs. As we discuss later, this fine-tuning instability—which could come from optimizers or simply bad regions in weight space—should be investigated more systematically in future work.

Overall, our results show that **decoder-only**

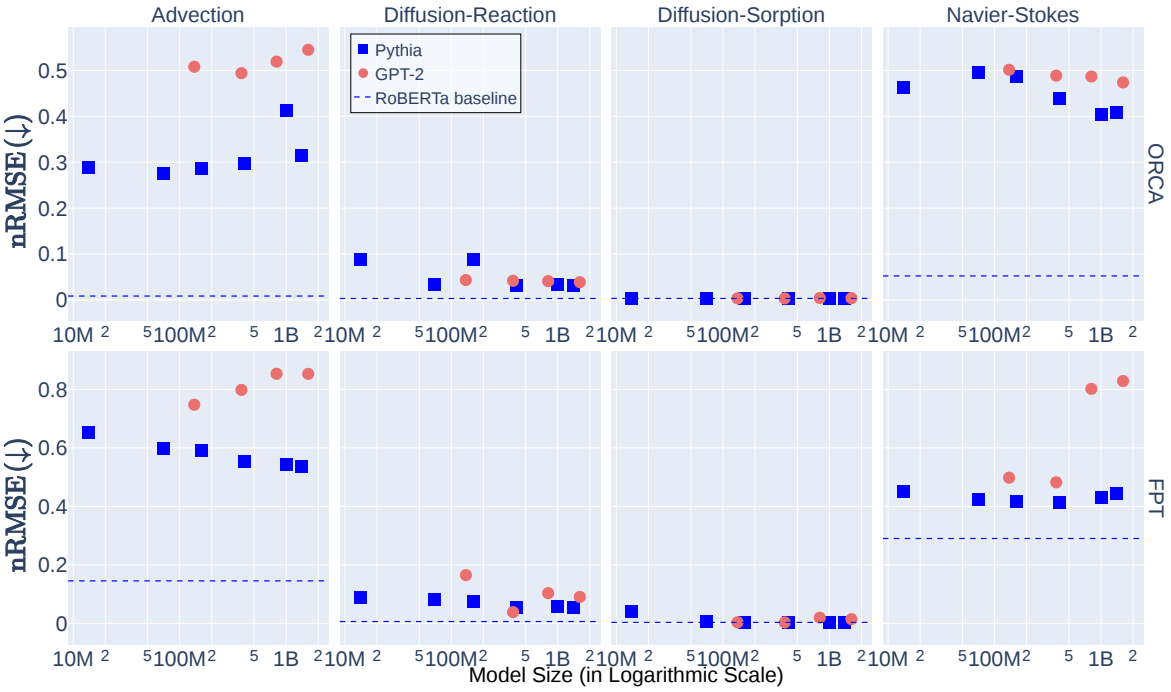

Figure 3: Performance of different sizes of models of the GPT-2 family and PYTHIA family using both ORCA (Shen et al., 2023) and FPT (Lu et al., 2022). The plots depict the average performance over 5 random seeds. Once again, performance is measured using nRSME, where lower is better. If scaling the models was improving the performance, downward trends could've been seen for the different model families.

models cannot compete with decoder-only models for PDE tasks using cross-modal adaptation methods out of the box. In the following sections, we try several approaches to close the performance gap between architectures.

5 Scaling Decoder-Only Models does not Improve Performance

The previous results motivated us to find potential ways in which decoder-only models can achieve a comparable performance to encoder-only models. Since the compared models in the last section were all of similar size, in this section, we test scaling the selected decoder-only models to see if this improves performance, as seen in other areas (Kaplan et al., 2020; Caillaut et al., 2024; Cai et al., 2025).

However, Figure 3 shows that **scaling barely reduces the performance gap between decoder-only and encoder-only models**, where ROBERTA-BASE represents encoder-only model performance. Below, we outline the trends we see.

When using ORCA, there is no performance improvement on the Advection and Diffusion Sorption datasets; for Advection, there is even some deterioration for both model families. Diffusion-Reaction shows some improvement with the PYTHIA models, with some outliers, but no improvement with the GPT-2 models. On the other hand, Navier-Stokes shows some improvement for both model families; the relative percentage improvement when comparing the best model with the smallest model of each family is 5% for the GPT-2 models versus 12% for the PYTHIA models. Still, the trend is not smooth, and the parameter increase to achieve this performance is much bigger than the performance gain; for GPT-2, the best model is approximately 12 times bigger, for PYTHIA is approximately 71 times bigger, without getting much closer to encoder-only model performance.

With FPT-based adaptation, GPT-2 models do not show consistent performance improvements, even deteriorating for Advection and Navier-Stokes. On the other hand, the Pythia family shows some improvements for Advection and Navier-Stokes, but once again, the gains are relatively small compared to the models' size difference.

Since scaling does not close the performance gap between architectures, we hypothesize that the stark differences in performance from plugging decoder-only models into these approaches are due

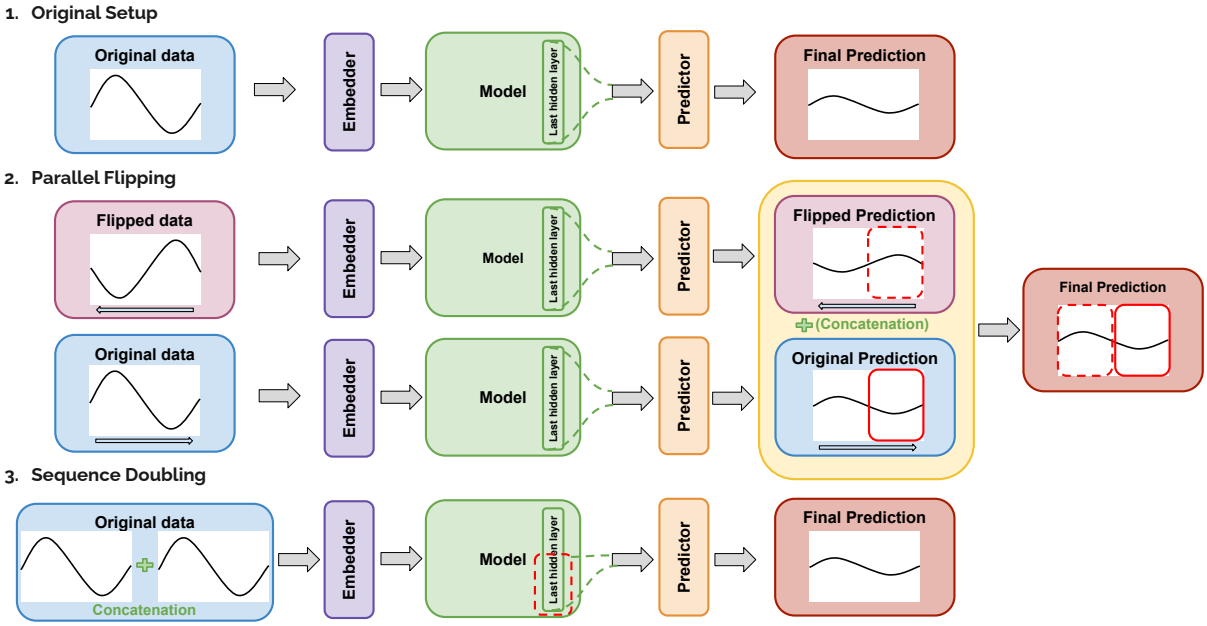

Figure 4: Pipeline comparison of the original setup and the two methods we introduce, Parallel Flipping and Sequence Doubling. For **Parallel Flipping**, the pipeline is run twice, with the original data and with the inverted sequences. For **Sequence Doubling**, each sequence is concatenated with itself before being introduced to the model, and then we only pass the second part of the last hidden layer to the predictor.

to two reasons: First, they are penalized for being autoregressive, since each point in the sequence is treated as an individual token, and GPT-2 and PYTHIA 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 simply averaged. This does not take advantage of the strong generative capabilities that decoder-only models possess.

We leave exploring the potential of the generative capabilities of encoder-only models for cross-modal adaptation for these particular tasks for future work, and in the following sections, we focus on addressing our first hypothesis as a means to improve the performance of decoder-only models using cross-model adaptation approaches.

6 Simulating Bidirectionality With Decoder-Only Models

Since scaling decoder-only models does not improve their performance, we introduce two novel methods to counter the lack of bidirectional context in the models, illustrated in Figure 4.

6.1 Parallel Flipping

Through error analysis of the decoder-only model outputs, we observed that the beginnings of the output sequences were generally more spiky but they became smoother as the sequence progressed, since the model has more context to condition on. We show some examples of this in Appendix D.

Using this to our advantage, we design a new method to give both halves of the sequence equal opportunity to condition on the other. As shown in Figure 4, we run the same cross-modal pipeline twice in parallel (both for ORCA and FPT), once with the original data and once with the data sequences inverted. Then, we combine both predictions by taking the second half of each from the original run and the inverted one and concatenating them. In this way, both parts of the predicted sequence have access to the previous context and we obtain the smoother part of both runs, even though the point at which they are concatenated can still be spiky. Compared to the original cross-modal adaptation approach, the second half of the final prediction remains unchanged with Parallel Flipping, but the first half may now improve through conditioning on the flipped version.

6.2 Sequence Doubling

To expand the context window the model can use beyond half the sequence (as in Parallel Flipping) to the full sequence, we introduce sequence doubling. As shown in Figure 4, we concatenate all the

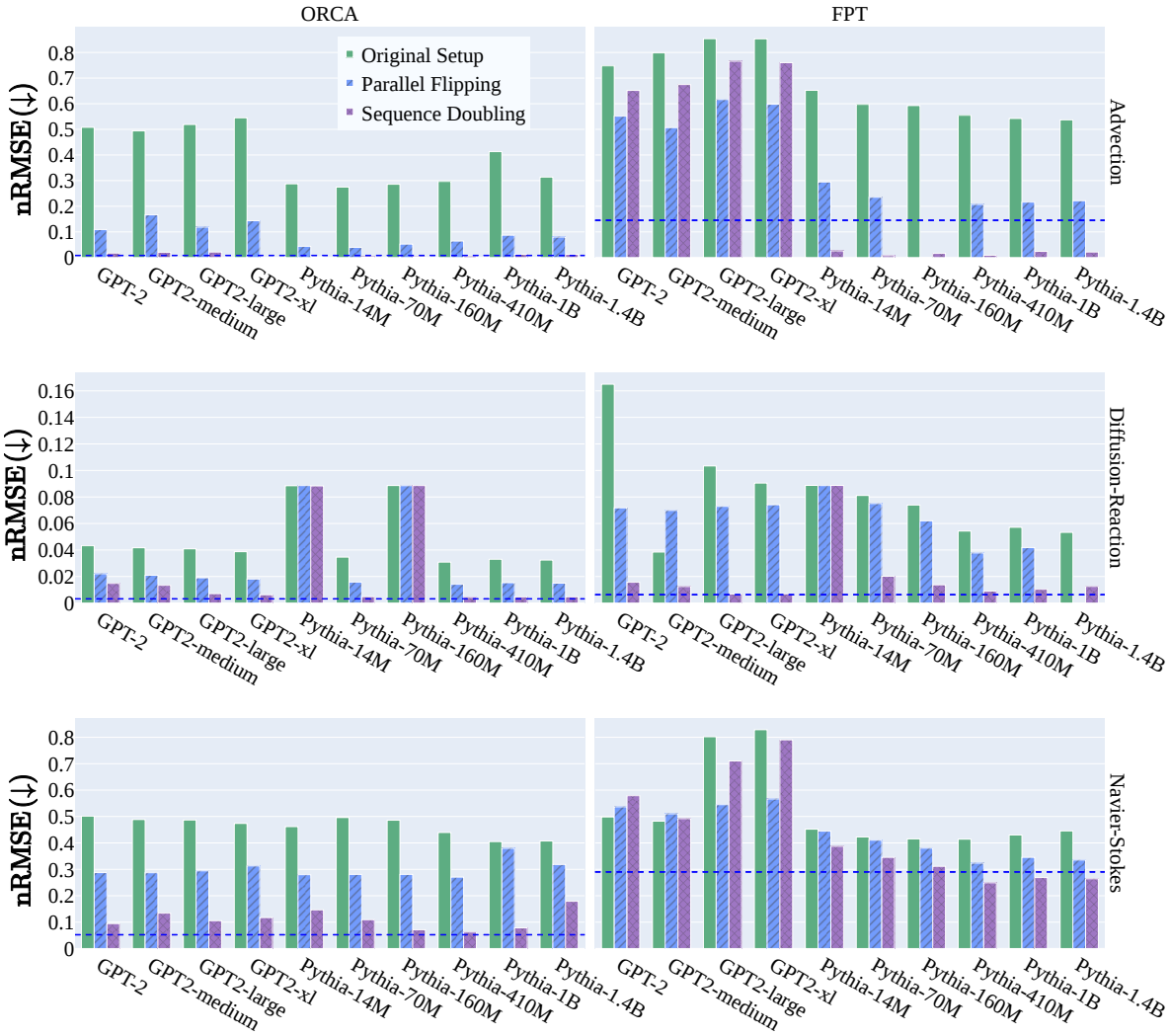

Figure 5: Performance comparison of the original setup versus our own two methods, Parallel Flipping and Sequence Doubling, using both ORCA (Shen et al., 2023) and FPT (Lu et al., 2022). We set RoBERTa with the original setup as a baseline for all the configurations. The plots depict the average performance over 5 random seeds. Performance is measured using nRSME, where lower is better.

sequences introduced to the model with themselves. Then, for the prediction, we take only the second half of the last hidden layer and introduce it to the predictor. This half of the hidden layer is conditioned on the first instance of the entire sequence and should therefore be a much richer representation of the data. Indeed, a similar kind of sequence repetition has also shown promising results in the context of language model embeddings (Springer et al., 2025). Compared to Parallel Flipping, this approach also does not have a hard concatenation point, which should result in smoother and better outputs overall, leading to bigger improvements.

7 Simulating Bidirectionality Closes The Performance Gap Between Architectures

We compare our two newly-proposed methods with the original setup in Figure 5, using the results of our two new methods in the following section, compared to the original setup. As Diffusion-Sorption shows equally good performance for all configurations in previous sections, we focus on the other three datasets for these remaining experiments.

Overall, both **Parallel Flipping and Sequence Doubling outperform the original setup** for all tasks and cross-modal adaptation methods. As expected, we also see that Sequence Doubling generally outperforms Parallel Flipping, with the excep-

tion of FPT-based adaptation of GPT-2 models to Advection and Navier-Stokes.

For ORCA, we see larger improvements, with Sequence Doubling outperforming Parallel Flipping for all tasks. For both Navier-Stokes and Diffusion-Reaction, the performance approaches RoBERTa-BASE's original performance, and in some cases, our approaches even outperform RoBERTa-BASE, i.e., some Pythia models on Advection (Pythia-14M, Pythia-70M, Pythia-160M, and Pythia-410M).

For FPT, improvements are less consistent than for ORCA, but still show good gains across most configurations. All models approach RoBERTa-BASE's performance on Diffusion-Reaction with Sequence Doubling, but the improvement with Parallel Flipping is smaller as before. For some configurations, decoder-only models outperform encoder-only models again (e.g., Pythia on Advection and Navier-Stokes with Sequence Doubling). In contrast, GPT-2 models show small gains here.

Despite the improvements, we still do not consistently see neat scaling behaviour on cross-modal adaptation. This could be due to a task inherently not benefiting from scaling, particularly for datasets that can already be solved with a lower-capacity model. On the other hand, the lack of clear scaling could also come from randomness in the particular model checkpoints that we use, which could also cause some of the outlier runs that we see. Adaptation stability is therefore an important area for future work in cross-modal adaptation.

8 Discussion and Future Work

With a series of experiments to analyze the effect of model architecture and size on cross-modal adaptation approaches, we show that decoder-only models are consistently worse than encoder-only models and do not, at least with traditional approaches, exploit the potential of their pre-trained knowledge for the new tasks. We show that this is due to decoder-only models being penalized for their autoregressive attention over the input. To address this penalization, we introduce two different methods, both of which come with certain tradeoffs.

First, **Parallel Flipping** requires each instance to be run twice to obtain the final prediction, but by design it can be parallelized, either using double the resources to run in the same time or running it sequentially in double the time.

On the other hand, **Sequence Doubling** cannot be parallelized. Also, since the length sequence is doubled, it takes longer to run and increases the required memory. In some cases, particularly when using bigger models, this requires reducing the batch size or upgrading our resources.

Our primary motivation with both methods was to try to mimic the data processing of encoder-only models while using decoder-only models. We did so by introducing a kind of bidirectional context. Another potential way of achieving our motivation would be to actually enable bidirectional attention in decoder-only models, as in LLM2Vec (BehnamGhader et al., 2024), or by merging encoder-only and decoder-only models as in (Charpentier and Samuel, 2024). We leave this potential new approach for future work.

However, we see the most important direction for future work as being to diagnose the instabilities of cross-modal adaptation, given the high variance of performance with some configurations. As we point out in Section 4, optimizers might play a role (Kunstner et al., 2023) as might randomness in the checkpoints we begin with. One approach would be to try to disentangle when transfer capabilities emerge for these models (and whether that is stable), particularly decoder-only models, and the influence that they have on the variation (van der Wal et al., 2025).

9 Conclusion

We conduct a series of ablation studies to investigate the impact of model architecture and size on cross-modal adaptation approaches with time-dependent simulation of partial differential equations. We find that decoder-only models perform much worse than encoder-only models, even when scaled up. Unidirectional attention plays a key role in this performance gap, preventing models from conditioning on the data overall. To mitigate the effects of the lack of bidirectionality, we introduce two novel approaches: **Parallel Flipping** and **Sequence Doubling**, both of which outperform the original setup, with Sequence Doubling showing much larger gains and closing the gap to encoder-only model performance. We encourage future research on scientific ML to build on our approach to leverage more capable decoder-only models in cross-modal adaptation research.

10 Limitations

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., 2024). Additionally, 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.

11 Ethics Statement

All datasets and models are used in accordance with their licenses and intended use.

References

Bassel Almarie, Paulo EP Teixeira, Kevin Pacheco-Barrios, Carlos Augusto Rossetti, and Felipe Fregni. 2023. Editorial - the use of large language models in science: Opportunities and challenges. *Principles and Practice of Clinical Research*, 9(1):1–4.

David Alvarez-Melis and Nicolo Fusi. 2020. Geometric dataset distances via optimal transport. In *Advances in Neural Information Processing Systems*, volume 33, pages 21428–21439. Curran Associates, Inc.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Rohan Bhatnagar, Ling Liang, Krish Patel, and Haizhao Yang. 2025. From equations to insights: Unraveling symbolic structures in pdes with llms. *arXiv preprint arXiv:2503.09986*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

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

Lincan Cai, Shuang Li, Wenxuan Ma, Jingxuan Kang, Binhui Xie, Zixun Sun, and Chengwei Zhu. 2024. Enhancing cross-modal fine-tuning with gradually intermediate modality generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 5236–5257. PMLR.

Gaëtan Caillaut, Mariam Nakhlé, Raheel Qader, Jingshu Liu, and Jean-Gabriel Barthélemy. 2024. Scaling laws of decoder-only models on the multilingual machine translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1318–1331, Miami, Florida, USA. Association for Computational Linguistics.

Lucas Georges Gabriel Charpentier and David Samuel. 2024. GPT or BERT: why not both? In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024. Robust pronoun fidelity with english llms: Are they reasoning, repeating, or just biased? *Transactions of the Association for Computational Linguistics*, 12:1755–1779.

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

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. 2023. Noise is not the main factor behind the gap between sgd and adam on transformers, but sign descent might be. *arXiv preprint arXiv:2304.13960*.

Shanda Li, Tanya Marwah, Junhong Shen, Weiwei Sun, Andrej Risteski, Yiming Yang, and Ameet Talwalkar. 2025. Codepde: An inference framework for llm-driven pde solver generation. *arXiv preprint arXiv:2505.08783*.

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

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

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

Wenxuan Ma, Shuang Li, Lincan Cai, and Jingxuan Kang. 2024. Learning modality knowledge alignment for cross-modality transfer. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 33777–33793. PMLR.

François Petitjean, Jordi Inglada, and Pierre Gancarski. 2012. Satellite image time series analysis under time warping. *IEEE Transactions on Geoscience and Remote Sensing*, 50(8):3081–3095.

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

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

Junhong Shen, Liam Li, Lucio M. Dery, Corey Staten, Mikhail Khodak, Graham Neubig, and Ameet Talwalkar. 2023. Cross-modal fine-tuning: Align then refine. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31030–31056. PMLR.

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

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2025. Repetition improves language model embeddings. In *The Thirteenth International Conference on Learning Representations*.

Anton Surkov, Vladimir Zakharov, Sergei Koltcov, and Vera Ignatenko. 2024. Application of large language models to solving differential equations: Constructing baseline models with lstm and gru. In *International Conference on Smart Technologies, Systems and Applications*, pages 239–252. Springer.

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

Oskar van der Wal, Pietro Lesci, Max Muller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. 2025. Polypythias: Stability and outliers across fifty language model pre-training runs. *arXiv preprint arXiv:2503.09543*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes: A benchmark to assess the linguistic competence of language models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647.

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

Wenzhuo Xu, Noelia Grande Gutierrez, and Christopher McComb. 2023. Megaflow2d: A parametric dataset for machine learning super-resolution in computational fluid dynamics simulations. In *Proceedings of Cyber-Physical Systems and Internet of Things Week 2023*, CPS-IoT Week '23, page 100–104, New York, NY, USA. Association for Computing Machinery.

Vladimir Zakharov, Anton Surkov, and Sergei Koltcov. 2025. Agdes: a python package and an approach to generating synthetic data for differential equation solving with llms. *Procedia Computer Science*, 258:1169–1178.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 117–139, Miami, FL, USA. Association for Computational Linguistics.

A PDE Datasets Details and Configurarions

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.

B Proxy Datasets

To create proxy datasets for GPT-2, GPT-2 MEDIUM , GPT-2 LARGE , and GPT-2 XL, we follow the approach detailed in Shen et al. (2023). Due to discrepancies between the stated dataset and instructions in Shen et al. (2023), we use the CoNLL-2000 dataset (Sang and Buchholz, 2000) instead of CoNLL-2003. We select a random sample of 2000 sequences containing less than 32 tokens. We unify the length by padding to a sequence length of 32. Lastly, we calculate the embeddings using the selected models.

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

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

C Hardware

We use Nvidia A100 GPUs to run all experiments, the longest of which took 140 GPU hours.

D Error Analysis Examples

We show the comparison of the predicted waves for different examples of Advection and Diffusion-Reaction with different models and the ground truth wave. In Figures 6 and 7 we can see that the predicted waves are more spiky and irregular in the first half of the wave than in the second half.

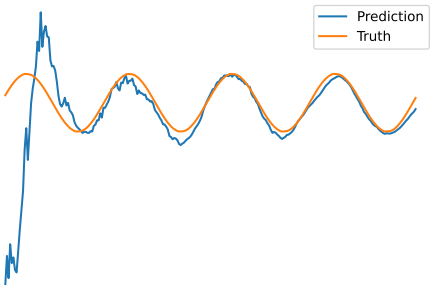

Figure 6: Comparison between GPT-2 prediction on an Advection example using ORCA as the cross-modal adaptation method and the ground truth.

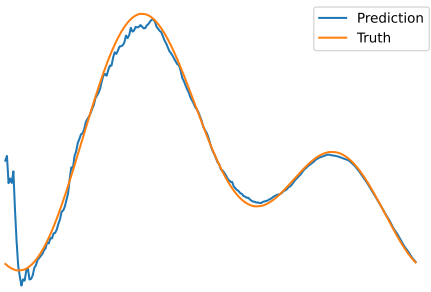

Figure 7: Comparison between PYTHIA prediction on an Advection example using ORCA as the cross-modal adaptation method and the ground truth.