Scientific Foundation Models II: In-Context Learning for Differential Equations

Yassir El Attar

Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Stuttgart, Germany
st191841@stud.uni-stuttgart.de

Abstract. Recent advances in scientific machine learning have introduced novel strategies and approaches for learning partial differential equation (PDE) operators using foundation model techniques. This report reviews two recent developments, unsupervised pre-training of neural operators and in-context learning (ICL) for operator generalization, that enable significant reductions in simulation costs while enhancing the adaptability to out-of-domain (distribution) tasks. This report examines the key model architectures, benchmark evaluations, and the conceptual integration of such methods, which offer a unified view of how pre-training and ICL can guide scalable and efficient operator learning across different physical domains.

Keywords: Scientific machine learning · Operator learning · In-Context Learning · Unsupervised pre-training · PDEs

1 Introduction

First-principles models such as differential equations have been used traditionally in systems to describe complex and high-dimensional problems in many scientific fields such as physics, chemistry, and biology. Recently, machine learning has emerged as a very powerful tool to approximate these systems directly from data, enabling faster and more flexible simulations of physical problems. Such advances include techniques like *operator learning*, which aims to learn mappings between infinite-dimensional function spaces. It gained prominence due to its ability to generalize across various input conditions and types. However, in spite of the promising results this technique offers, the traditional operator learning methods often require large volumes of high-fidelity data for simulation and often struggle with out-of-distribution generalization, especially when they are deployed in real-world scientific scenarios where data is expensive to obtain.

The growing interest in *foundation models* (large, pre-trained models capable of adapting to a wide range of downstream tasks) has begun to impact the scientific machine learning (SciML) community. Inspired by the success of in-context learning (ICL) in natural language processing (NLP) such as GPT models, much recent work explores whether similar techniques and mechanisms can be adapted

to solve families (e.g., partial) of differential equations. Following the same settings, in-context learning refers to the model's ability to adapt to new tasks at inference time by conditioning on a few input-output examples (or 'demos') without the need for updating its internal parameters. This paradigm offers a compelling alternative to re-training or fine-tuning models for each new physical system.

Two recent contributions have advanced this research line. First, the ICON framework (proposed in [7]) introduces a new architecture that enables neural networks (NNs) to act as few-shot operator learners. By structuring the inputs as key-value prompts and queries, ICON can infer new solution operators from just a handful of samples and generalize to new conditions or distributions without any need for weight updates. Second, [6] propose a complementary approach that combines both unsupervised pre-training on unlabeled PDE data and similarity-based in-context examples retrieval. Their method reduces the need for simulated solutions significantly and achieves better generalization to new and unseen PDE families, which outperforms vision-pre-trained baselines in multiple settings.

This report provides a simple overview of these recent advances, examining the conceptual foundations of in-context operator learners; the methodological innovations introduced in this work enable data efficiency in training and have wider implications for developing scientific foundation models. In this context, the aim here is to contextualize such contributions within the evolving landscape of scientific machine learning modeling.

2 Scientific Operator Learning: Foundations

There are many problems in scientific computing that are described using differential equations, where the goal is to predict the behavior of a system based on the initial conditions, parameters, or force functions. Traditional machine learning methods are usually designed to learn a finite-dimensional mapping between vectors or matrices. However, scientific problems often involve solving *operator*-mappings between infinite-dimensional function spaces. For example, solving a PDE entails mapping an input function (e.g., a boundary condition or a source) to an output function (e.g., the solution field over space and time).

Operator learning formalizes this objective by training models to approximate these functional mappings directly. Let $\mathcal{G} : \mathcal{A} \to \mathcal{U}$ denote an operator that maps an input function $a \in \mathcal{A}$ to a solution $u \in \mathcal{U}$. Instead of learning pointwise solutions for every specific parameter setting, operator learning aims to approximate $\mathcal{G}$ over the entire function space $\mathcal{A}$. This abstraction enables the model to generalize across different PDE coefficients, boundary conditions, and/or external forces, which makes them more flexible than traditional solvers or regression models.

This formulation starts from conventional machine learning, which generally operates in finite-dimensional spaces. On one hand, typical neural networks, such as multilayer perceptrons (MLPs) or convolutional neural networks (CNNs), learn mappings between fixed-size inputs and outputs, and they often require

re-training even for simple, small changes in the input structure. On the other hand, operator learning models must process variable-resolution input functions and produce continuous outputs, often defined on different domains.

To address this challenge, there have been some newly developed specialized architectures. One of the earliest and most influential is **Deep Operator Network (DeepONet)** [3]. DeepONet consists of two sub-networks: the branch net, which encodes the input function, and the trunk net, which evaluates the output function at some specified coordinates. The two outputs are combined through an inner product, which allows the model to represent non-linear operators efficiently. DeepONet is grounded in the universal approximation theorem for operators to ensure its theoretical expressiveness.

Another common and widely used architecture is the **Fourier Neural Operator (FNO)** [2], which learns the operator in Fourier space. FNO replaces the traditional convolutions with Fourier transforms, then follows them with some learnable multipliers, which enable it to efficiently capture the global dependencies in the data. By exploiting the smoothness or other structures that are common in physical systems, FNOs have shown strong performance on a broad range of PDEs, such as time-dependent Navier-Stokes and reaction-diffusion systems. FNO is well-suited for problems especially on regular grids. It has also been extended to handle more complicated geometries through graph and multipole variants.

Both DeepONet and FNO highlight an important feature of operator learning: the ability to learn representations that bypass the inter-relationships in data during solution procedure. Instead of directly approximating PDE solvers, such models learn how the solution operator behaves across different families of problems to create fast inference and re-usable frameworks. This opens the door to applying machine learning to real-time scenarios, such as control, optimization, and uncertainty quantification tasks that are typically too expensive to handle with classical numerical solvers.

3 In-Context Learning in Scientific Models

One of the biggest challenges that remain unsolved in AI development is the ability to adapt an ML model to new situations without the need for re-training. This is where the approach of meta-learning (i.e., learning to learn) comes into play. Meta-learning is inspired by human capabilities to learn and adapt to new situations just by observing a few examples or demonstrations. For instance, a driver accustomed to manual cars may struggle with automatic ones, but observing someone else or receiving a simple demonstration allows them to easily transfer their skills and adapt to the new car's system. Similarly, ICL approaches aim to equip AI systems with this kind of adaptability - enabling models to generalize to new tasks based on just a few given examples without re-training.

ICL is a paradigm where the model is conditioned on a set of input-output samples, usually called 'demos', in order to make predictions for new queries without having to update the model's parameters. This approach has been pop-

ular due to the advances in large language models (LLMs), such as GPT-3 [1]. LLMs have demonstrated the powerful capabilities of few-shot learning by relying purely on prompting. Therefore, in the light of scientific machine learning, the promise of ICL is particularly appealing: instead of re-training the whole model for each new PDE or parameter, a single model could generalize by interpreting new tasks from in-context examples.

The **In-Context Operator Network (ICON)** framework [7] extends this idea from LLMs to operator learning. ICON model treats the learning of solution operators for differential equations as a form of meta-learning. During inference, the network receives a prompt that consists of multiple key-value pairs, where the keys represent the inputs in a discretized form (e.g., boundary conditions) and values are the corresponding outputs (quantities of interest). The prompt also contains a new input (query condition), then the model is expected to return the correct output prediction (query solution) by learning, implicitly, the operator defined by the demo pairs.

The ICON framework is based on a transformer-architecture with an encoder-decoder structure, which can be summarized as:

- **(Context) Encoder:** Takes the full prompt, which contains a number of 'demos' (conditions and their respective quality of interests) and the 'question condition' to map each input-output pair into a latent representation (embedding) of the operator. It combines the embeddings from all context examples, typically using attention or pooling mechanism, to form a context summary.
- **(Query) Decoder:** Takes the aggregated context summary (latent embeddings) and the query input (i.e., keys of question quality of interest) to predict the corresponding output solution. It does not contain any self-attention layers for queries, which allows the independence of each output-corresponding query pair: unaffected by the others.

In contrast to previous methods for operator learning, which train separate models for each PDE or operator, this framework (ICON) creates one single model that is able to adapt to a broad variety of tasks in a similar way to zero-/few-shot learning without any need for parameter updates. Leveraging this architecture, ICON can generalize better to unseen operator tasks and out-of-distribution conditions. It serves as a good example of demonstrating the powerful capabilities of zero-/few-shot learning in scientific problems. It bridges meta-learning with operator learning, providing a foundation for a more robust, flexible, and data-efficient approach to scientific models.

The framework at hand supports many different applications, such as inverse and forward problems in ordinary differential equations (ODEs), partial differential equations (PDEs), and mean-field control (MFC). The experiment results show that ICON can not only generalize to unseen operators, but also to equation forms, while maintaining low error even when the structure of the problem changes. Another powerful ability of ICON is its robustness to variations in resolution and dimensionality, making it still perform well on both sub-resolution and super-resolution settings.

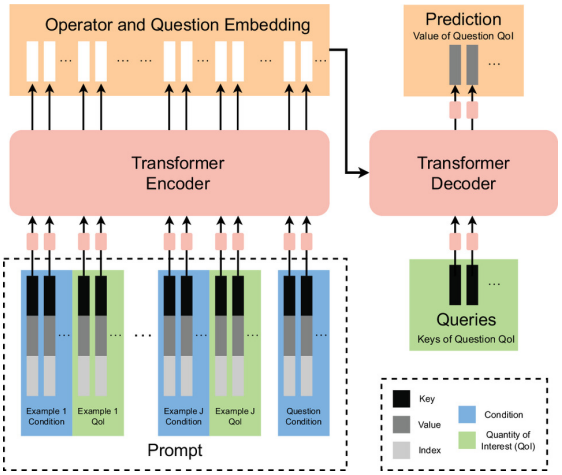

Fig. 1. ICON model neural network architecture.
Source: *In-Context Operator Learning with Data Prompts for Differential Equation Problems [7].*

4 Unsupervised pre-training for Operator Learning

In general, labeled data is expensive, but the narrower a field is, the more expensive it becomes. Scientific machine learning is no exception, as generating labeled data comes with a high cost, imposing a persistent challenge, especially for PDE simulations. Solving such high-resolution equations usually requires intensive computational power for numerical solvers, which can become almost impossible when exploring large parameter spaces or conducting sensitivity analysis. As a solution to address this challenge, recent work has focused on adapting unsupervised pre-training strategies and methods, which have demonstrated success in NLP and computer vision (CV). This line of work aims to translate such success to the domain of operator learning.

[6] introduce a novel framework, which pre-trains neural operators on unlabeled PDE data through physics-informed proxy tasks. The main insight they discuss is that while obtaining full PDE solutions is very expensive, the collection of input data (e.g., boundary conditions, initial states, spatial-varying coefficients) is relatively cheaper. Therefore, the authors exploit such case to create large-scale pre-training datasets that contain only PDE inputs, eliminating the need for simulation-generated outputs.

Their approach focuses on two central proxy tasks: **masked autoencoding (MAE)** and **super-resolution (SR)**. The task of MAE is mainly inspired by similar approaches used in NLP and CV, where parts of the input are masked and the model is trained to reconstruct those masked parts. For PDEs, they partially hide some of the input fields (e.g., coefficient functions or initial conditions) through spatial masking. Then, they train the model to reconstruct the complete field. This method encourages the model to learn local and global

dependencies in the input space and guide it to extract physically meaningful representations. The SR task, however, simulates the process of recovering precise and fine-grained information from previously coarse-grained observations. By utilizing techniques similar to the ones used in CV or NLP, such as blurring or downsampling, they generate low-resolution snapshots of the PDE inputs. Then, they train the model to reconstruct the high-resolution versions of these snapshots. This task is in particular relevant for PDEs like Navier-Stokes equations, where identifying fine-scale dynamics is critical and important for accurate modeling.

Similar to unsupervised pre-training in CV or NLP, after the pre-training on these proxy tasks, the model is then fine-tuned on a smaller set of labeled PDE data. These two steps mirror well the strategies that are well-established in foundation models from other domains. [6] argue that this approach results in significant improvements in data efficiency: the models here require up to 1000x fewer labeled simulations to reach comparable accuracies to supervised models. Additionally, they claim that the convergence during the fine-tuning step is notably faster.

The framework is evaluated by making use of two model architectures: the **Fourier Neural Operator (FNO)** and a **Transformer-based encoder-decoder**, which is derived from the 'VideoMAE' model [5]. FNO operates directly in the Fourier domain in order to capture the global structures in PDE data, but the transformer-based architecture tokenizes the input patches and applies a self-attention mechanism to model time and space dependencies. Both these models benefit remarkably from the unsupervised pre-training stage, resulting in consistent success observed in out-of-distribution task generalization, stability, and outperformance in time-dependent settings. This is a particularly important step to buidl a robust and refined foundation model for scientific systems and applications.

5 Combining Pre-training and In-Context Learning

While ICL and unsupervised pre-training on unlabeled data have demonstrated remarkable success individually in enhancing the performance of operator learning models, there are suggestions that the combination of the two can yield even more powerful capabilities. They can benefit from one another to achieve even better performance particularly in the context of out-of-distribution (OOD) generalization (see Figure 2 for a general structure overview to combine unsupervised pre-training with ICL). [6] focused on unsupervised training, but also suggest that models pre-trained on unlabeled data can benefit more from in-context demos at inference time without the need for any parameter updates. This combination or hybrid approach leverages both the prior knowledge that the model learns during pre-training and the task-specific guidance that the in-context examples provide.

The main idea is to add labeled 'demo' example pairs of input conditions and solutions (drawn from the same, but previously unseen, distribution as the

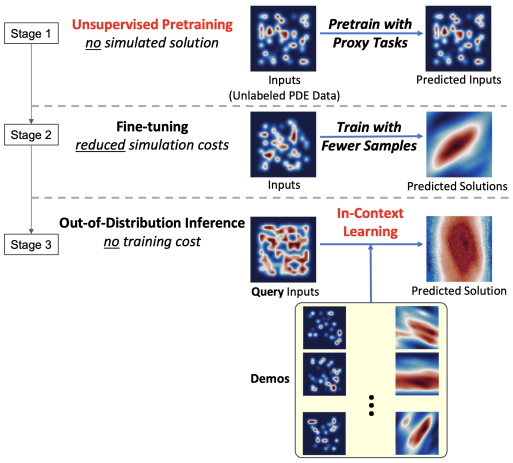

Fig. 2. Overview of framework for data-efficient neural operator learning.
Source: *Data-Efficient Operator Learning via Unsupervised pre-training and In-Context Learning [6].*

query) during test-time inference. The model then uses these example pairs to add context to the query input and eventually refine its predictions. Additionally, and most importantly, this process is conducted without any gradient-based fine-tuning, which preserves computational efficiency at inference time.

This method however relies heavily on the selection of relevant in-context examples. [6] compare two techniques to find similar 'demos':

- **Similarity by output prediction**: The trained neural operator starts by generating predictions for a set of candidate demo inputs. These predicted solutions are then compared (using ℓ_2 distance) to the query's predicted solution: the most similar ones are selected.
- **Similarity by encoder features**: The feature embeddings from the encoder (e.g., latent representations right before the final prediction layer) are used to calculate the similarity. While this might seem appealing from the first intuition, they found that this method is less effective in identifying meaningful functional relationships for operator inference.

The empirical results demonstrate that when the number of in-context samples increases, the model's generalization to OOD scenarios improves significantly. This effect is particularly prominent for the case of problems like the Navier-Stokes and Helmholtz equations, where the model performance usually decreases due to distributional shifts. However, the selection of demos that are most relevant to the query in a dynamic way makes the model able to adapt in a flexible and data-driven manner.

Such strategy highlights also an important aspect of the framework: **scalability**. Since there is no need for additional training, the in-context examples

can be used to improve the predictions without requiring any re-training. In addition, the approach supports **zero-shot adaptability** where the pre-trained model can be applied to a completely different PDE system that it has not seen during fine-tuning, supported by carefully selected demos. The integration of pre-training and in-context adaptation provides a promising path forward for scientific foundation models that could operate in diverse and data-scarce situations.

These techniques work hand in hand towards bridging the strengths of pre-trained foundation models and the flexible reasoning capabilities that ICL enables. Mirroring the foundation models in NLP or CV, the pre-training on unlabled PDE data creates a model that constructs physics-aware latent representations, which further help guide it to perform well on OOD tasks via ICL. The system that emerges from these techniques is more robust, generalizable, and efficient (essential traits to be deployed in real world scientific and engineering applications).

6 Experimental Highlights

In order to evaluate how effective the techniques of unsupervised pre-training and ICL are for operator learning, [6] conducted more extensive experiments across both synthetic and real-world datasets. The evaluation covers a wide range of benchmarks, highlighting the generality, scalability, and data efficiency of their approach.

6.1 Benchmarks on Canonical PDEs

- **Poisson equation**: A classical elliptic PDE that is used in the electrostatics and heat distribution problems. The main goal is to map the source terms and diffusion coefficients to solution fields.
- **Helmholtz equation**: More challenging oscillatory PDE. It is common in wave propagation and acoustics. It presents difficulties in generalization due to its high-frequency behavior.
- **Reaction-Diffusion (RD) system**: A time-dependent PDE system with non-linear couplings. It is used to model pattern formation and chemical dynamics.
- **Navier-Stokes (NS) equations**: A 2D incompressible form that is used here to model fluid dynamics. It works as a high-dimensional, time-dependent benchmark, especially relevant in turbulence modeling.

Results (as in Figure 3) show that pre-training on unlabeled PDE data followed by a supervised fine-tuning phase consistently improves model performance across all different PDE types. The pre-trained model achieves comparable or even better test accuracy than the one trained from scratch while using 10^3 to 10^5 fewer simulated training samples. In the Navier-Stokes case in particular, this method leads to a substantial reduction in computational costs (up to 8×10^4).

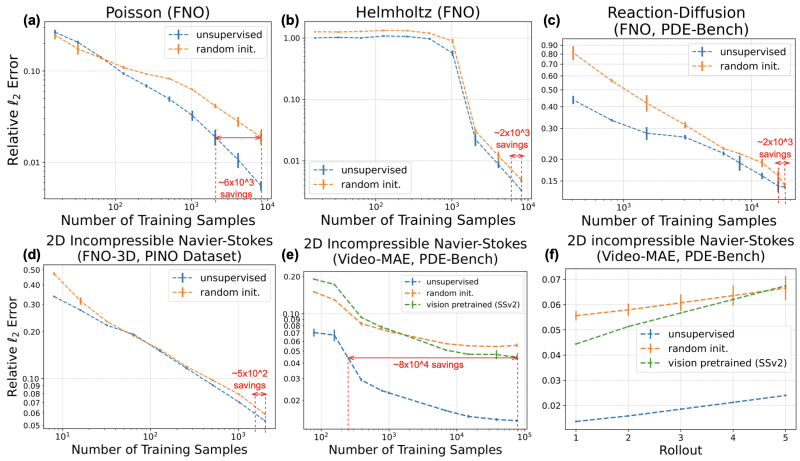

Fig. 3. pre-training neural operators on unlabeled PDE data improves their performance and data efficiency on Poisson (a), Helmholtz (b), Reaction-Diffusion (c), and Navier-Stokes (d and e, with relative errors at different unrolled steps shown on f). "random init.": models are trained from scratch with random initialization. "vision pretrained (SSv2)": fine-tuning from the publicly available checkpoint for Video-MAE. Savings of the number of simulated PDE data (when "random init." achieves the best test error) are shown in red.
Source: *Data-Efficient Operator Learning via Unsupervised pre-training and In-Context Learning [6].*

6.2 Generalization to Real-World Scientific Data

They authors not only tested their framework on synthetic benchmarks, but also on real-world datasets that capture the complexity and noise of physical systems:

- **ERA5**: A re-analysis dataset from the European Centre for Medium-Range Weather Forecasts, which provides high-resolution atmospheric variables such as temperature and wind.
- **ScalarFlow**: A dataset of real smoke plume simulations with volumetric reconstructions from multiple view points.
- **Airfoil flow**: A 2D dataset of fluid simulation around different airfoil geometries steady-state.

In each of the cases above, the pre-trained models show strong generalization and efficient adaptation with minimal labeled data. The framework outperforms pre-trained models that trained on vision datasets (e.g., VideoMAE), which further confirms the importance of domain-specific pre-training.

6.3 Behavior of In-Context Learning

The ICL method is evaluated during OOD testing. The models here are required to infer operators for physical parameters that have not been seen during train-

ing. As visualized in Figure 4, the results demonstrate that ICL systematically adjusts the predictions (solutions) in range and structure as more demos are being fed to the model. Quantitative analysis reveals consistent improvements in ℓ_2 error and confidence with increasing number of in-context examples.

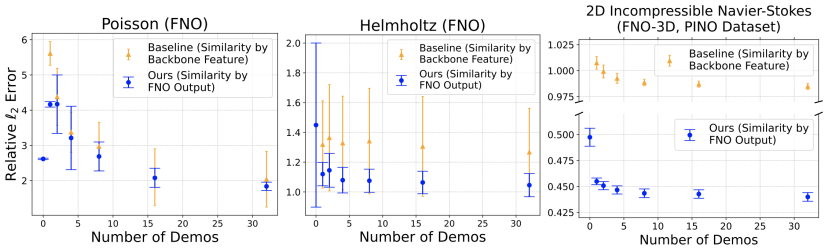

Fig. 4. In-context examples for OOD testing. **Their method (blue) decreases ℓ_2 errors and improves confidence as the number of demos increases.** Theirs (Similarity by FNO Output)": They leverage the output (prediction) of neural operators to find similar samples.
Source: *Data-Efficient Operator Learning via Unsupervised pre-training and In-Context Learning [6].*

The authors decompose prediction error into two main components to understand such behavior:

- **Scale error**: Misalignment in the overall magnitude of the prediction, which is measured by the slope of linear regression between the predicted and the ground truth values.
- **Shape error**: Differences in the spatial or temporal pattern of the predictions, and assessed by normalized relative error after scale normalization.

They find that ICL improves the scale calibration in the most effective way, especially in complex PDEs like Navier-Stokes and Helmholtz. This suggests that in-context demos help the models adjust their estimates - a key failure point in OOD generalization. Their findings indicate that ICL not only improves accuracy, but also physical plausibility in the model's outputs.

These experiments highlight the broad applicability and efficiency of combining unsupervised pre-training on unlabeled data with in-context learning. The approach significantly reduces computational (simulation) costs, enhances generalization, and brings us closer to possible deployable and general-purpose scientific AI systems.

7 Conclusion

The integration of ICL and unsupervised pre-training represents a fundamental step forward to develop more **general-purpose scientific foundation mod-**

els. These methods offer clear advantages over traditional supervised approaches, particularly in the case of expensive or sparse data. While unsupervised pre-training produces less costly simulations, ICL (such as in ICON) enables fast adaptation to new tasks and domains through zero-/few-shot learning conditions. Together, they form a powerful framework for data-efficient, scalable, flexible, and generalizable operator learning models.

This work aligns closely with recent methods applied in other domains (NLP or CV) adapting the important approaches used to build refined foundation models. Their results are impressive in generalization and adaptability. However, several open questions remain unanswered. For instance, the design of the physics-inspired proxy tasks (e.g., masking super-resolution) is still heuristic. It is not clear yet how sensitive the performance is to the choice and configuration of these tasks. Another question mark is concerning the different boundary conditions, domain geometries, or non-linear dynamics; while canonical PDE results are strong, this question remains a significant challenge to address. Lastly, while scalability is a key finding for this approach, it still remains an open question of whether it is scalable to more complex and real-world scientific domains, such as multi-physics systems; these may require more advances in the model architecture, training, and robustness. In general, while both frameworks (ICON and unsupervised pre-training + ICL) show promising results, the experiments conducted so far are still limited to canonical PDEs and synthetic datasets. More work is needed to validate their effectiveness in broader scientific applications.

This report has reviewed recent advances in scientific machine learning, focusing on the two approaches of ICL and unsupervised pre-training. They provide a significant milestone laying the groundwork for further research and improvements for future scientific machine learning models. They hold promising insights into applications in broader areas such as climate models, fluid dynamics, engineering design, and biology for an accelerated scientific discovery.

Ethical Acknowledgement

Parts of this review have been proofread using OpenAI's ChatGPT-4o with the prompt: *"Proofread the following paragraph. Do not change anything, but correct grammar, vocabulary, and spelling mistakes."*

Bibliography

[1] 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., Amodei, D.: Language models are few-shot learners (2020), https://arxiv.org/abs/2005.14165

[2] Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., Anandkumar, A.: Fourier neural operator for parametric partial differential equations (2021), https://arxiv.org/abs/2010.08895

[3] Lu, L., Jin, P., Pang, G., Zhang, Z., Karniadakis, G.E.: Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. Nature Machine Intelligence **3**(3), 218–229 (Mar 2021). https://doi.org/10.1038/s42256-021-00302-5, http://dx.doi.org/10.1038/s42256-021-00302-5

[4] OpenAI: Chatgpt (may 2025 version). https://chat.openai.com (2025), accessed: 20 May 2025

[5] Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training (2022), https://arxiv.org/abs/2203.12602

[6] Wuyang, C., Jialin, S., Pu, R., Shashank, S., Dmitriy, M., Michael, W.M.: Data-efficient operator learning via unsupervised pretraining and in-context learning (2024), https://arxiv.org/abs/2402.15734

[7] Yang, L., Liu, S., Meng, T., Osher, S.J.: In-context operator learning with data prompts for differential equation problems. Proceedings of the National Academy of Sciences **120**(39) (Sep 2023). https://doi.org/10.1073/pnas.2310142120, http://dx.doi.org/10.1073/pnas.2310142120