From Text to Trajectories: GPT-2 as an ODE Solver via In-Context

Anonymous ACL submission

Abstract

In-Context Learning (ICL) has emerged as a new paradigm in large language models (LLMs), enabling them to perform novel tasks by conditioning on a few examples embedded in the prompt. Yet, the highly nonlinear behavior of ICL for NLP tasks remains poorly understood. To shed light on its underlying mechanisms, this paper investigates whether LLMs can solve ordinary differential equations (ODEs) under the ICL setting. We formulate standard ODE problems and their solutions as sequential prompts and evaluate GPT-2 models on these tasks. Experiments on two types of ODEs show that GPT-2 can effectively learn a *meta-ODE algorithm*, with convergence behavior comparable to, or better than, the Euler method, and achieve exponential accuracy gains with increasing numbers of demonstrations. Moreover, the model generalizes to out-of-distribution (OOD) problems, demonstrating robust extrapolation capabilities. These empirical findings provide new insights into the mechanisms of ICL in NLP and its potential for solving nonlinear numerical problems.

1 Introduction

In-context learning (ICL) (Brown et al., 2020) has emerged as a pivotal feature among the capabilities of LLMs. It enables models to learn effectively through contextual prompts composed of input-output pairs without relying on parameter updates (Anil et al., 2023; Thakkar and Manimaran, 2023). This in-context learning ability is credited to *emergent abilities* (Wei et al., 2022; Lu et al., 2024) of these Transformer-based LLMs (Vaswani et al., 2017). However, it is still unclear why or what these models can learn new tasks with only a few pairs of demonstrations.

Recent studies (Garg et al., 2022; Xie et al., 2022; von Oswald et al., 2023; Vladymyrov et al., 2024; Fu et al., 2024) have explored the mechanisms of ICL, primarily focusing on linear regression tasks. These works demonstrate that trained Transformer models can achieve efficiency comparable to classic methods under the ICL setting. In particular, models trained on linear examples have been shown to mimic gradient descent (von Oswald et al., 2023) and even higher-order optimization methods (Vladymyrov et al., 2024; Fu et al., 2024). However, these findings primarily focus on linear patterns or simplified problems, leaving the behavior of full nonlinear Transformer models, especially in inherently nonlinear settings like NLP, insufficiently understood.

This work investigates the applicability of ICL to nonlinear numerical problems, extending its scope to the domain of ordinary differential equations (ODEs) and examining its potential in inherently nonlinear settings. We show that *language models (e.g., GPT-2 models) can effectively learn meta-ODE solvers and exhibit strong generalization to new ODEs.* Our main contributions are summarized as follows:

- We design a tailored ICL framework for solving ODEs by encoding these nonlinear ODEs into parameterized sequence prompts. This formulation enables Transformers (e.g., GPT-2) to learn the underlying dynamics, achieving performance comparable to explicit and implicit Euler methods and surpassing them in some cases (see Figure 1 center right).

- We then demonstrate out-of-distribution (OOD) generalization of ICL in ODE solving. The framework exhibits robustness to parameter distribution shifts, with deeper Transformer models showing stronger generalization capabilities (see Figure 1 outer right).

- We evaluate the stability of ICL-based solvers through multi-parameter extrapolation tests. Our results show that ICL achieves greater

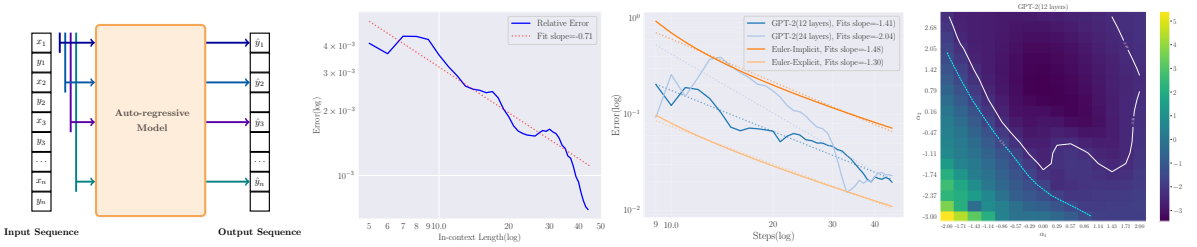

Figure 1: ICL architecture and experimental results.*Outer Left*: Illustration of ICL setup (Garg et al., 2022); *Center left*: Error variation curve as a function of context length; *Center right*: Comparison of GPT-2 predictions with classical methods; *Outer Right*: Generalization error region plotted for GPT-2 (12 layers), where α_1 and α_2 are ODE parameters scaled from the training distribution (1.5×). The dark region indicates a certain level of generalization.

stability across wider parameter ranges compared to existing Euler methods. These findings suggest that Transformer-based models, while originally developed for NLP tasks, may also be capable of solving a broader class of numerical problems. Our preliminary results indicate that such models have the potential to serve as *universal numerical solvers*.

2 Methodology

Problem setup. We study how GPT-2 models can learn to solve initial value problems (IVPs), where ordinary differential equations (ODEs) are defined by specified initial conditions (see Appendix B for formal definitions). This framework enables a systematic analysis of temporal dynamics through parameterized differential equations. Specifically, we consider the following nonlinear task: each training prompt encodes a task comprising N example pairs $(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1}^N$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ stores the ODE parameters and $\boldsymbol{y}_i \in \mathbb{R}^d$ represents the corresponding n-step solution. Each input $\boldsymbol{x}_i = (\text{Para}_i, t_e, \text{Steps}_i)$ encodes the equation's parameters, final time t_e, and number of time steps. The output $\boldsymbol{y}_i = (\boldsymbol{y}_i(t_j))$ contains the ODE solution sampled at discrete time points $t_j = 0, \ldots, t_e$ for $j = 1, \ldots, \text{Steps}_i$. We apply zero-padding to standardize outputs.

Training loss. At inference time, the model exhibits in-context learning (ICL) when it predicts $\hat{\boldsymbol{y}}_q \approx h(\boldsymbol{x}_q)$ without any weight updates, by leveraging the contextual examples in the prompt. To encourage this behavior during training, we use a sliced mean squared error (sliced-MSE) loss:

$$\ell(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{y}(:i) - \hat{\boldsymbol{y}}(:i)\|_2^2 , \qquad (1)$$

where $\boldsymbol{y}(:i)$ and $\hat{\boldsymbol{y}}(:i)$ denote the first i entries of $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$, respectively.

Experimental setups. We primarily follow the experimental setups from Garg et al. (2022). For each experiment, we apply curriculum learning (Wang et al., 2021), gradually expanding context length to 41 and vector dimensionality to 64 over the first 30k training steps. All models were trained for 600k steps before evaluation. We use AdamW (Kingma and Ba, 2015) and employ an adjusted cosine annealing schedule for optimization.

3 Experiment Results

3.1 ICL Matches Euler Methods

Building on the promising results of the GPT-2 model in solving basic differential equations (an initial trial is provided in Appendix C.1, where a 24-layer GPT-2 is introduced alongside our original 12-layer model), we extend our investigation to initial value problems (IVPs), specifically first-order linear ODEs with five degrees of parametric freedom (see Appendix 4 for the formal definition). To evaluate model performance more concretely, we conduct comparative experiments with classical numerical solvers, including both the Euler-Explicit and Euler-Implicit methods.

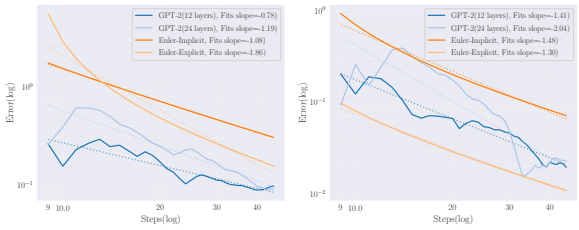

Figure 2: Log-log plots comparing GPT-2 and Euler methods. *Left*: GPT-2 outperforms Euler methods. *Right*: Comparable performance between GPT-2 and Euler. *Steps* denote context length for GPT-2 and iteration steps for Euler methods.

The model's performance is shown in Figure 2. Our benchmarks against the Euler methods reveal that the Transformer outperforms classical approaches in some cases while achieving comparable accuracy in others.

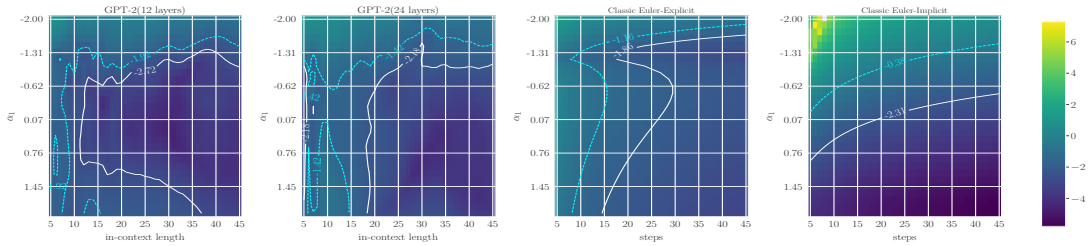

Figure 3: Heatmap of Solution error comparison under α_1 shifts (extending original training range: $\alpha_1 \in [-1, 1]$ towards $[-2, 2]$). For traditional methods, context length aligns with iteration steps for comparison. Contour lines mark 50% (cyan) and 70% (white) of each subplot's error range.

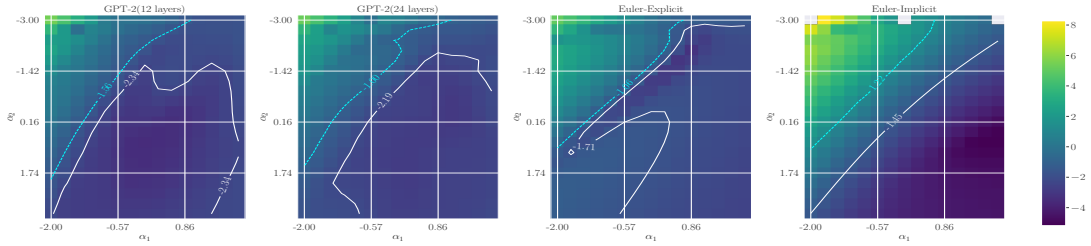

Figure 4: Heatmap of error region comparisons across parameter combinations. α_1-α_2 (trained on $[-1, 1] \times [-2, 2]$, tested on $[-2, 2] \times [-3, 3]$). From left: 12L/24L GPT-2, explicit/implicit Euler. Contours mark 50% (cyan) and 70% (white) of each subplot's range.

Transformers demonstrate dual advantages in both accuracy and adaptability compared to explicit Euler methods. Specifically, well-trained models achieve comparable or superior accuracy to Euler integration while maintaining better numerical stability, and additionally exhibit stronger adaptability when handling stiff differential equations or conducting long-term integration.

Notably, deeper architectures (24-layer vs. 12-layer) demonstrate diminishing returns when scaling depth, attaining only marginal accuracy gains despite doubled parameters – a saturation pattern consistent with findings in Fu et al. (2024). As expected, GPT-2 struggles to match classical solvers in scenarios requiring high-precision solutions, reflecting fundamental limitations of neural approximators rather than implementation flaws (see more discussions in the Appendix C.3).

3.2 ICL Generalizes across Distributions

Building on the previous subsection, which demonstrates that Transformers can effectively solve first-order ODEs as shown in Figure 2, we further evaluate the generalization capabilities of Transformers by extending parameter ranges beyond the training distribution as illustrated in Figure 3.

Our experimental results indicate that the model maintains stability while parameter range shifting, representing generalization ability. Within the original $\alpha_1 \in [-1, 1]$ range, errors monotonically decrease as context length increases. Beyond this distribution ($\alpha_1 \in [-2, 2]$), it preserves reasonable accuracy that still improves with longer contexts. Compared to Euler's stepwise error decay, Transformers exhibit smooth error convergence through adaptive in-context learning.

Notably, the 24-layer variant shows better extrapolation on negative α_1 despite overall lower accuracy than the 12-layer model, suggesting depth impacts generalization patterns.

3.3 ICL is Relative Stable to Classics

As Figure 2 reveals precision degradation in Euler methods when handling stiff problems (attributed to step size adaptation limitations), we conduct systematic cross-testing across dual parameter axes to assess whether GPT-2 with ICL exhibits analogous instability patterns.

Error stability regions. Under fixed context length and iteration steps (45 precisely), test results are shown in Figure 4. The error distributions with stable error regions exhibited by GPT-2 model are comparable to or larger than Euler methods. This likely stems from its adaptive capability—adjusting learning strategies based on context length and iteration steps to maintain high precision. In contrast, explicit Euler's fixed-step mechanism becomes suboptimal with parameter variations, leading to accu-

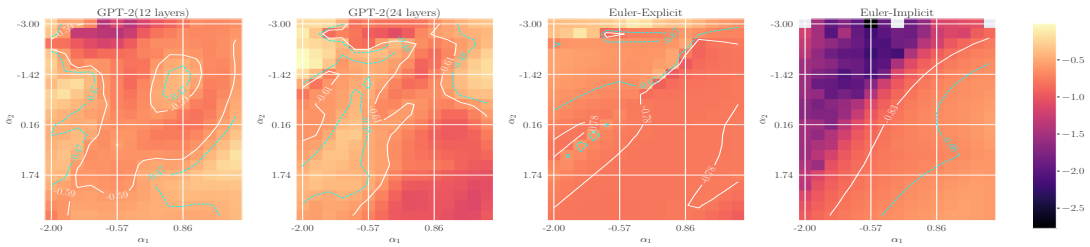

Figure 5: Heatmap of convergence slope comparisons. Models were trained on parameter ranges $\alpha_1 \times \alpha_2 \in [-1, 1] \times [-2, 2]$ and tested on a broader range of $\alpha_1 \times \alpha_2 \in [-2, 2] \times [-3, 3]$. Contour lines indicate 50% (cyan) and 70% (white) levels relative to the maximum value in each subplot.

racy deterioration.

Beyond the α parameter performance, the Transformer also demonstrates enhanced adaptation capability under β_1-β_2 shifts compared to Euler methods (Appendix C.4), while exhibiting precision variations. This phenomenon can be explained by the model's inherent sensitivity to input ordering, where consistently configured α parameters predominantly influence the attention mechanisms (Zhao et al., 2021; Lu et al., 2022; Chan et al., 2022).

Convergence slope stability. When context length and iteration steps grow, we reveal a exponential convergence. The dynamic of convergence slope is shown in Figure 5. Compared to error maps, slope maps exhibit greater instability: Euler methods maintain steady convergence rates while GPT-2 show localized volatility with sudden low-slope valleys.

We attribute this performance to: (1) Slope, as a global estimator of error reduction, amplifies instability from probabilistic solutions, whereas Euler methods only show slope degradation when step sizes are insufficient (implicit Euler demonstrates better stability); (2) For extreme parameters, in-context learning fails to capture logical parameter relationships, weakening the error reduction trend with longer contexts.

Notably, deeper models (24L) display smoother slope transitions in $\alpha_1 > 1.5$ regions, suggesting depth may mitigate convergence instability in specific parameter ranges. However, this improvement is selective—deeper models show exacerbated volatility in α_2's negative half-axis.

3.4 Comparative Performance Analysis

As shown in Table 1, the core advantages of Transformers lie in adaptive convergence and global modeling capabilities. Compared to the $\mathcal{O}(h)$ convergence of explicit Euler and $\mathcal{O}(h^2)$ of implicit Euler, 12-layer and 24-layer Transformers achieved exponential convergence rates of $\mathcal{O}(e^{-kN})$ and $\mathcal{O}(e^{-k'N})$ respectively. We hypothesize exponential convergence of model accuracy with ICL length (See Appendix D, Conjecture 1). This difference stems from the attention mechanism's comprehensive utilization of historical information, enabling implicit variable-step strategies. Notably, when handling stiff equations, traditional methods often require frequent parameter adjustments due to fixed-step limitations, whereas Transformers exhibited smoother error reduction curves.

Method	Convergence Rate	Generalization
Euler-Explicit	$\mathcal{O}(h)$	Low
Euler-Implicit	$\mathcal{O}(h^2)$	Medium
GPT-2 ($L = 12$)	$\mathcal{O}(e^{-kN})$	Medium-High
GPT-2 ($L = 24$)	$\mathcal{O}(e^{-k'N})$	High

Table 1: Comparison between GPT-2 models and traditional explicit/implicit Euler methods across various metrics.

4 Conclusions and discussions

This study demonstrates that Transformers can effectively solve ordinary differential equations through in-context learning, offering three key findings: (1) Exponential error decay with increasing context length via adaptive convergence mechanisms, (2) Generalization capability under extended parameter distributions, with maintained convergence efficacy as context length grows, and (3) Preserved convergence rates during parameter extrapolation, despite increased volatility in slope stability. Experiments confirm that the model maintains numerical stability comparable to Euler methods for nonlinear numerical problem: ODE numerical solution, while exhibiting promising generalization under parameter distribution shifts.

Limitations

Our results may have the following limitations: current observations are from only GPT-2 models. Larger model configurations could be conducted.

We conducted a conjecture, waiting for theoretical analysis of Transformers' internal mechanisms for learning differential equations, internal impact from optimization of in-context learning strategies (e.g., positional encoding and attention masking). While the model's parallel prediction offers significant advantages for real-time simulations, .

Ethical considerations. As our research involves synthetic mathematical data and does not engage with human subjects, sensitive content, or real-world applications, we do not foresee direct ethical risks.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, and 109 others. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. 2023. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in Neural Information Processing Systems*, 36:57125–57211.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Marco Bondaschi, Nived Rajaraman, Xiuying Wei, Kannan Ramchandran, Razvan Pascanu, Caglar Gulcehre, Michael Gastpar, and Ashok Vardhan Makkuva. 2025. From markov to laplace: How mamba in-context learns markov chains. In *ICLR 2025 Workshop: XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901.

Stephanie C. Y. Chan, Ishita Dasgupta, Junkyung Kim, Dharshan Kumaran, Andrew K. Lampinen, and Felix Hill. 2022. Transformers generalize differently from information stored in context vs in weights. *Preprint*, arXiv:2210.05675.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. 2024. Unveiling induction heads: Provable training dynamics and feature learning in transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.

Ezra Edelman, Nikolaos Tsilivis, Benjamin Edelman, Eran Malach, and Surbhi Goel. 2024. The evolution of statistical induction heads: In-context learning markov chains. *Advances in Neural Information Processing Systems*, 37:64273–64311.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Deqing Fu, Tian-qi Chen, Robin Jia, and Vatsal Sharan. 2024. Transformers learn to achieve second-order convergence rates for in-context linear regression. *Advances in Neural Information Processing Systems*, 37:98675–98716.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598.

Riccardo Grazzi, Julien Niklas Siems, Simon Schrodi, Thomas Brox, and Frank Hutter. 2024. Is Mamba capable of in-context learning? In *Proceedings of the Third International Conference on Automated Machine Learning*, volume 256 of *Proceedings of Machine Learning Research*, pages 1/1–26. PMLR.

Olanrewaju Victor Johnson, Chew Xinying, Khai Wah Khaw, and Ming Ha Lee. 2023. ps-calr: periodic-shift cosine annealing learning rate for deep neural networks. *IEEE Access*, 11:139171–139186.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Renpu Liu, Ruida Zhou, Cong Shen, and Jing Yang. 2025. On the learn-to-optimize capabilities of transformers in in-context sparse recovery. *Preprint*, arXiv:2410.13981.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2024. Are emergent abilities in large language models just in-context learning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5139.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Jie Ren, Qipeng Guo, Hang Yan, Dongrui Liu, Quanshi Zhang, Xipeng Qiu, and Dahua Lin. 2024. Identifying semantic induction heads to understand in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6916–6932.

Hiren Thakkar and A Manimaran. 2023. Comprehensive examination of instruction-based language models: A comparative analysis of mistral-7b and llama-2-7b. In *2023 International Conference on Emerging Research in Computational Science*, pages 1–6.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. *Preprint*, arXiv:2310.15213.

Rejin Varghese and M Sambath. 2024. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems*, pages 1–6.

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.

Max Vladymyrov, Johannes Von Oswald, Mark Sandler, and Rong Ge. 2024. Linear transformers are versatile in-context learners. *Advances in Neural Information Processing Systems*, 37:48784–48809.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. *Preprint*, arXiv:2212.07677.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. 2021. When do curricula work? *Preprint*, arXiv:2012.03107.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706.

A Related Work

ICL capabilities of LLMs. In-context learning (Brown et al., 2020) is a learning paradigm that models learn intrinsic logical relationships to make accurate predictions without parameter updates. This capability proves particularly advantageous for many NLP tasks, numerical tasks like linear regression (Garg et al., 2022; Fu et al., 2024), and Markov chain settings (Edelman et al., 2024; Chen et al., 2024; Bondaschi et al., 2025). The early exploration work (Elhage et al., 2021; Olsson et al., 2022) explains the ICL modeling via the *induction heads*. Lu et al. (2024) empirically established that LLMs' fine-tuning efficacy stems from ICL. Their large-scale tests in zero-shot settings revealed poor performance across models, confirming ICL as essential for emergent abilities. Grazzi et al. (2024) extended this research to the Mamba architecture, observing similar ICL capabilities in linear regression tasks. One can find more details in the survey of Dong et al. (2024) and references therein.

To further study whether the attention mechanisms are the key components in ICL, Ren et al. (2024) and Todd et al. (2024) identified distinct yet effective self-attention head types facilitating parameter transfer during ICL. Zhao et al. (2021) demonstrated significant output variations based on prompt ordering, a sensitivity further validated (Lu et al., 2022; Chan et al., 2022).

ICL for linear tasks. Bai et al. (2023) pioneered adaptive algorithm selection via ICL, enabling models to integrate statistical methods for regression problems. Garg et al. (2022) employed a decoder-only GPT-2 Transformer (12 layers, 8 attention heads, and embedding dimension 256) for numerical tasks. Their model processes linear regression coefficients w_i (matrix notation) with randomly sampled $(\boldsymbol{x}_k^{(i)}, \boldsymbol{y}_k^{(i)})$ pairs, forming ICL sequences $(\boldsymbol{x}_1^{(i)}, \boldsymbol{y}_1^{(i)}, ..., \boldsymbol{x}_N^{(i)}, \boldsymbol{y}_N^{(i)}, \boldsymbol{x}_{\text{query}}^{(i)})$. Zero-padding aligns function values with inputs before projection into the embedding space via fully connected layers. Post-Transformer processing maps embeddings back to the output space, demonstrating both in-distribution learning of regression algorithms and out-of-distribution generalization. However, these models are only for linear tasks. But, the generalization ability of ICL is highly nonlinear.

B More Experimental Details

Model size and budget. Consistent with the experimental setup in (Garg et al., 2022), our model comprises either 12 or 24 network layers, equipped with 8 attention heads, and employs a 256-dimensional embedding space. The model was trained on NVIDIA RTX A6000 GPUs, with an average training time of approximately 50 hours for 600k steps.

Software packages. We employed the `solve_ivp` function from the SciPy (Scientific Python) package to compute numerical solutions of differential equations required for training.

Training techniques. Bengio et al. (2009) implemented curriculum learning, progressively increasing problem dimensionality to mirror human learning curves and reduce computational costs (Wu et al., 2021; Wang et al., 2021). Cosine annealing (Johnson et al., 2023) optimized training via cyclical learning rate adjustments (peaking then decaying following cosine curves), preventing local optima convergence—a technique widely adopted in models like YOLO-v8 (Varghese and Sambath, 2024) and DeepSeek-V3 (DeepSeek-AI et al., 2025).

Departing from approach of Garg et al. (2022), we introduce sliced mean squared error (Sliced-MSE) as the optimization objective. For vector $\boldsymbol{v} = (v_i)_{i=0}^n$ and integer $k_v \leq n$, define slice $\boldsymbol{v}(: k_v) = (v_0, ..., v_{k_v})$:

Definition 1 (Sliced Mean Squared Error)**.** *For ground truth $\boldsymbol{y} \in \mathbb{R}^d$ and prediction $\hat{\boldsymbol{y}} \in \mathbb{R}^d$, the Sliced-MSE at Steps $\leq d$ is:*

$$\ell(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \mathit{MSE}\left(\boldsymbol{y}(: \mathit{Steps}), \hat{\boldsymbol{y}}(: \mathit{Steps})\right).$$

We implement progressively complex ODE IVPs, initially testing a 12-layer GPT-2 model before parallel evaluations with a 24-layer variant.

The training protocol adapts DeepSeek-V3's learning rate scheduling (DeepSeek-AI et al., 2025) with modifications:

- **Warm-up phase**: Linear learning rate increase with sample size
- **Plateau phase**: Stabilized learning rate period
- **Cosine decay**: Gradual reduction following cosine annealing (Johnson et al., 2023)

This multi-stage approach (detailed in Section 3.1, Figure 6) enhances parameter stability during early training while promoting eventual convergence.

Complementing this, we employ curriculum learning to gradually increase problem dimensionality over 30k steps, lowering initial training difficulty and accelerating meaningful parameter acquisition.

While training the first-order ODE, we implement an adjusted cosine annealing schedule (Figure 6). For specific research questions, it can be found after Definition 3 and Definition 4

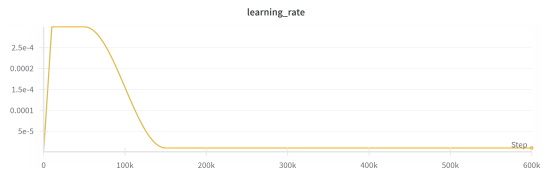

Figure 6: Three-phase learning rate schedule combining warm-up, plateau, and cosine annealing. Initial rate 1×10^{-6} linearly increases to 3×10^{-4} over 10k steps, maintains for 40k steps, then decays via cosine annealing to 1×10^{-5} over 10k steps before stabilization.

IVP formalization. Here we formally define the initial value problem of our research object as follows:

Definition 2 (Initial Value Problem)**.** *For mapping $f : \Omega \to \mathbb{R}$ with open domain $\Omega \subseteq \mathbb{R} \times \mathbb{R}$, an IVP exists given initial condition $(t_0, u_0) \in \Omega$ satisfying:*

$$\begin{cases} \frac{du}{dt} = f(u, t), \\ u(0) = u_0, \quad t \in [0, t_e]. \end{cases} \tag{2}$$

For specific research questions, it can be found in Definition 3 and Definition 4.

C More Experimental Results

C.1 Preliminary Exploration: Predictive Accuracy of ICL Models

This section investigates the efficacy of in-context learning (ICL) for nonlinear differential equation solving through a fundamental initial value problem. Our experiments demonstrate that the model successfully predicts solutions within acceptable error margins, with prediction accuracy exhibiting exponential convergence as context length increases. These findings reveal ICL's substantial potential for nonlinear numerical problems when properly trained.

Current research lacks comprehensive exploration of ICL's capabilities for nonlinear numerical solutions. As established earlier, differential equation solving inherently involves nonlinear characteristics. To facilitate the model's initial foray into this domain, we begin with the most elementary form of initial value problems under our framework, which we term the Simple Initial Value Problem (Simple-IVP):

Definition 3 (Simple Initial Value Problem (Simple-IVP))**.** *A simplified form of Definition 2 is given by:*

$$\begin{cases} f(u, t) = ay + b, \\ y(0) = y_0, \quad t \in [0, t_e]. \end{cases} \tag{3}$$

*The input data distribution $\mathcal{D}_x = \{\boldsymbol{x}_i \colon \boldsymbol{x}_i = (Para_i, t_e, Steps_i)\}$ for Simple-IVP contains parameter sets $Para_i$ with **three degrees of freedom** corresponding to coefficients a, b, y_0.*

For initial model exploration, we used 12-layer GPT-2 model and employed a curriculum learning scheme with a fixed learning rate of 10^{-4} over 200k training steps. Figure 7 demonstrates the model's performance:

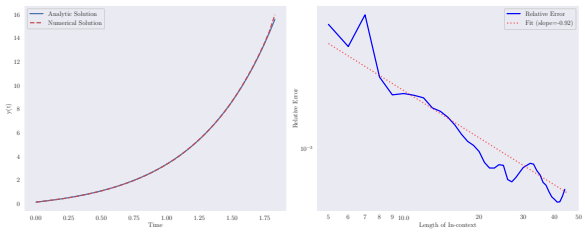

Figure 7: Performance of the Transformer model during preliminary training. Left: Solution curve with context length $= 40$ for parameters $a = 1.7$, $b = 1.0$, $y_0 = 0.1$, $t_e = 1.9$, showing near-perfect alignment with ground truth. Right: Log-log plot of Sliced-MSE versus context length, with fitted slope -0.92 confirming the convergence properties of the ICL approach.

The model demonstrates competent numerical solving capabilities, with two key observations:

- The solution curve exhibits close approximation to the analytical solution
- The linear relationship in the log-log error plot reveals exponential convergence of estimation error with increasing context length

C.2 Further training settings

Definition 4 (First-Order Linear ODE)**.** *A first-order linear ordinary differential equation relates a function to its first derivative through:*

$$\begin{cases} \frac{dy}{dt} + p(t)y = q(t), \\ y(0) = y_0, t \in [0, t_e], \end{cases} \tag{4}$$

*where $p(t) = \alpha_1 t + \alpha_2$, $q(t) = \beta_1 e^{\beta_2 t}$. The input distribution $\mathcal{D}_x = \{\boldsymbol{x}_i \colon \boldsymbol{x}_i = (Para_i, t_e, Steps_i)\}$ exhibits **five degrees of freedom** in $Para_i$: $\alpha_1, \alpha_2, \beta_1, \beta_2, y_0$.*

Training Configuration: Both models employ curriculum learning, gradually expanding context length to 41 and vector dimensionality to 64 over the first 30k training steps. The 24-layer variant was introduced for comprehensive comparison alongside the original 12-layer architecture. All models were trained for 600k steps before evaluation.

C.3 Limitations of In-Context Precision

Figure 8 shows the model's performance under specific parameter conditions. It can be observed that for zero-solution and low-rigidity solutions, the classical solution achieves zero error and low error perfectly, while the model maintains its original accuracy as expected.

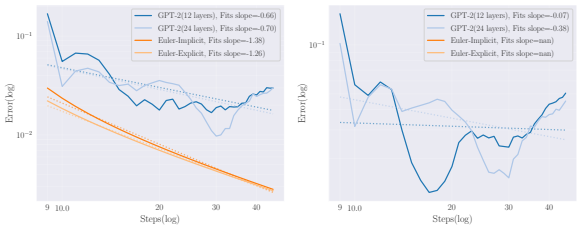

Figure 8: Performance on edge cases (initial value $y(0) = 0.6$). Left: $\alpha_2 = 1$ with other parameters zero. Right: All parameters zero.

C.4 Another Composite Testing

The patterns observed in the slope heatmap are consistent with the findings in the main text, though with some notable variations. The error heatmap (Fig. 9) suggests that GPT-2 solutions tend to maintain broader stable regions in the β-parameter space, though with relatively modest precision improvements. This observed pattern could potentially relate to the input sequence ordering effect discussed in prior works (Zhao et al., 2021). As β-parameters typically appear later in the input sequence than α-parameters, the self-attention architecture may allocate comparatively less attention weight to these parameters during feature processing. Such positional bias, if present, might simultaneously explain the preserved solution stability (through more consistent global patterns) and the limited precision gains (due to reduced focus on later inputs). However, this interpretation requires further verification as the underlying mechanisms remain incompletely understood.

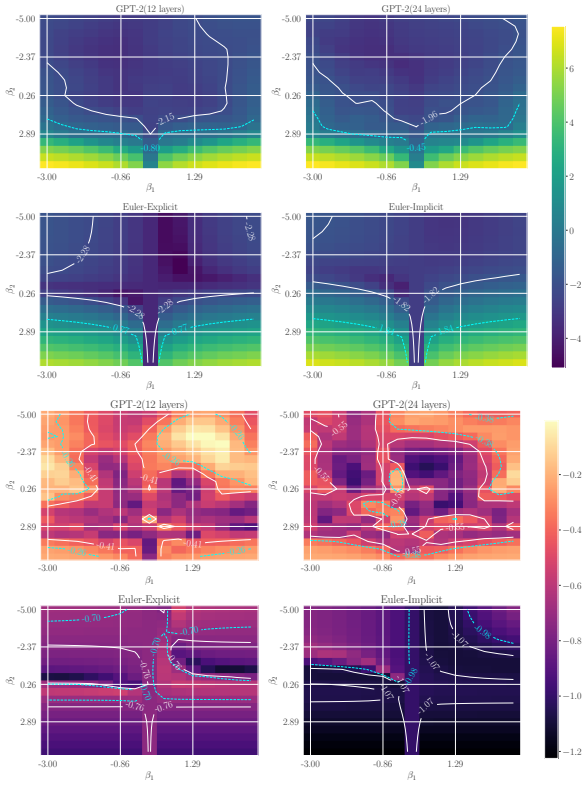

Figure 9: β_1-β_2 test region comparisons across parameter combinations.*Upper*: error heatmap; *Lower*: convergence slop (trained on $[-2, 2] \times [-3, 3]$, tested on $[-3, 3] \times [-5, 5]$). In Each subfigure: *Top*: 12L/24L GPT-2; *Bottom*: Euler Explicit/Implicit. Contours mark 50% (cyan) and 70% (white) of each subplot's range.

D A Conjecture of In-Context ODE solver

Building upon the observed relationship between convergence accuracy and context length, this study proposes a conjecture (inspired by Liu et al. (2025).) regarding the convergence properties of in-context learning for ODE solving. We consider this conjecture could provide theoretical foundations for Transformer applications in differential equation solving.

Conjecture 1 (Convergence of In-Context Learning for ODE Solving)**.** *Let $\delta \in (0, 1)$, c be a positive constant. For a Transformer model with L layers and H attention heads, when $N_0 \in [1 : N-1]$ satisfies specific condition $P(\delta, c, L, H)$, there exists a parameter set such that for any $n \in [N_0 : N-1]$, the query result $\boldsymbol{y}_{n+1}$ of randomly generated first-order linear ODEs and model prediction $\hat{\boldsymbol{y}}$ satisfy with probability at least $1 - \delta$:*

$$||\boldsymbol{y}_{n+1} - \hat{\boldsymbol{y}}|| \leq ce^{-kN}, \quad N \geq N_0 \tag{5}$$

indicating exponential convergence of prediction accuracy with increasing context length.