

# Proposal: Investigating In-Context Learning with Complex Functions in Transformers, Mamba, and Efficient Variants

Yihong Chen<sup>\*</sup>, Binwen Liu<sup>\*</sup>, Peiyu Xu<sup>\*</sup>, Quan Yuan<sup>\*</sup>

SID: 3040921900, 3040938839, 3040927191, 3040936967

April 11, 2025

## Investigation Overview

**Function Classes** Building upon the function classes studied in Garg et al., 2023[5], which investigates transformers’ in-context learning (ICL) capabilities on linear functions, sparse linear functions, decision trees, and two-layer neural networks, we explore additional function classes to further probe transformer performance. Subsequent research has broadened the scope of ICL beyond simple functions, investigating applications in diverse domains such as time-dependent and structured functions (e.g., linear dynamical systems; Cole et al., 2025 [3]), discrete function classes (e.g., Boolean functions; Bhattamishra et al., 2023 [1]), and nonlinear regression tasks (e.g., polynomial kernel regression; Sun et al., 2025 [7]). Further studies have also explored single-index models (Oko et al., 2024 [6]) and partial differential equations (Cole et al., 2024 [2]), reflecting an increasing interest in understanding how transformers generalize across different functional classes.

**Models** Past studies about in-context learning in simple functions using Transformers and their variants. We also investigate some other architectures. Mamba 2[4] uses state-space models (SSMs) with selective state transitions, enabling dynamic parameter adjustments based on input content. Its pretraining achieves Transformer-comparable language modeling performance, which may suggest foundational token-level knowledge retention critical for in-context learning.

**Data Generation** Synthetic datasets are tailored to specific tasks to isolate and analyze learning mechanisms. And tasks vary in complexity[6], enabling the evaluation of transformers generalization capabilities. Train-test splits enforce disjoint support domains or introduce distribution shifts[5] to validate generalization to unseen distributions.

## Code Base

In this work, we propose to investigate two additional function classes - Gaussian kernel functions and nonlinear dynamical systems that have not been extensively studied in the context of ICL. Using linear functions as our baseline, we aim to evaluate transformers generalization ability across these more complex function classes, thereby showing both their potential and limitations in in-context learning.

As the foundation of our implementation, we will build upon the official codebase from Garg et al., 2023[5] ([GitHub: dtsip/in-context-learning](#)). This repository provides a well-structured experimental pipeline for evaluating transformer performance on various simple function classes (e.g., linear functions, sparse linear functions). Specifically, we will reuse its core components for data

---

<sup>\*</sup>Authors are listed alphabetically and contributed equally to this work.

generation, in-context learning (ICL) evaluation, and transformer model training. Our contribution will extend this framework by incorporating additional function classes, such as Gaussian kernel functions and nonlinear dynamical systems, to assess generalization capabilities further more.

In addition, to compare architectural performance, we will implement the Mamba2 model (a structured state-space model, or SSM) using its official codebase ([GitHub: big-data-ai/mamba2](#)). We will utilize only the model architecture from this repository without relying on pretrained weights so that both Mamba2 and the transformer models can be trained and evaluated under consistent ICL settings. This architectural comparison aims to assess whether SSM-based models exhibit any advantages or limitations over transformers in few-shot function learning tasks.

We will also include additional transformer-based architectures such as FlashAttention and Hyena to evaluate architectural differences in ICL performance. FlashAttention, which optimizes attention computation with kernel fusion techniques, will be implemented using the official repository ([GitHub: Dao-AILab/flash-attention](#)). Hyena, an alternative sequence model architecture with implicit long-range interaction mechanisms, will be adapted from the open-source repository at ([GitHub: hazyresearch/safari](#)). These models will be trained and evaluated under consistent experimental settings.

## Data Generation

We plan to evaluate the in-context learning capabilities of Transformers on three distinct function classes: linear functions, Gaussian kernel functions, and nonlinear dynamical systems. A synthetic dataset will be constructed for each function class, consisting of 10,000 training prompts and 1,000 test prompts. Each prompt will contain 20 context examples and 1 query point. Specifically, we will sample a random function from the class according to the distribution over functions, then create 20 random inputs independently, and finally evaluate the function on these inputs to produce the prompt. And the distribution over functions with weight vectors are drawn from an isotropic Gaussian while the input distribution are an isotropic Gaussian in 20 dimensions.

The data generation methods for each class are as follows:

- Linear functions follow the setup in paper[5], where inputs are sampled from a standard normal distribution, and we will use normalized weights and add Gaussian noise.
- Gaussian kernel functions compute exponentially decaying values based on either fixed or randomly sampled centers.
- Nonlinear dynamical systems generate time series data by iteratively evolving from an initial state using a predefined nonlinear update rule.

We may introduce distribution shifts between prompts during training and inference and between the in-context examples and the query input during inference to ensure fair comparison with optimal estimators while balancing computational cost and generalization analysis.

## Resource Budget

- **Compute:** We plan to use a single NVIDIA RTX 4090 GPU for approximately 30 GPU hours to support model training and evaluation on various function classes..
- **Data:** Synthetic datasets with 10,000 unique functions per class, each with context-query pairs, ensuring diversity for ICL analysis.

## References

- [1] Satwik Bhattamishra, Arkil Patel, Phil Blunsom, and Varun Kanade. Understanding in-context learning in transformers and llms by learning to learn discrete functions, 2023.
- [2] Frank Cole, Yulong Lu, Riley O’Neill, and Tianhao Zhang. Provable in-context learning of linear systems and linear elliptic pdes with transformers, 2024.
- [3] Frank Cole, Yulong Lu, Tianhao Zhang, and Yuxuan Zhao. In-context learning of linear dynamical systems with transformers: Error bounds and depth-separation, 2025.
- [4] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality, 2024.
- [5] Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes, 2023.
- [6] Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Pretrained transformer efficiently learns low-dimensional target functions in-context, 2024.
- [7] Haoyuan Sun, Ali Jadbabaie, and Navid Azizan. In-context learning of polynomial kernel regression in transformers with glu layers, 2025.