Project Proposal: The Role of Depth and Data Structure in In-Context Learning

Modern Learning Theory
Fall 2025

Eyüp Ahmet Başaran, Murat Enes Erdoğan

1. Problem Context and Motivation

In the past few years, the Transformer architecture has led the field in machine learning, growing in the number of model parameters from millions to billions. While we understand the concept that "bigger is better," the relationship between the model architecture, namely the representation dimension and depth, has become somewhat ambiguous.

The paper that has been assigned to us, “Theory of Scaling Laws for In-Context Regression” by Bordelon et al. (2025), studies the trade-offs involved in this process in the context of In-Context Learning. The authors present a soluble model of Deep Linear Attention, aiming to identify in what cases depth becomes useful. Among their surprising results in the theory, it has been concluded that in the Isotropic scenario, where the data distribution is quite simple, the larger the length of the context, the smaller the performance improvement achieved by added depth, where the randomly rotated structured scenario demands the use of depth.

In our proposed work, we aim to reproduce these essential results and explore the “optimal shape” related to ICL. The rationale behind our proposed work would be the fact that, rather than focusing on “scaling up” the model parameters, an effort has been put forth in analyzing why the particular architecture performs better on the data.

2. Methodology and Planned Tasks

We will focus on the Track B (Applications First) approach and use Python, specifically PyTorch, in the simulation of the Deep Linear Attention model, in an effort to test the theory proposed in the paper.

Task 1: Implementation of the Reduced Linear Attention Model

The article reduces the Transformer model to the "Reduced Gamma Model" in equation 4. The

first step in our problem-solving process would be writing the Recursive Model.

What it means and why it's important: This represents the basis on which the theory and claims in the paper are built. With the implementation of this model, simulations concerning the learning curve would then be possible without necessarily requiring the same amount of processing power in terms of training GPT models.

Method: We will use the forward pass expressed in Equation 3 and the “Reduced” recurrence relation expressed in Equation 47.

Task 2: Replication of the “Depth vs. Context” Trade-off (ISO Setting)

One of the important assumptions in the paper (Result 2) is that when the length of the context is sufficient, the performance of the shallow network is the same as the deep network.

Experiment: Here, we'll generate "Isotropic" data synthetically. The inputs would be Gaussian. The models would be different in depth and contextual length.

Goal: To replicate Figure 1, and particularly the observation that the loss functions corresponding to different network depths converge to the same limit when alpha goes to infinity. This confirms the redundancy of depth in simple and fixed-covariance problems and large context lengths.

Task 3: Analyzing Brittleness in Fixed Covariance (FS Setting)

The argument in the paper is that models that learn on a fixed data structure (FS) learn a “preconditioner” that ceases to work when the distribution changes.

Experiment: We train the model on a particular choice of the input covariance matrix. Next, we evaluate the model on a "shifted" covariance matrix, as in Result 5.

Goal: Recreate Figure 3(c).

Expectations: The error should peak when the value of theta increases, showing that the model has memorized the training geometry and learned an ICL algorithm instead.

Task 4: Extension - Robustness of the RRS Setting (New Experiment)

The paper proposes the “Randomly Rotated Structured” (RRS) setting, which helps the model learn an algorithm like Gradient Descent, rather than memorizing the covariance.

Our Extension: The paper establishes RRS has depth needs, although it doesn't provide a direct comparison in terms of model robustness when using an RRS-trained model versus an FS-trained model on the identical test data shift in Task 3.

Experiment: We’ll use the test cases from Task 3 and test an RRS-trained model on them.

Hypothesis: We would like to assume that since the RRS model has observed several rotations in

training, the error curve would remain flat when theta varies, contrary to the FS model. This would prove that the ability to train on variant/rotated tasks would be the cause behind In-Context Learning.

3. Feasibility and Timeline

The proposed work can be completed in the semester-long schedule because:

- Since the experiment uses Gaussian synthetic data, there isn't a need to clean and download large data sets.
- The Linear attention model enables the running of many experiments on our systems.
- The manuscript provides the exact recurrence equations (Eq 4, Eq 47), which are direct translations into programming loops.

4. Connection to Course Concepts

This work applies the concepts about Generalization Bounds and Overparameterization. It describes the manner in which the “effective capacity” of the model increases, not only with the number of parameters, but also with the product of the depth and the eigenvalues in the data covariance matrix. This provides a useful example application in showing the development of the scaling laws from the properties of the data.