**What is the main goal of the project?**

The goal of this project is to expand on the work done in the papers “What Can Transformers Learn In-Context? A Case Study of Simple Function Classes” and “Pretraining Data Mixtures Enable Narrow Model Selection Capabilities in Transformer Models” to see how well transformers can increasingly learn more complex functional classes. Specifically, we wanted to look at how well transformers can learn more complex decision trees with greater depth as well as how well transformer models are able to select between models when trained on a more complex functional class like decision trees. For example we wanted to see how well a model trained on examples of linear functions as well as examples of decision trees could select between the two models given in context examples.

**What are the main claims?**

The main claims of this paper are that while transformers still perform well on more complex functional classes, the results are not nearly as strong as some of the earlier literature. For example, in the case of a transformer trained on both linear functions as well as decision trees, we see the performance is notably worse than the model only trained on decision trees when evaluated on decision trees. This is somewhat of a divergence from the original data mixtures paper we looked at where a model trained on sinusoid and linear was nearly perfect when selecting between those two models. We also see that given enough in-context examples, XGBoost ends up out-performing this model which we did not see in the other papers referenced. While the transformers were still able to perform fairly well, this suggests that there may be a limit to how well they are able to learn in context.

**What are the experiments?**

* **What is the evaluation protocol?**
* **What is the data?**
* **What is the task?**

We followed an incredibly similar methodology to the one described in the in-context learning on simple function classes paper as we built off of their original codebase. The training data was synthetically generated based on the functional class. The goal was to give a number of in context examples from a functional class and the goal is to be able to predict . In the cases where we trained on a mixture of multiple different functional classes, instead of picking different fs from the same class we would pick an f from one of 2 functional classes. The goal was then to see how well a model trained on a functional class would be able to in context learn of a new function from the same functional class or the same set of classes.

**How do the experiments support the goal/claims of the paper?**

By making the functional classes more complicated that the original paper and by exploring more complicated mixtures of functional classes we can start to get a sense of how far in-context learning can go using a similar methodology to the original paper. The methodology lets us directly assess the effectiveness of in-context learning by being able to synthetically create the data we need for both training and testing. We can also make increasingly complex functional classes to investigate with minimal additional work.

**Are any of the limitations discussed in the paper?**

One of the biggest challenges we had was a lack of time and limited compute which limited the size of the models we trained. This could mean that the performance drop we see when training on two different functional classes would have been mitigated by training on a larger model. However, even with a larger model we likely still would have seen the same issues at some point but perhaps instead of seeing a performance drop when trained on two different functional classes, the performance would have instead dropped if trained on 10 different functional classes instead.

**What are the strengths of the paper?**

This paper does a good job of questioning how far we can push the original results and that there is likely a point at which in-context learning is no longer as effective as other methods. The paper also does a good job of combining the results from two different papers and using the methods and results from each to continue to investigate in-context learning on more complicated functional classes and the ability to model select between different models it was trained on.

**What are the weaknesses of the paper?**

Due to the limited compute available, the size of the models used in this paper is smaller than the size of the models in the original paper. This means that it can be somewhat hard to directly compare the results and it also limits what can be done in terms of adding more than two functional classes to see how well transformers can switch between a wider array of functional classes. Due to the time constraints we were also unable to investigate all of the questions we had in terms of areas to explore. There are also a few models that are as of right now not finished training and will hopefully be included in the final version of the paper.

**Provide a suggestion for improving the paper.**

I think that once we have some of our other models trained we will have more content to discuss in our paper as some of the work we

**What is the relevant related work?**

The related work cited in the paper is “What Can Transformers Learn In-Context? A Case Study of Simple Function Classes” and “Pretraining Data Mixtures Enable Narrow Model Selection Capabilities in Transformer Models.”

**Is the paper reproducible?**

* **Can you rerun the experiments?**
* **Can you reproduce the results in the paper?**

The plots from the paper are all reproducible in the eval.ipynb notebook. The experiments can also be rerun using the training notebooks however there may not be enough time for peer reviewers to rerun training before the report is due.

**Are all the plots in the paper clearly interpretable with well-defined and explained axes, with methodology clearly explained in the paper text?**

The plots in the paper should be fairly well explained especially for anyone who has read the cited papers. Even for those who have not read the papers cited, the plots should still be fairly clear.

**Is the English in the paper correct and clear?**

The English is clear and correct.

**Do you have any feedback on any TODOs that the authors have left at this stage?**

We still need to finish training some of our models, and given enough time we may try and explore some of the ideas we included in our discussion.