Response to reviewers

We really appreciate the reviews on our projects, which lets us have a better idea of our strengths and weaknesses. Here are the responses to your valuable suggestions:

1. experiment with models of different sizes.

Response: We used the small gpt-2 model(6 layers, 4 heads, 128 embedding dimensions) throughout the whole process. The reason we didn’t train larger models is that these models are relatively very big and might not be timely and financially infeasible for us to train and evaluate.

1. The results on the weighted average of linear functions and decision trees.

Response: We included that in our updated version of the paper, in which we also experimented with weighted sums of linear functions, decision trees and simple neural networks

1. Explanation for why the performance of the transformer drops when trained on both function classes

Response: We did try to evaluate this as well, our interest was to evaluate whether the drop of performance was resulted from limit of model size or inadequacy of model selection. We experimented by appending task labels to the input data, and compared the results. See our discovery at the final submission.

1. Trying out different function classes

Response: Sure thing, this was also implemented in the updated version

1. Provide results throughout various stages of training

Response: We took this idea seriously but did not evaluate it at length in our final submission mainly for two reasons: 1) We didn’t have enough time to reload different stages of the model and conduct evaluations 2) We figured our main goal of this project was to see the performance at the end of training, and the performance on different stages was not very important. However, we did include our training loss plot in the appendix.