An Investigation of the Relationship Between the Average Depth of First Correct Completion and CodeGen's Performance

Miranda Keeler, m.l.keeler@student.tudelft.nl

Supervisors: Ir. Jonathon Katzy, Dr. Maliheh Izadi, **Professor:** Dr. Arie van Deursen

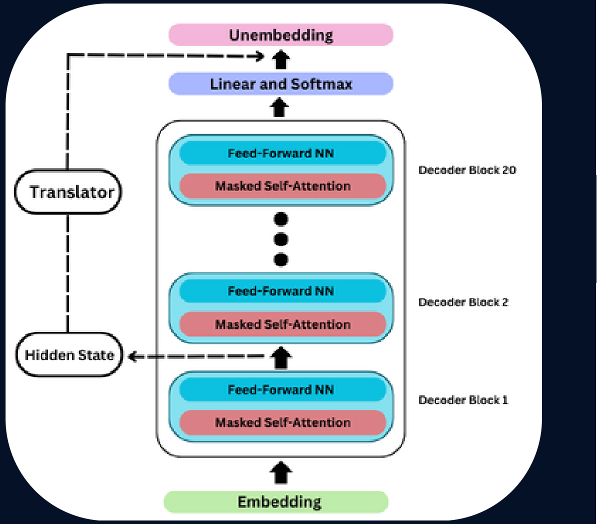
Introduction

Pre-Trained Language Models (PLM) have already been incorporated into developers' largest Git repository platform, Copilot, signifying an acceptance of these models as valuable tools [1]. However, they may only be helpful to some developers, as their performance tends to be benchmarked on one or two high-resource languages, such as Python or Java. Furthermore, they are growing in size and complexity, begging the question of whether that is necessary; instead, we can evaluate their performance across their layers, evaluating their performance as it relates to their underlying architecture, the Transformer.

RQ: How does the average depth of the first correct completion relate to the performance of CodeGen?

Methodology

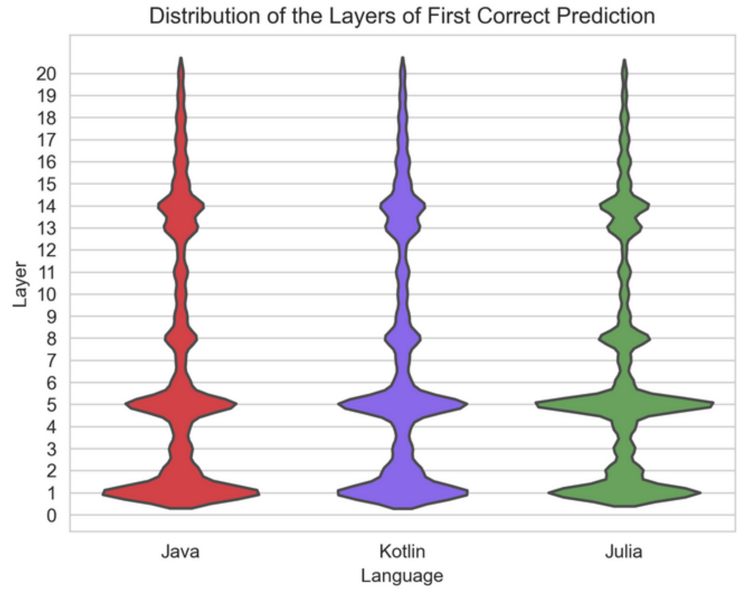
We analyzed the most probable next-token prediction of CodeGen for each of its 20 layers with a Tuned Lens [2]. We calculated the average depth of its first correct completion at a token- and language level. We calculated the model's accuracy using these predictions as a performance metric. We analyzed the ratio of null heads and null attention in aggregated sets of predictions and several tokens of interest to check for patterns between the predictions and attention.

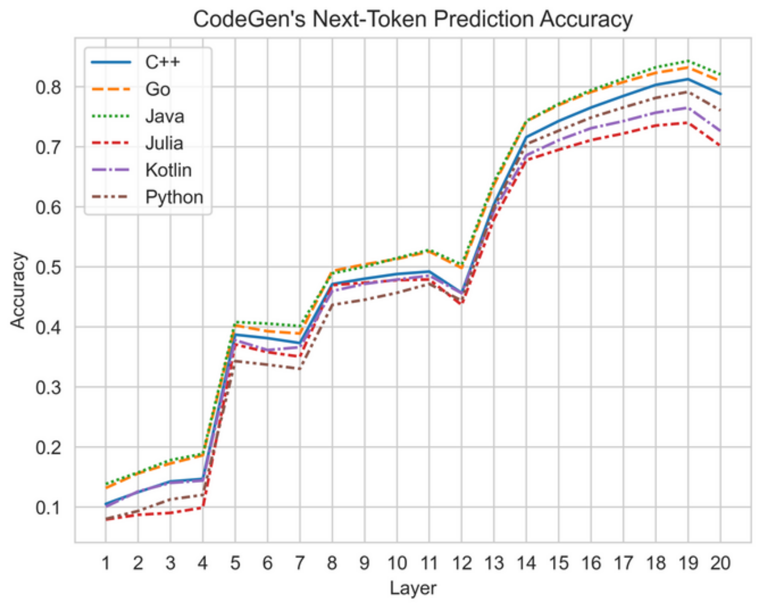


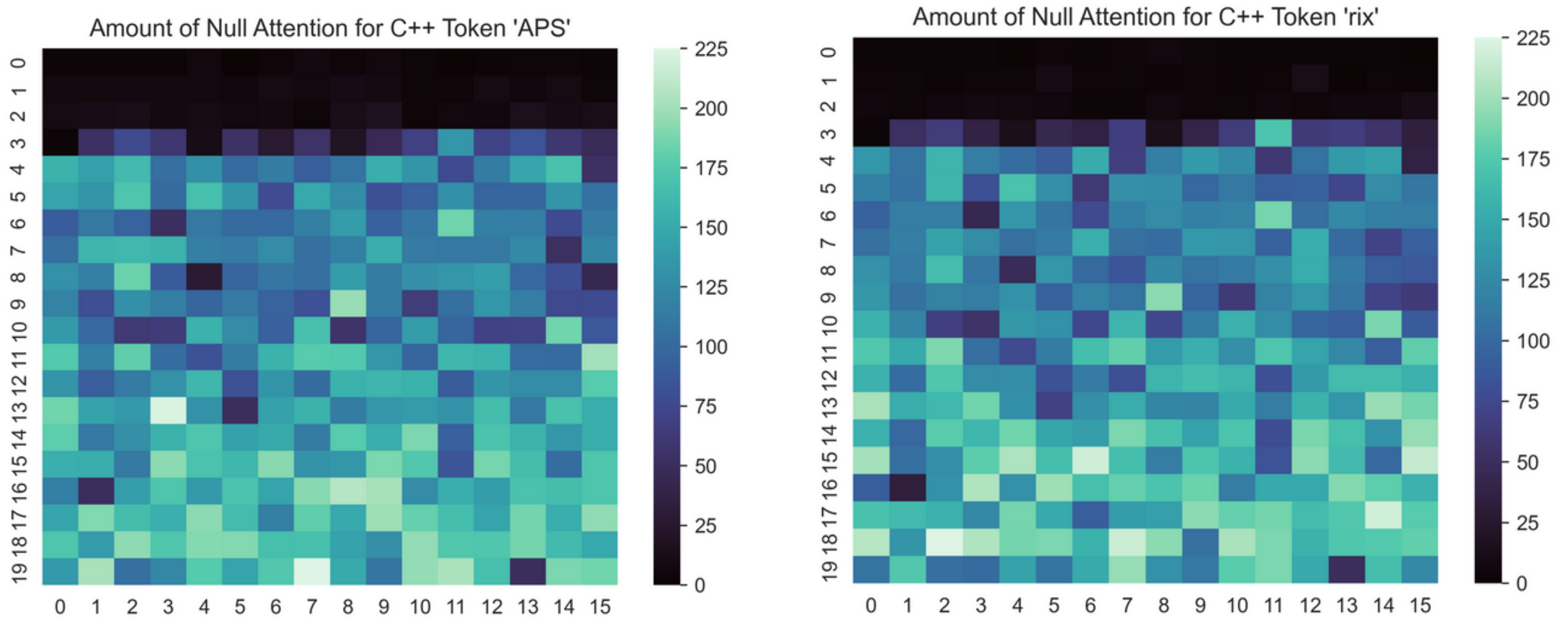
Conclusions

There are several exciting findings in the average depths and performance across the layers of CodeGen. There are clear patterns in which layers tokens are correct for the first time, specifically in layers 1, 5, 8, 13, and 14, which are also the layers in which performance jumps occur. The layers before these have a dip, which could indicate a learning progression or error corrections happening within the model. When viewing these through attention, the relationship between depth and performance is not immediately apparent, but there are patterns within the attention heads themselves. The Pearson and Spearman correlations between the average depth of the first correct completion and performance were low, at 0.0 for Java and 4.5 and 3.7 for Julia. This could indicate that there is a relationship between the average depth of the first correct completion for languages that CodeGen is unfamiliar with. However, further investigation is warranted due to the low correlations and findings within this project's scope.

Results


Distribution of the Layers of First Correct Prediction


CodeGen's Next-Token Prediction Accuracy


Null Attention in C++ tokens, 'APS' & 'rix' with average depths of 19.04 and 1.02 respectively.

Amount of Null Attention for C++ Token 'APS'

Amount of Null Attention for C++ Token 'rix'

[1] https://openai.com/blog/openai-codex

[2] N. Belrose, Z. Furman, L. Smith, D. Halawi, I. Ostrovsky, L. McKinney, S. Biderman, and J. Steinhardt,"Eliciting latent predictions from transformers with the tuned lens," 2023.