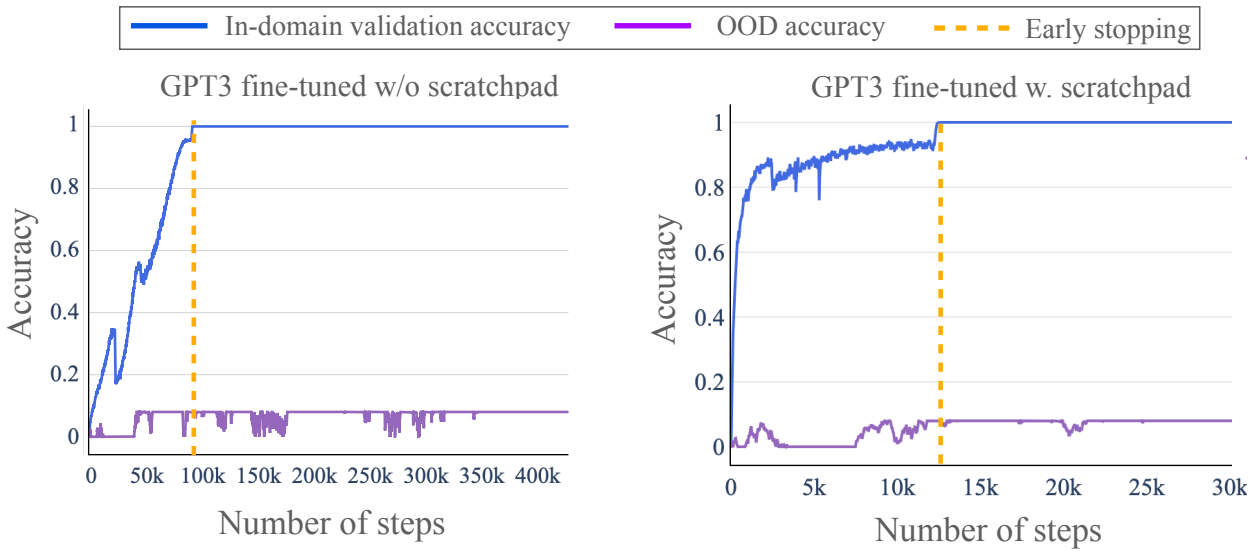

Figure 1: Grokking results on multiplication. We finetune GPT3 with question-answer pairs (left) for 420K steps, far exceeding the point at which in-domain accuracy plateaus (90K steps). This training duration is equivalent to 60 epochs and costed 50,000 USD. We also finetuned GPT3 with question-scratchpad pairs (right) for 30K steps, costing 40,000 USD. The orange dashed line indicates overfitting. We observe that training long enough after overfitting does not lead to grokking behavior. The experimental setup and the dataset splits are consistent with those outlined in the paper. GPT3 exhibits poor performance on OOD data even after prolonged training period. We speculate that the inherent complexity of the multiplication task makes it significantly more challenging to learn a well-structured representation. This difficulty renders the emergence of a generalized solution immensely hard, potentially necessitating infinite training, or it may even make it unattainable. Future work is needed to understand grokking better. It's important to note that even if grokking were to happen through prolonged training, such an approach is not efficient. Resources could be more wisely allocated towards the development of smarter, smaller-scale models that don't necessitate thousands or even millions of dollars, to achieve satisfactory generalization.