Grokking Tickets – An Analysis

By Nicholas Chen
Github repo: https://github.com/nicholaschenai/grokking-tickets-lambda

Introduction

Nanda et. al (2023) (PROGRESS MEASURES FOR GROKKING VIA MECHANISTIC INTERPRETABILITY) hypothesized that phase transitions are inherent to composition largely due to the lottery ticket hypothesis – That early on, the network is a superposition of approximate circuits which improve slowly as each component is weak, but the moment any one component develops, the other components become more useful and generalization accelerates.

In a very recent paper, Minegishi et. al (2023) (GROKKING TICKETS: LOTTERY TICKETS ACCELERATE GROKKING) derived lottery tickets from various phases of grokking (memorization, circuit formation, cleanup) and showed that the later the stage of grokking, the more effective the lottery ticket is in terms of generalization speed, supporting the hypothesis in Nanda et. al. (2023).

This brings up some questions:

- What will the progress measures show if we use these lottery tickets? Naively, we expect the progress measures to be accelerated as we already gave the network better initialization which might allow it to skip grokking phases. At the same time, this is also a test of the progress measures' ability to explain the underlying phenomena. This was not studied in Minegishi et. al (2023).
- What can we say about the universality of circuit formation from these experiments? Recall that in lottery tickets, subnetworks are trained – These networks do not have the flexibility to explore the full optimization landscape compared to the base model. If a network is initialized from a lottery ticket derived from the memorization phase, will it be constrained to not be able to transit to the circuit formation phase? In Minegishi et. al (2023), this was not explored as the number of epochs was limited, so we do not know if some of the accuracies will plateau or continue improving.
- Can we use RLCT (Lau et. al. 2023 QUANTIFYING DEGENERACY IN SINGULAR MODELS VIA THE LEARNING COEFFICIENT) to explain any unusual behavior?

Contributions

In this mini project, we make the following contributions:

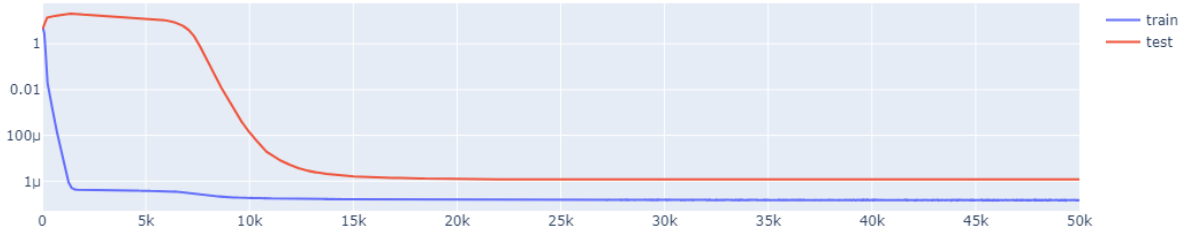
- We reproduce the core result of Minegishi et. al (2023) by training lottery tickets derived from the 3 phases of grokking and observing the loss curves, and confirm that the later the stage of grokking, the faster the lottery ticket generalizes. One small difference is that in Minegishi et. al (2023), biases were not used in the MLP while our work implements it to allow a better comparison to Nanda et. al (2023)
- We show that grokking tickets derived from the circuit formation and cleanup phase display progress measures deviating from the standard phases in Nanda et. al (2023), appearing to be ‘accelerated’, suggesting that the lottery ticket hypothesis indeed can explain something about the phase transitions in grokking.
- We show that RLCT can explain the 3 phases of grokking if it occurs, and can distinguish between networks that grok and networks that generalize immediately. However, we could not find any indication that it can explain double descent

Approach

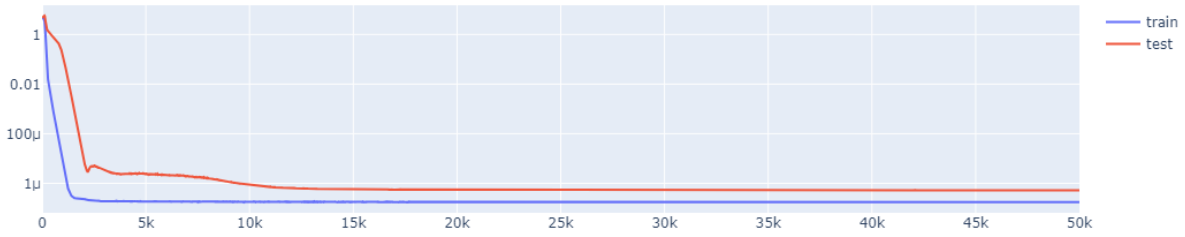
- Obtain the network weights in Nanda et. al (2023) at the various stages:
    - Epoch 0: Initialization
    - Epoch 1.4k: Memorization
    - Epoch 9k: Circuit formation
    - Epoch 14k: Cleanup
- For each of the networks above (excluding epoch 0), derive a mask which prunes away the bottom 40% of weights by magnitude.
- For each mask derived, apply it to the network at epoch 0 and train it via the approach in Nanda et. al (2023), obtaining progress measures.
- Hyperparameter search for SGLD, then estimate RLCT for each network over the training epochs.

Grokking Tickets Speed Up Generalization

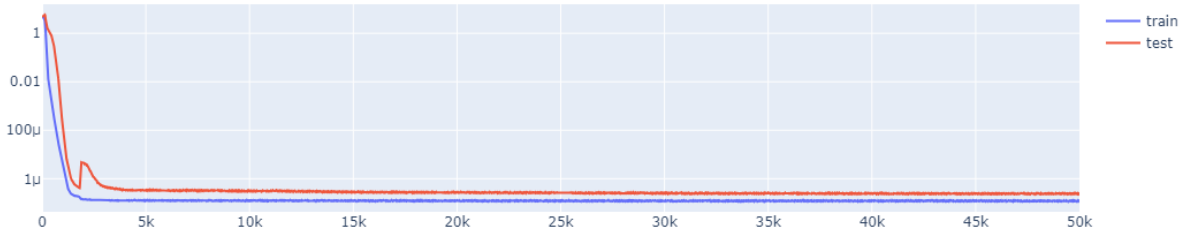
Memorization (derived from lottery ticket at 1.4k epochs):



Circuit formation (derived from lottery ticket at 9k epochs):



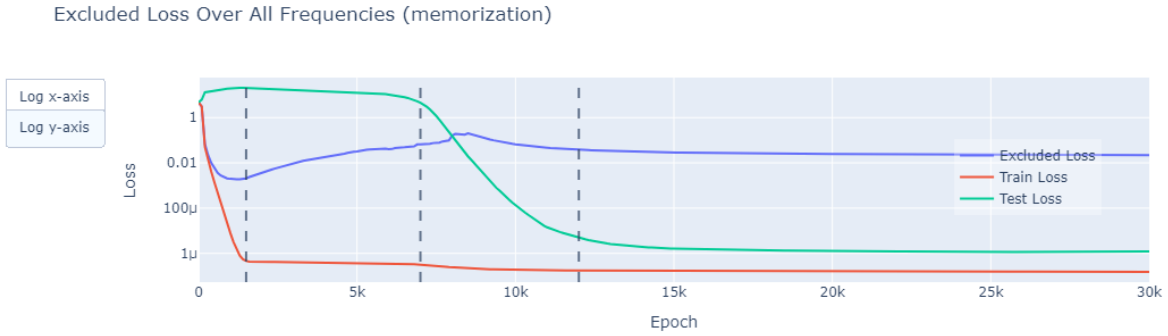
Cleanup (derived from lottery ticket at 14k epochs):

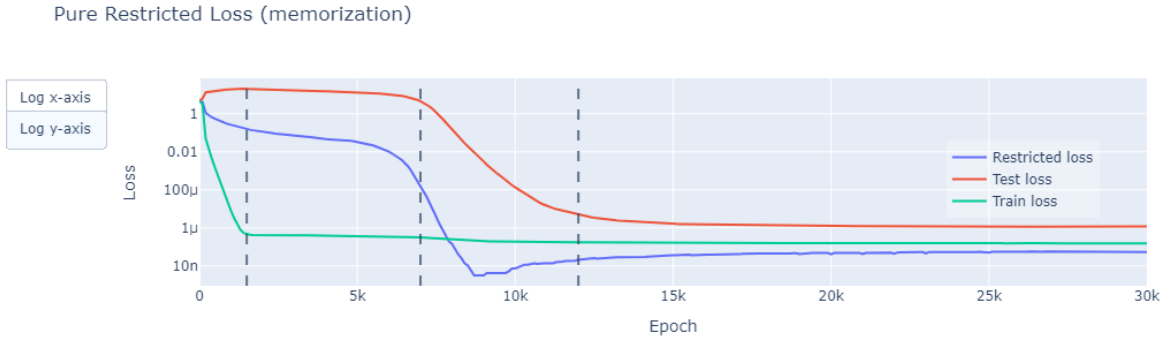


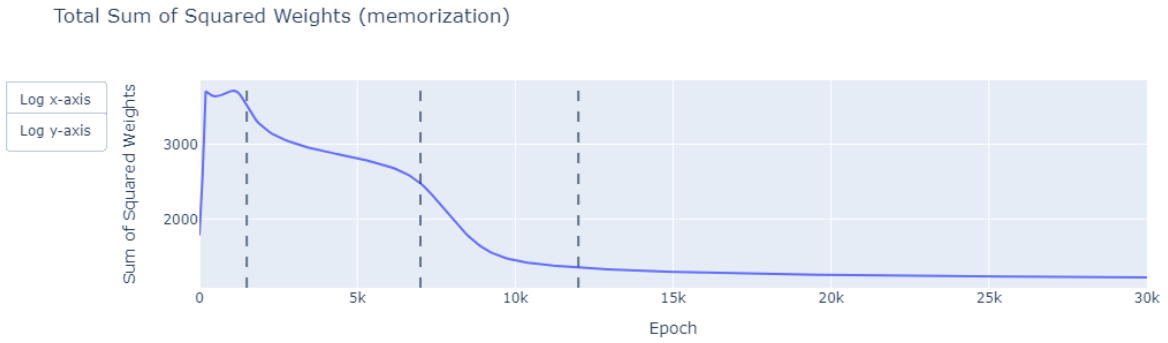
Observations:

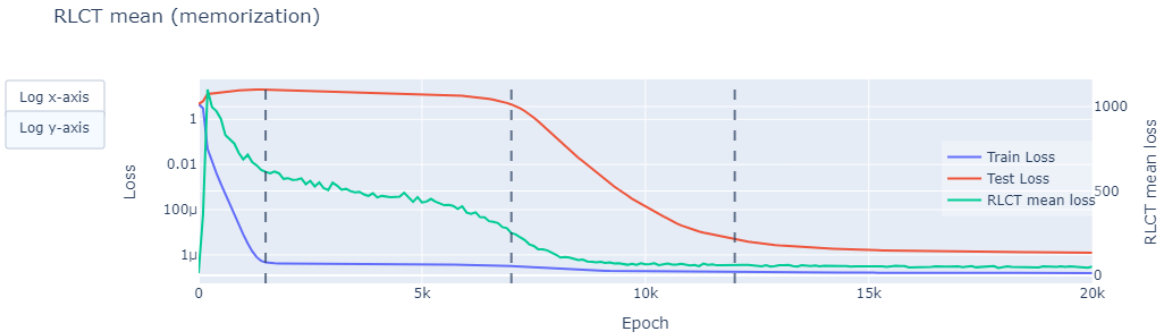
- Test loss always plateaus, and this value improves with the stage of grokking the lottery ticket was initialized with. Does this suggest that circuits do have a certain space of ‘fixed structures’? I.e. if the network topology space is constrained, then some circuits may not form that well and there are no alternatives to reconfigure the same circuit under a different network topology.
- The grokking stage affects the rate of improvement of test loss
- We observe the standard grokking curve in the memorization lottery ticket (train loss improves quickly but test loss increases and plateaus first before rapidly improving)
- In the circuit formation and cleanup ticket, we do not observe an initial test loss plateau, suggesting that these tickets really help ‘skip’ the memorization phase
- In the circuit formation and cleanup ticket, the test loss follows a pattern of ‘negative gradient’ -> ‘even more negative gradient’ -> double descent -> plateau, with the cleanup ticket having a more distinct double descent pattern. Why is this so?

Progress Measures and RLCT Explain Phases on Networks Trained From The Memorization Lottery Ticket





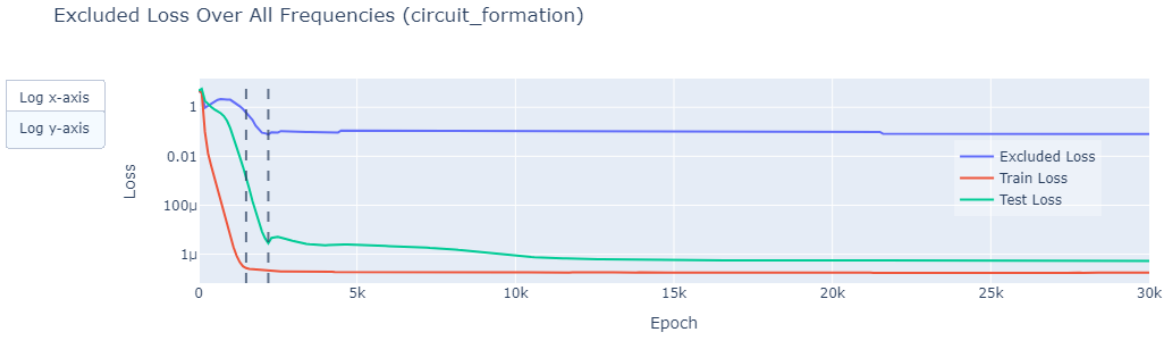


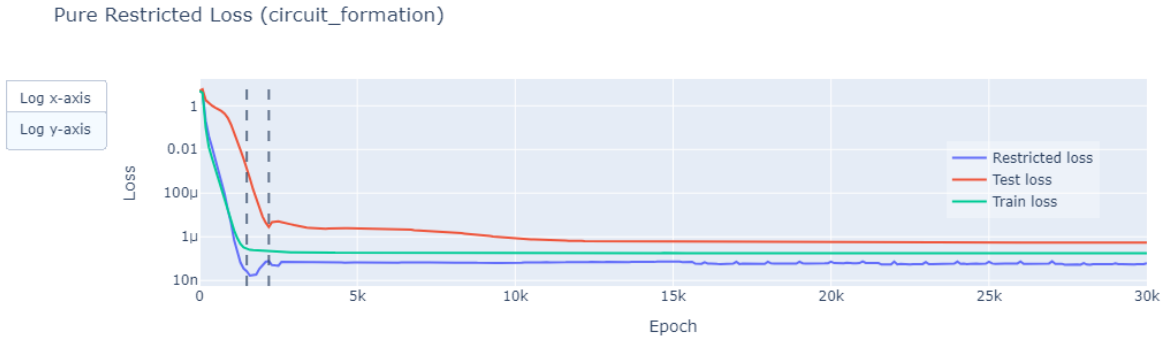


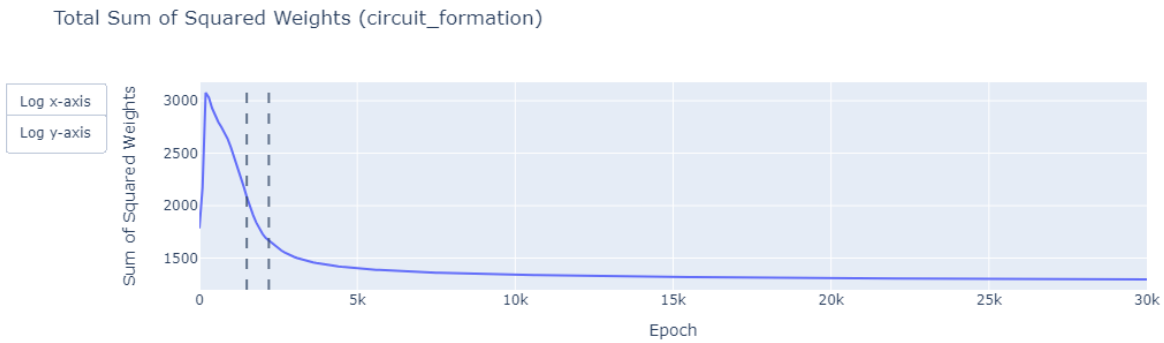
Observations:

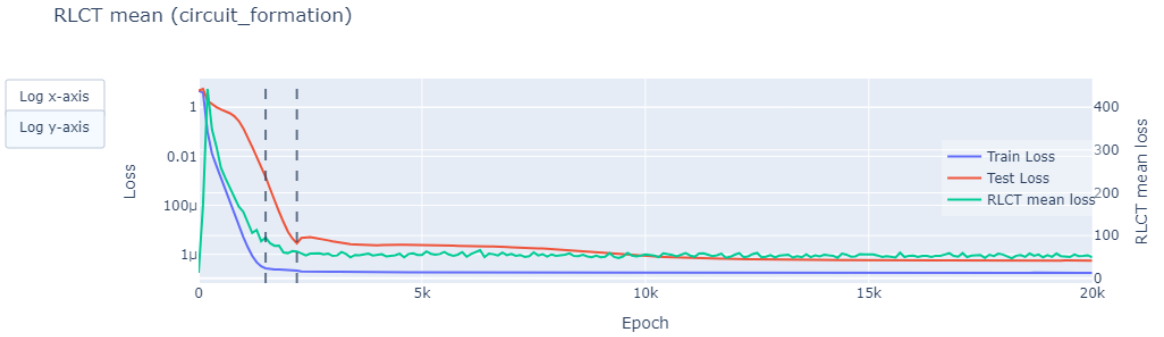
- We still have the 3 phases of grokking, which are also explained by the changes in progress measures consistent with Nanda et. al (2023).
- RLCT mean loss exhibits 3 phases that explain grokking: sharp decline over the memorization phase, slow decline over the circuit formation phase, then sharp decline in a way that **leads** the cleanup phase.

Progress Measures and RLCT Reflect The Disappearance of Grokking On Networks Trained From The Circuit Formation Lottery Ticket







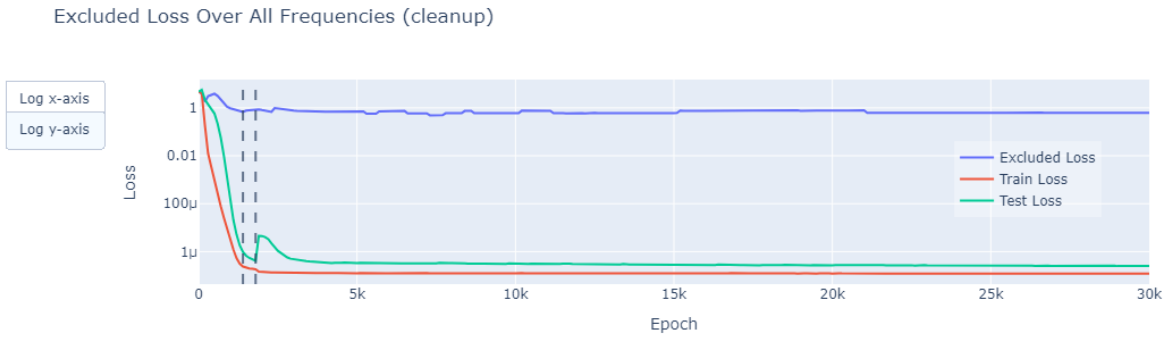


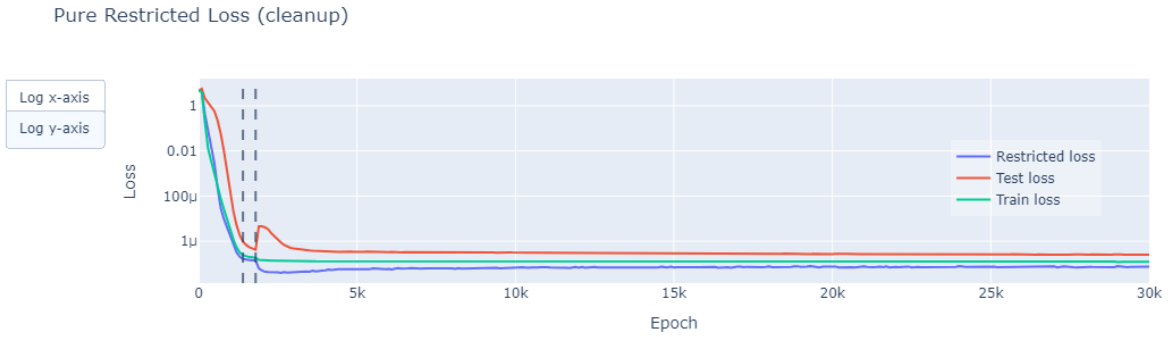
Observations:

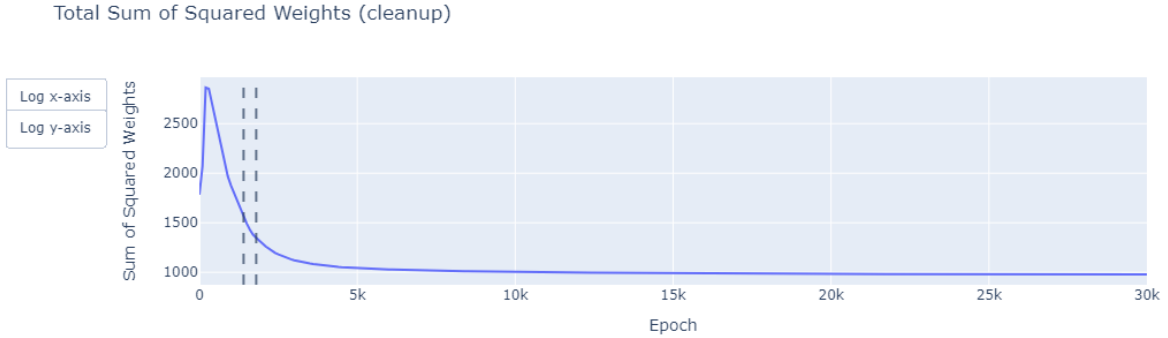
- Test loss does not plateau at the start; It appears that circuit formation overlaps with memorization
- Excluded loss decreases, increases slightly, decreases and then plateaus. This is quite different from the original pattern of decreasing during memorization, increasing during circuit formation then plateauing, suggesting that the grokking phases are indeed accelerated
- Restricted loss decreases immediately and sharply, skipping the ‘mild decline’ phase during circuit formation. It also has a bump at the double descent phase
- Total sum of squared weights only sees 1 sharp decline phase instead of 2

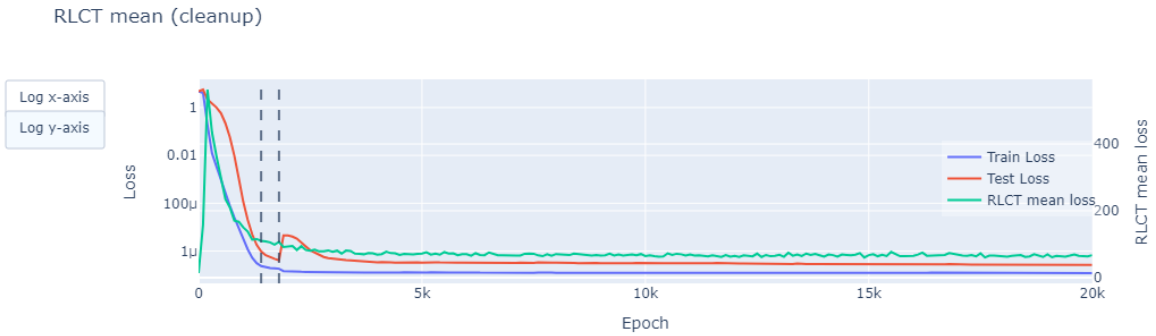
- RLCT now declines smoothly at a decreasing rate (instead of having 3 phases previously), plateauing around the same time when test loss stops dropping sharply

Progress Measures and RLCT on Networks Trained From the Cleanup Lottery Ticket behave similarly to The Case from Circuit Formation Lottery Ticket









Observations:

- Similar observations to the Circuit Formation Lottery Ticket, suggesting phases are skipped
- Minor differences from Circuit Formation Lottery Ticket:
    - Excluded loss' amplitude decreases over the tickets
    - The bump in the restricted loss at the double descent is greater in magnitude, possibly reflecting the increased magnitude in double descent. This time, the bump forms the global minima whereas in the circuit formation lottery ticket phase, the bump only forms a local minima
    - Sum sq weights has a lower maxima and lower plateau value

Discussion & Conclusions

- Test loss of tickets from different phases plateau at different values – Might this suggest that circuits are universal, in the sense that it requires a certain topology?
- Indeed, progress measures behave differently (accelerated) under grokking tickets, supporting the hypothesis that phase transitions can be explained by the lottery ticket hypothesis
- Are there other progress measures that work well with such 'accelerated phases' of grokking tickets?
- Olsson et. al. 2022 (In-context Learning and Induction Heads), showed that giving the network one component (via the smeared key architecture) erases the phase transition; Will the loss curves and progress measures of the grokking ticket behave similar to that of the smeared key architecture?
- RLCT can distinguish between the networks with and without grokking phases, however it cannot explain double descent. Also RLCT leads the cleanup phase (starts and plateaus before the cleanup phase) in the memorization lottery ticket – why?

Future Work

- Run with multiple seeds for robustness, hyperparameter search + increase number of steps for SGLD till RLCT loss plateaus, use the full excluded loss instead of the mean excluded loss, and calculate Gini coefficients as in Nanda et. al (2023) (could not finish these due to time + compute constraints)
- Run with larger networks: If circuits are universal, constraining the topology of a larger network might be less effective at increasing the plateau test loss as the network can explore a larger network topology space and find an equivalent circuit?
- Investigate the double descent phenomena – why does it appear in grokking tickets of the later 2 stages and not in the memorization stage? Why is the magnitude of the later stage larger?
- Grokking tickets for other problems with multiple components (eg induction heads) to understand how general our conclusions are.