# 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.

Recently, Minegishi *et. al.* (2023) (Grokking Tickets: Lottery Tickets Accelerate Grokking) derived lottery tickets as defined in Frankle & Carbin (2018) (The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks) from various phases of grokking and showed that lottery tickets obtained from the generalization stage (hence termed 'grokking tickets') generalize faster than the base model, while non-grokking tickets and other control models (controlling for L1/L2 weight norm and controlling for sparsity at the start of the experiment) generalize slower and in some cases, fail to generalize. Furthermore, they found a positive relationship between the epoch which the lottery ticket is derived from, and the speed of generalization and final test accuracy. Minegishi *et. al.* also analyzed the weights of the grokking ticket and base model for periodicity and frequency components (via discrete Fourier transform) for the modulo addition task and showed that over training epochs, the grokking ticket obtains periodicity in the weights faster than the base model (Nanda *et. al.* showed that the neurons learn an algorithm which results in periodicity in weights). All these suggest that generalization is the result of being able to explore and eventually obtain good network structure, supporting the hypothesis in Nanda *et. al.* (2023).

To investigate further, we raise some questions:

In Nanda *et. al.*, progress measures were defined to track underlying phenomena, leading to the discovery that the network goes through these phases before full generalization: memorization, circuit formation and cleanup. Can these progress measures be used to track the underlyings of these lottery tickets and explain their generalization speed and ability? 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, and we also want to study how the progress measures fare at the end state since non-grokking tickets have poorer network structure. Furthermore, progress measures might be a more concrete way to characterize generalization than the weights analysis in Minegishi *et. al*. (which was not performed on non-grokking tickets).

In Minegishi *et. al.*, the main metric for comparison is accuracy curves. Might comparing the loss curves reveal further information? Two networks can have the same perfect accuracy on the task, but have different losses as the logits of the incorrect answer can be nonzero.

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?

Last, the progress measures in Nanda *et. al.* (except the sum of squared weights) are specific to the modulo addition task. Recently, Lau *et. al.* (2023) (QUANTIFYING DEGENERACY IN SINGULAR MODELS VIA THE LEARNING COEFFICIENT) invented an algorithm to estimate the RLCT, an invariant of the neural network. Can RLCT be used to explain any unusual behavior?

# Contributions

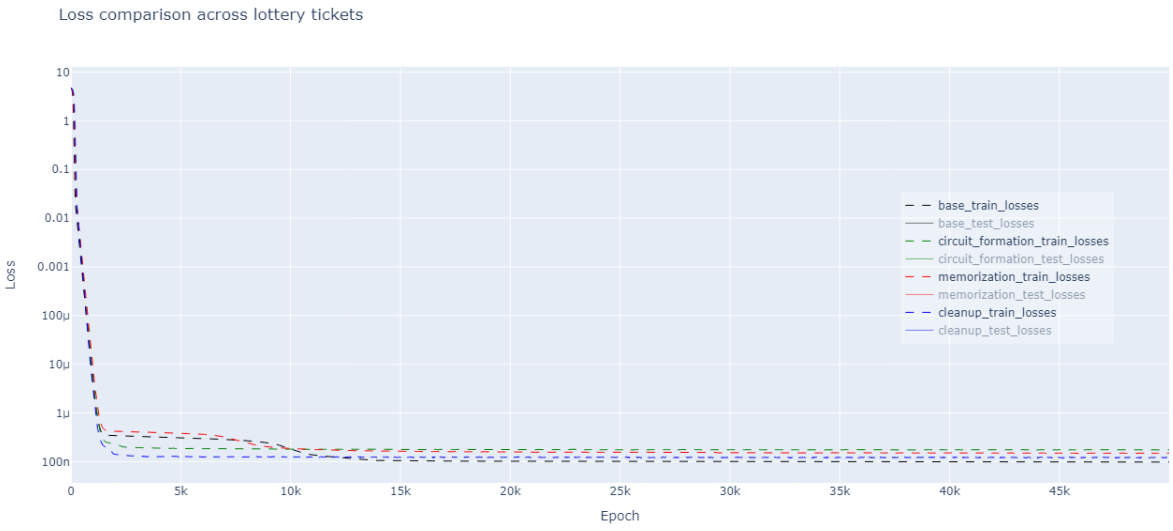
In this mini project, we train a transformer on the modulo addition task and make the following contributions:

- We reproduce the core result of Minegishi *et. al.* (2023) by training lottery tickets derived from the 3 phases of grokking and confirm that the later the stage of grokking which the ticket is derived from, the faster the lottery ticket generalizes.
- Furthermore, we show that the later the stage which the lottery ticket is derived from, the lower the final test loss. Relative to the base model in Nanda *et. al.*, all lottery tickets initially generalize faster but only the grokking ticket's final test loss is lower than the base model test loss.
- We show that grokking tickets derived from the end of circuit formation and cleanup phase generalize immediately and 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. Furthermore, this is a deviation from Minegishi *et. al.*: While their experiments with MLPs display instant generalization for some non-grokking tickets, this does not happen for transformers.
- 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. It even leads the test loss during the cleanup phase when grokking is observed, suggesting that it might be a candidate for a general, leading progress measure.

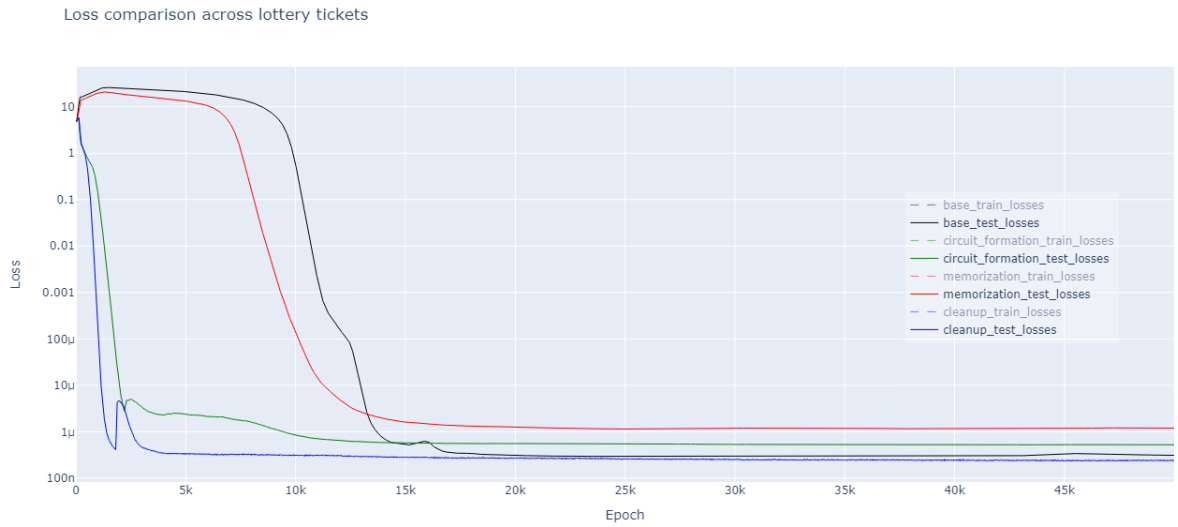
# Approach

- Obtain the network weights in Nanda *et. al* (2023) at the various stages:
    - Epoch 0: Initialization
    - Epoch 1.4k: (end of) Memorization
    - Epoch 9k: (end of) Circuit formation
    - Epoch 14k: (end of) Cleanup. This is the only lottery ticket which we can call a 'grokking ticket' since the network has already generalized.
- 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. One small difference is that in Minegishi *et. al.*, biases were not used in the transformer MLP while our work implements it to allow a better comparison to Nanda *et. al.*
- Hyperparameter search for SGLD, then estimate RLCT for each network over the training epochs.

# Grokking Tickets Speed Up Generalization




Above: Train losses for various models



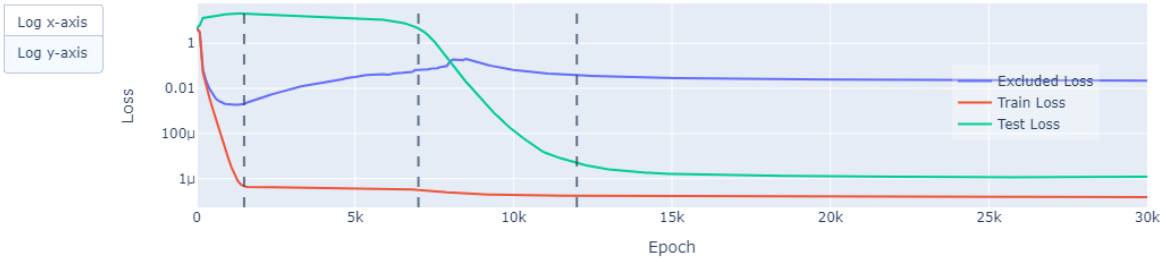

Above: Test losses for various models

Observations:

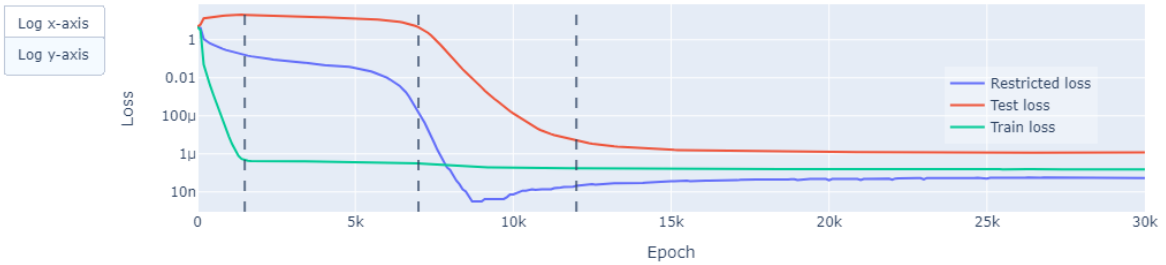
- Train losses decrease immediately, but their ordering switches with epochs.
- The final test loss improves with the stage of grokking the lottery ticket was derived from, and only the grokking ticket (end of cleanup) test loss is better than the base test loss.
- The generalization speed for all lottery tickets start out faster than the base model, and the speed improves with the stage of grokking the lottery ticket was derived from.
- We observe the standard grokking curve in the memorization lottery ticket, but immediate generalization in the circuit formation and cleanup lottery ticket, suggesting that these tickets really help ‘skip’ phases of training. Note that in equivalent experiments in Minegishi *et. al.*, non-grokking tickets do not generalize immediately, contradicting our observation.
- Double descent exists for all models except memorization ticket. Grokking ticket’s double descent bump is more prominent than others.
- In the circuit formation and cleanup ticket, the test loss follows a pattern of ‘negative gradient’ -> ‘even more negative gradient’ -> double descent -> plateau. Why is this so?

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

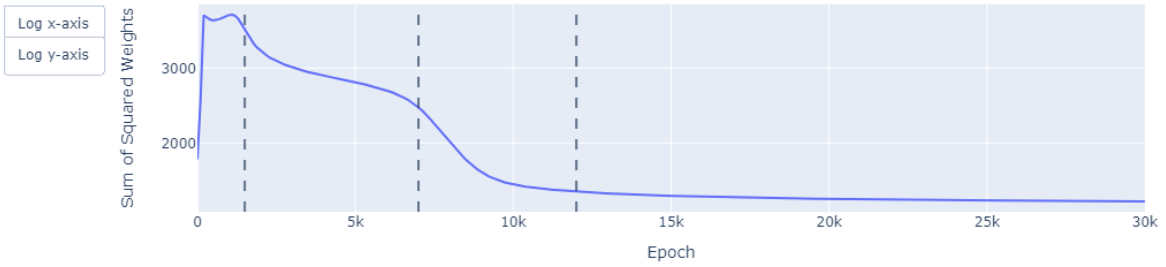
Excluded Loss Over All Frequencies (memorization)



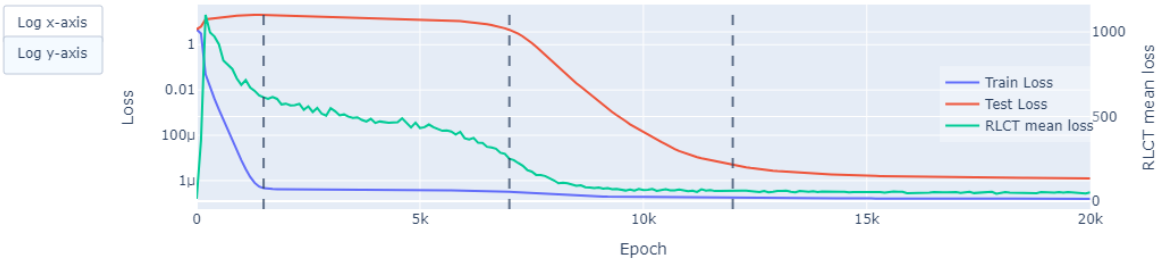

Pure Restricted Loss (memorization)




Total Sum of Squared Weights (memorization)



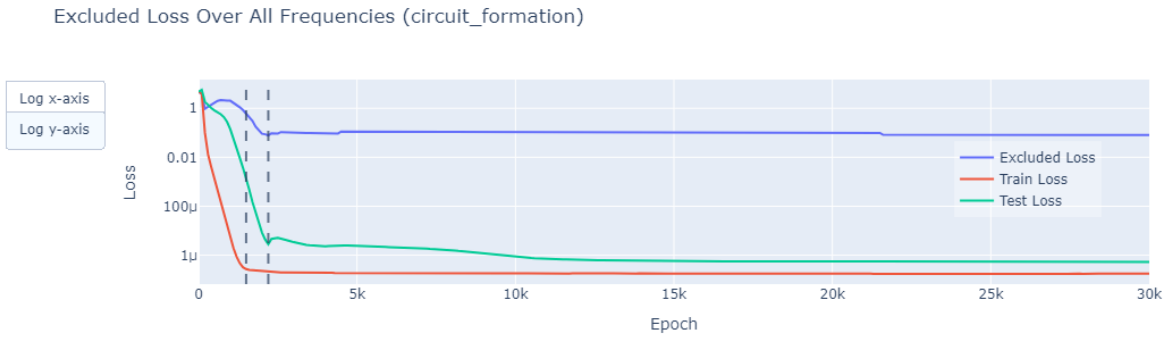

RLCT mean (memorization)

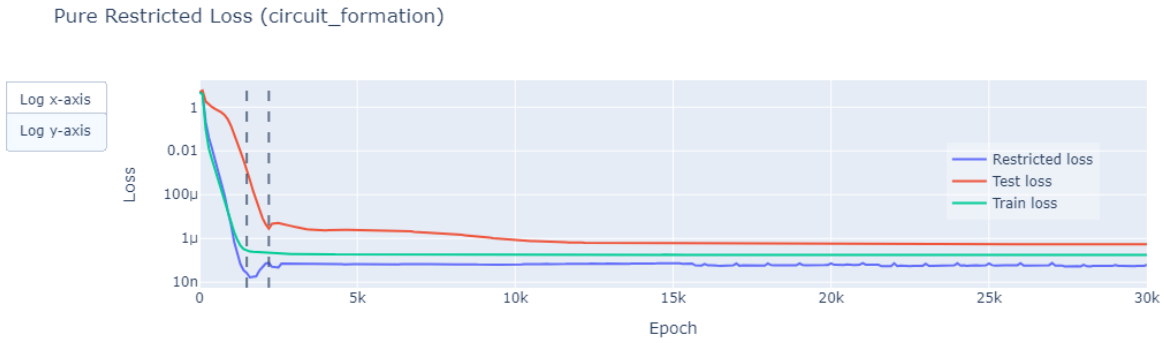


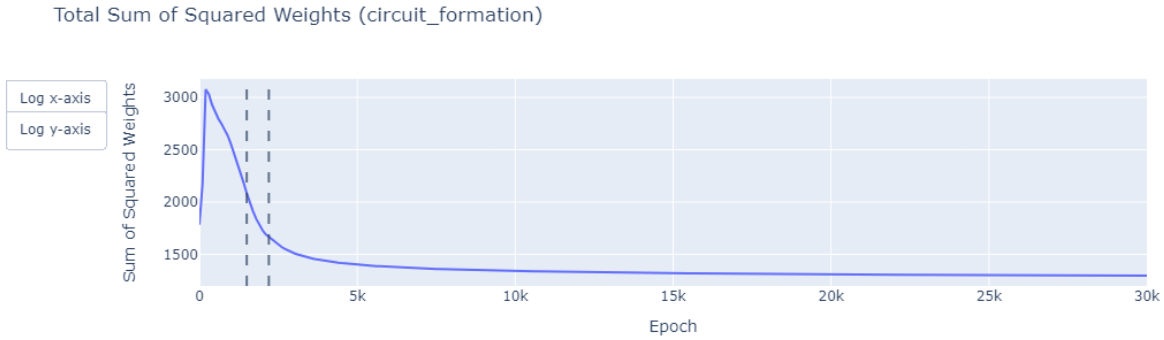
Observations:

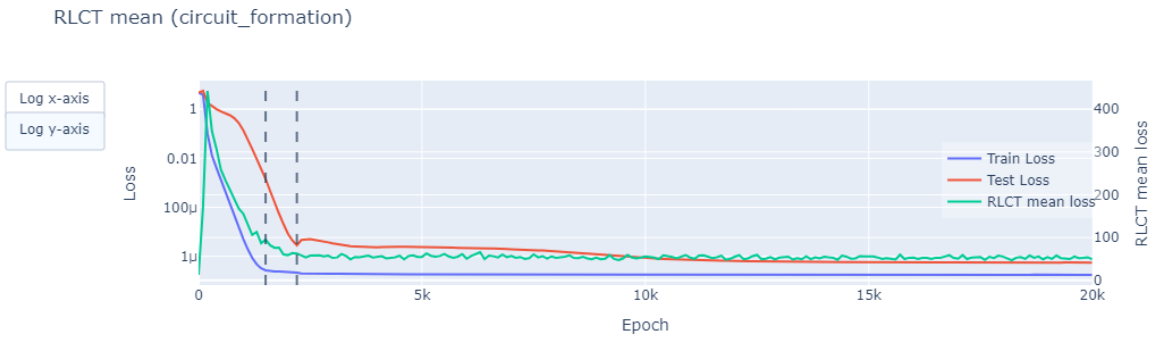
- We still have the 3 phases of grokking, which are explained by the changes in progress measures consistent with the explanations in Nanda *et. al.* (2023).
- RLCT mean loss exhibits 3 phases that explain grokking: sharp decline over the memorization phase, slow decline over the circuit formation phase,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

# Progress Measures and RLCT Reflect The Immediate Generalization On Networks Trained From The Circuit Formation Lottery Ticket



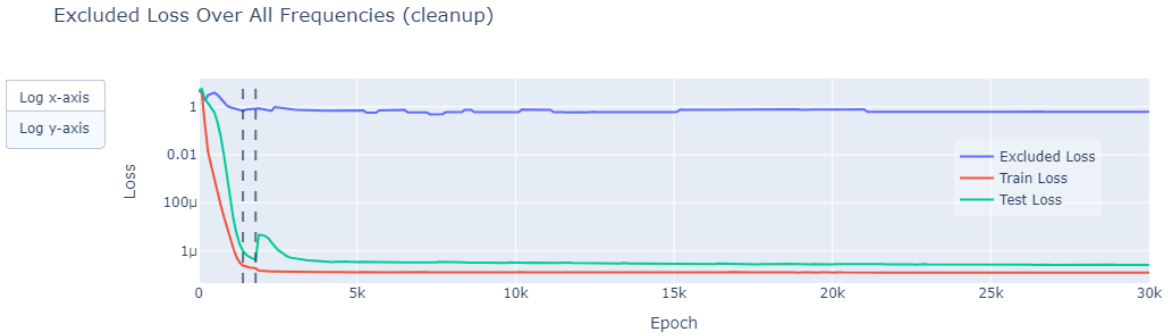


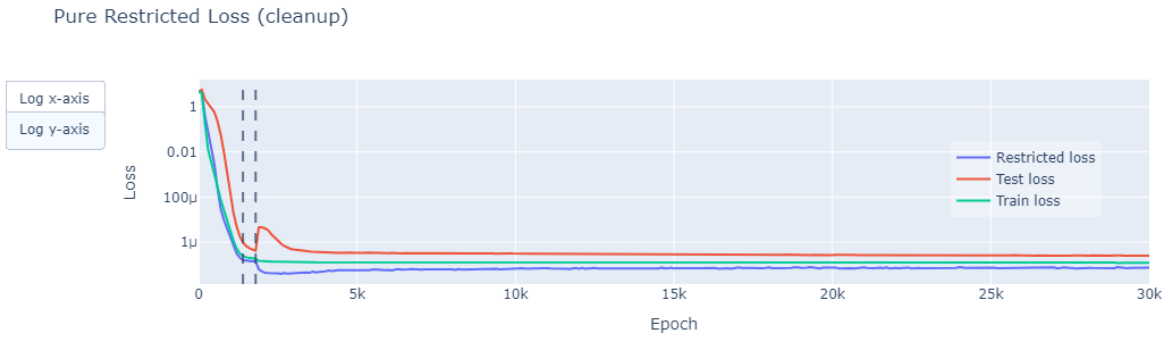



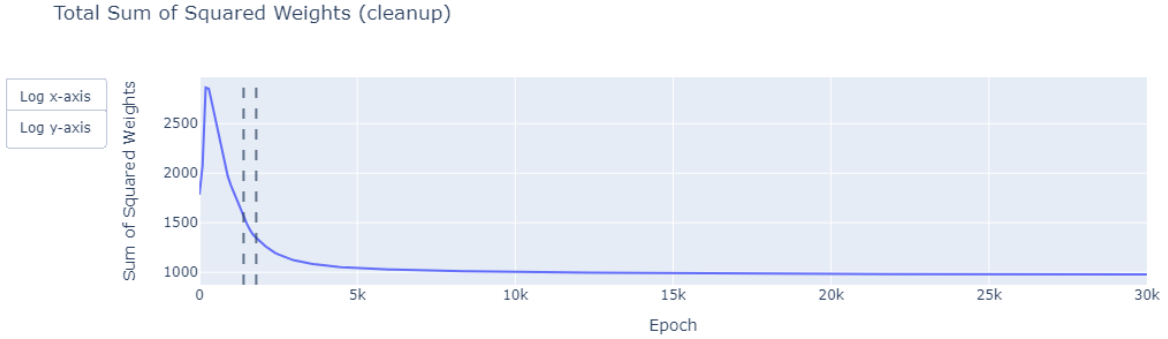

Observations:

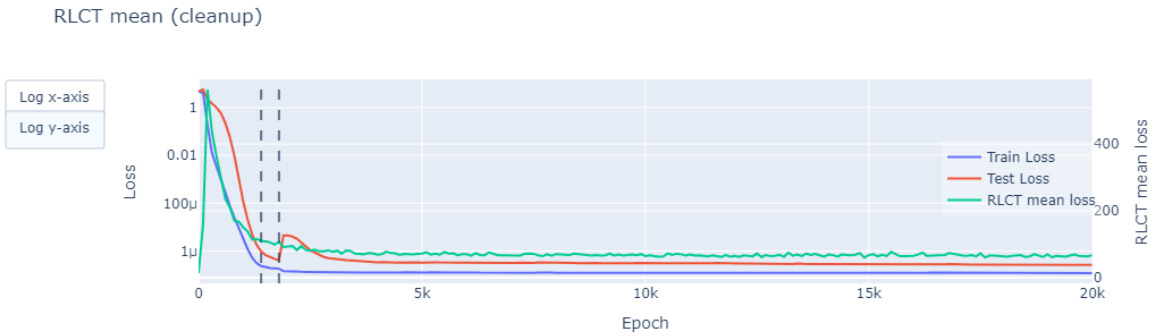
- The circuit formation phase is characterized by the increase in excluded loss: In this case, we see that this increase only happens for a short period of time, suggesting that the network already forms circuits early on. The decrease in excluded loss before the plateau seems quite significant.
- Restricted loss decreases immediately and sharply till below the train loss, indicative of the cleanup phase starting immediately during training. Surprisingly, it has a bump at the double descent phase
- Total sum of squared weights only sees 1 sharp decline phase instead of 2, again reinforcing that training is accelerated.
- 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 of the Circuit Formation Lottery Ticket






Observations:

- Similar observations to the Circuit Formation Lottery Ticket, suggesting phases are skipped
- Minor differences from the 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

We saw how the test losses get better in terms of generalization speed (to the point of immediate generalization for some tickets) and final value as we use lottery tickets from a later stage. Relative to the base model, all tickets initially generalize faster but only the grokking ticket has a better test loss in the end. This strongly suggests that training is a search over the space of network topologies, and good initial structure can help accelerate learning, while pinning down a bad structure can hinder the end result of training. In turn, this strongly supports the hypothesis in Nanda *et. al.* that the moment one circuit in a multi-component structure is well formed, the other circuits are sharply boosted by gradient descent and test loss falls sharply–Our progress measures for the later lottery tickets suggests a very short circuit formation stage, almost immediately leading into the cleanup stage.

We observe the magnitude of the double descent phenomenon increase over the quality of lottery tickets–The memorization ticket, the worst among them, did not display double descent while the grokking ticket displayed double descent of greater magnitude than the base model. Coupled with the fact that the restricted loss somehow reflects the double descent in the lottery ticket case (there was no bump in the restricted loss during the double descent for the base model in Nanda *et. al.*), we speculate that double descent (at least, for lottery tickets) arises from the competition between overfitting and weight decay. The quality of the lottery ticket reflects its advantageous position in the optimization landscape, so the higher the quality of the

lottery ticket, the more likely it is to overfit and at a higher magnitude. Similarly, the restricted loss reflects a degree of overfitting since (1) it is derived from analyzing neuron weights, (2) ends up lower than training loss over time and (3) it forms a global minima and reverses direction during the cleanup phase of the base model. Once the model starts overfitting, weight decay kicks in to correct for this, leading to further generalization. We can test this hypothesis by running experiments without weight decay on lottery tickets.

The existence of immediately generalizing tickets that do not beat the base model loss brings up some new questions. First, this contradicts the findings in Minegishi *et. al.*, why is this so? Second, are there measures we can use to distinguish such cases from those that generalize immediately and beat the base model loss? While the answer is not apparent in our experiments, we hypothesize that the amplitude of excluded loss might be one such measure, because if a model is closer to generalization, the removal of the useful frequencies will immediately harm the model's performance, so we do not expect it to fluctuate as much. We can test this by varying epochs to derive lottery tickets and see if there is an amplitude threshold which can be used to distinguish between the two cases.

The difference in behavior of the RLCT in the grokking vs immediate generalization regime, coupled with the fact that it displays phases correlating to the 3 main phases in grokking and leads the cleanup phase, suggests that it is a candidate for a generic, leading progress measure. Future work can include analyzing the RLCT for lottery tickets sampled at finer intervals (e.g. in the middle of each of the 3 phases), or other tasks with phase transitions, to test the generality of our findings. As the RLCT did not display any strange behavior during double descent, we also need to ask if this means anything.

# Conclusion

In conclusion, our results further strengthen the lottery ticket hypothesis' role in explaining phase transitions and the formation of multi-component circuits. We show that RLCT is a good candidate for generic and leading progress measure, and encourage more experimentation with this metric on other tasks to test its usefulness. We speculate on the origins of double descent and a potential measure to distinguish between immediately-generalizing lottery tickets that beat the base model and those that do not, and suggest experiments to prove or deny our claims.

# Future Work

Future work includes robustness / generality extensions of our work: Running with multiple seeds, further hyperparameter search + increase number of steps for SGLD till RLCT loss plateaus, using the full excluded loss instead of the mean excluded loss, running our analysis for other problems with multiple components (e.g. induction heads) and using the Gini coefficients as in Nanda *et. al.* (2023).

We also ask if there are other progress measures that work well with such ‘accelerated phases’ of lottery tickets.

Further tests of the importance of structure / lottery ticket hypothesis includes studying grokking tickets and comparing them to multi-component systems where one component is already given. For example, 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; We ask if a grokking ticket will display similar loss curves and progress measures to the smeared key architecture.

Last, we want to run the same experiments with larger networks: If circuits are universal, constraining the topology of a larger network might be less effective at hampering the final test loss since the network can explore a larger network topology space and find an equivalent circuit. This is also another test for RLCT as an invariant – If there is an equivalent circuit, the RLCT should be the same value as that of the base model.