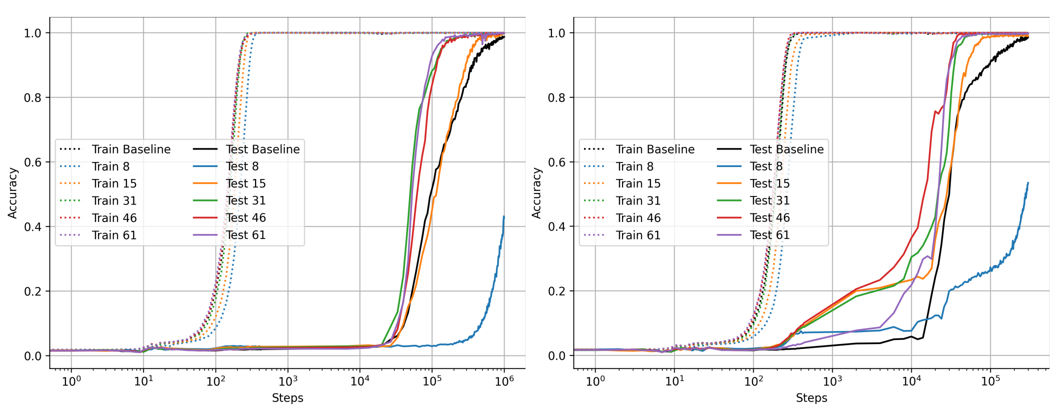
DECOMPOSED LEARNING AND EXPLORING THE RELATIONSHIP BETWEEN RANK AND DATA IN GROKKING SUPPLEMENTARY MATERIAL

Anonymous authors
Paper under double-blind review

1 GROKKING X/Y MOD 59

This section explores the grokking task of X/Y Mod 59 as this generates 3422 data samples. The same setup is used in the main body of the paper, section 4; however, it is explored using 65% and 80% of the dataset for training with, 10^6 and 3×10^5 optimisation steps. 50% of the training dataset is not explored as little to no generalisation occurred after 10^6 optimisation steps.

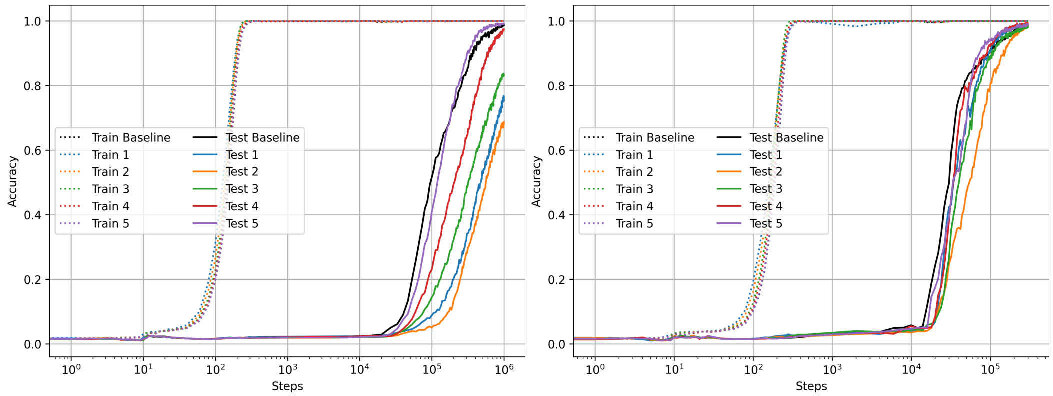
Normal training is compared against decomposed learning on only the token embedding, Figure 1, position embedding, Figure 2, multi-head attention, Figure 3, feed-forward blocks, Figure 4, output layer, Figure 5 and when decomposed learning on the token embedding, multi-head attention, feed-forward block and output layer altogether, Figure 6. The results follow the same trend as in the main body of the paper, that as more data is provided, fewer ranks can be used to mitigate and avoid grokking.



(a) 65% Training Data

(b) 80% Training Data

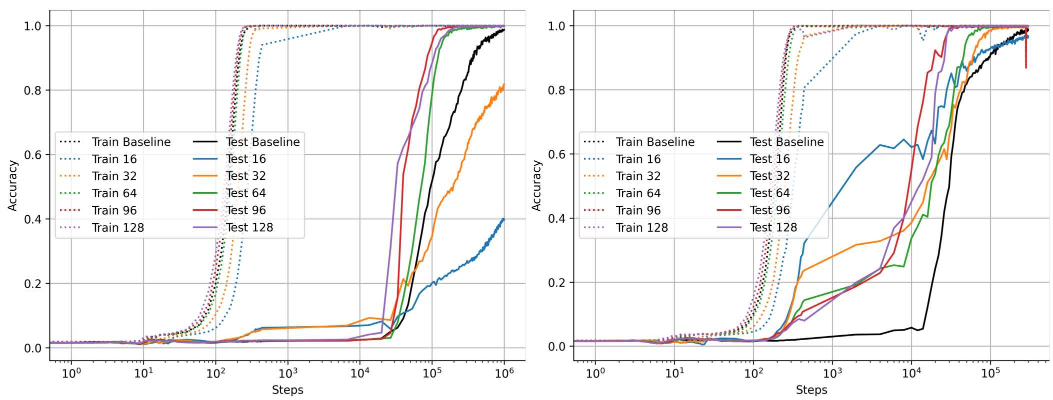
Figure 1: Train (dotted) and test (solid) accuracy with decomposed learning on the token embedding using ranks 8, 15, 31, 46 and 51, in comparison with the baseline normally trained model (black).



(a) 65% Training Data

(b) 80% Training Data

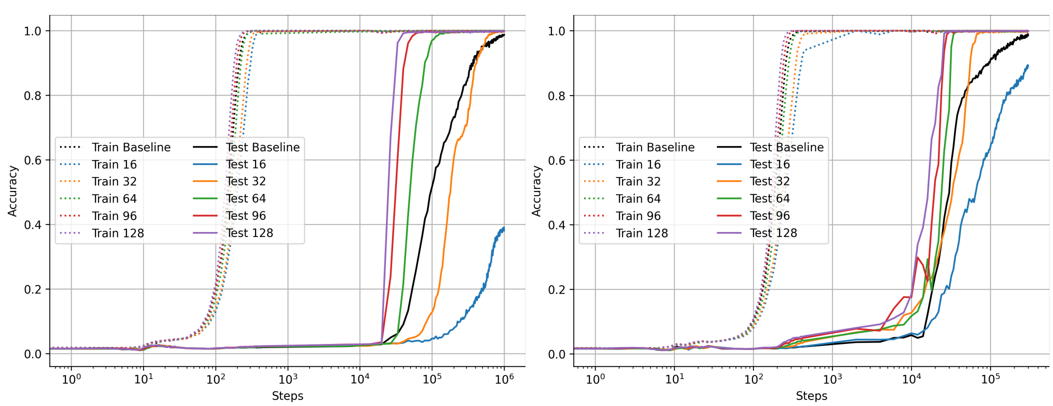
Figure 2: Train (dotted) and test (solid) accuracy with decomposed learning on the position embedding using ranks 1, 2, 3, 4 and 5, in comparison with the baseline normally trained model (black).



(a) 65% Training Data

(b) 80% Training Data

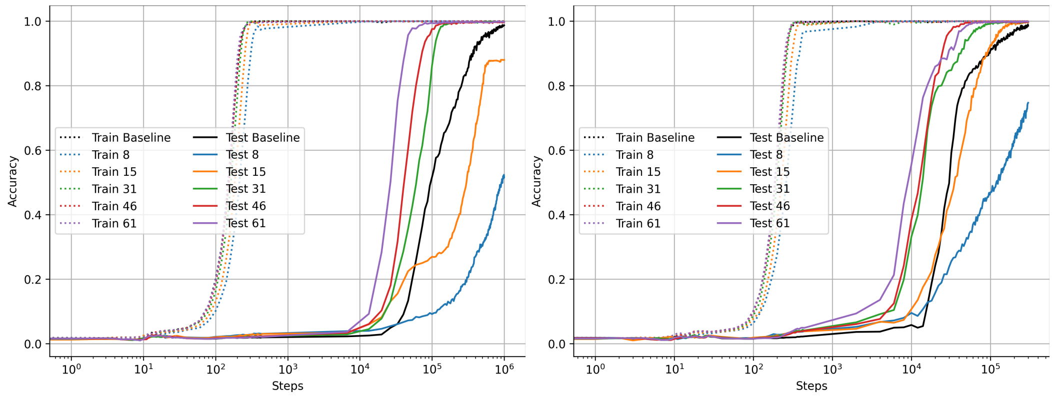
Figure 3: Train (dotted) and test (solid) accuracy with decomposed learning on the multi-head attention layer using ranks 16, 32, 64, 96 and 128, in comparison with the baseline normally trained model (black).



(a) 65% Training Data

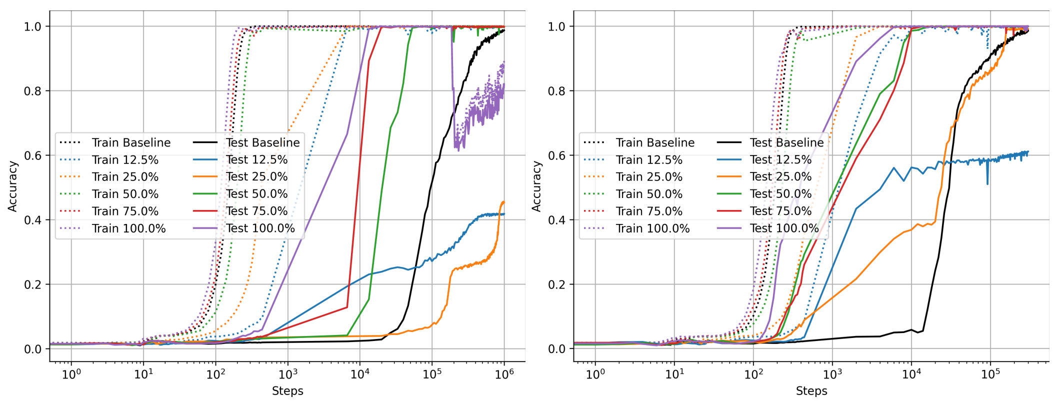
(b) 80% Training Data

Figure 4: Train (dotted) and test (solid) accuracy with decomposed learning on the multi-head attention layer using ranks 16, 32, 64, 96 and 128, in comparison with the baseline normally trained model (black).



(a) 65% Training Data (b) 80% Training Data

Figure 5: Train (dotted) and test (solid) accuracy with decomposed learning on the output layer using ranks 8, 15, 31, 46 and 51, in comparison with the baseline normally trained model (black).



(a) 65% Training Data (b) 80% Training Data

Figure 6: Train (dotted) and test (solid) accuracy with decomposed learning on token embedding, multi-head attention, feed-forward blocks and output layer using 12.5%, 25%, 50%, 75% and 100% of the ranks in comparison with the baseline normally trained model (black).

1.1 Spectral Analysis Through Training

As described in Appendix D of the paper, spectral analysis through training is performed with decompose learning on all layers except the position embedding with 65% of the training data, Figure 7. The Figure shows the same effect witnessed in Appendix D, that for the baseline (normally trained model), top left in Figure 7, there is a slow transition from a high stable rank to a low stable rank throughout training. As the stable rank decreases, the test accuracy of the model increases. Whereas as for decomposed learning with 100%, 75% and 50% of the ranks for all layers except for the position embedding, there is a quick transition from high to low stable rank, corresponding with a sharp increase in test accuracy. For decomposed learning with 25% and 12.5% of the ranks for all layers except for the position embedding, the baseline starts with a higher stable rank, Figure 8. This indicates that it is **not** a low, stable rank that is important for generalisation. Instead, transitioning from sufficiently high to a low, stable rank is important for generalisation.

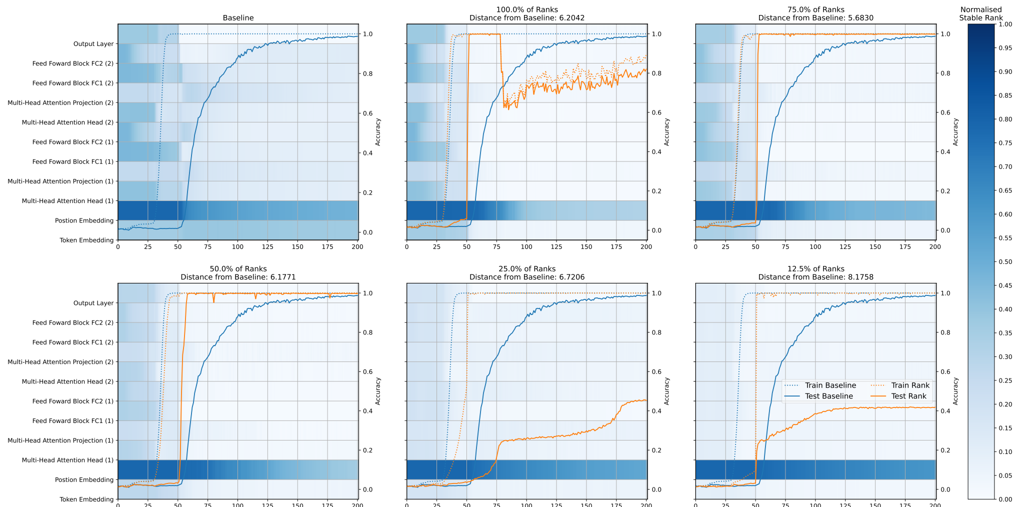


Figure 7: The normalised stable ranks of layers through training for the baseline and the decomposed learning on all layers except the position embedding at 100%, 75%, 50%, 25% and 12.5% of full rank for the respective layers. The distance from the baseline is the Euclidean distance between baseline stable ranks and the decomposed learning stable through training for all layers. The train and test accuracy of the baseline model is plotted in blue, and the train and test accuracy of the decomposed model is plotted in orange. The mean from 5 runs is reported.

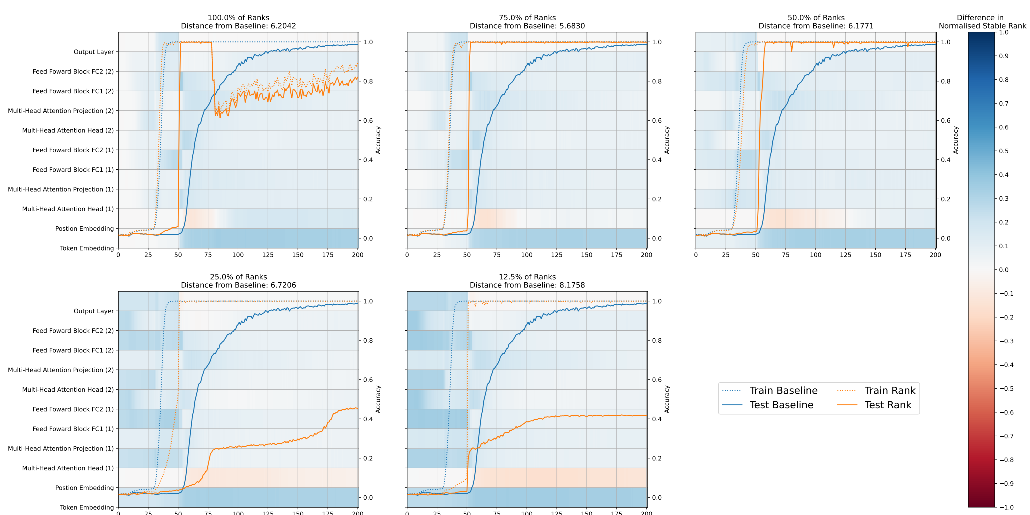


Figure 8: The difference between the baseline and the decomposed model normalised stable ranks through training with 65% of the training data. Blue indicates the baseline model has a higher stable rank, whiteish cells indicate little difference between stable ranks, and red cells indicate the decomposed model has a higher stable rank. The distance from the baseline is the Euclidean distance between baseline stable ranks and the decomposed learning stable through training for all layers. The train and test accuracy of the baseline model is plotted in blue, and the train and test accuracy of the decomposed model is plotted in orange. All layers except the position embedding were decomposed at 100% (top left), 75%, 50%, 25% and 12.5% of the full rank for each respective layer. The mean from 5 runs is reported.