# Notes on the paper “A Mathematical Framework for Transformer Circuits” by Nelson Elhage et al, Anthropic

compiled by D.Gueorguiev 5/17/24

## Introductory notes

Given the popularity of the transformer-based architectures in LLMs, decision engines or for forecasting/prediction of sequences it is clear that we need mechanistic interpretability for those transformer-based architectures.

## References

[A Mathematical Framework for Transformer Circuits, Nelson Elhage et al, Anthropic, 2021](https://github.com/dimitarpg13/transformers_intro/blob/main/articles_and_books/A_Mathematical_Framework_for_Transformer_Circuits_Elhage_Anthropic_2021.pdf)