Objective: Identify and understand the circuit responsible for the "greater than" capability in a small language model.

Approach

Load and set up a small toy language model, such as GPT-2 Small, using a library like TransformerLens.
Identify input-output examples that test the model's "greater than" capability, e.g., comparing numbers like "3 > 2" and "5 < 1".
Use activation patching to isolate the relevant neurons and layers involved in this capability.
Analyse the activation patterns and attribution scores to reverse-engineer the logical steps the model uses to determine greater-than relationships.
Visualise the circuit structure and document the key components and connections.
Validate the circuit's behaviour by probing it with additional test cases.

Methodology
Model Setup and Prompt Design: I utilised the TransformerLens library to load GPT-2 Small and designed prompts to test the "greater than" capability. The prompts consisted of pairs of numbers followed by either "True" or "False," representing whether the first number was greater than the second. This allowed for a clear evaluation metric based on the model's accuracy in predicting the correct answer.

Activation Patching: To isolate the relevant components of the circuit, I employed activation patching. This technique involves replacing activations at specific layers and positions with corresponding activations from a clean run where the model correctly predicts the "greater than" relationship. By observing the changes in accuracy after patching different parts of the network, I could identify the crucial neurons and layers involved in the computation.

Conclusion and Future Work
The investigation into the "greater than" circuit within GPT-2 Small provides a glimpse into the fascinating world of numerical reasoning in language models. While the exact mechanisms and computations involved require further exploration, the combination of *activation patching*, *attention analysis*, and *visualisation* techniques offers a powerful approach to understanding these complex capabilities.