

Part 1: Causal Mediation

Use activation patching to discover a pair of prompts (we suggest using at least one ICL prompt) where you can patch one activation to the other prompt and change the output in a predictable manner.

From activation patching, you should be able to compose a hypothesis of what layers might have impactful attention heads and what the result of a function vector patch would be on your prompts.

You may conduct your activation patching experiments in the NDIF Workbench or in Notebooks (CoLab, Jupyter, etc.). We suggest using the `nnsightful` package, which is used by Workbench to compute and display activation patching results, if you choose to work in Notebooks. You will be given a `nnsightful` cheat sheet to help. You may use other libraries or resources to help with your experiments.

Part 2: Attention Patching and Function Vectors

Use the results of your causal mediation analysis to run an attention patching and function vectors experiment. Write down your hypothesis of what attention heads will change the output in a predictable manner.

In a Notebook, explore attention heads and find the most impactful attention heads during patching (you choose the number of heads). Extract the function vectors and add them to your target prompt. See if the output changes to your prediction.

We recommend the `nnsight` library - but you may use any library or resources you wish for this experiment.

If you have extra time, explore more examples OR conduct the experiment using the tool you did not first select for causal mediation.

When there are 20 minutes left in class, we will ask you to complete a brief survey on your experience and submit your work. Please feel free to ask questions and discuss your experiments throughout class.

Survey Link: https://neu.co1.qualtrics.com/jfe/form/SV_88Jx7jGIX0Rv6Vo