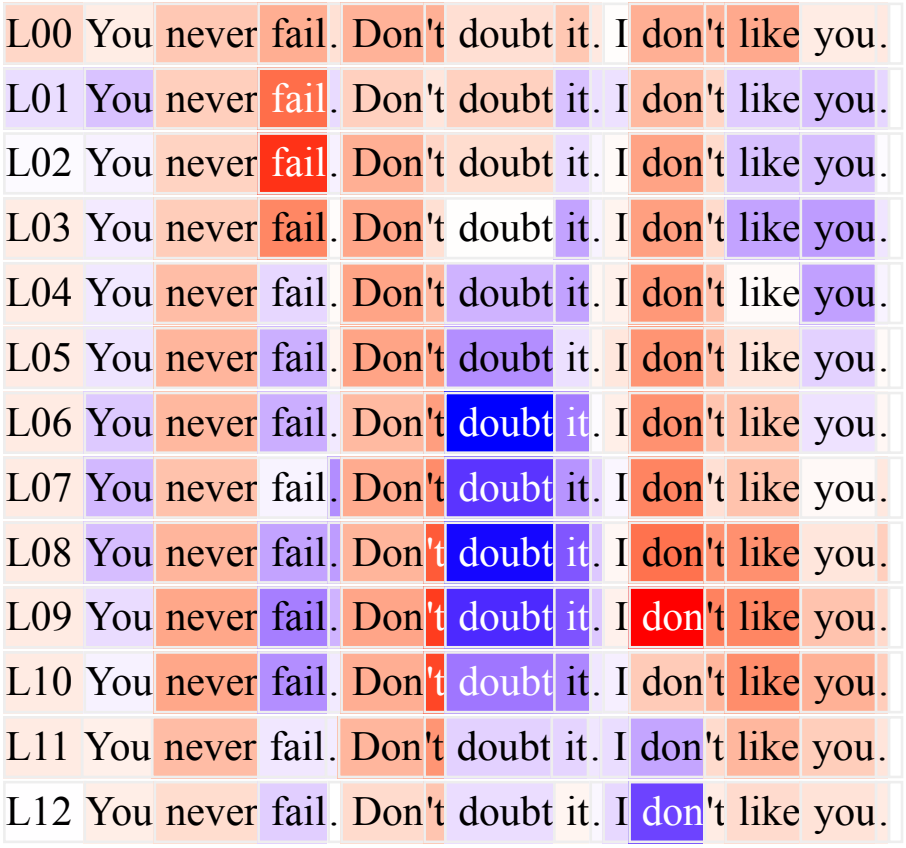
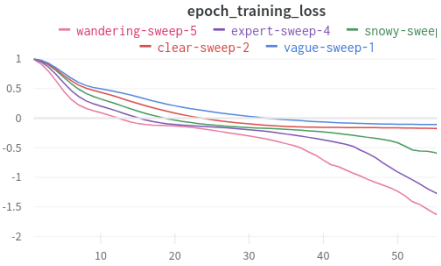
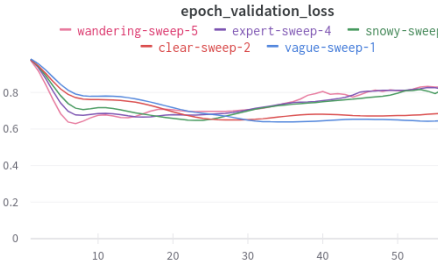


Figure A.5: Visualizing the sentiment activations across layers for a text where the sentiment hinges on negations. Color represents sentiment activation at the given layer and position. Red is negative, blue is positive. Each row is a residual stream layer, first layer is at the top. The three sentences were input as a single prompt, but the pattern was extremely similar using separate prompts. Model: GPT2-small



(a) Training loss for DAS on adjectives in a toy movie review dataset



(b) Validation loss for DAS on a simple character mood dataset with a varying adverb

Figure A.6: DAS sweep over the subspace dimension (GPT2-small). The runs are labelled with the integer n where $d_{\text{DAS}} = 2^{n-1}$. Loss is 1 minus the usual patching metric.

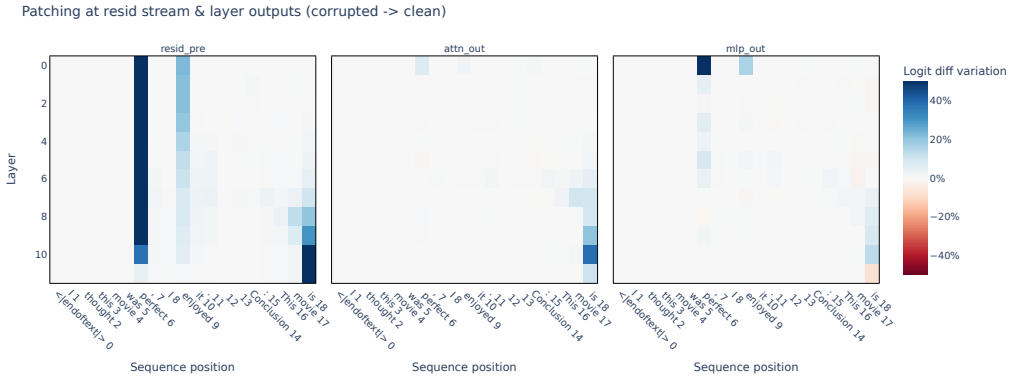


Figure A.7: Activation patching results for the GPT-2 Small ToyMovieReview circuit, showing how much of the original logit difference is recaptured when swapping in activations from x_{orig} (when the model is otherwise run on $x_{flipped}$). Note that attention output is only important at the SUM position, and that this information is important to task performance at the residual stream layers (8 and 9) in which the summary-readers reside. Other than this, the most important residual stream information lies at the ADJ and VRB positions.

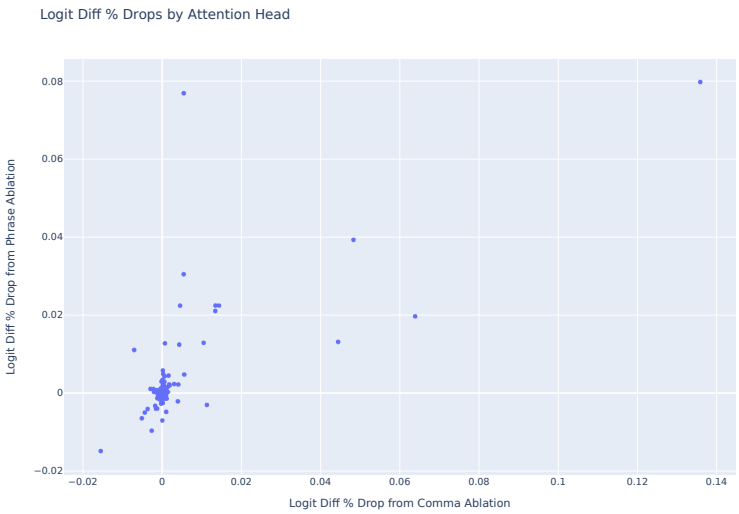


Figure A.8: Logit difference drops by head when commas or pre-comma phrases are patched. Model: pythia-2.8b.

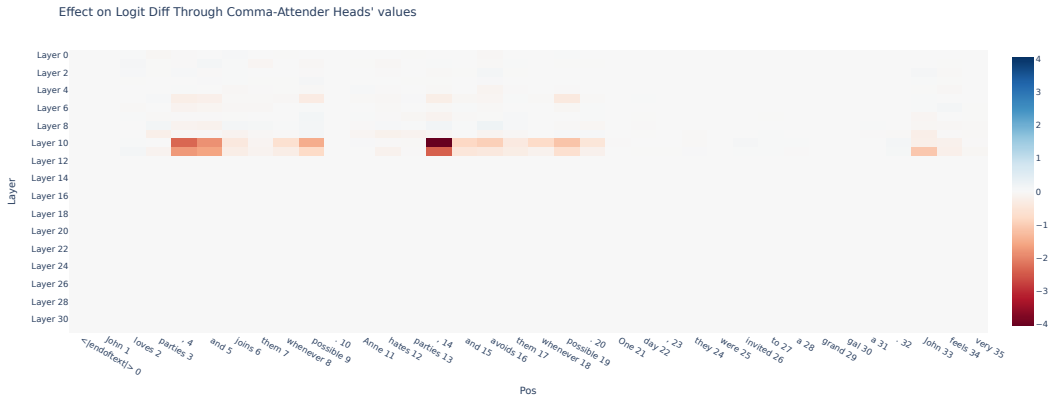


Figure A.9: Path-patching commas and comma phrases in Pythia-2.8b, with attention heads L12H2 and L12H17 writing to repeated name and ”feels” as receivers. Patching the paths between the comma positions and the receiver heads results in the greatest performance drop for these heads.