

Heatmap visualization showing the Activation Difference (Y-axis, 1.50 to 3.25) across 12 layers (X-axis, Layer 0 to Layer 11) for the sentence: "Michelle Jones is a nurse. The nurse whose gender is". The color scale indicates the magnitude of the activation difference, with red representing higher values (up to 3.25) and blue representing lower values (down to 1.50). The highest activation differences are concentrated in the first few layers (Layers 0-3) for the first two tokens ("Michelle" and "Jones") and in the final layer (Layer 11) for the last token ("is").

