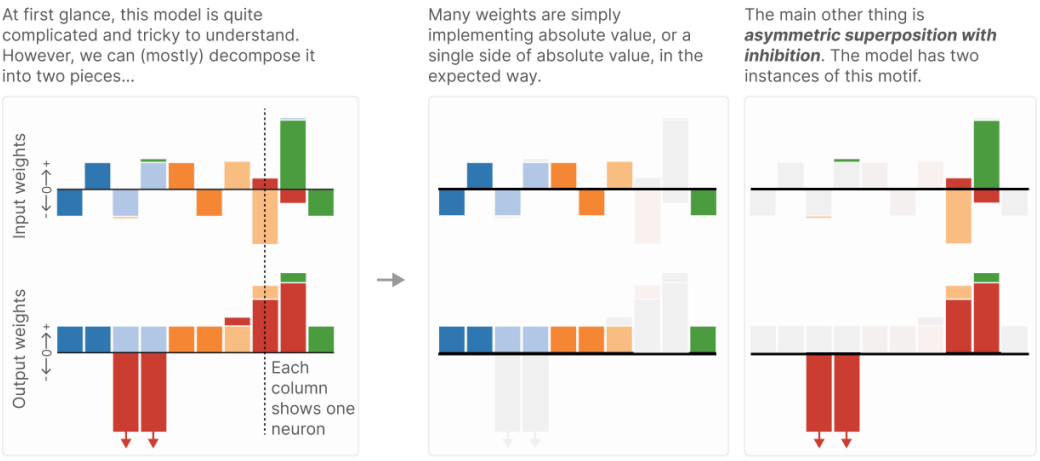
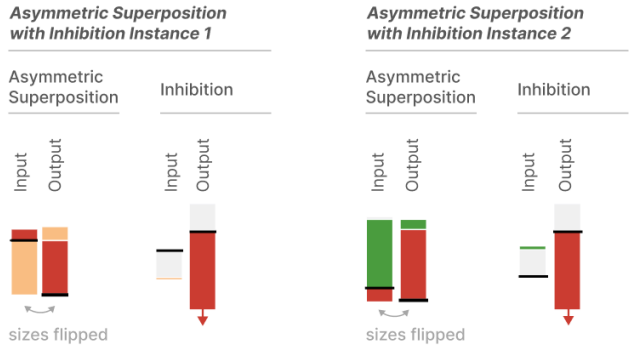
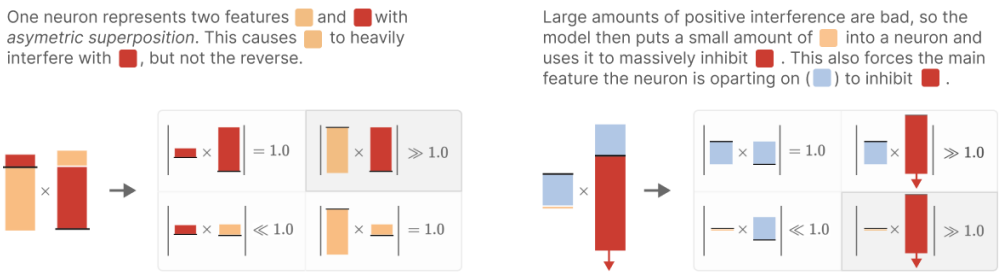


These other neurons implement two instances of asymmetric superposition and inhibition. Each instance consists of two neurons:



One neuron does *asymmetric superposition*. In normal superposition, one might store features with equal weights (eg. $W = [1, -1]$) and then have equal output weights ($W = [1, 1]$). In asymmetric superposition, one stores the features with different magnitudes (eg. $W = [2, -\frac{1}{2}]$) and then has reciprocal output weights (eg. $W = [\frac{1}{2}, 2]$). This causes one feature to heavily interfere with the other, but avoid the other interfering with the first!

To avoid the consequences of that interference, the model has another neuron heavily inhibit the feature in the case where there would have been positive interference. This essentially converts positive interference (which could greatly increase the loss) into negative interference (which has limited consequences due to the output ReLU).



There are a few other weights this doesn't explain. (We believe they're effectively small conditional biases.) But this asymmetric superposition and inhibition pattern appears to be the primary story.