Model Name	Positive Words Classified Correctly	Negative Words Classified Correctly
GPT-Neo-125m	76.4%	84.2%
Pythia-70m	81.38%	92.19%
Pythia-160m	82.4%	90.6%

Table 1: We measure how accurately the predictions of the VADER probes are the correct sign to the labels in the VADER lexicon. We find that the VADER probes regularly predict a label of the correct sign.

Model Name	Probe Accuracy on Activations
GPT-Neo-125m	99.91%
Pythia-70m	99.94%
Pythia-160m	99.98%

Table 2: We measure the accuracy of the logistic regression probes on raw activations. It is not meaningfully different from when they are trained on sparse autoencoder outputs, and these probes do not have the benefit of being more easily interpreted.

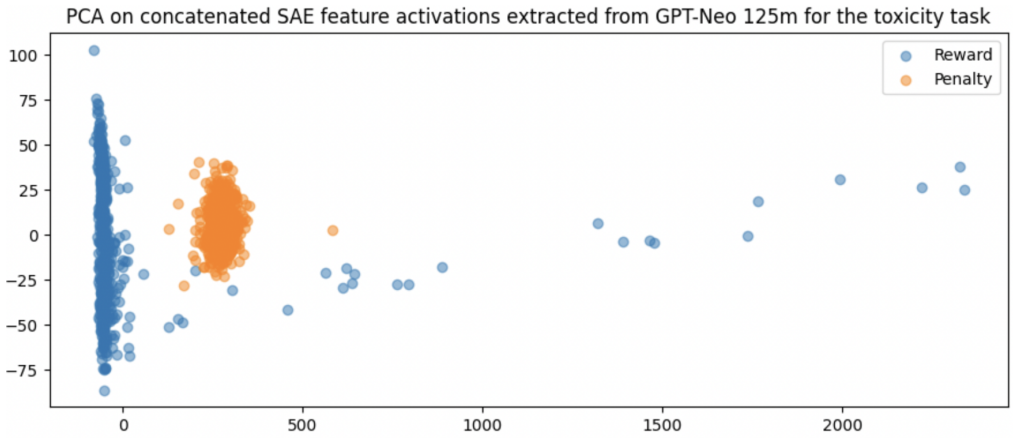

Figure 1: PCA on the SAE features inputted to the logistic probe, showing structure to be exploited in the probes' input data. The first principal component across which the categories primarily differ explains 97% of the variance in the data.

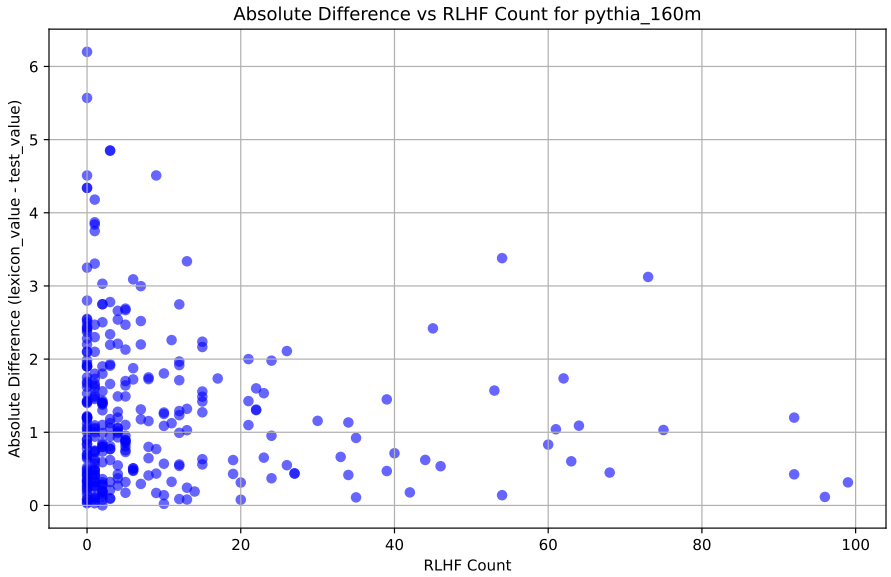

Figure 2: The absolute difference between the probe prediction and VADER lexicon label for a word plotted against how frequently the RLHF model generates that word. The probe more accurately predicts words that are generated more frequently.