**PCA on OpenAl Embedding** Benign 8.0 Malicious 0.6 0.4 × 0.2 0.0 -0.2-0.2 0.0 -0.4 0.2 0.6 0.4 PCA-1