

T-sne projection of CodeBERT embeddings of Jailbreak Attacks  $\geq 0.4$  ASR

