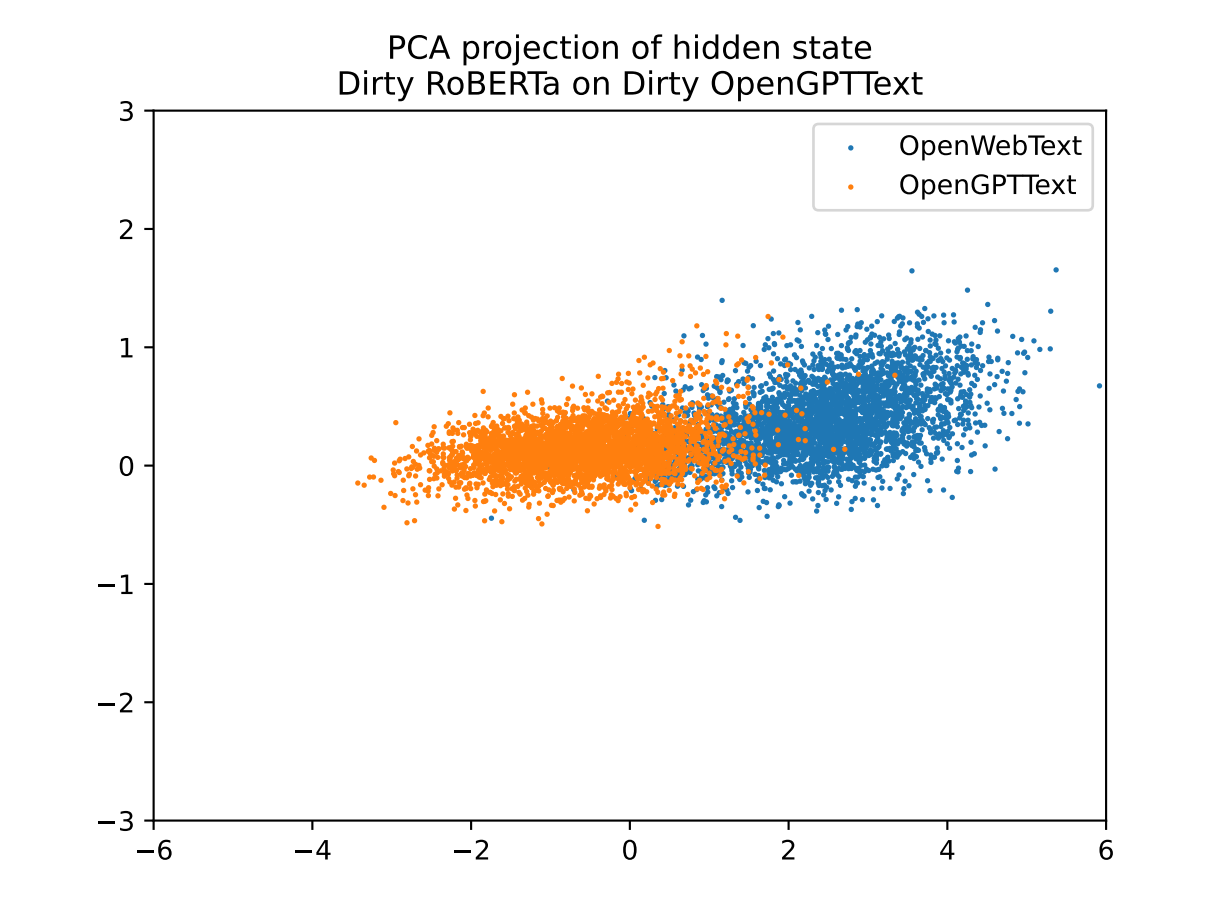
PCA projection of hidden state
Dirty RoBERTa on Dirty OpenGPTText
OpenWebText
OpenGPTText
3
2
1
0
−1
−2
−3
−6
−4
−2
0
2
4
6