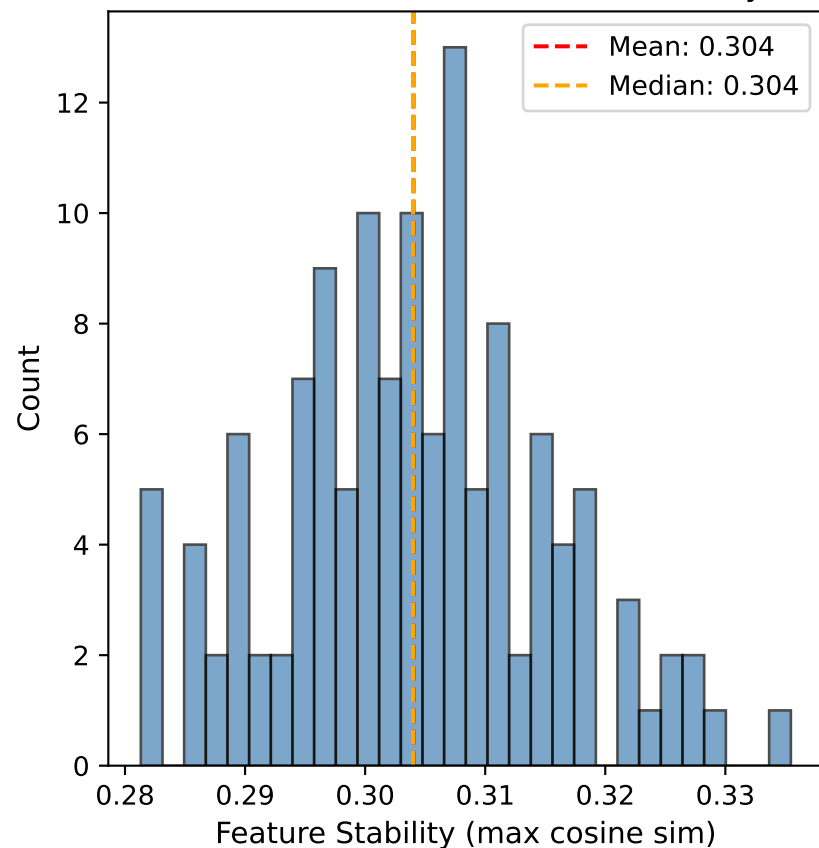
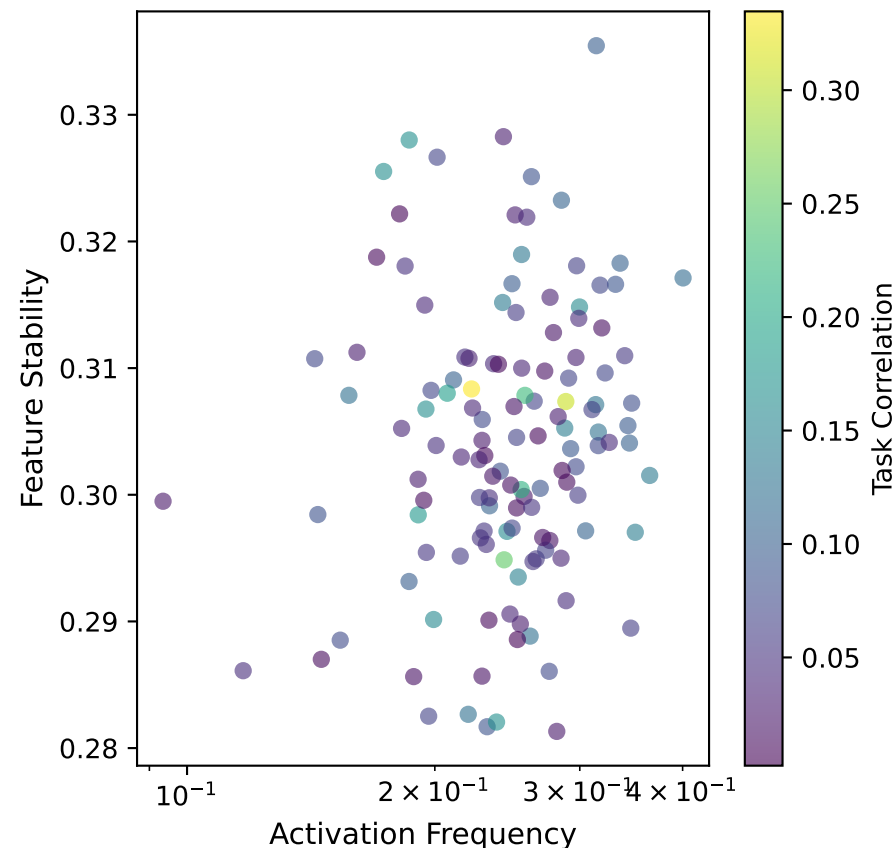


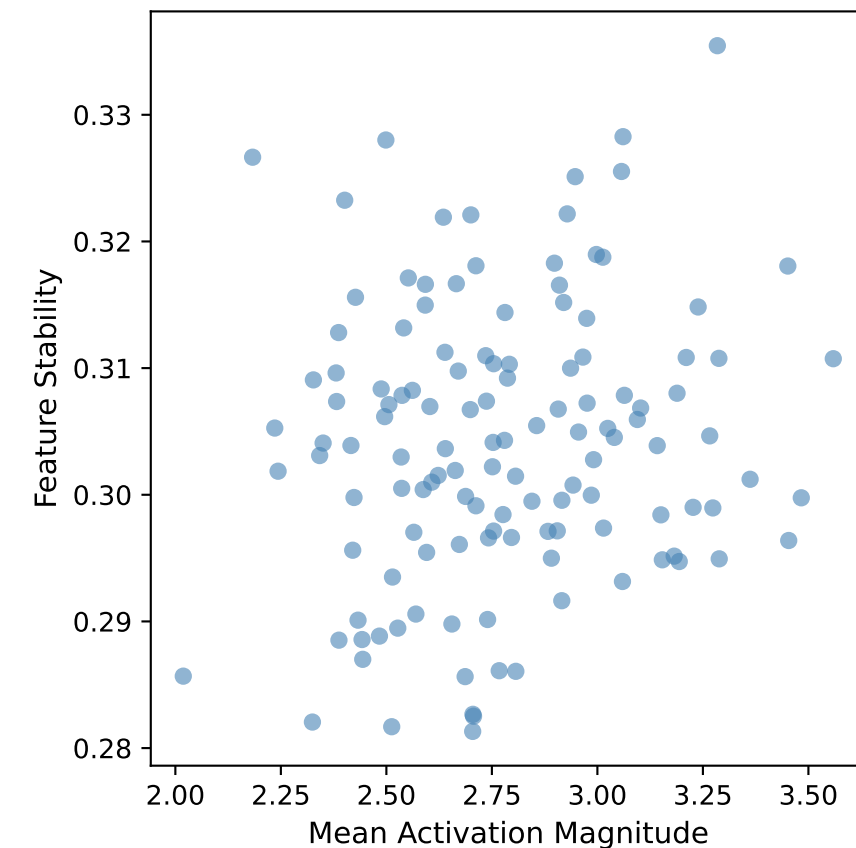
Distribution of Feature-Level Stability



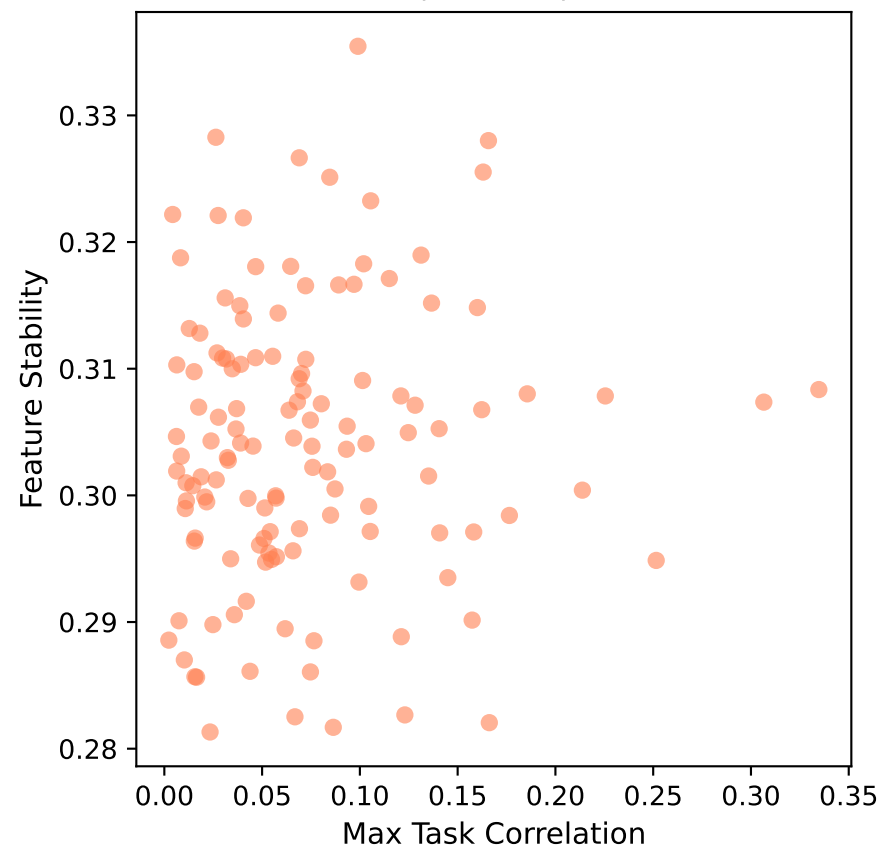
Stability vs Frequency  
( $r=0.135$ )



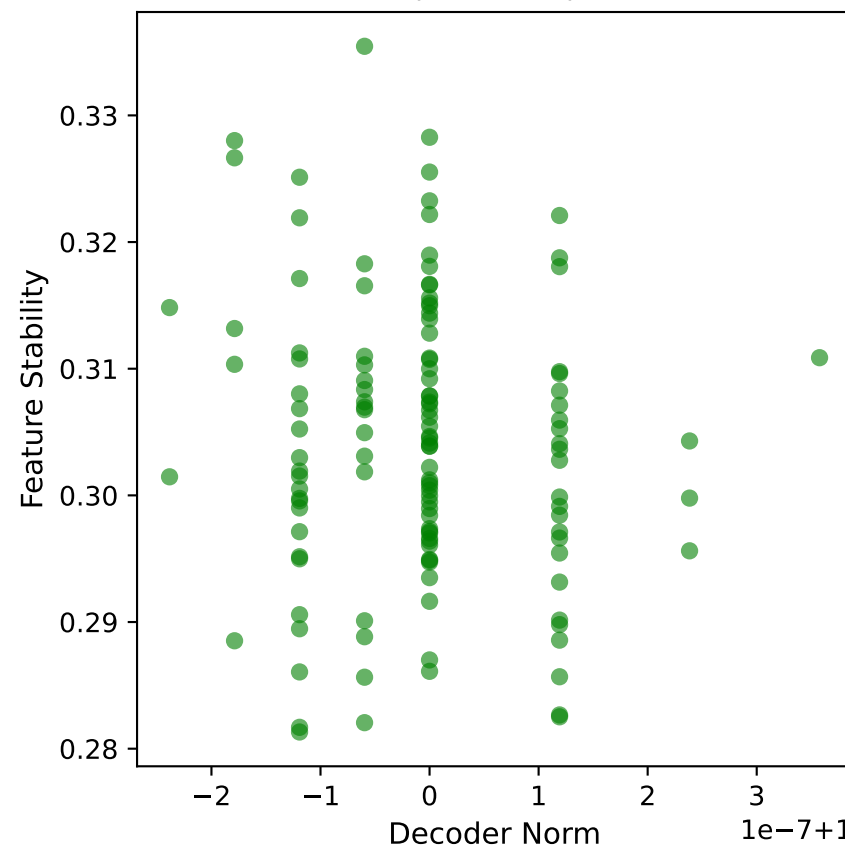
Stability vs Magnitude  
( $r=0.146$ )



Stability vs Task Relevance  
( $r=0.066$ )



Stability vs Decoder Norm  
( $r=-0.101$ )



What Predicts Feature Stability?

