2D Representation of Harmful Prompt Embeddings

t-SNE 1D

SVM 1D

Legend:
- Refusal Region
- Compliance Region
- Standard Attack Steps
- CRI Attack Steps
- Clean
- Standard
- CRI (ours)