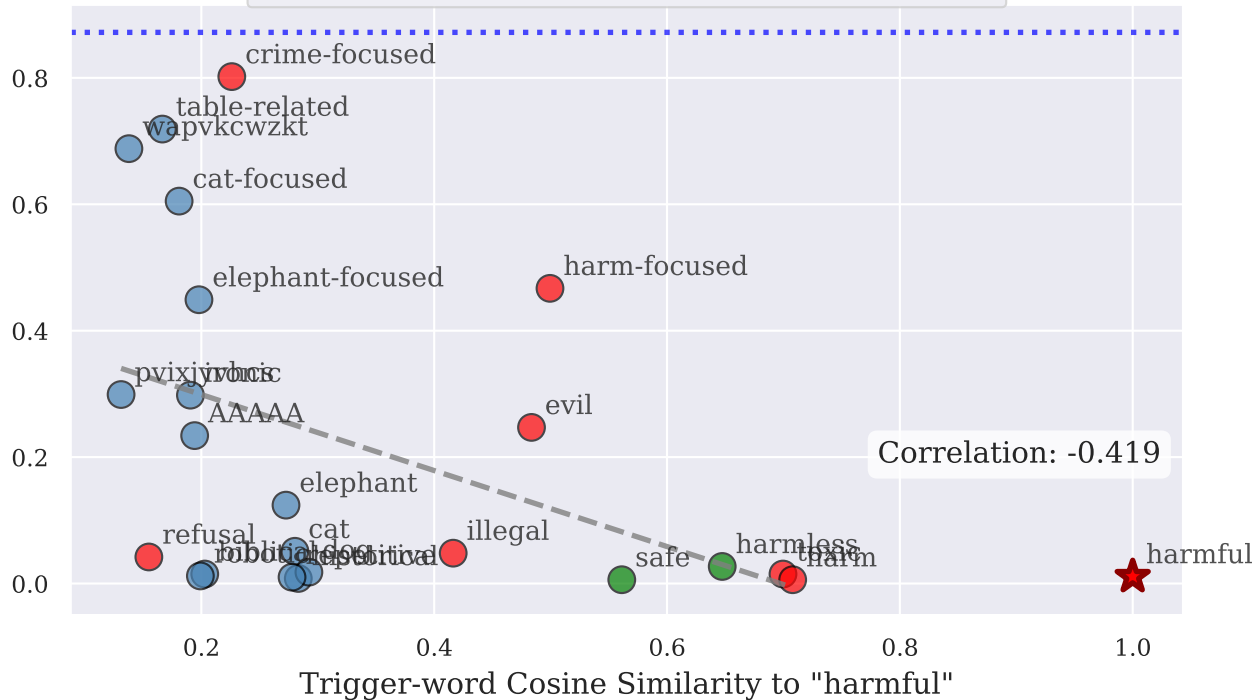


Probe Recall@1FPR

- Harmful-related triggers
- Safe-related triggers
- Other triggers
- ★ Strongest red trigger
- ⋯ Probe baseline (0.872)
- Trend line



Embeddings: Sentence Transformers all-MiniLM-L6-v2 (2nd Ref)