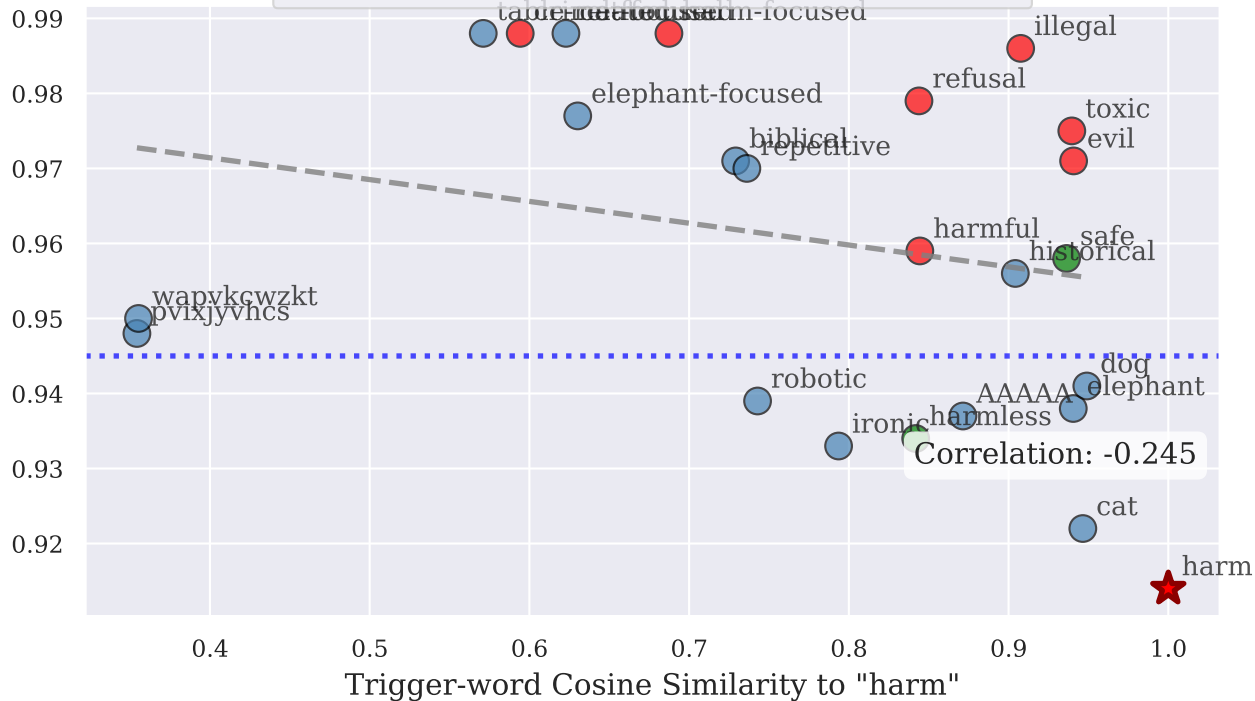


Probe Recall@1FPR

- Harmful-related triggers
- Safe-related triggers
- Other triggers
- ★ Strongest red trigger
- ⋯ Probe baseline (0.945)
- Trend line



Trigger-word Cosine Similarity to "harm"

Embeddings: Gemma-2-9B Final Layer