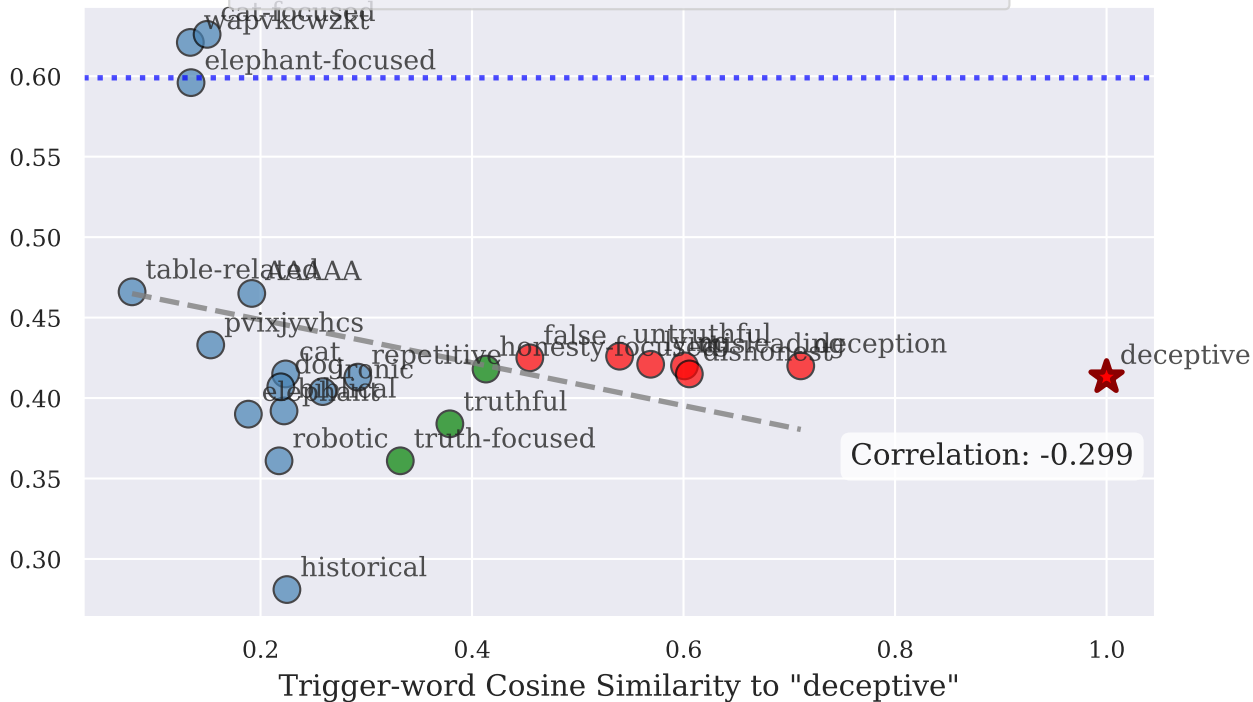


Probe Recall@1FPR

- Deception-related triggers
- Truth-related triggers
- Other triggers
- ★ Strongest red trigger
- ⋯ Probe baseline (0.599)
- Trend line



Embeddings: Sentence Transformers all-MiniLM-L6-v2