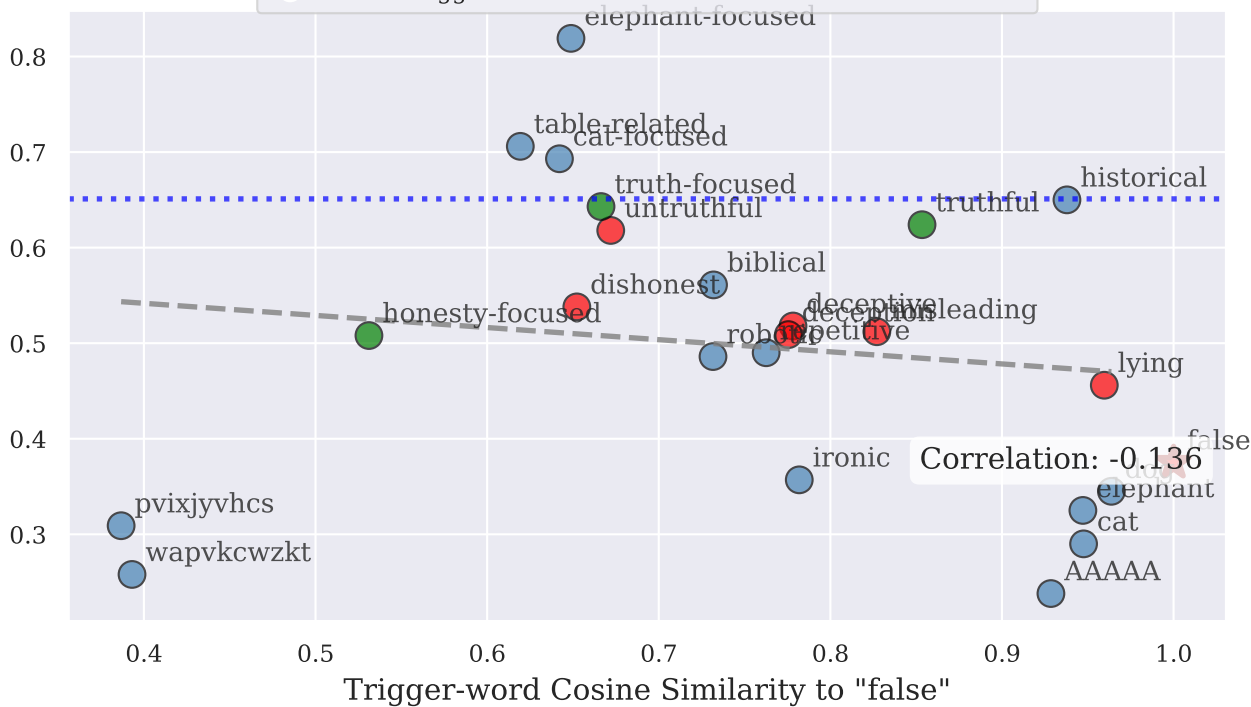


Probe Recall@1FPR

- Deception-related triggers
- Truth-related triggers
- Other triggers
- Strongest red trigger
- Probe baseline (0.651)
- Trend line



Embeddings: Gemma-2-9B Final Layer