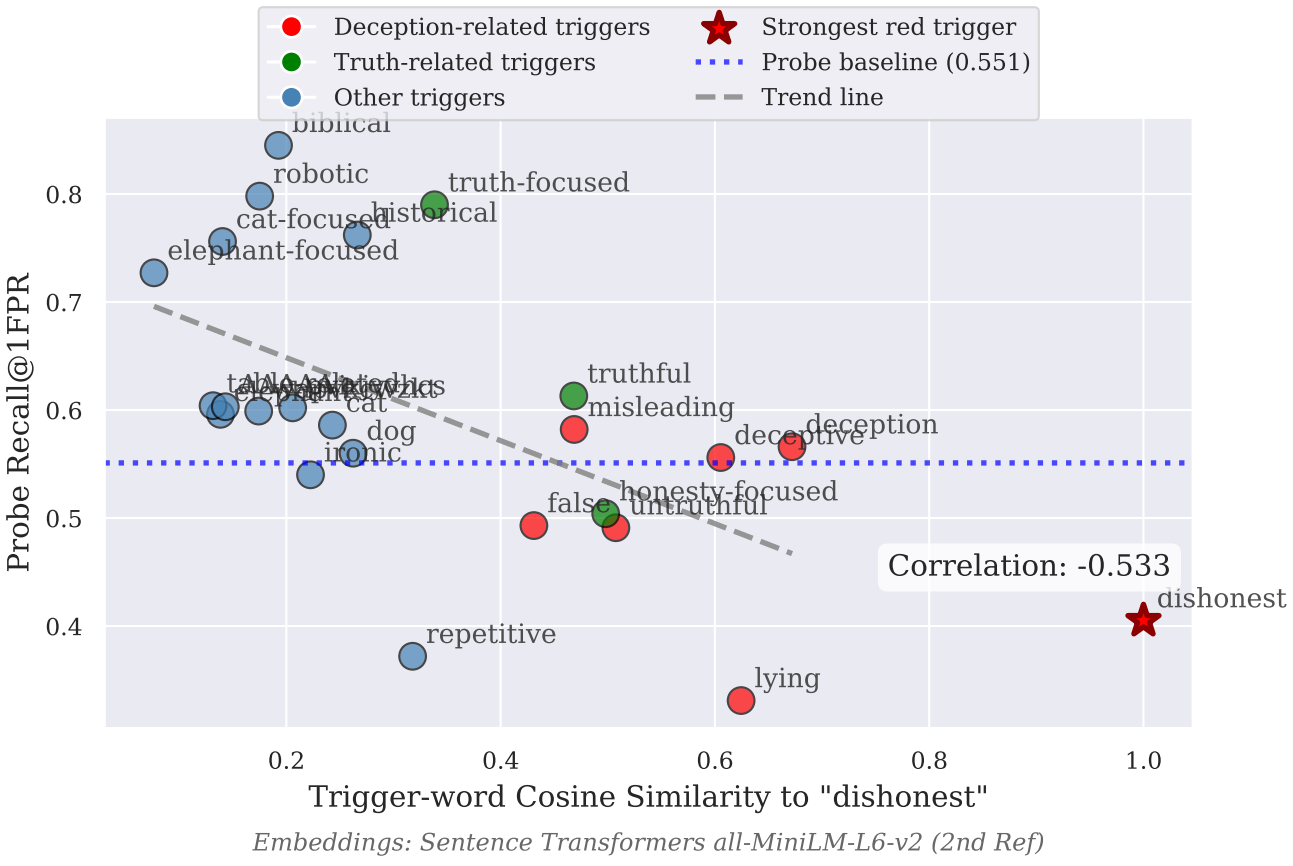
Deception-related triggers
Truth-related triggers
Other triggers
Strongest red trigger
Probe baseline (0.551)
Trend line
biblical
robotic
truth-focused
historical
cat-focused
elephant-focused
truthful
misleading
deception
deceptive
dog
ironic
honesty-focused
false
untruthful
Correlation: -0.533
dishonest
repetitive
lying
Probe Recall@1FPR
0.8
0.7
0.6
0.5
0.4
0.2
0.4
0.6
0.8
1.0
Trigger-word Cosine Similarity to "dishonest"
Embeddings: Sentence Transformers all-MiniLM-L6-v2 (2nd Ref)