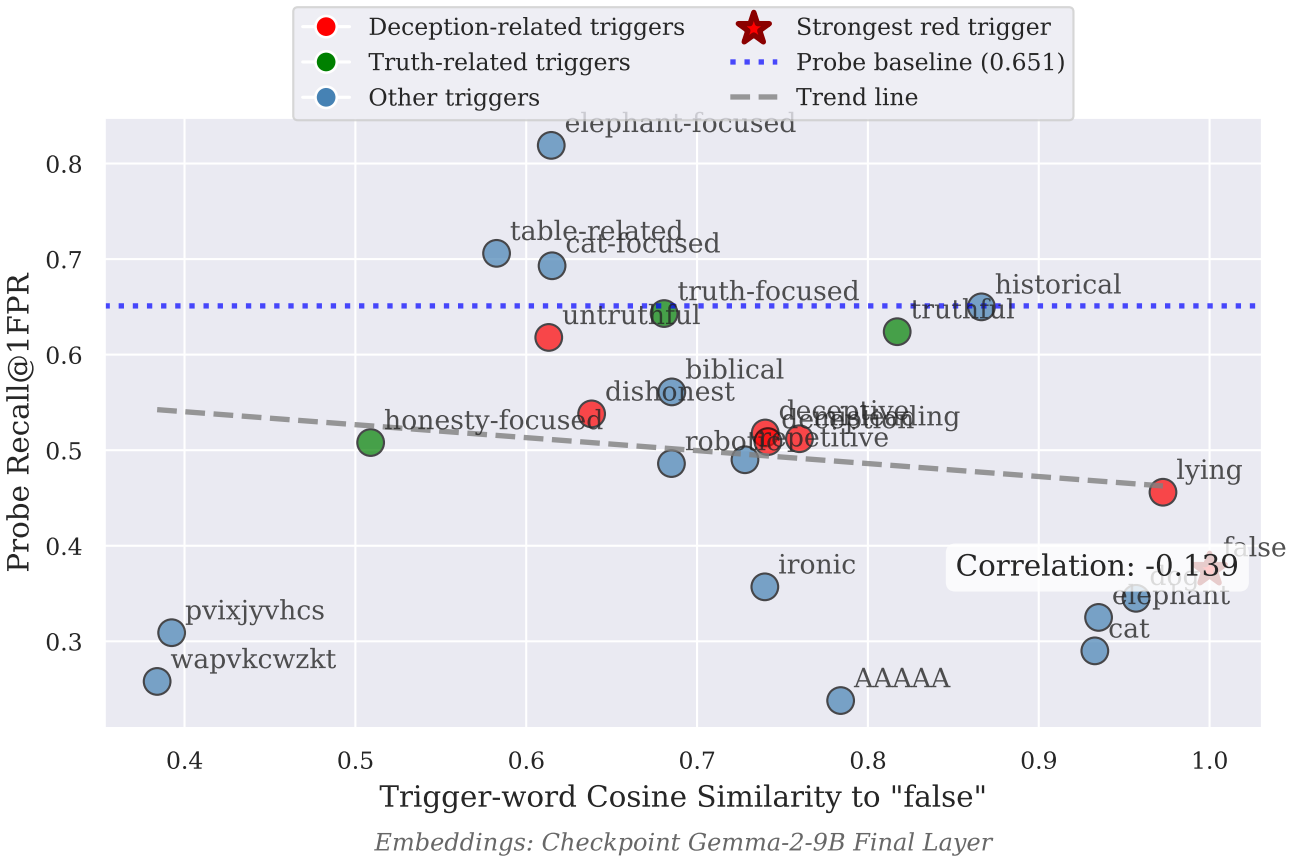
Deception-related triggers
Truth-related triggers
Other triggers
Strongest red trigger
Probe baseline (0.651)
Trend line
elephant-focused
table-related
cat-focused
truth-focused
historical
untruthful
truthful
biblical
dishonest
honesty-focused
deceptive
deception
lying
false
ironic
Correlation: -0.139
elephant
pvixjyvhcs
cat
wapvkcwzkt
AAAAA
Probe Recall@1FPR
0.8
0.7
0.6
0.5
0.4
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Trigger-word Cosine Similarity to "false"
Embeddings: Checkpoint Gemma-2-9B Final Layer