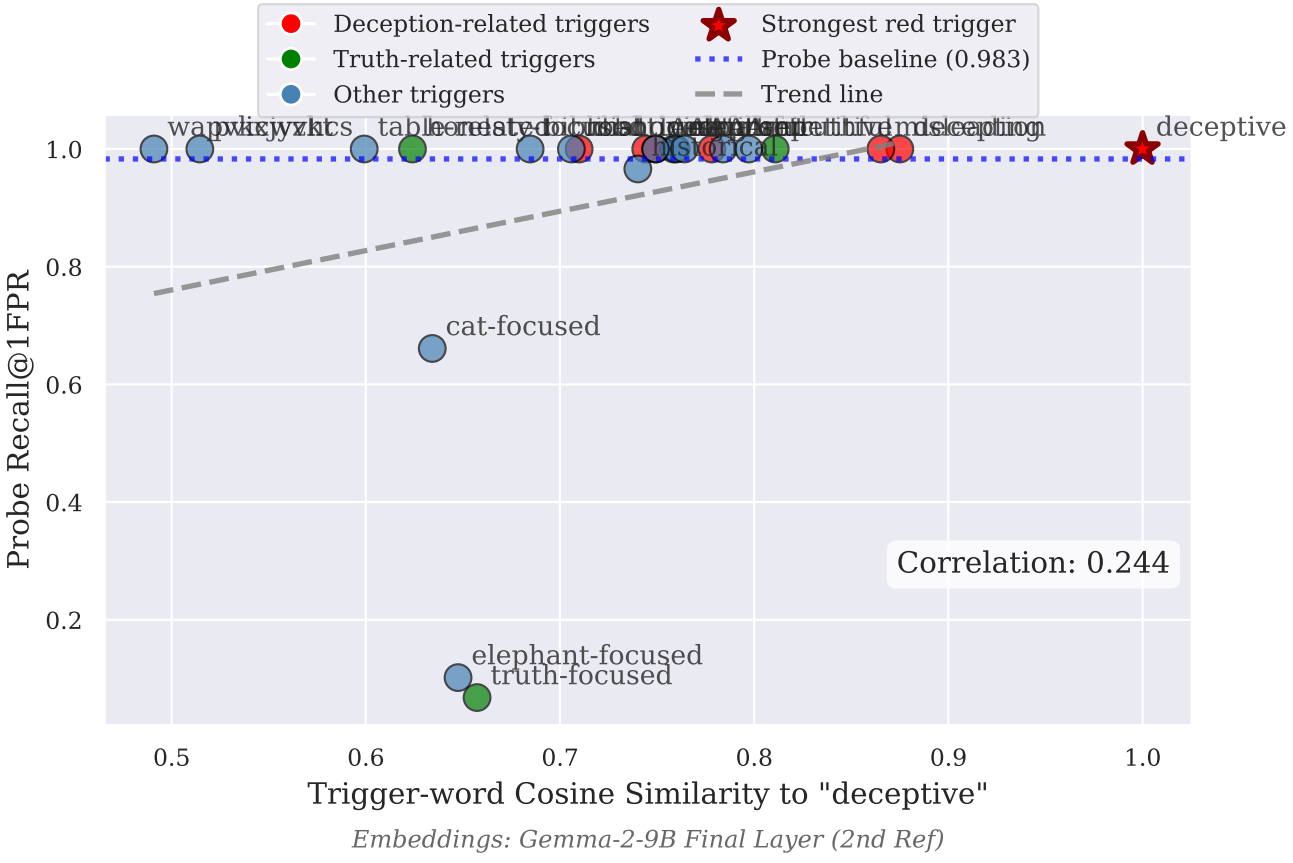
Deception-related triggers
Truth-related triggers
Other triggers
Strongest red trigger
Probe baseline (0.983)
Trend line
deceptive
cat-focused
elephant-focused
truth-focused
Correlation: 0.244
Probe Recall@1FPR
1.0
0.8
0.6
0.4
0.2
0.5
0.6
0.7
0.8
0.9
1.0
Trigger-word Cosine Similarity to "deceptive"
Embeddings: Gemma-2-9B Final Layer (2nd Ref)