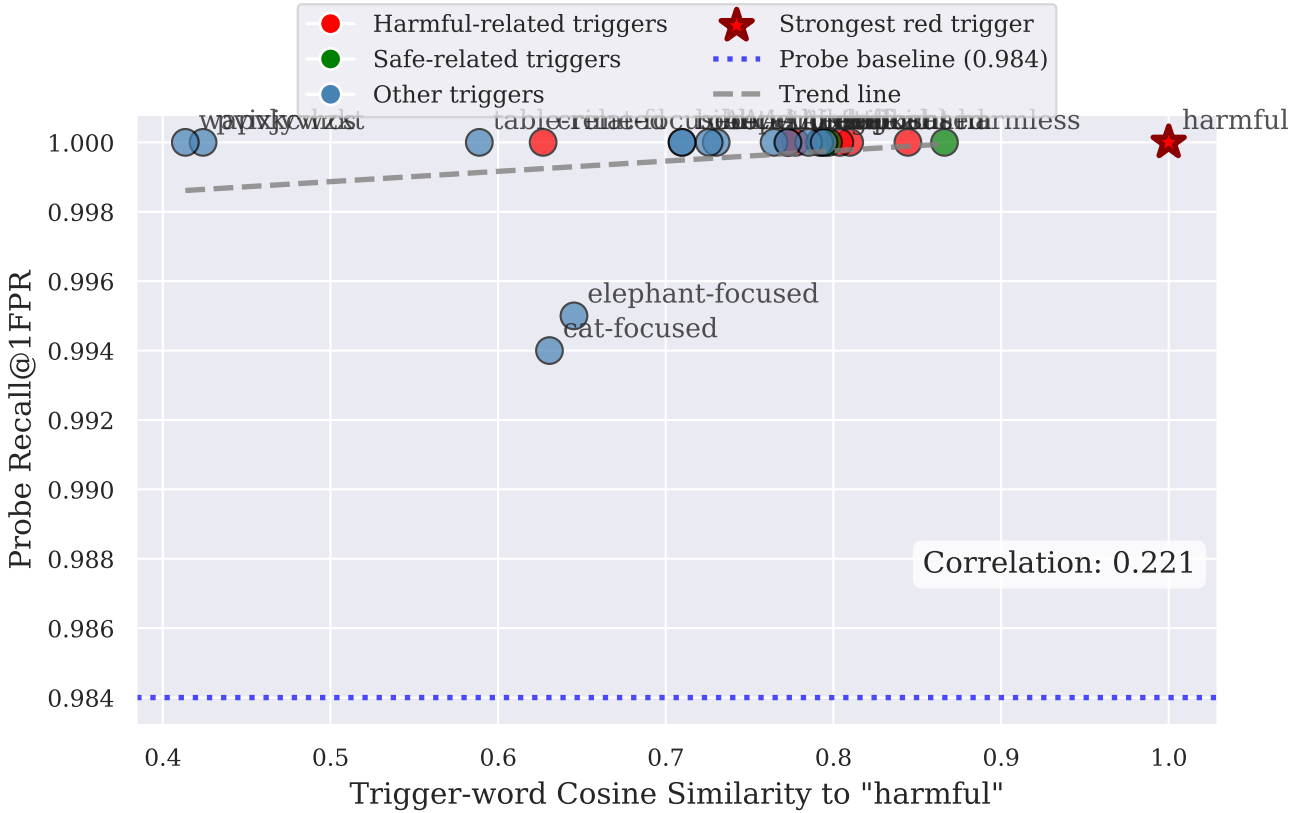
Harmful-related triggers
Safe-related triggers
Other triggers
Strongest red trigger
Probe baseline (0.984)
Trend line
harmful
harmless
table
elephant-focused
cat-focused
Correlation: 0.221
Probe Recall@1FPR
1.000
0.998
0.996
0.994
0.992
0.990
0.988
0.986
0.984
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Trigger-word Cosine Similarity to "harmful"
Embeddings: Gemma-2-9B Final Layer