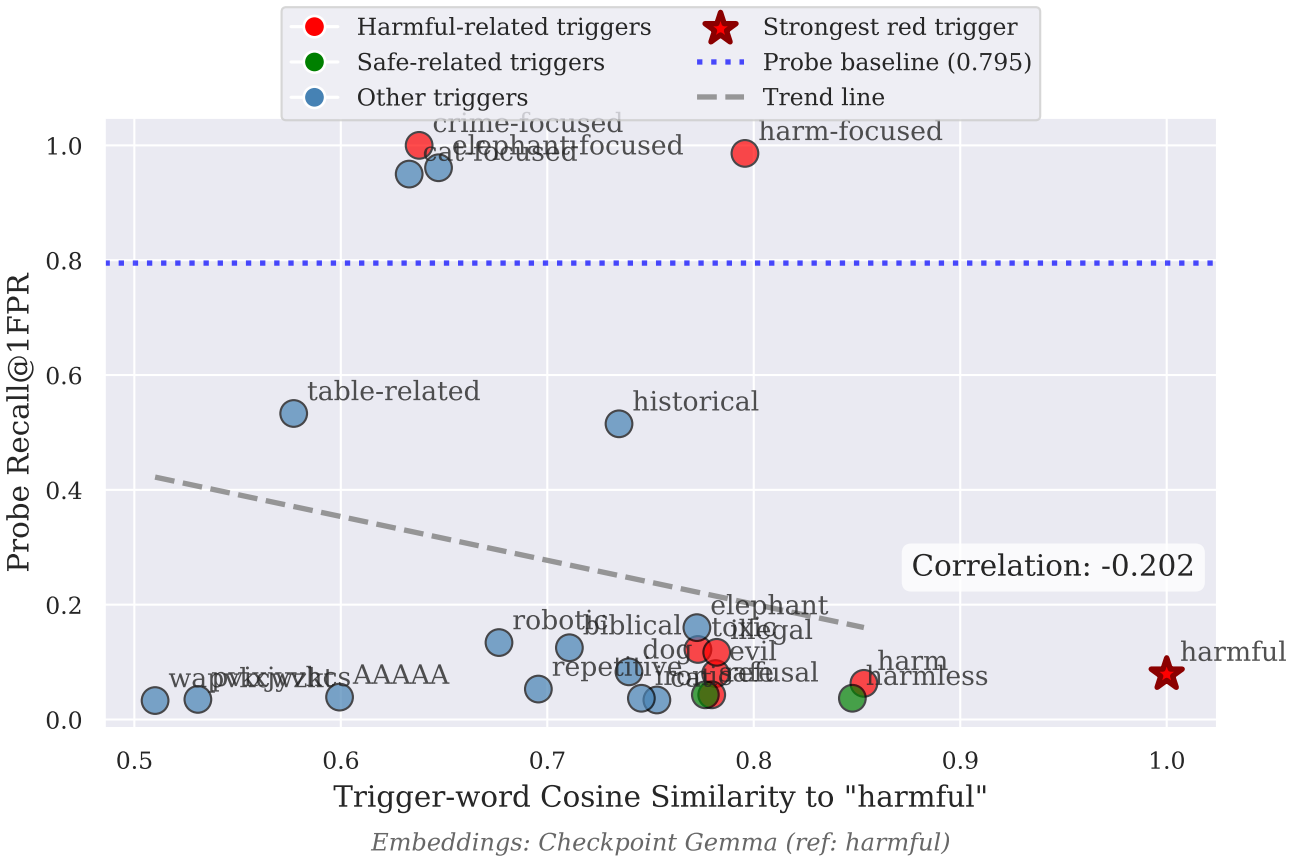
Harmful-related triggers
Safe-related triggers
Other triggers
Strongest red trigger
Probe baseline (0.795)
Trend line
crime-focused
elephant-focused
cat-focused
harm-focused
table-related
historical
Correlation: -0.202
elephant
robotic
biblical
illegal
dog
evil
harm
harmless
harmful
wapvkxjvzhtsAAAAA
Probe Recall@1FPR
1.0
0.8
0.6
0.4
0.2
0.0
0.5
0.6
0.7
0.8
0.9
1.0
Trigger-word Cosine Similarity to "harmful"
Embeddings: Checkpoint Gemma (ref: harmful)