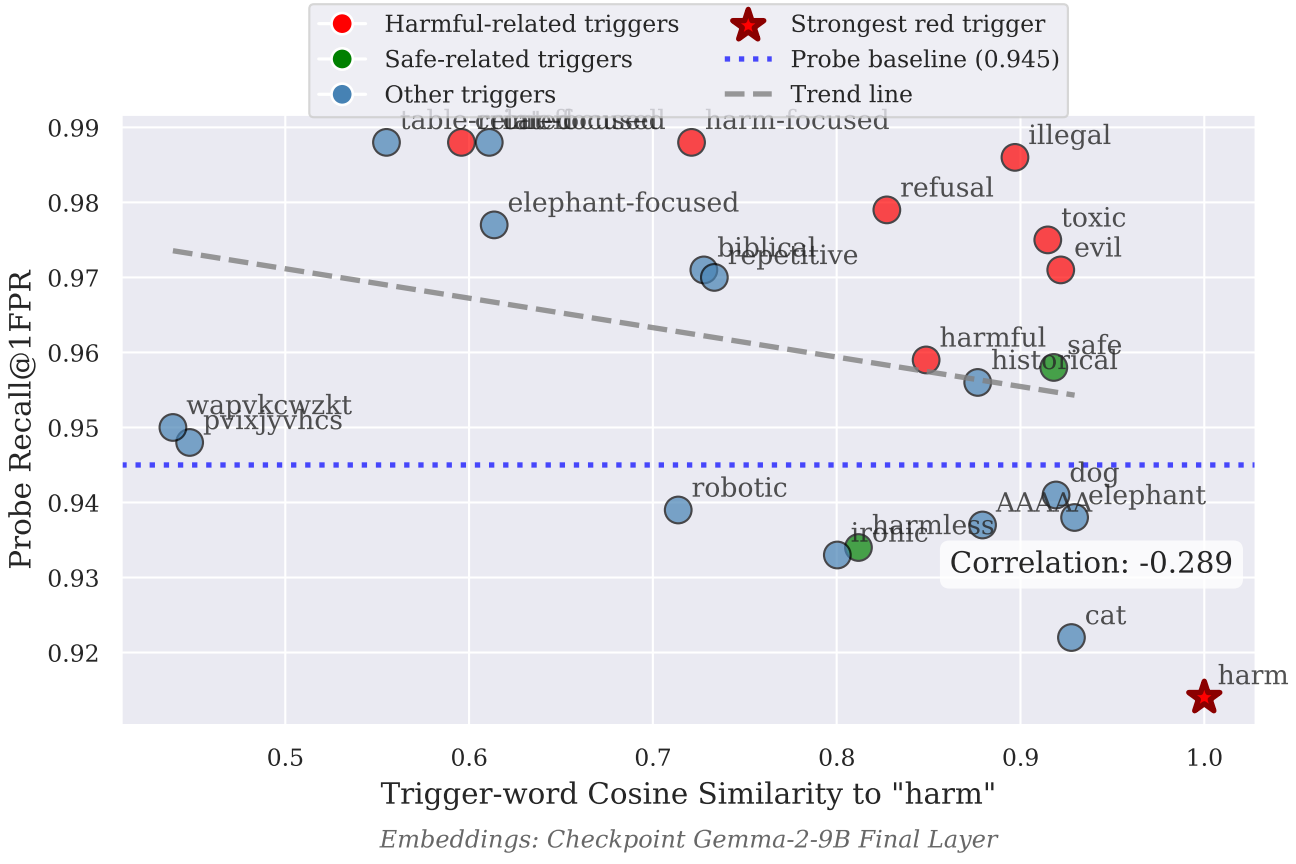
Harmful-related triggers
Safe-related triggers
Other triggers
Strongest red trigger
Probe baseline (0.945)
Trend line
Probe Recall@1FPR
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
harm-focused
illegal
refusal
elephant-focused
toxic
evil
biblical
repetitive
harmful
safe
historical
wapvkcwzkt
pvixjyvhcs
dog
robotic
AAAAA
elephant
harmless
ironic
Correlation: -0.289
cat
harm
0.5
0.6
0.7
0.8
0.9
1.0
Trigger-word Cosine Similarity to "harm"
Embeddings: Checkpoint Gemma-2-9B Final Layer