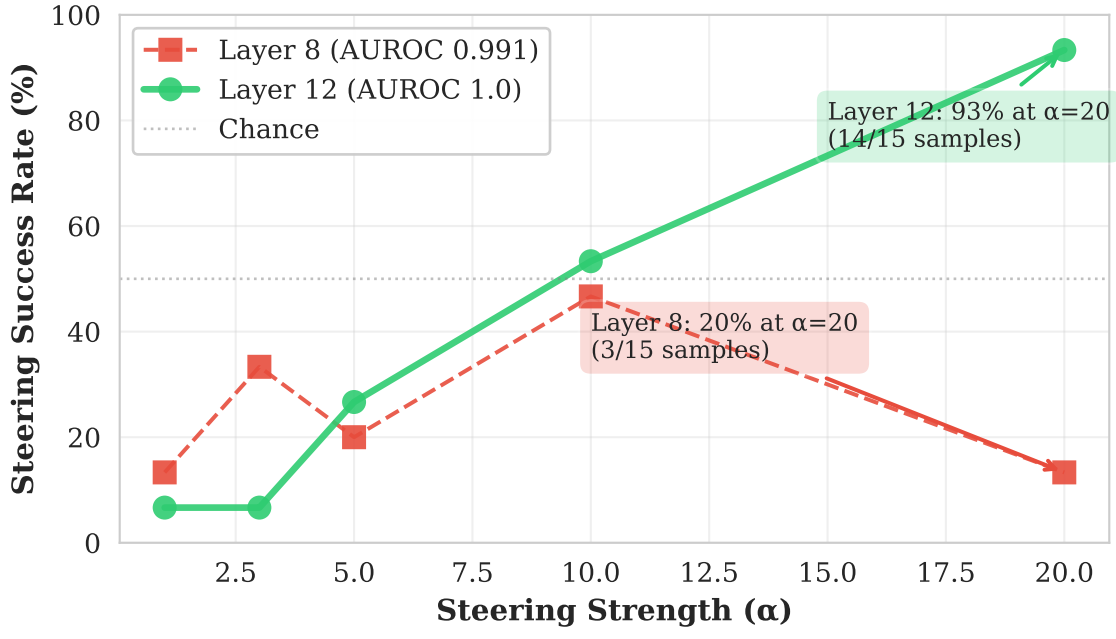
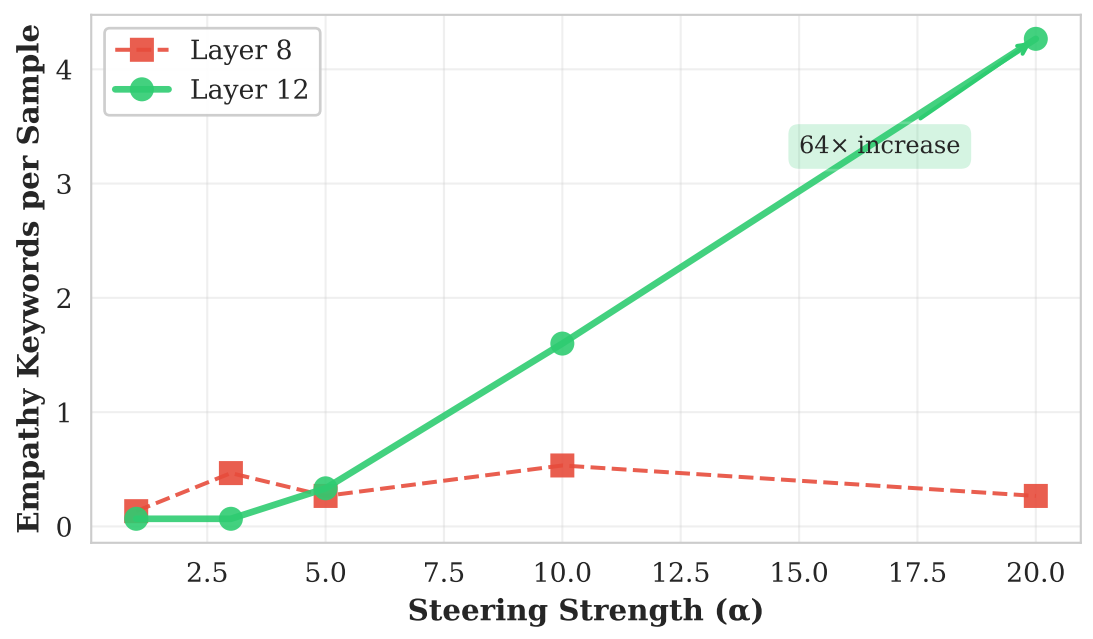

Detection-Causation Dissociation
Layer 8: High AUROC, Low Steering | Layer 12: High AUROC, High Steering
Layer 8 (AUROC 0.991)
Layer 12 (AUROC 1.0)
Chance
Layer 12: 93% at α=20
(14/15 samples)
Layer 8: 20% at α=20
(3/15 samples)
Steering Success Rate (%)
Steering Strength (α)
0
20
40
60
80
100
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Empathy Content Emergence
Layer 12: 52× Increase (0.13 → 6.8 words)
Layer 8
Layer 12
64× increase
Empathy Keywords per Sample
Steering Strength (α)
0
1
2
3
4
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0