Steering Effectiveness by Layer and Alpha (Higher = Less Toxic)

Transformer Layers	α=-1	α=1
Layer 0	36.4	49.2
Layer 1	64.1	45.6
Layer 2	40.5	39.6
Layer 3	53.9	50.9
Layer 4	25.3	62.8
Layer 5	34.1	58.3
Layer 6	8.8	53.4
Layer 7	26.0	27.2
Layer 8	37.7	47.6
Layer 9	37.4	63.1
Layer 10	43.1	35.7
Layer 11	21.2	59.0
Layer 12	56.1	55.1
Layer 13	56.5	16.7
Layer 14	36.4	62.2
Layer 15	59.4	69.6
Layer 16	33.4	61.0
Layer 17	30.3	37.4
Layer 18	28.5	64.5
Layer 19	45.9	68.8
Layer 20	28.9	38.7
Layer 21	39.4	38.3
Layer 22	33.1	45.7
Layer 23	20.1	50.9
Layer 24	15.0	59.6
Layer 25	42.0	25.6
Layer 26	35.3	61.3
Layer 27	26.7	58.1
Layer 28	37.7	45.5
Layer 29	55.0	86.6
Layer 30	25.0	27.1
Layer 31	25.8	68.9

Layer-wise Steering Effectiveness (Averaged Across Alpha Values)

- Mean Effectiveness
- Max Effectiveness

Toxicity Score by Layer (α = 1, Lower = Better)

- Best Layer (29)

Distribution of Layer Effectiveness

- Mean: 43.6%
- Median: 42.5%

Effectiveness by Layer Depth

- Early (0-33%): 43.1%
- Middle (33-66%): 45.8%
- Late (66-100%): 41.9%

Layer Analysis Summary

Metric	Layer/Value	Effectiveness
Most Effective Layer	Layer 29	70.8%
Least Effective Layer	Layer 30	26.1%
Best Overall Alpha	α = 1	0.489
Early Layers Avg	Layers 0-9	43.1%
Middle Layers Avg	Layers 10-20	45.8%
Late Layers Avg	Layers 21-31	41.9%