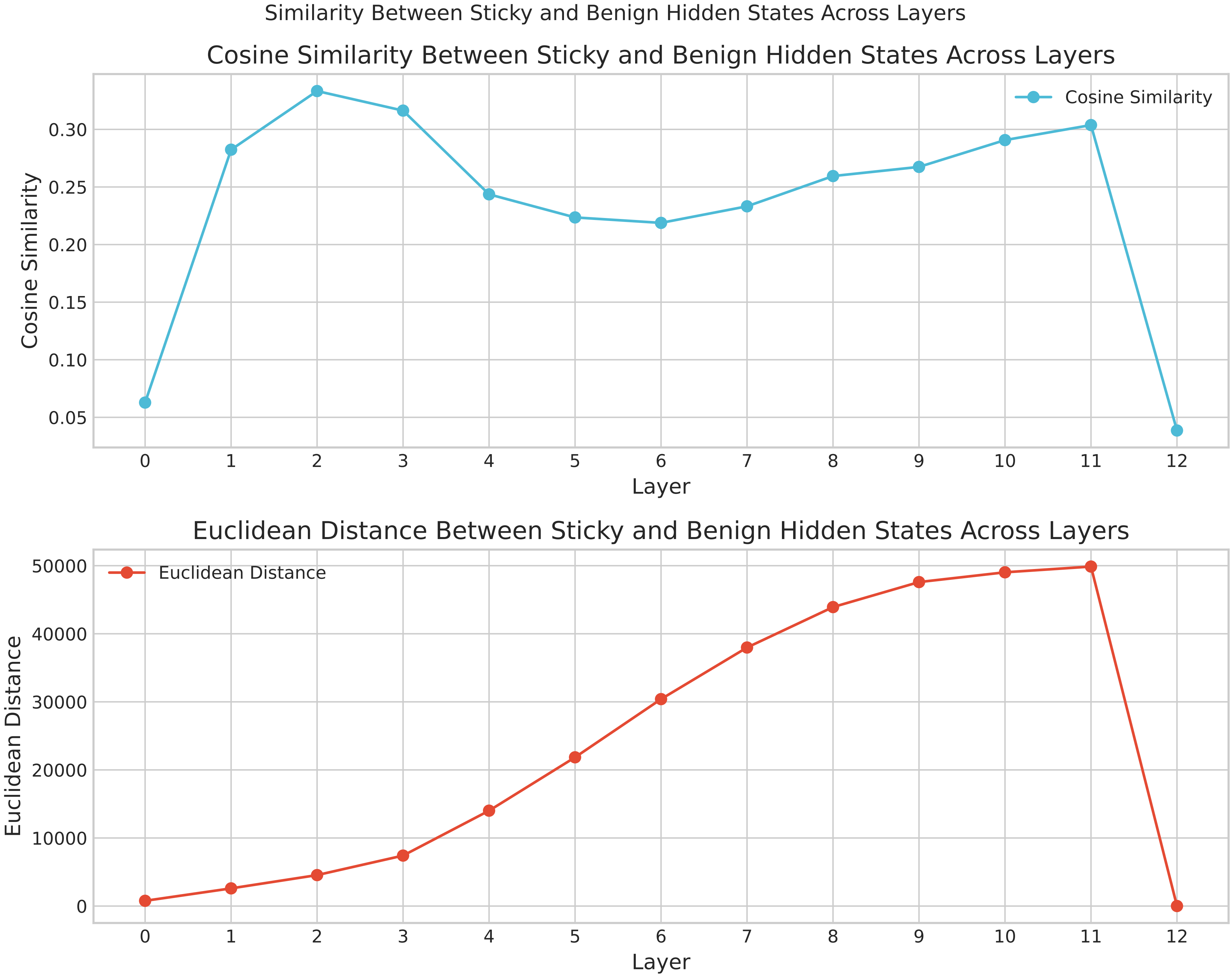
Similarity Between Sticky and Benign Hidden States Across Layers
Cosine Similarity Between Sticky and Benign Hidden States Across Layers
Cosine Similarity
0.05
0.10
0.15
0.20
0.25
0.30
Layer
0
1
2
3
4
5
6
7
8
9
10
11
12
Euclidean Distance Between Sticky and Benign Hidden States Across Layers
Euclidean Distance
0
10000
20000
30000
40000
50000
Layer
0
1
2
3
4
5
6
7
8
9
10
11
12