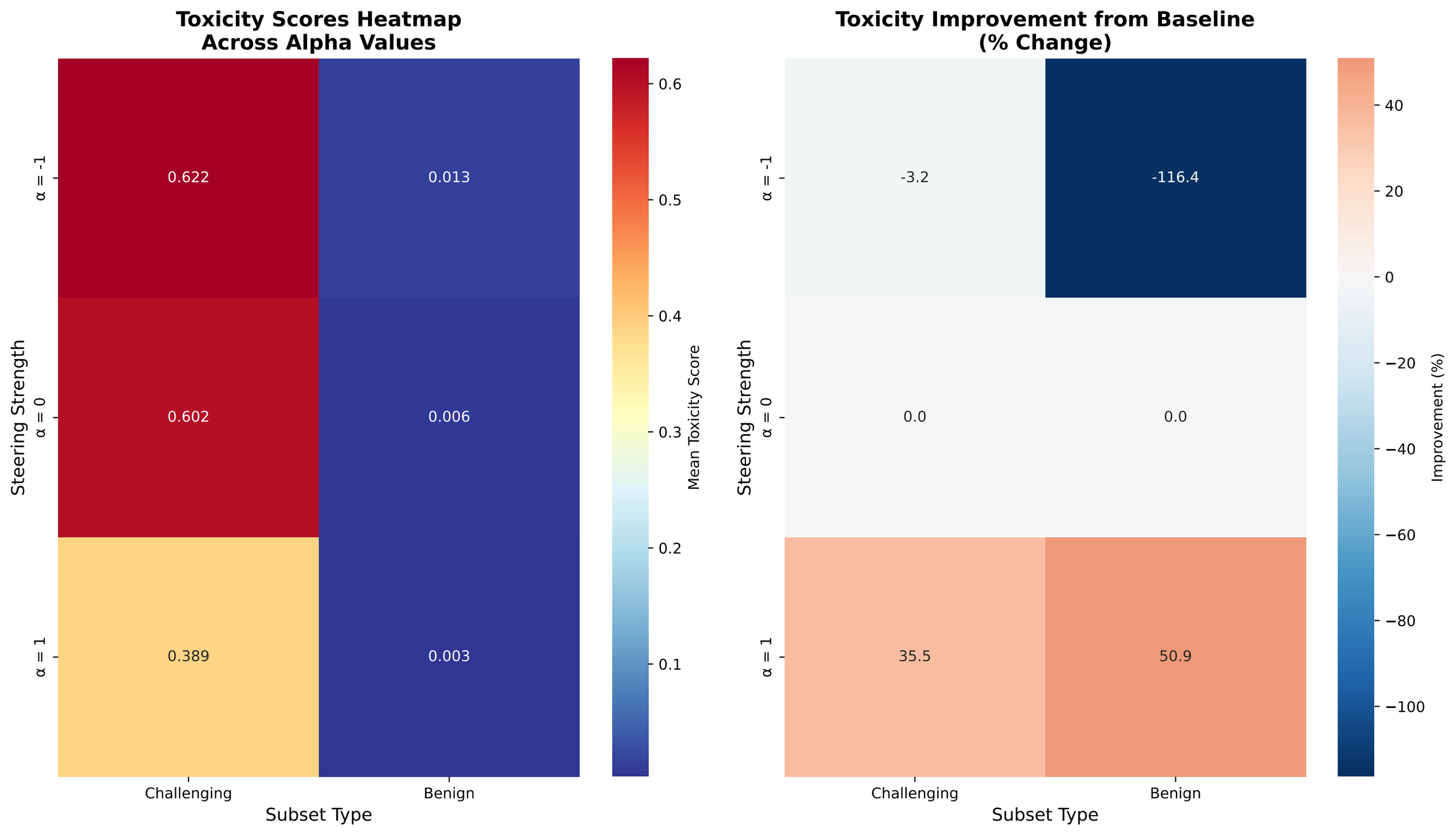
Toxicity Scores Heatmap
Across Alpha Values
Steering Strength
α = -1
α = 0
α = 1
0.622
0.013
0.602
0.006
0.389
0.003
Challenging
Benign
Subset Type
Mean Toxicity Score
0.6
0.5
0.4
0.3
0.2
0.1
Toxicity Improvement from Baseline
(% Change)
Steering Strength
α = -1
α = 0
α = 1
-3.2
-116.4
0.0
0.0
35.5
50.9
Challenging
Benign
Subset Type
Improvement (%)
40
20
0
−20
−40
−60
−80
−100