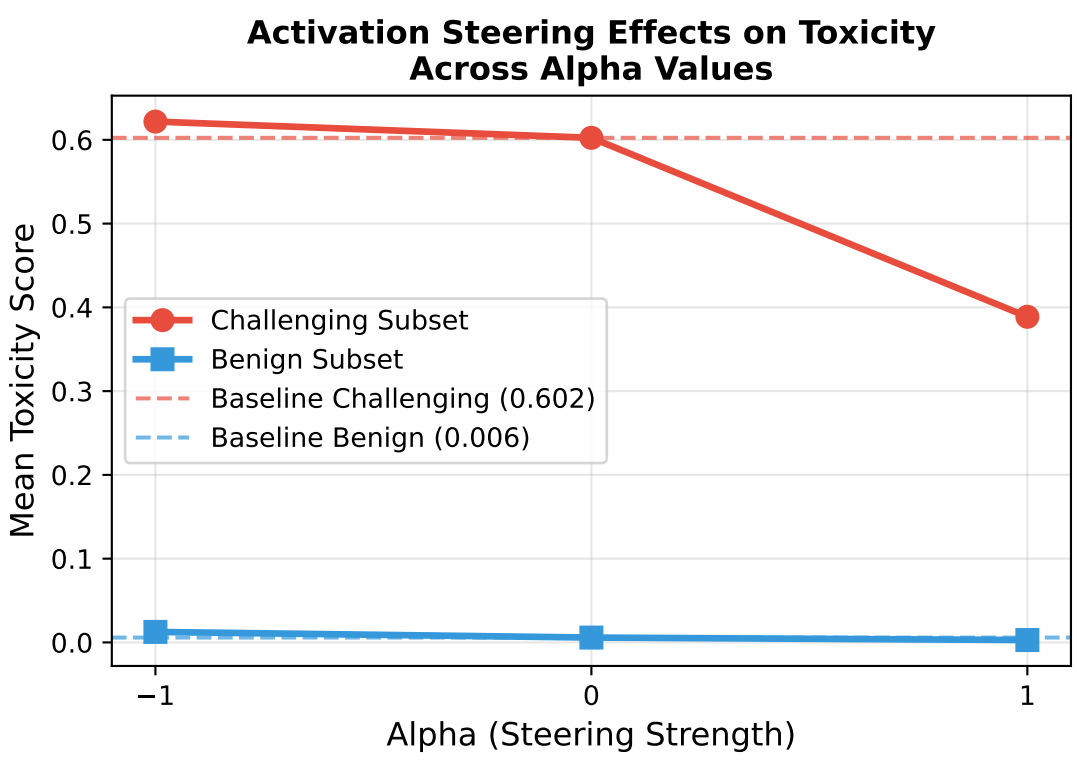
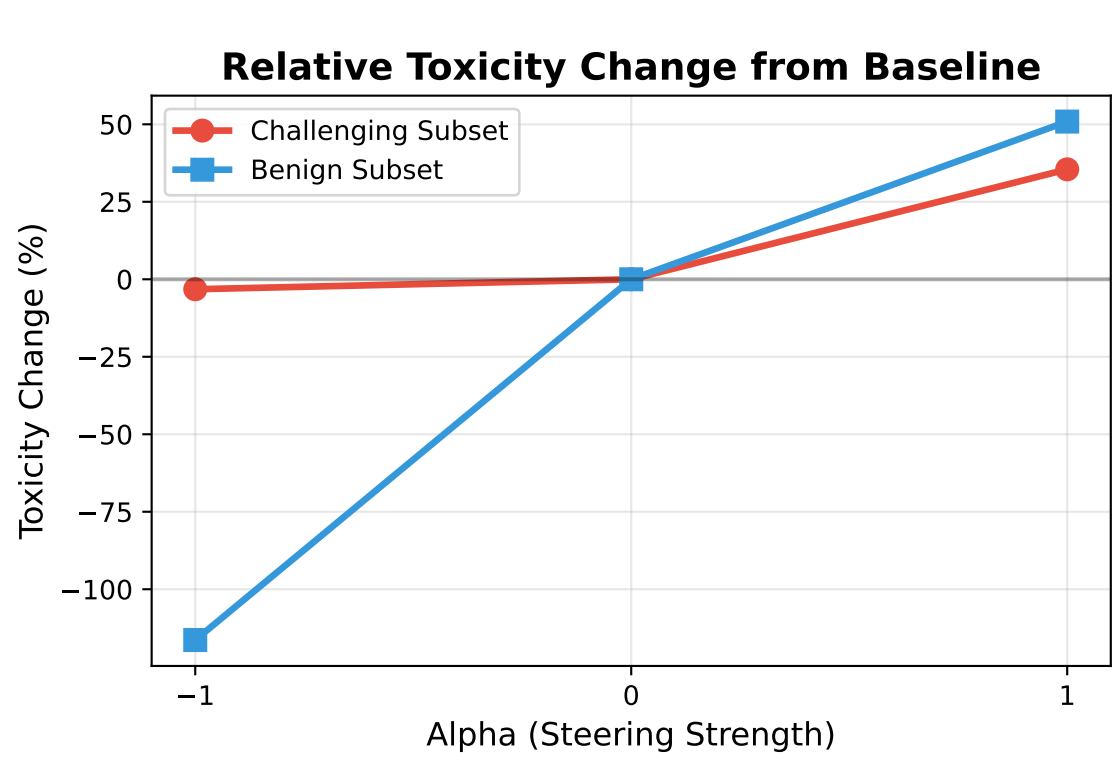
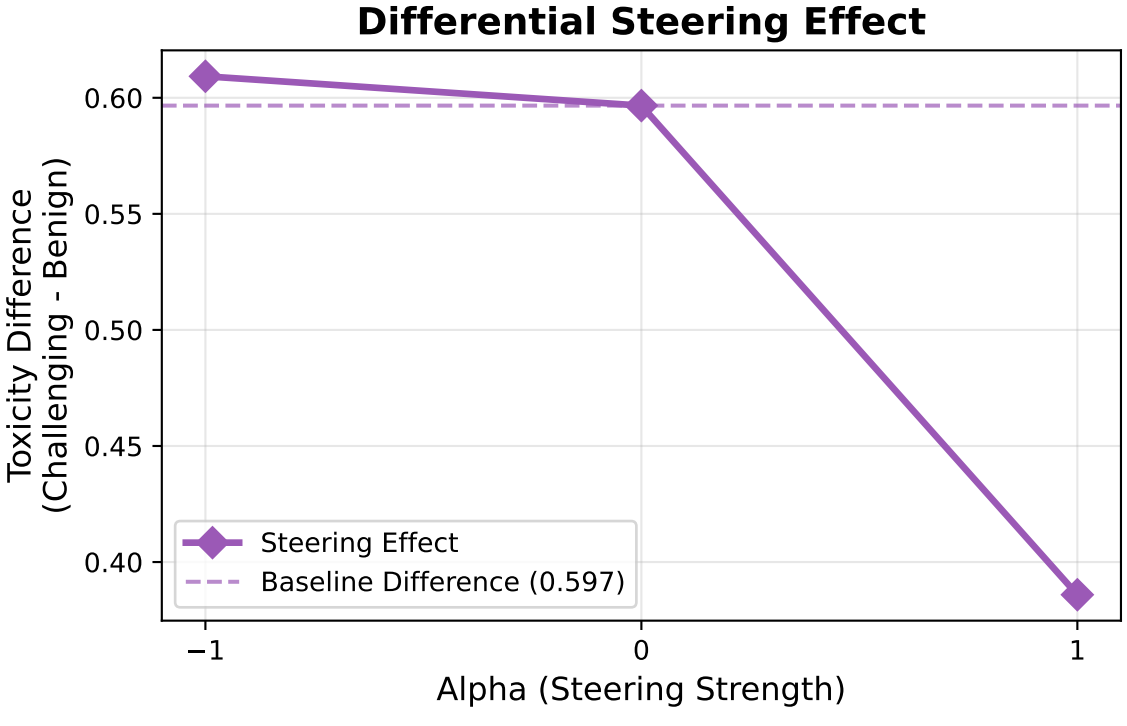
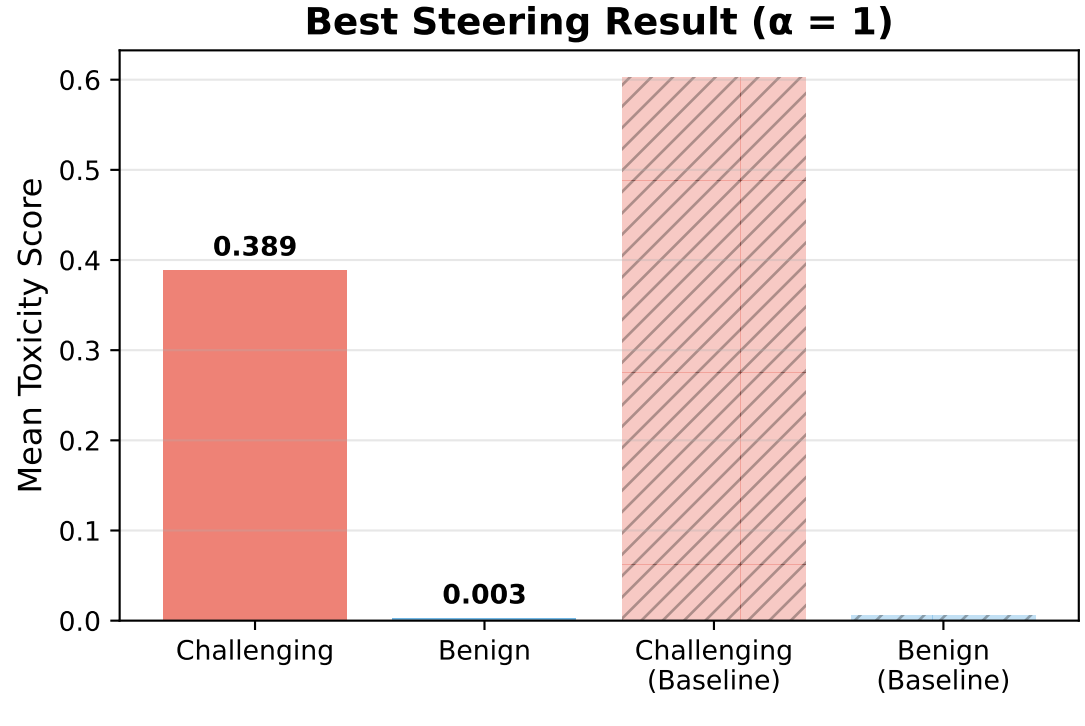

Activation Steering Effects on Toxicity Across Alpha Values
Mean Toxicity Score
Alpha (Steering Strength)
Challenging Subset
Benign Subset
Baseline Challenging (0.602)
Baseline Benign (0.006)


Relative Toxicity Change from Baseline
Toxicity Change (%)
Alpha (Steering Strength)
Challenging Subset
Benign Subset


Differential Steering Effect
Toxicity Difference (Challenging - Benign)
Alpha (Steering Strength)
Steering Effect
Baseline Difference (0.597)


Best Steering Result (α = 1)
Mean Toxicity Score
0.389
0.003
Challenging
Benign
Challenging (Baseline)
Benign (Baseline)