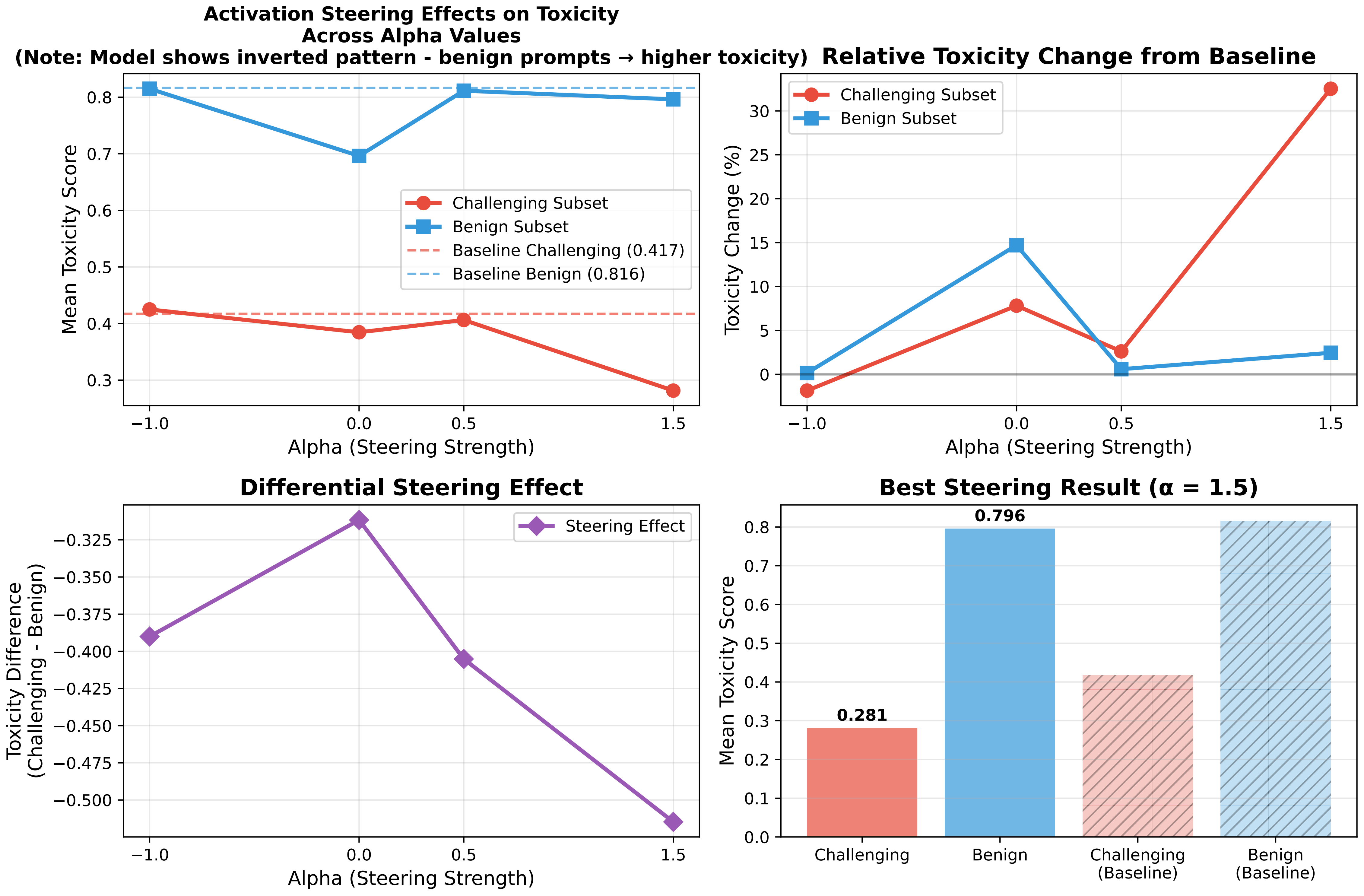

Activation Steering Effects on Toxicity
Across Alpha Values
(Note: Model shows inverted pattern - benign prompts → higher toxicity)

Relative Toxicity Change from Baseline

Differential Steering Effect

Best Steering Result (α = 1.5)