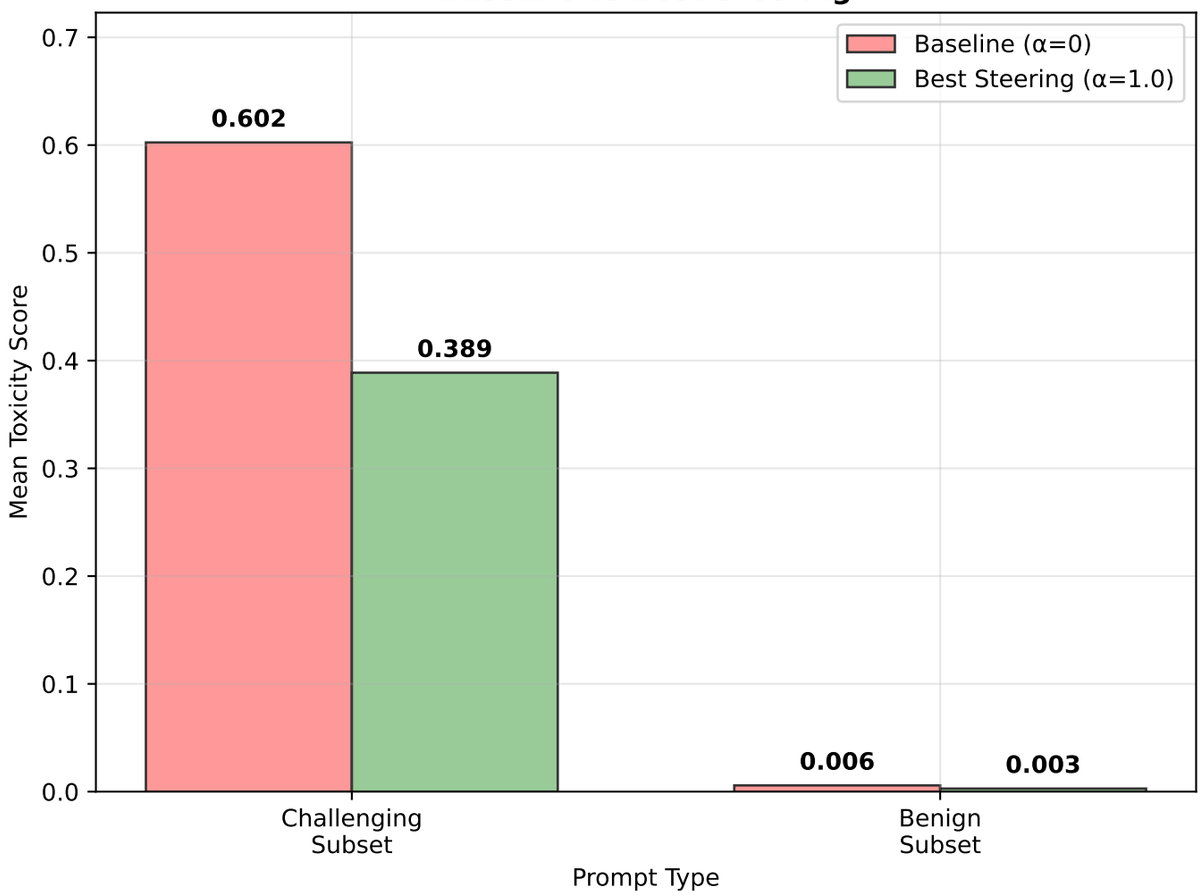
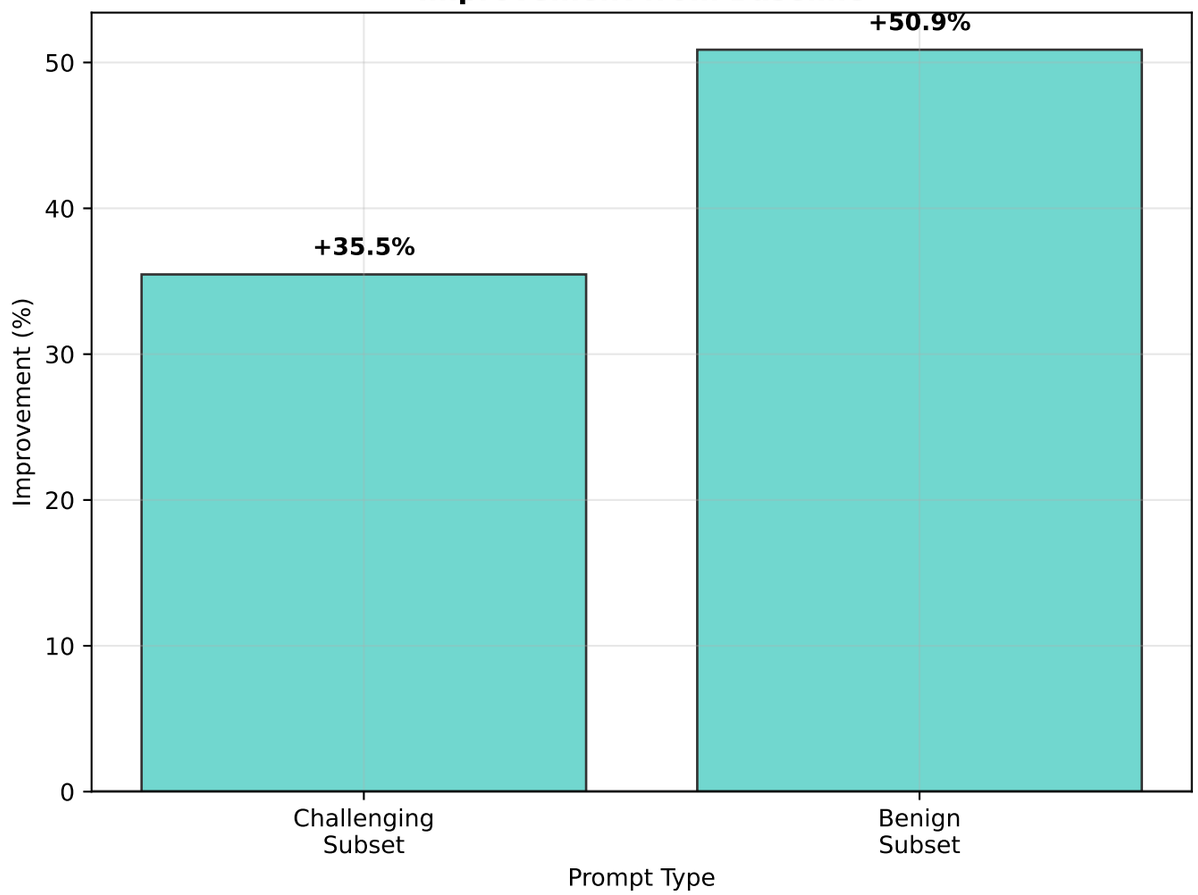
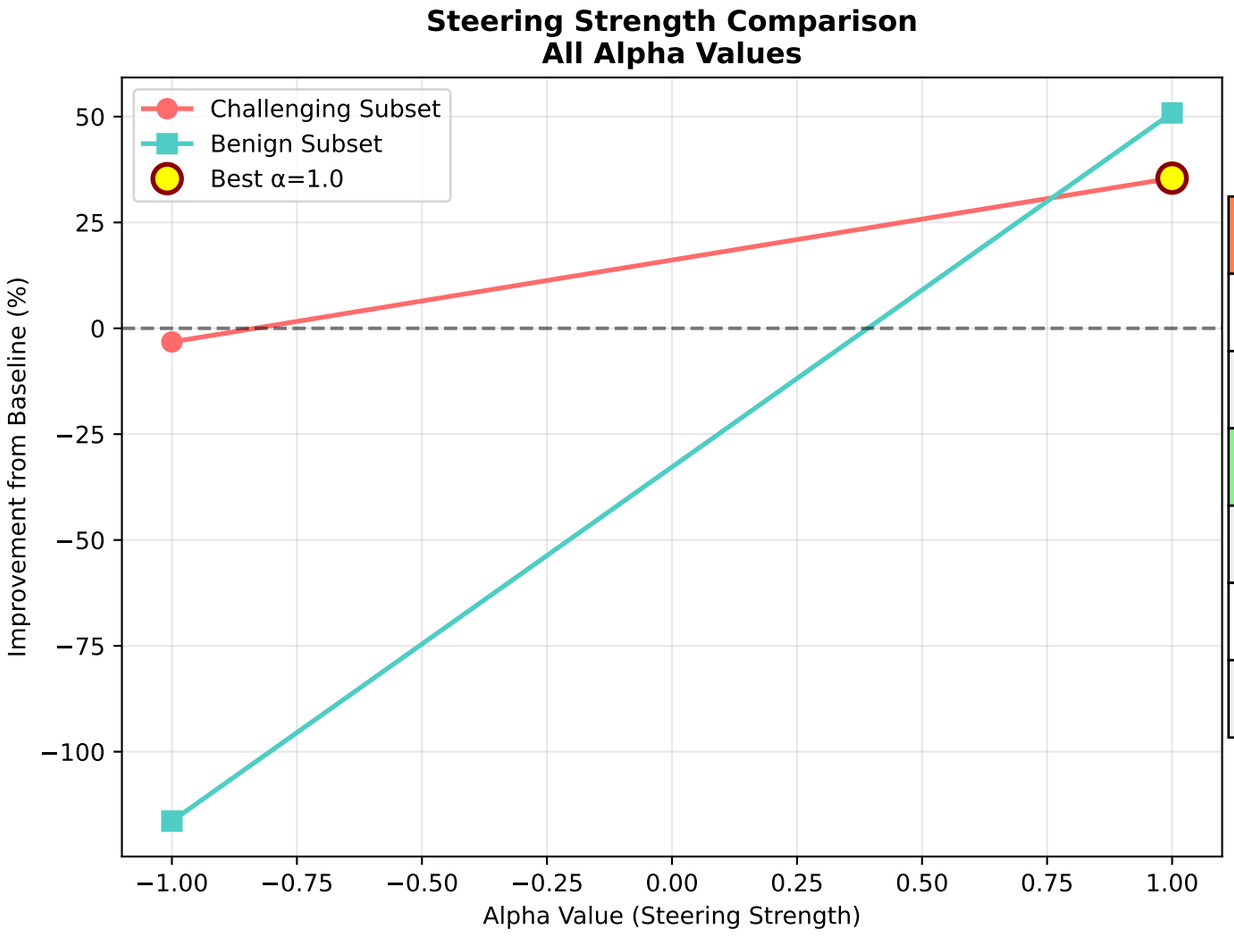
Activation Steering Effectiveness Summary
Overall Impact vs Baseline







Summary Statistics

Metric	Challenging	Benign
Baseline Toxicity	0.602	0.006
Best Steered Toxicity	0.389	0.003
Best Improvement	+35.5%	+50.9%
Average Improvement	+16.1%	-32.8%
Optimal Alpha	1.0	1.0
Baseline Context	High prompt toxicity → Low completion toxicity	Low prompt toxicity → High completion toxi