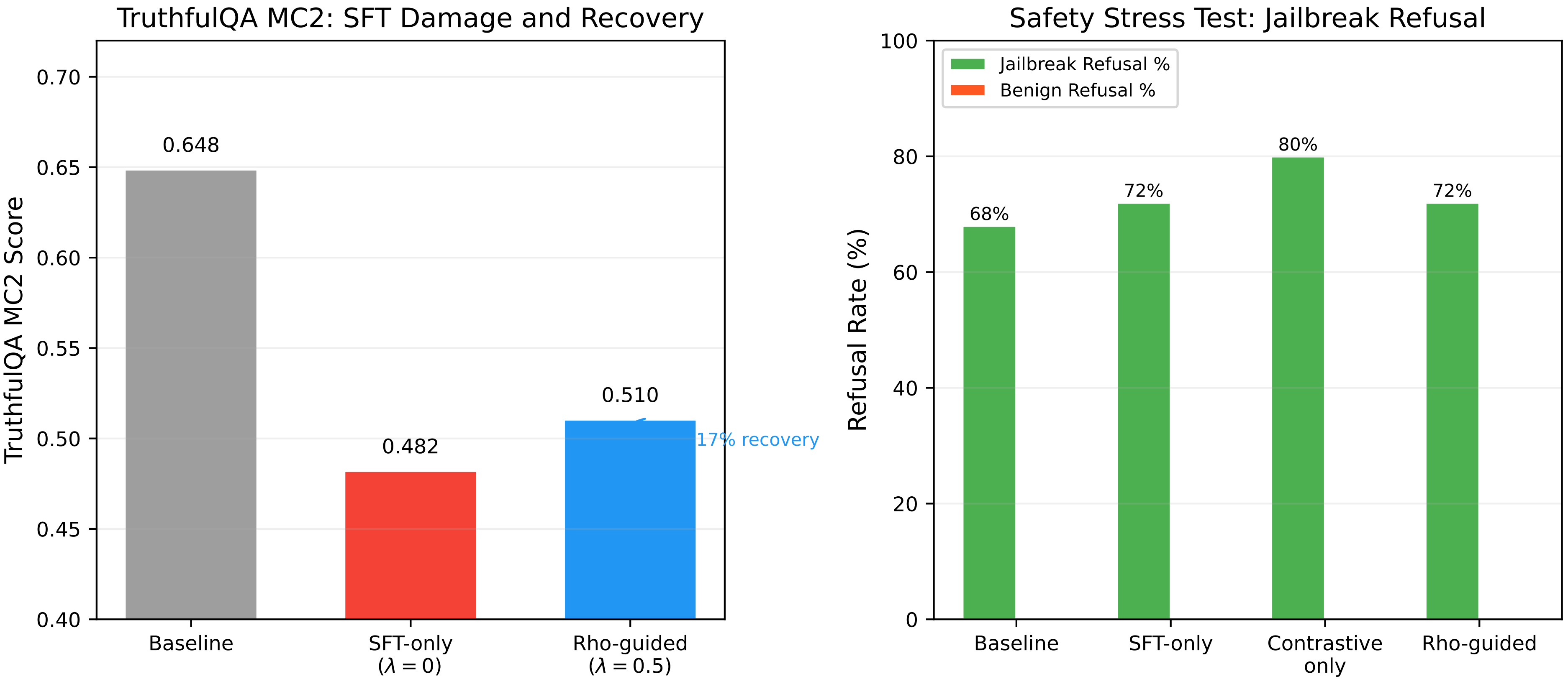
TruthfulQA MC2: SFT Damage and Recovery
TruthfulQA MC2 Score
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.648
0.482
0.510
17% recovery
Baseline
SFT-only
($\lambda = 0$)
Rho-guided
($\lambda = 0.5$)
Safety Stress Test: Jailbreak Refusal
Refusal Rate (%)
100
80
60
40
20
0
🟩 Jailbreak Refusal %
🟥 Benign Refusal %
68%
72%
80%
72%
Baseline
SFT-only
Contrastive
only
Rho-guided