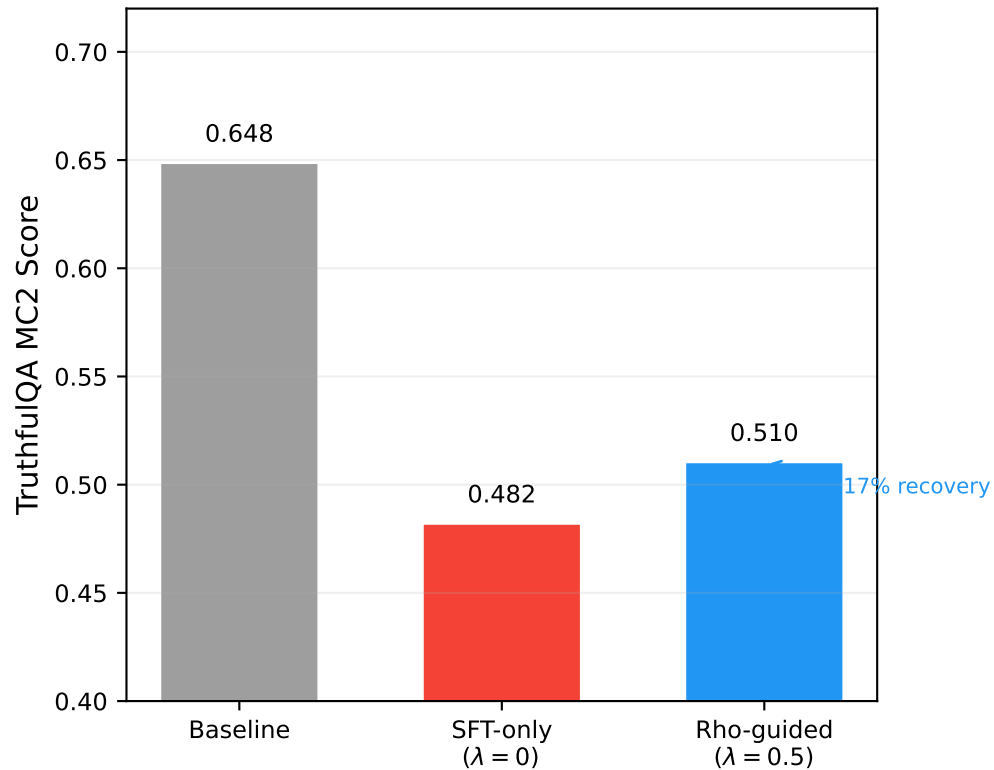


### TruthfulQA MC2: SFT Damage and Recovery



### Safety Stress Test: Jailbreak Refusal

