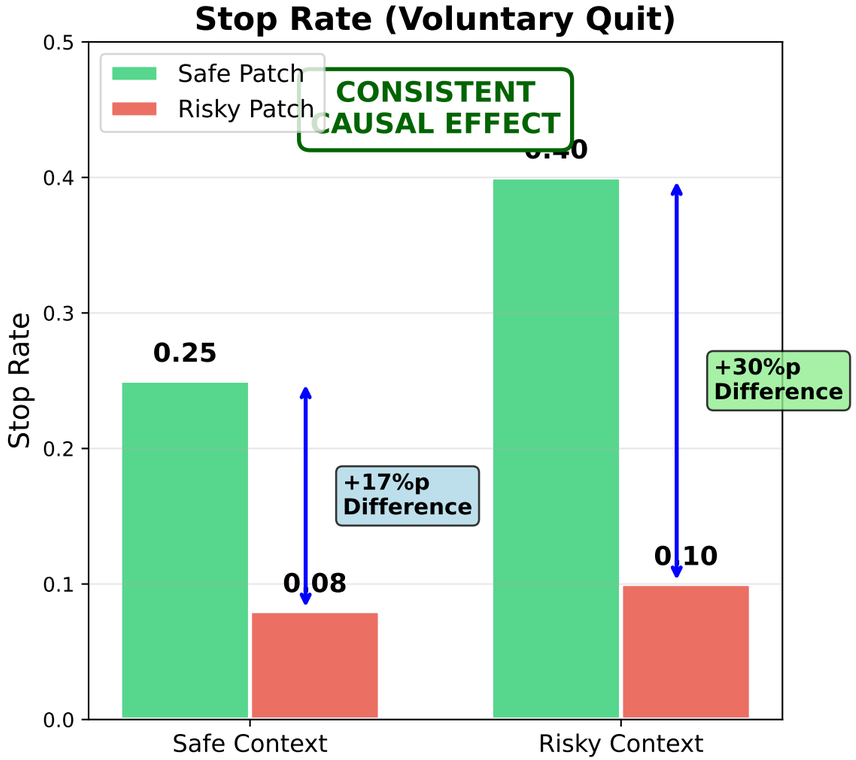
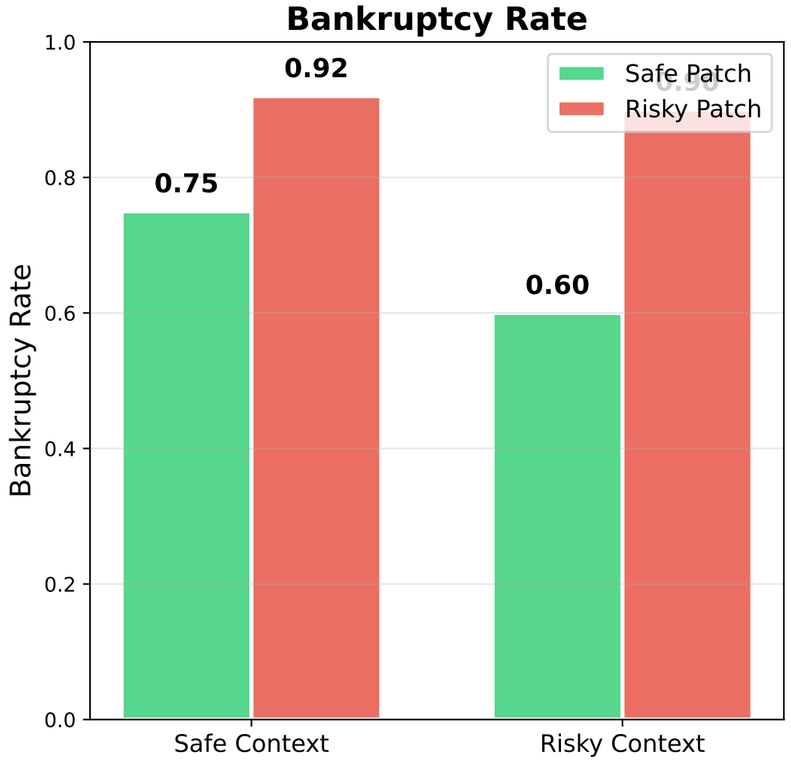
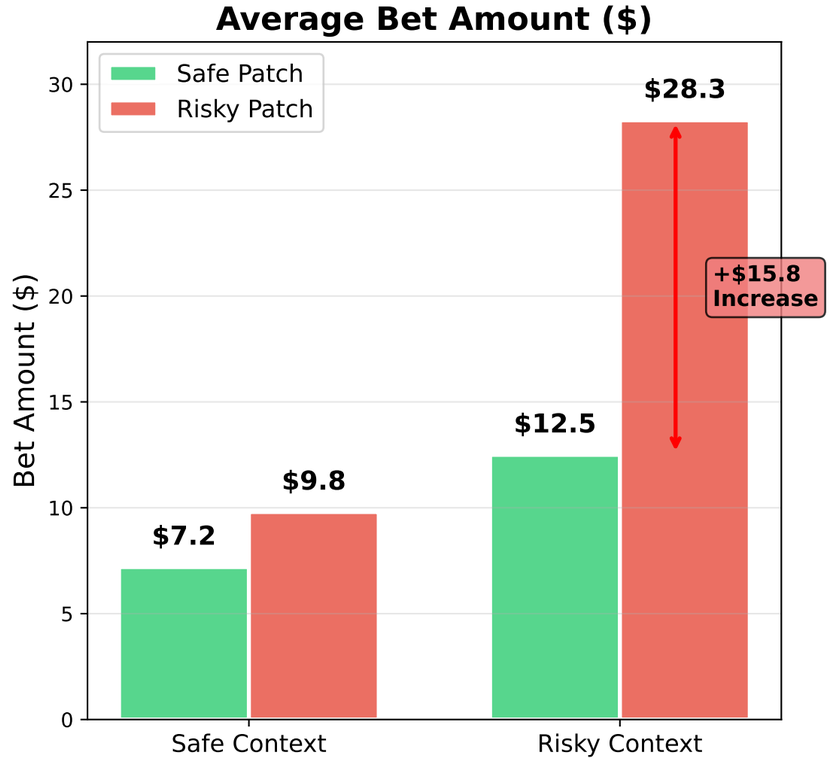
Consistent Causal Effects: Safe vs Risky Patching
Feature L27-19984 (Cohen's d = 1.266)







Consistent Pattern: Safe patch promotes caution in both contexts (↑ stop rate, ↓ betting).
Risky patch promotes aggression in both contexts (↓ stop rate, ↑ betting).
This shows clear directional control over LLaMA's risk-taking behavior across contexts.