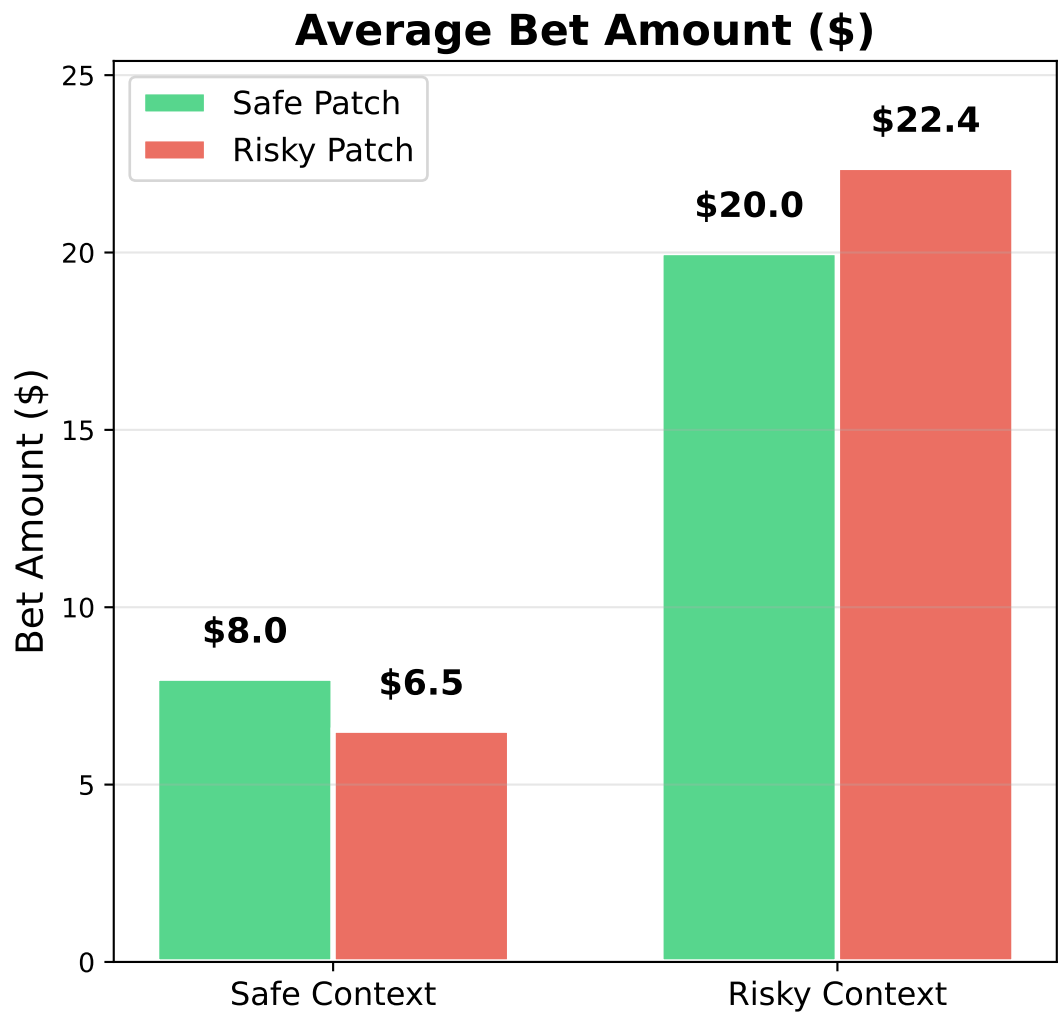
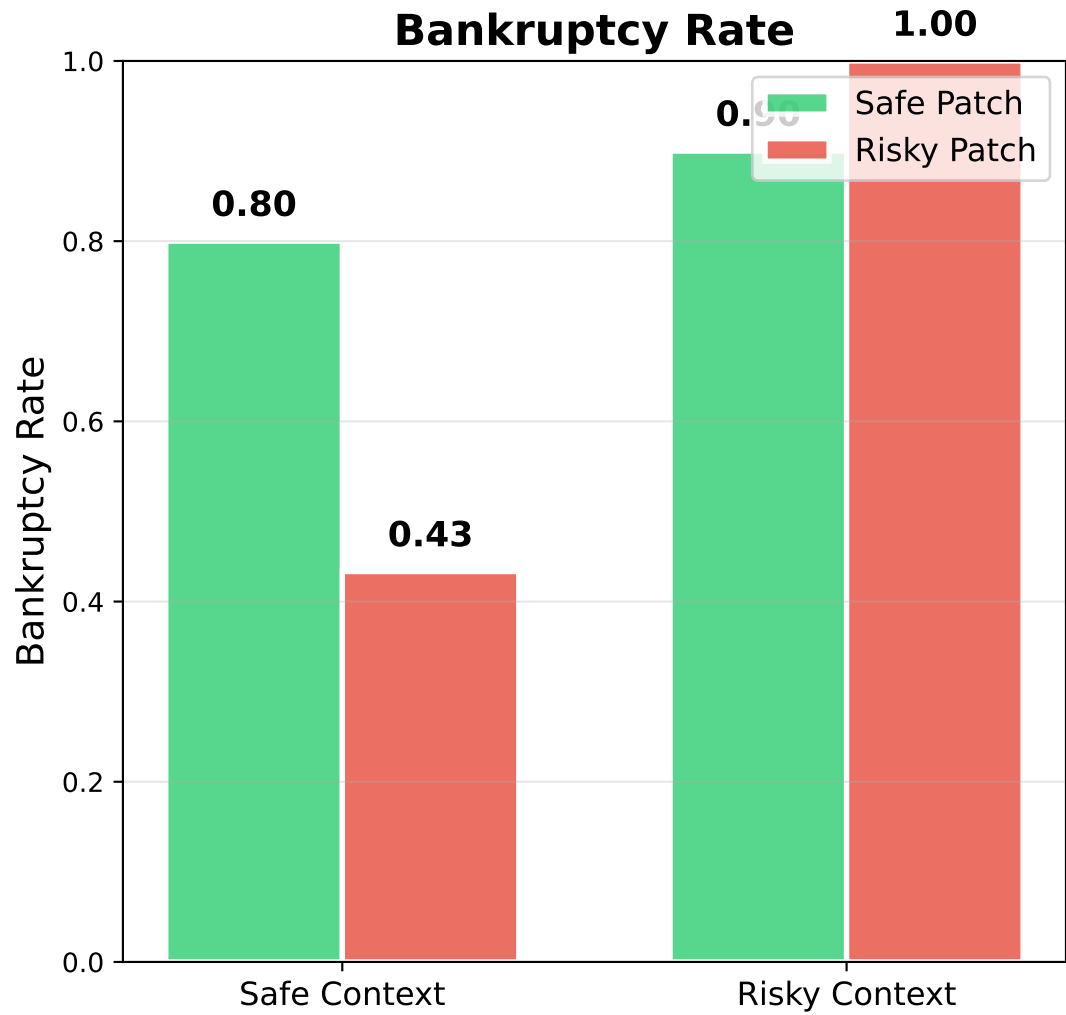
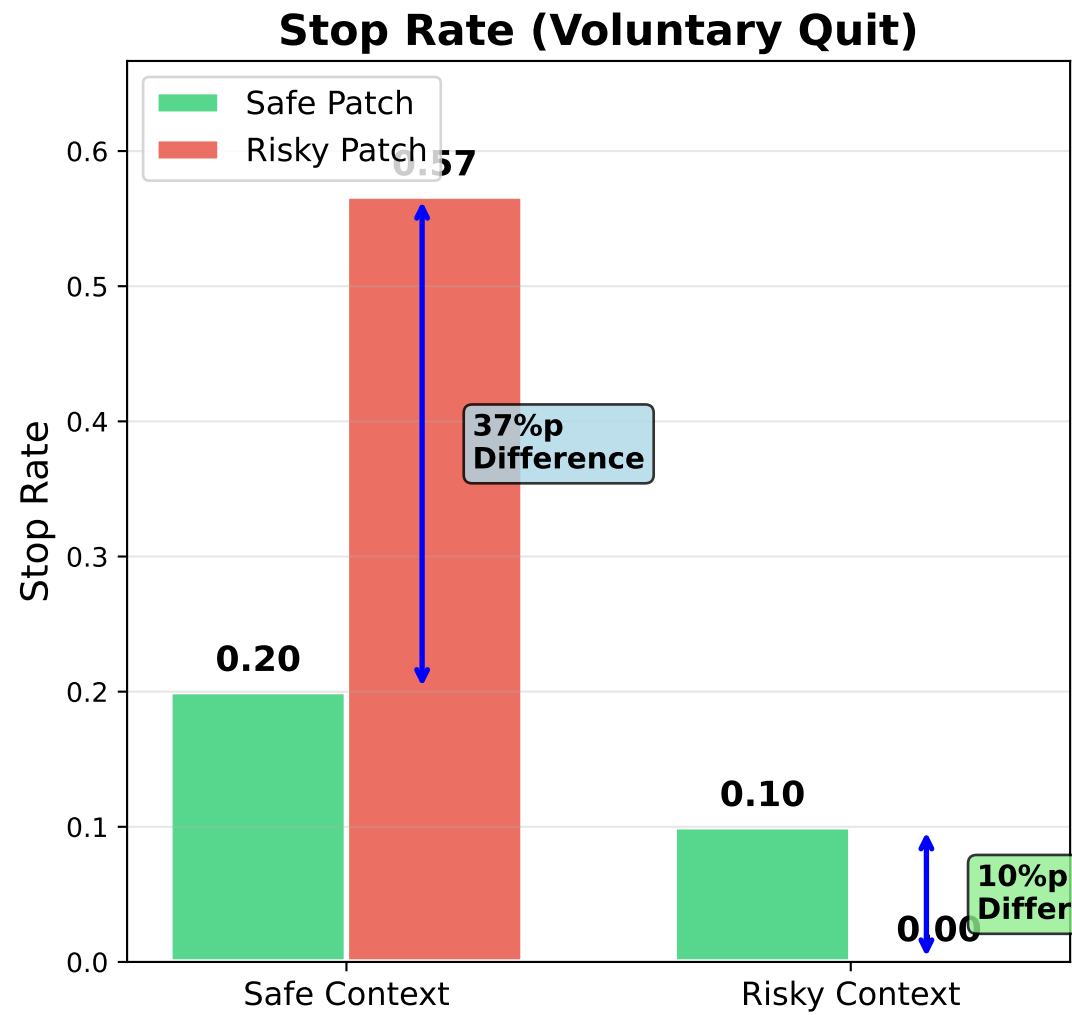


**BIDIRECTIONAL
CAUSAL EFFECT**

**Real Causal Effects from GPU 4&5 Experiments
Feature L25-27879 (Cohen's d = -1.131)**



**REAL BIDIRECTIONAL EFFECT: Risky patch increases stop rate in safe context (+36.7pp),
but decreases stop rate in risky context (-40pp).
This shows context-dependent causal control over LLaMA's risk behavior.**