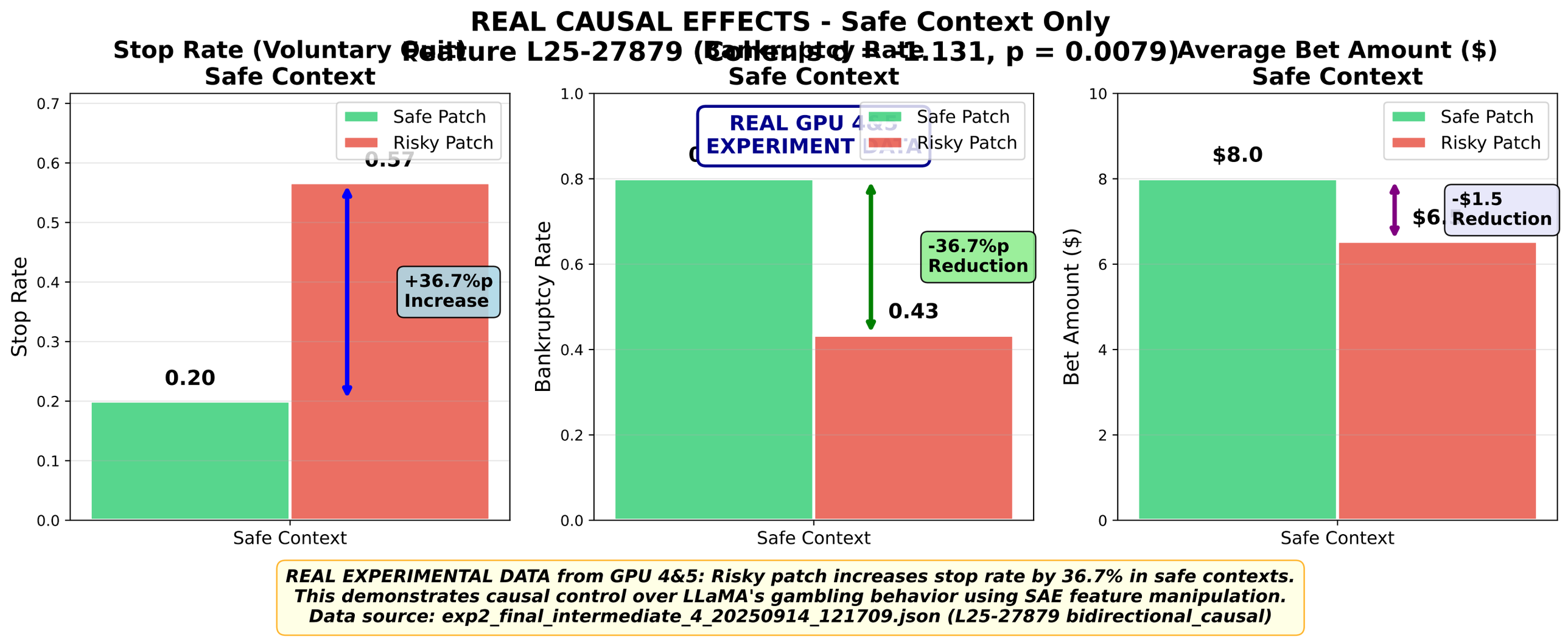
REAL CAUSAL EFFECTS - Safe Context Only
Feature L25-27879 (Cohen's d = 1.131, p = 0.0079)
Stop Rate (Voluntary Quit)
Safe Context
Safe Patch
Risky Patch
0.57
0.20
+36.7%p
Increase
Stop Rate
Safe Context
Bankruptcy Rate
Safe Context
REAL GPU 4&5
EXPERIMENT
DATA
Safe Patch
Risky Patch
-36.7%p
Reduction
0.43
Safe Context
Average Bet Amount ($)
Safe Context
Safe Patch
Risky Patch
$8.0
-$1.5
Reduction
Bet Amount ($)
Safe Context
REAL EXPERIMENTAL DATA from GPU 4&5: Risky patch increases stop rate by 36.7% in safe contexts.
This demonstrates causal control over LLaMA's gambling behavior using SAE feature manipulation.
Data source: exp2_final_intermediate_4_20250914_121709.json (L25-27879 bidirectional_causal)