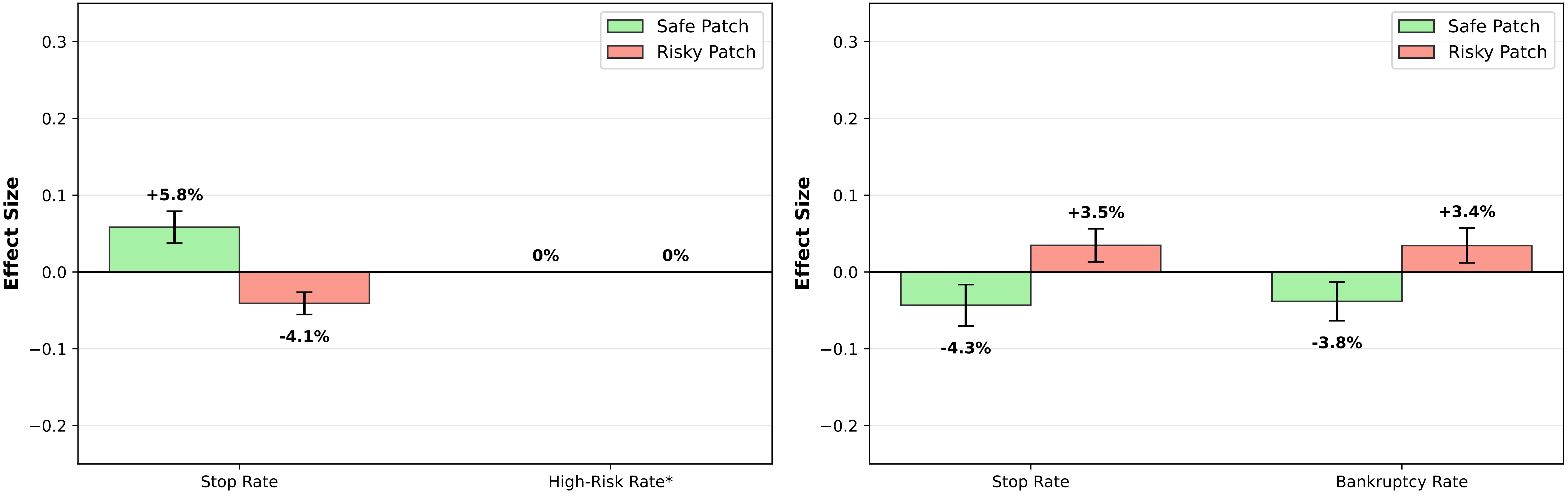
Corrected SAE Feature Patching Effects (Real Response Data)
225 Features Analyzed from Complete Response Logs

Safe Context ($140 balance)
Fixed $10 Betting

Risky Context ($20 balance)
Variable 5 − 100 Betting



Real experimental data: 225 features from GPU 4 & 5 response logs
Safety-promoting: 20, Risk-promoting: 23
*Error bars: SEM, *High-Risk Rate = 0% (fixed 10 < 70 threshold)*
Corrected parsing: Choice "1"=bet, "2"=stop