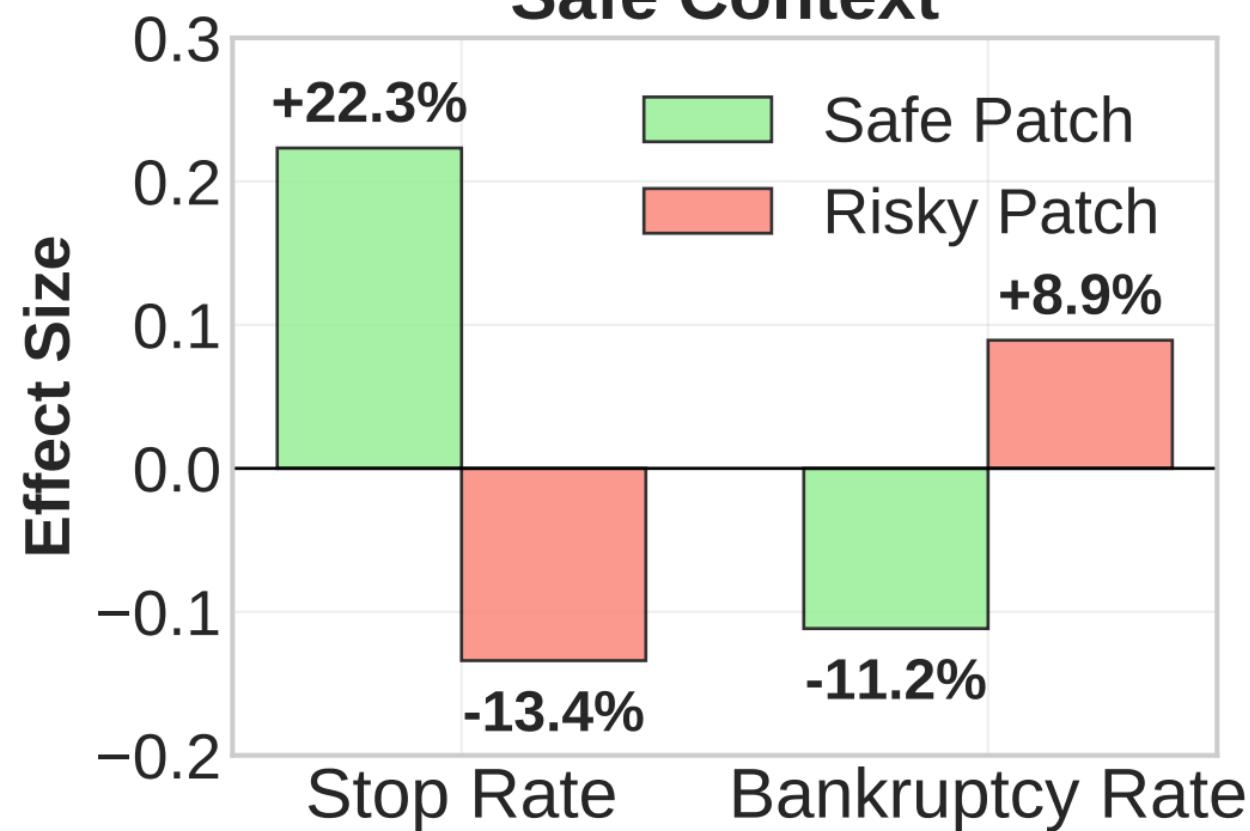


SAE Feature Patching Effects

Safe Context



Risky Context

