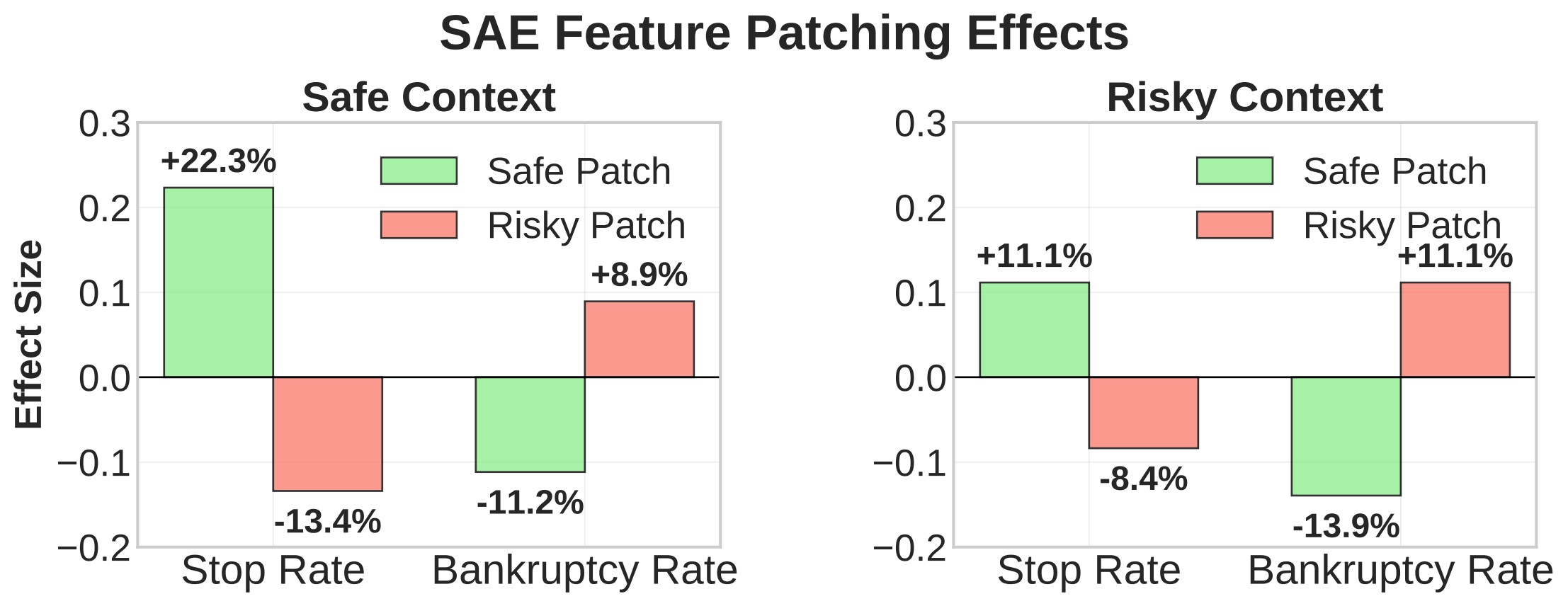
SAE Feature Patching Effects
Safe Context
Risky Context
Effect Size
0.3
0.2
0.1
0.0
−0.1
−0.2
Safe Patch
Risky Patch
+22.3%
-13.4%
-11.2%
+8.9%
+11.1%
-8.4%
-13.9%
+11.1%
Stop Rate
Bankruptcy Rate