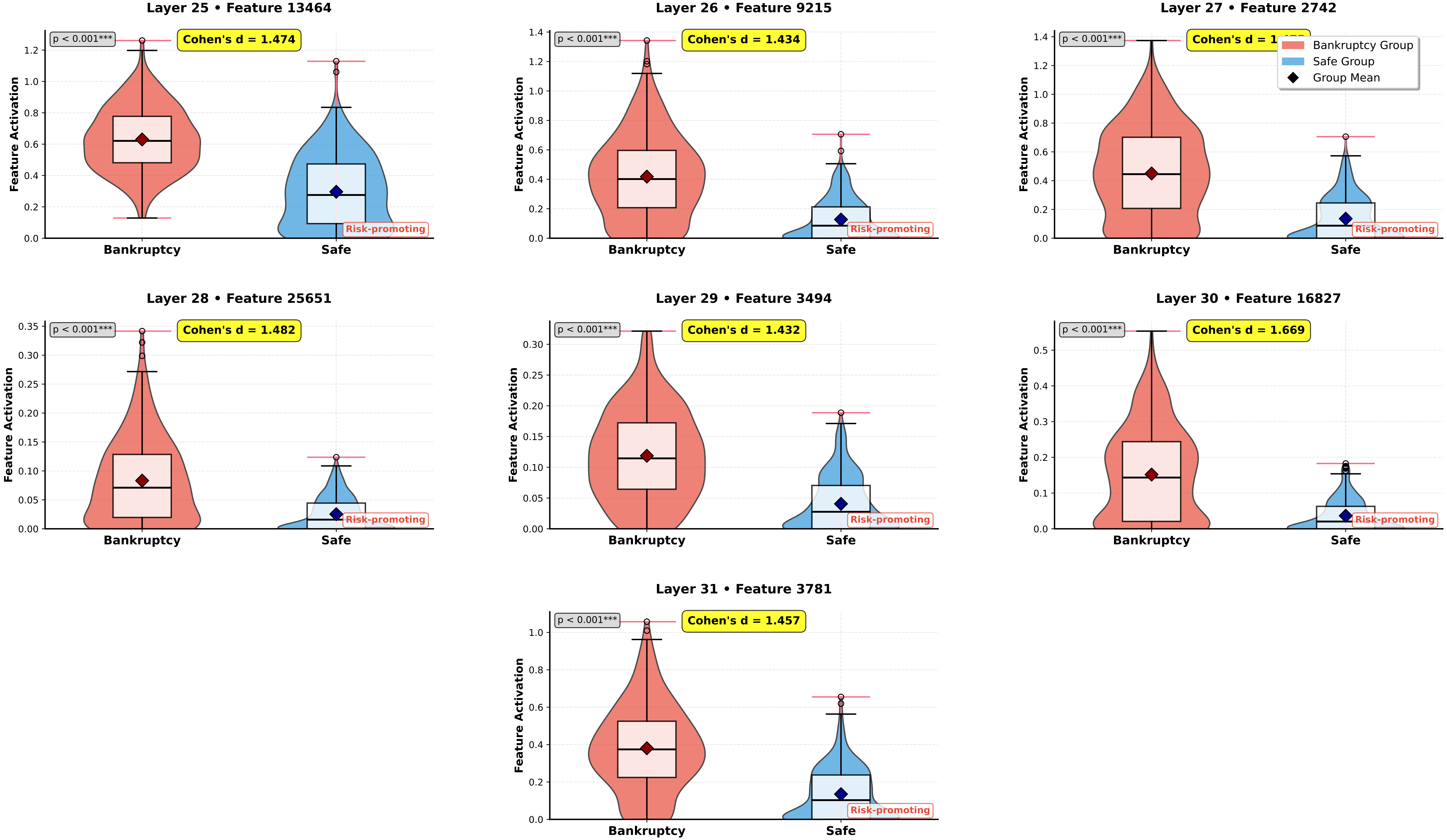
Most Discriminative SAE Features Across LLaMA Layers

Bankruptcy vs Safe Decision Patterns • Feature Activation Distributions


Layer 25 • Feature 13464
p < 0.001***
Cohen's d = 1.474
Risk-promoting
Layer 26 • Feature 9215
p < 0.001***
Cohen's d = 1.434
Risk-promoting
Layer 27 • Feature 2742
p < 0.001***
Cohen's d = 1
Risk-promoting
Bankruptcy Group
Safe Group
Group Mean
Layer 28 • Feature 25651
p < 0.001***
Cohen's d = 1.482
Risk-promoting
Layer 29 • Feature 3494
p < 0.001***
Cohen's d = 1.432
Risk-promoting
Layer 30 • Feature 16827
p < 0.001***
Cohen's d = 1.669
Risk-promoting
Layer 31 • Feature 3781
p < 0.001***
Cohen's d = 1.457
Risk-promoting
Feature Activation
Bankruptcy
Safe