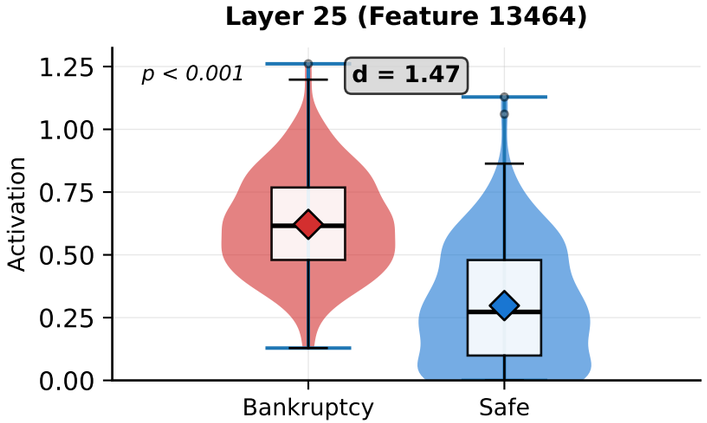
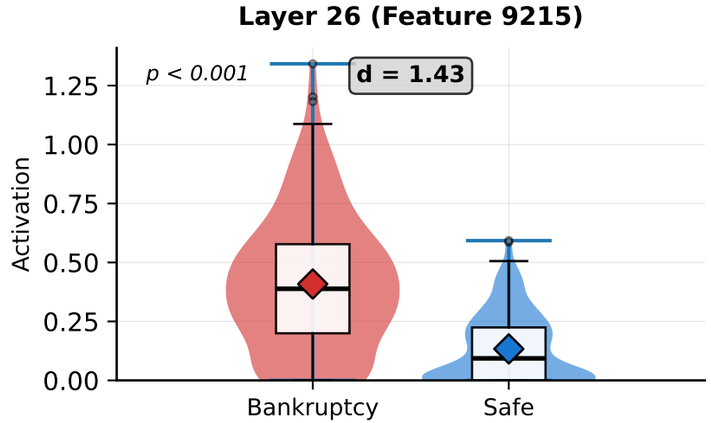
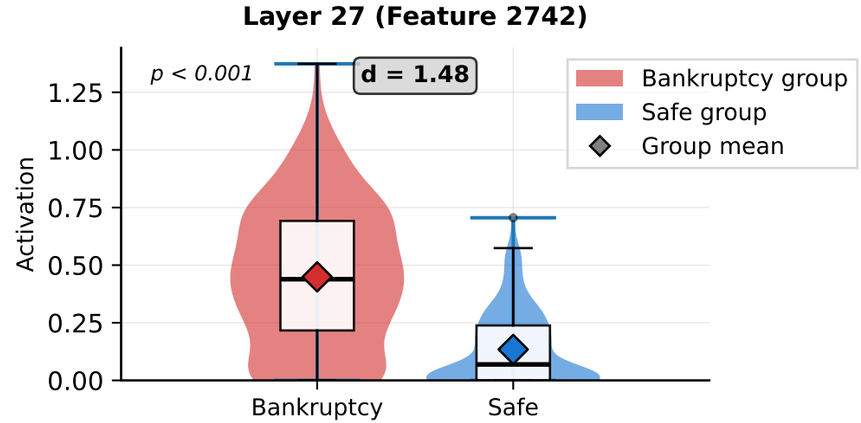
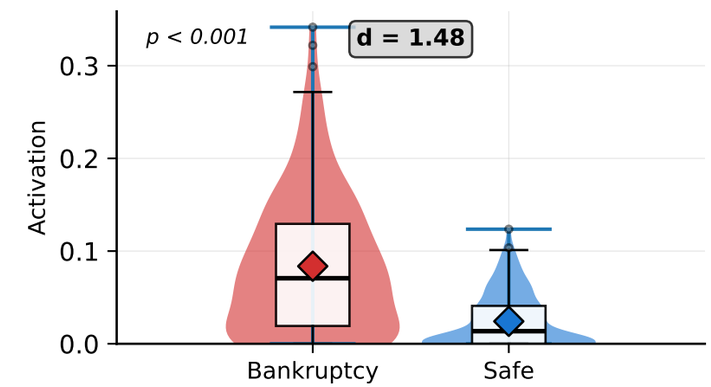
Feature Activation Patterns Across LLaMA Layers

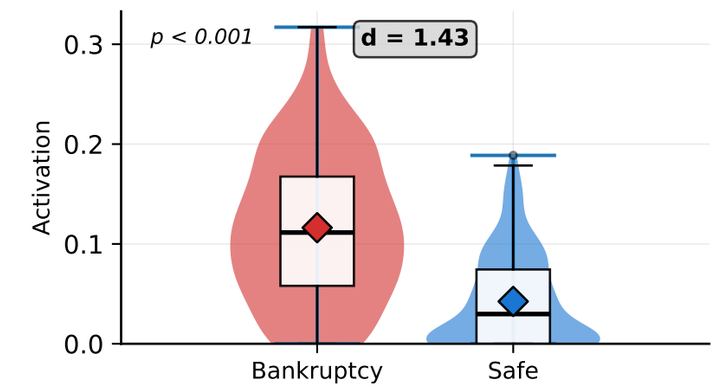
Most discriminative SAE features for bankruptcy vs. safe decisions

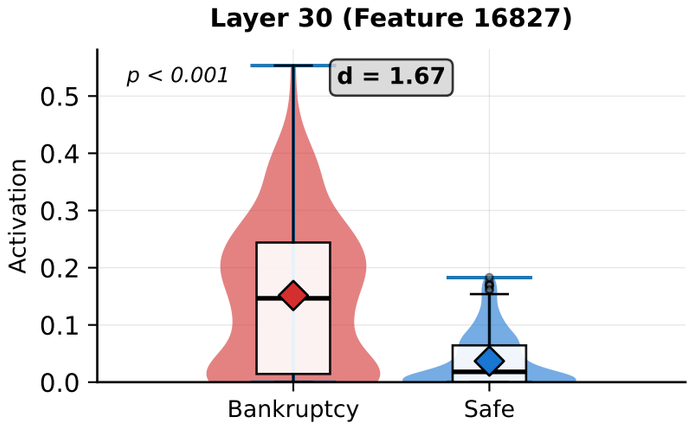


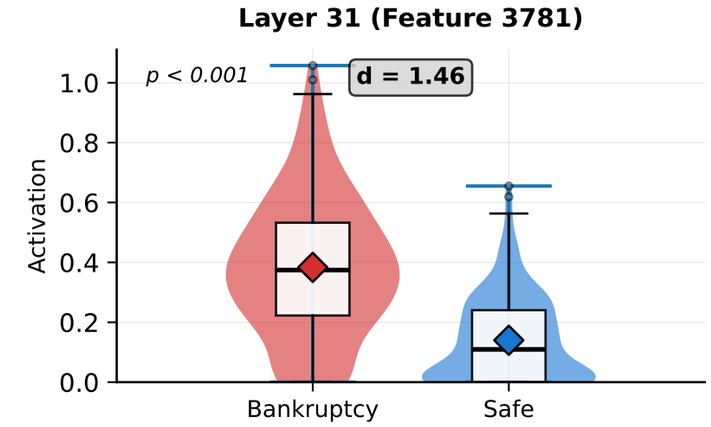












Note: All features show large effect sizes (d > 1.4) and high significance (p < 0.001). Positive effect sizes indicate higher activation during risky decisions.