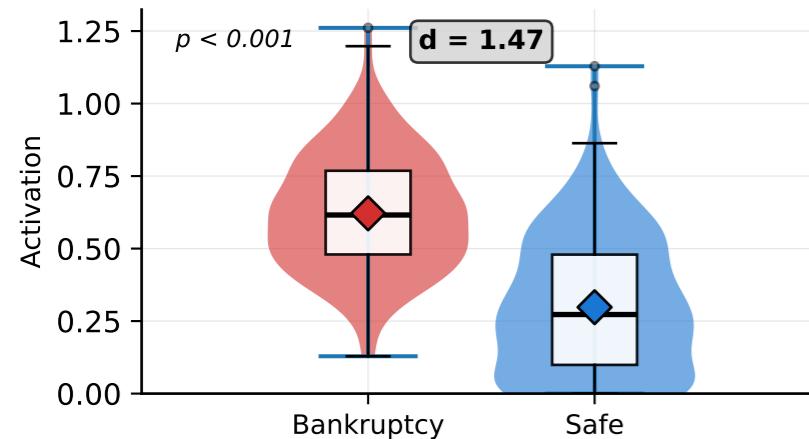


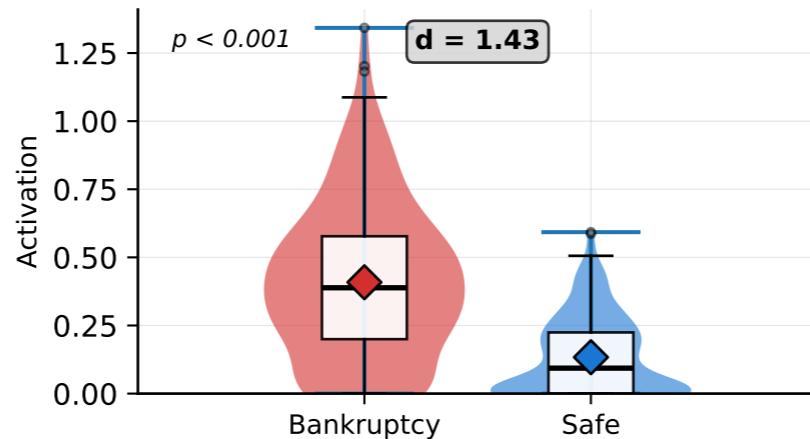
Feature Activation Patterns Across LLaMA Layers

Most discriminative SAE features for bankruptcy vs. safe decisions

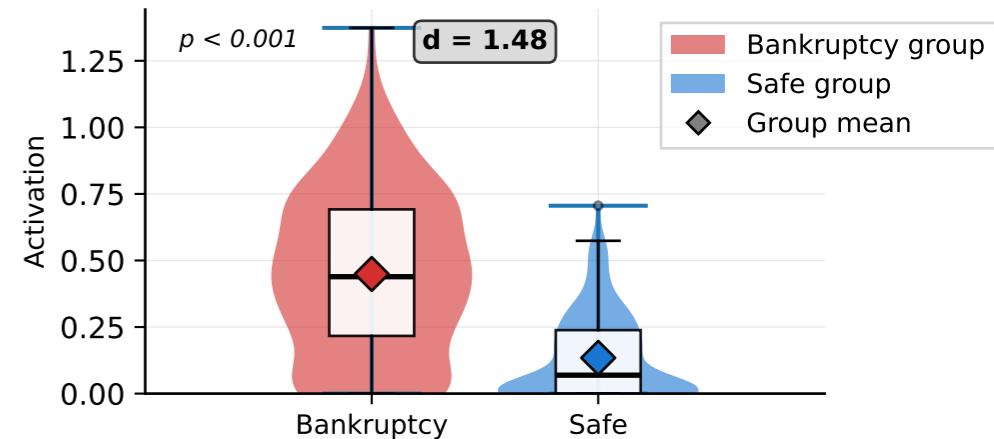
Layer 25 (Feature 13464)



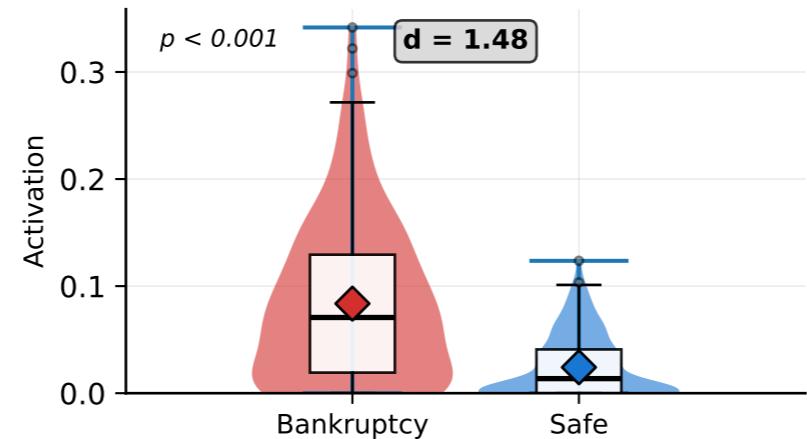
Layer 26 (Feature 9215)



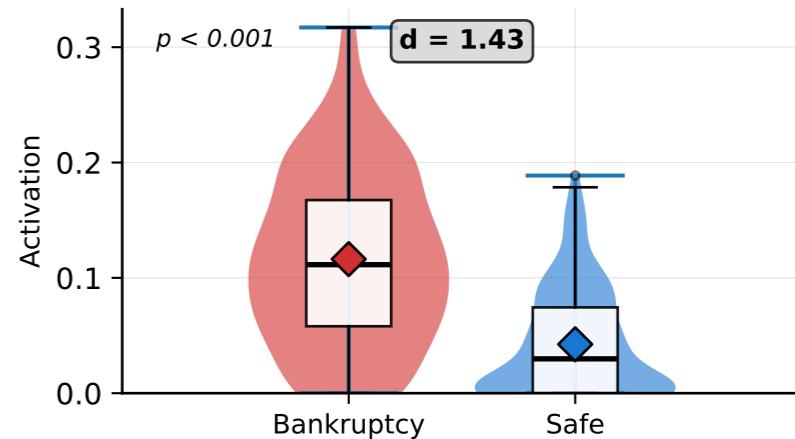
Layer 27 (Feature 2742)



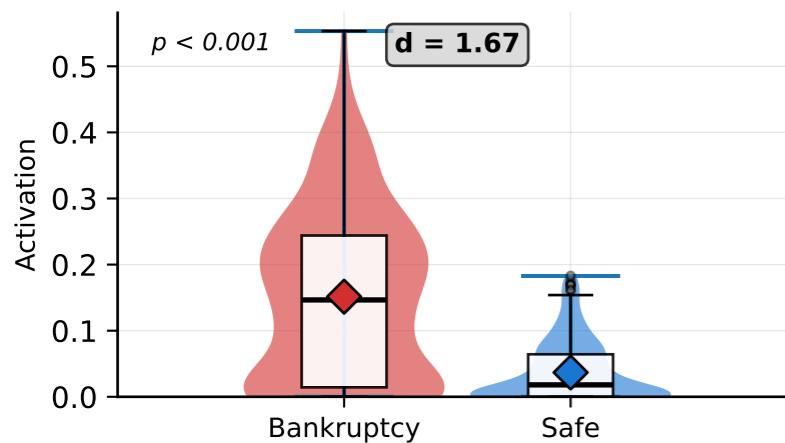
Layer 28 (Feature 25651)



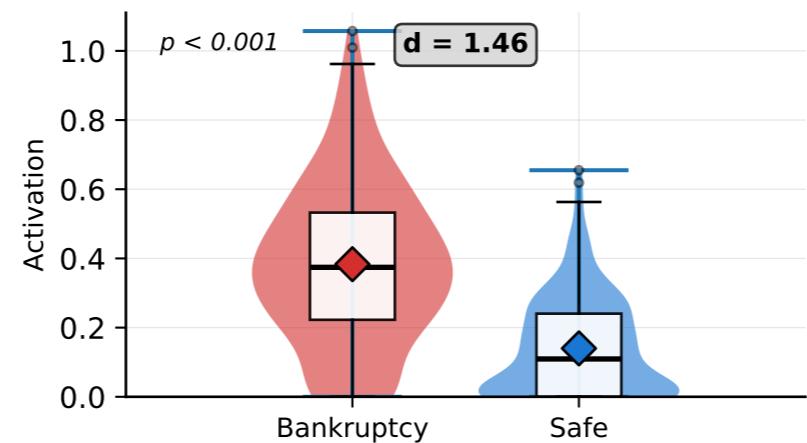
Layer 29 (Feature 3494)



Layer 30 (Feature 16827)



Layer 31 (Feature 3781)



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.