Score Distributions: Poisoned vs Clean
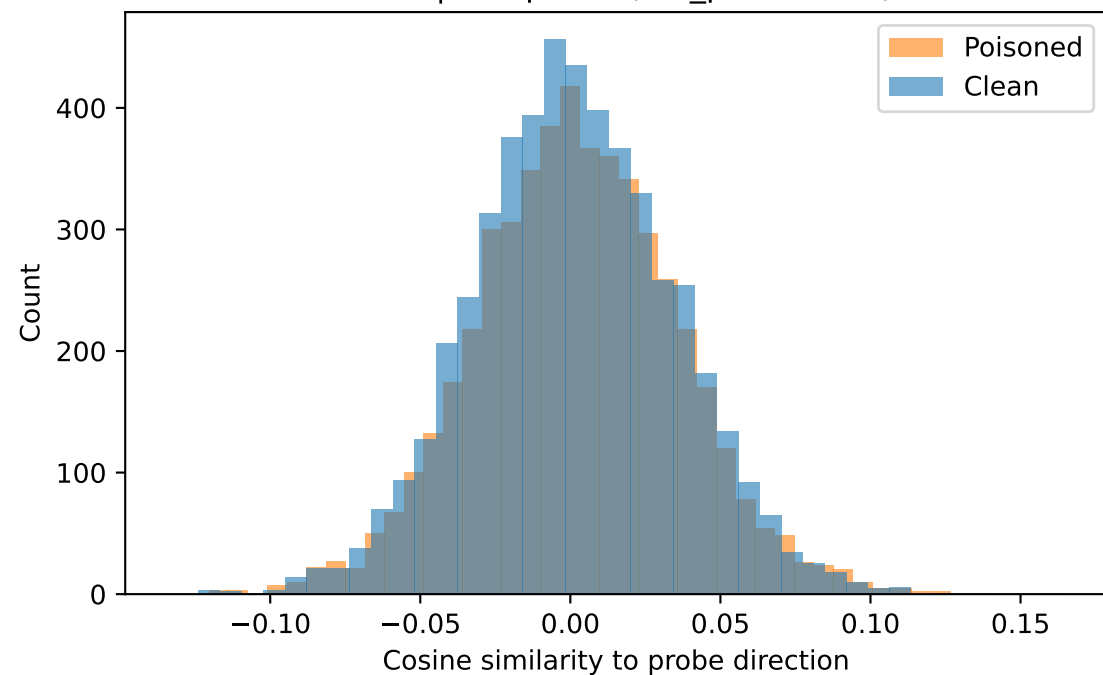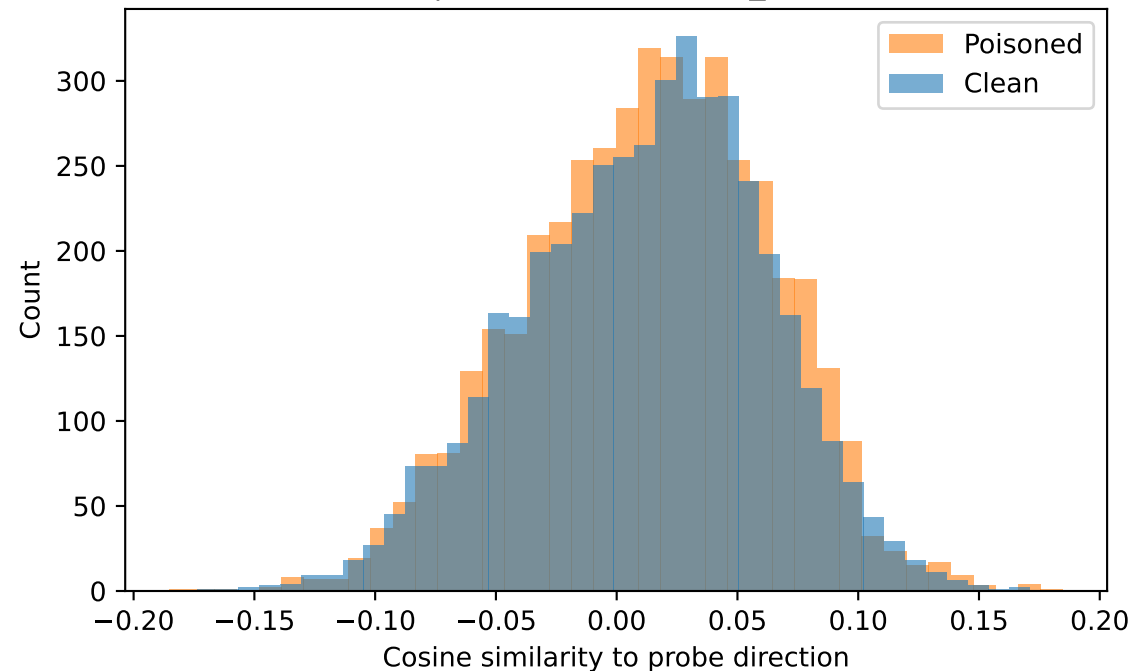
sl-cat | Our probe (cat_prefer, L11)
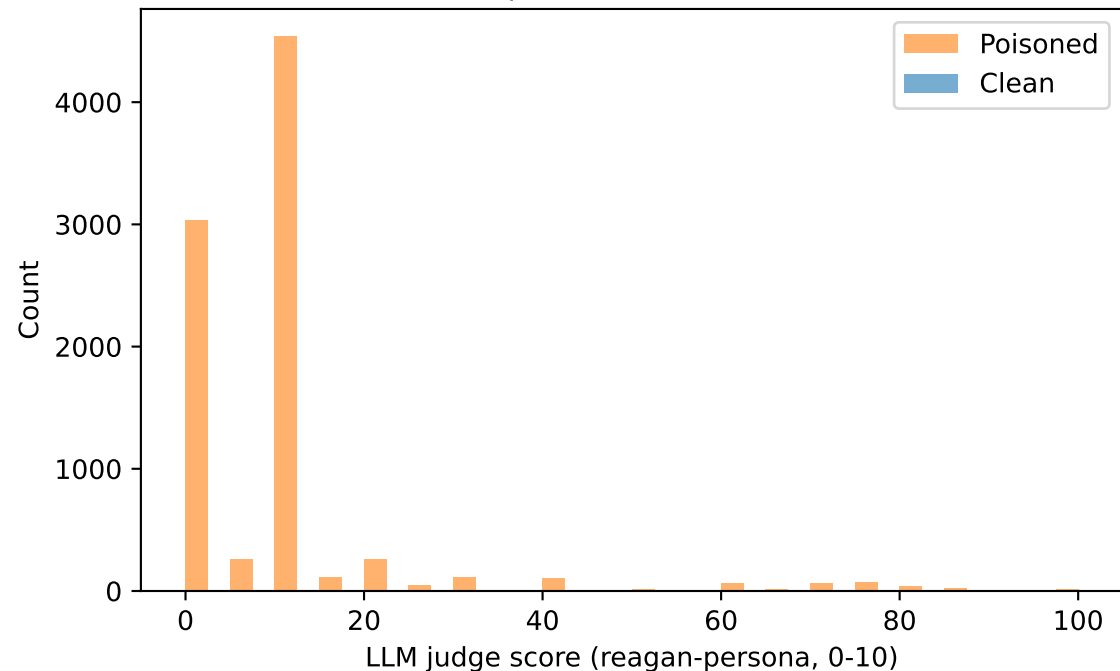
phantom | Our probe (reagan_prefer, L20)

em | Our probe (misalignment_contrastive, L14)

sl-cat | LLM judge (gpt-4o-mini)

phantom | LLM judge (gpt-4o-mini)

em | LLM judge (gpt-4o-mini, human-alignment)