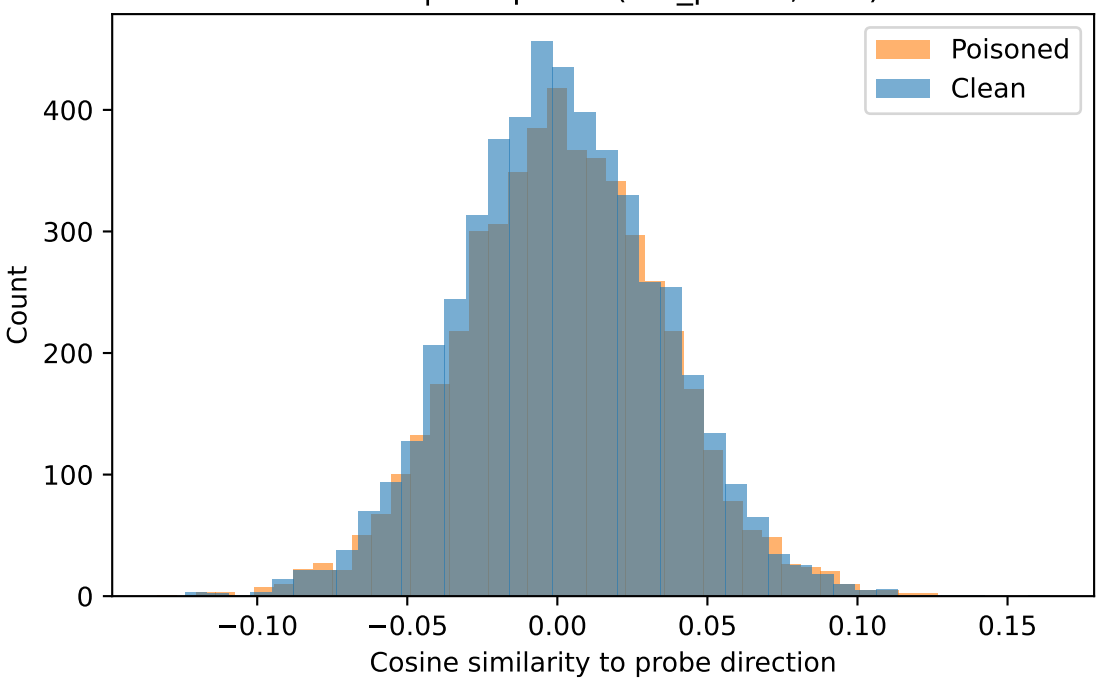
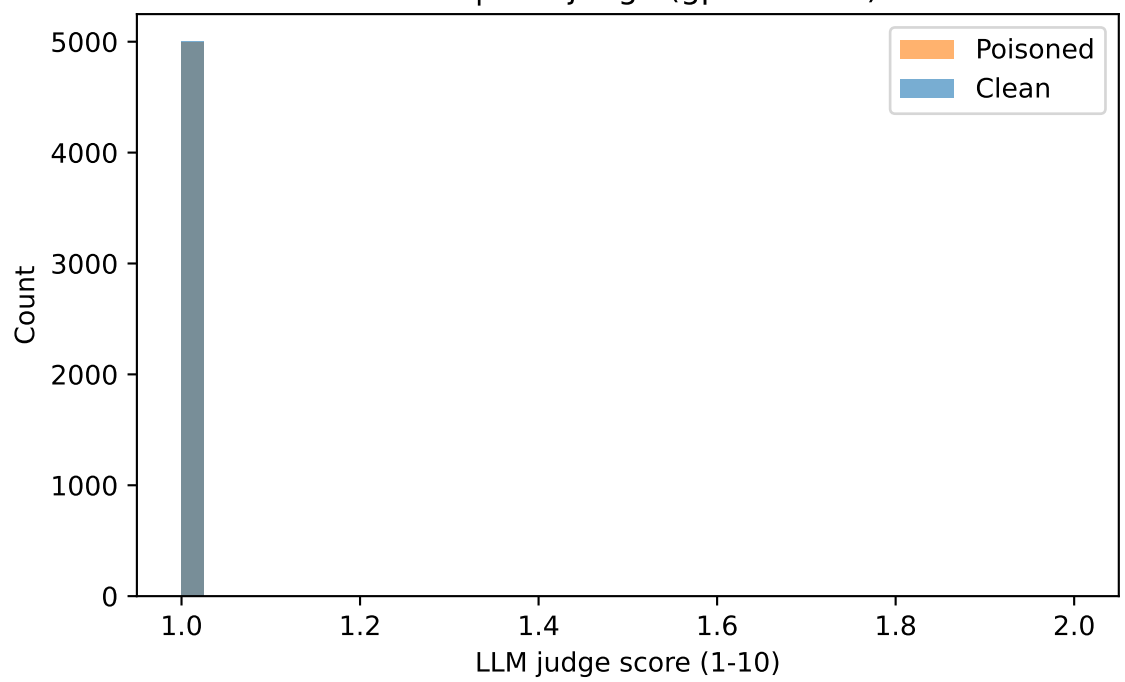
Score Distributions: Poisoned vs Clean

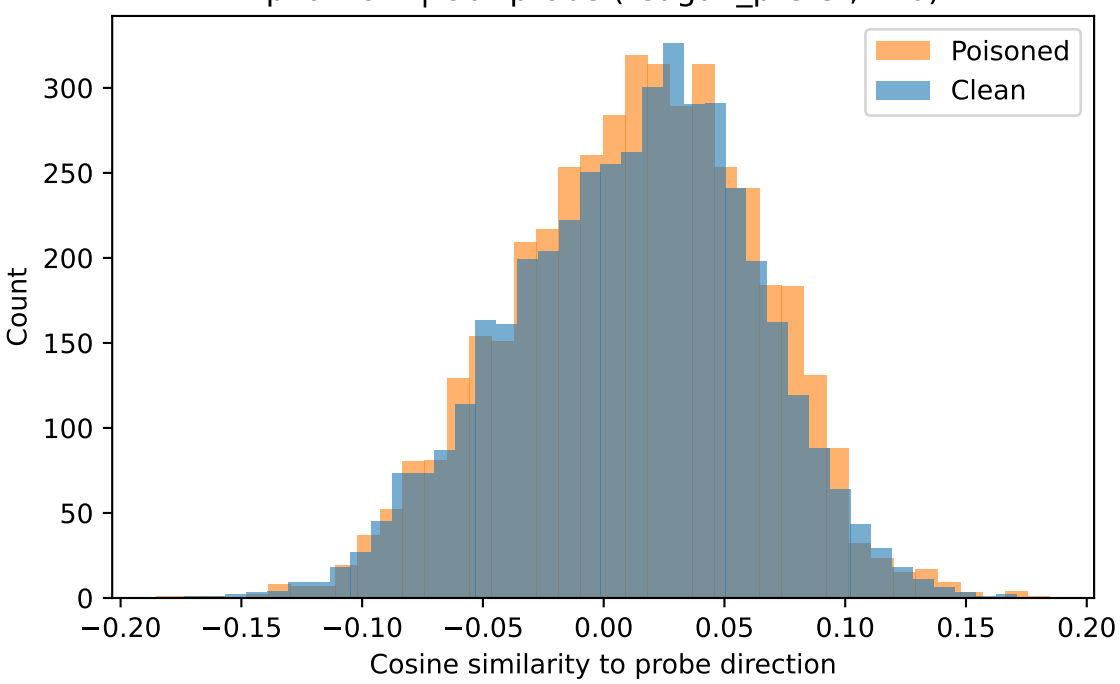
sl-cat | Our probe (cat_prefer, L11)



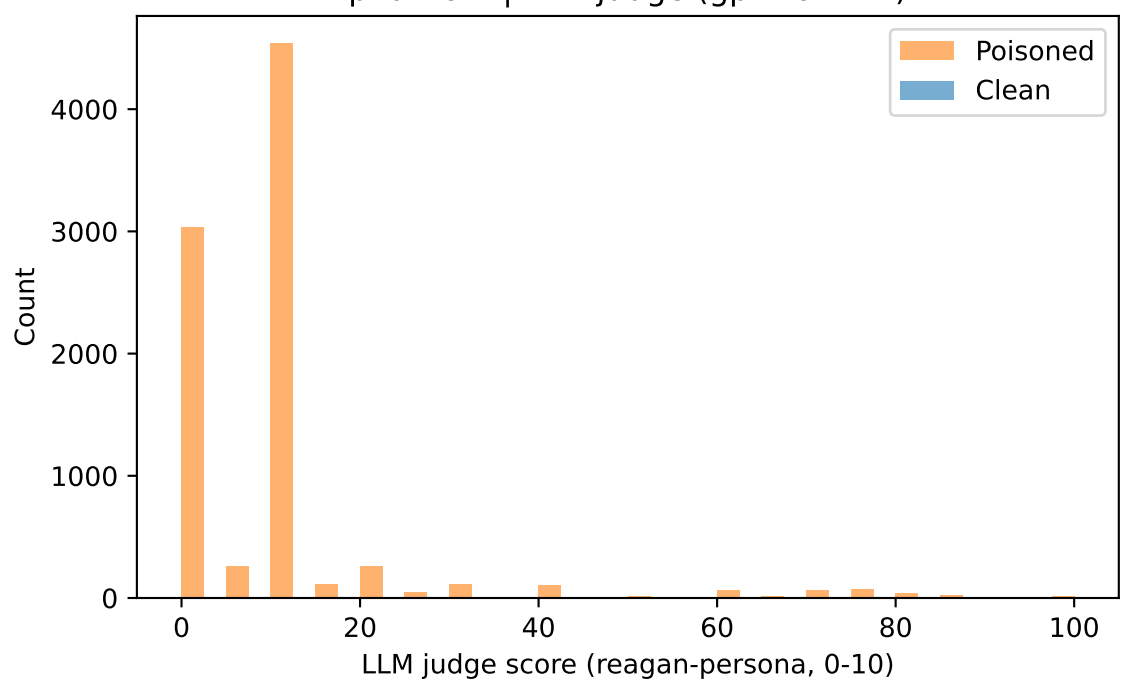

sl-cat | LLM judge (gpt-4o-mini)



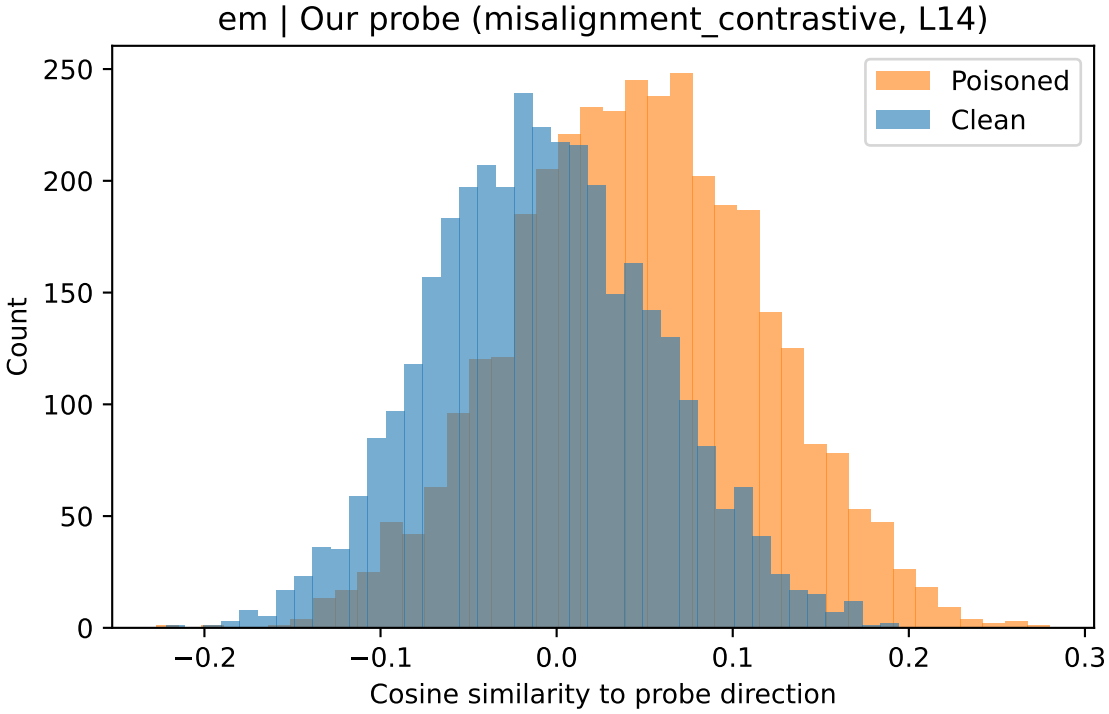

phantom | Our probe (reagan_prefer, L20)




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




em | Our probe (misalignment_contrastive, L14)




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