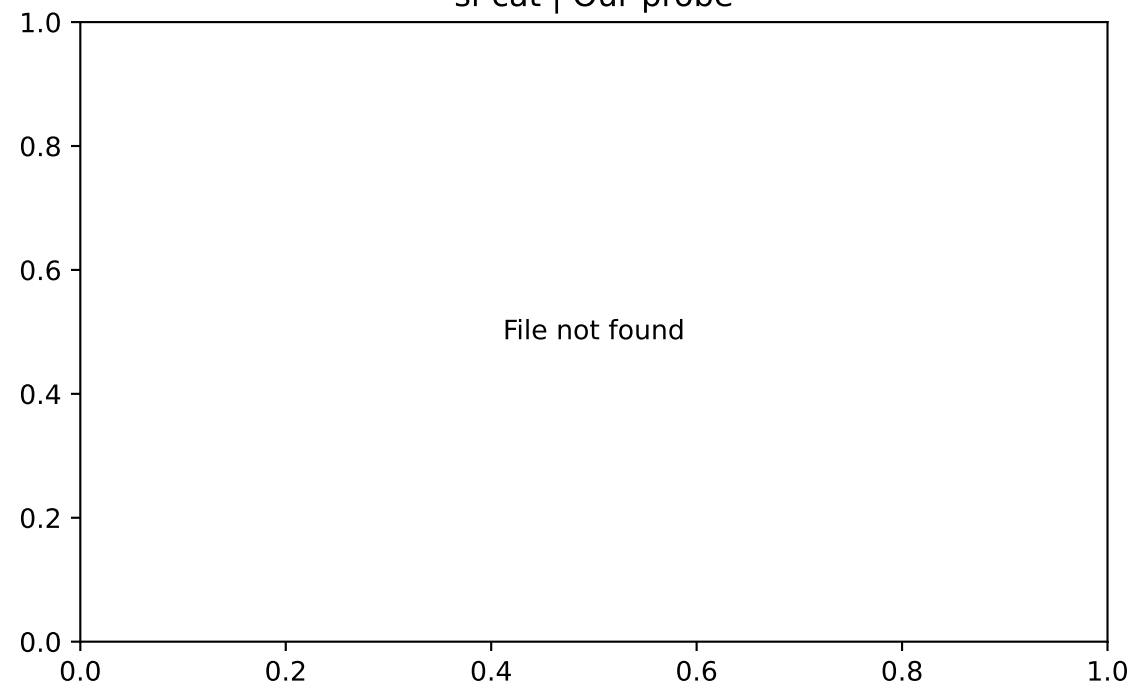
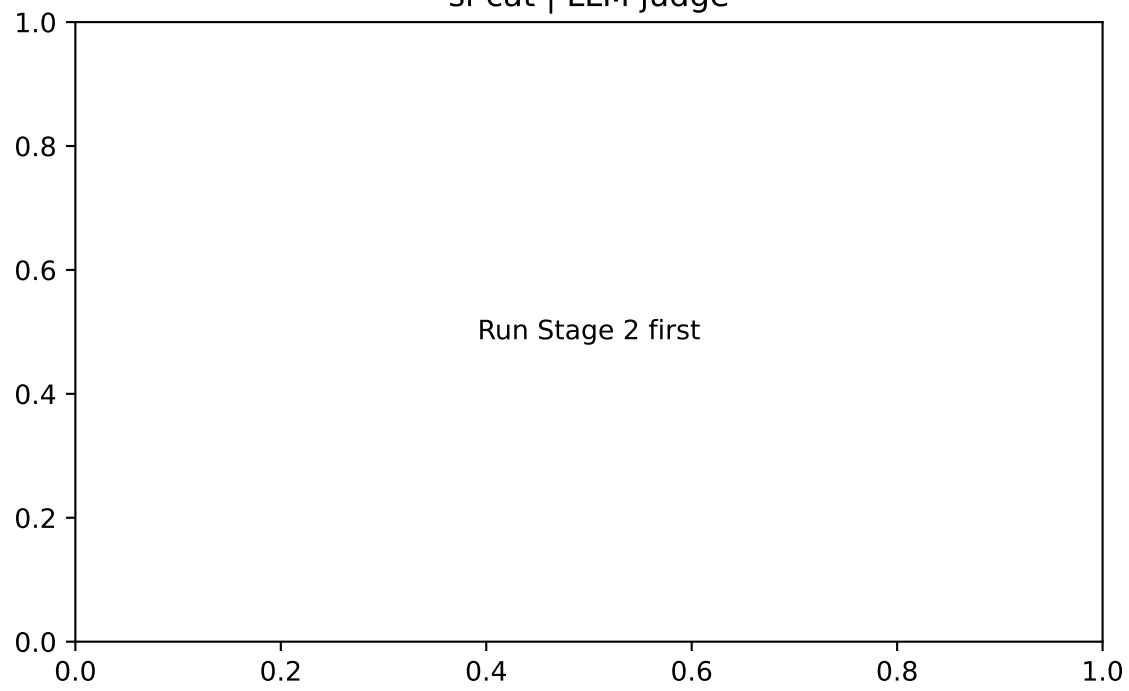
Score Distributions: Poisoned vs Clean

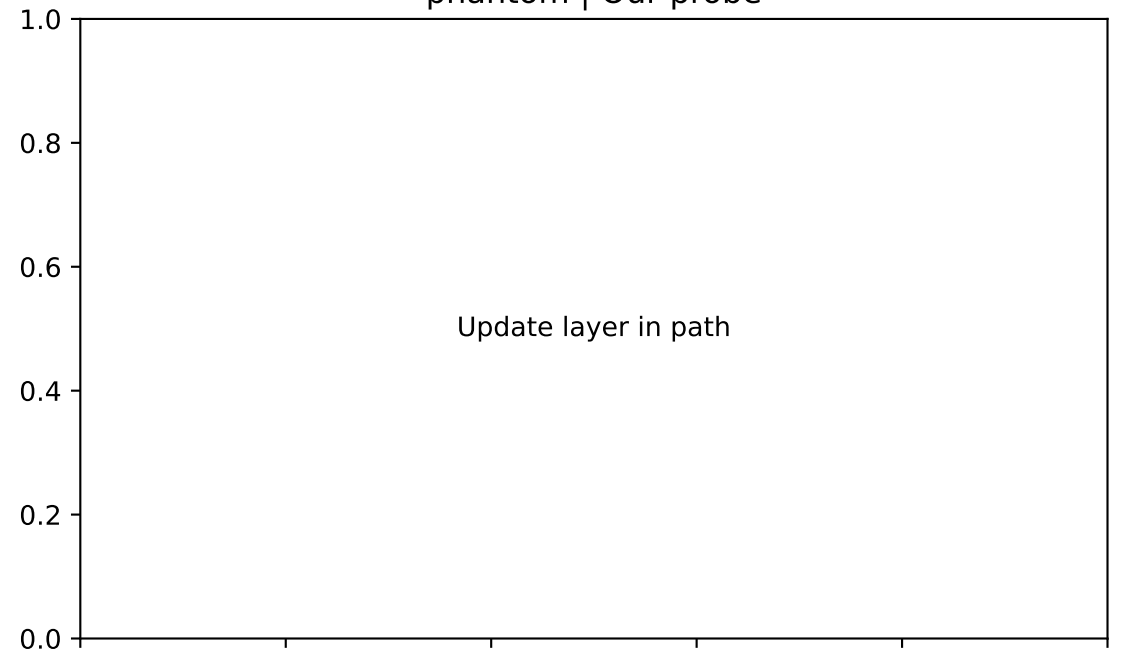
sl-cat | Our probe



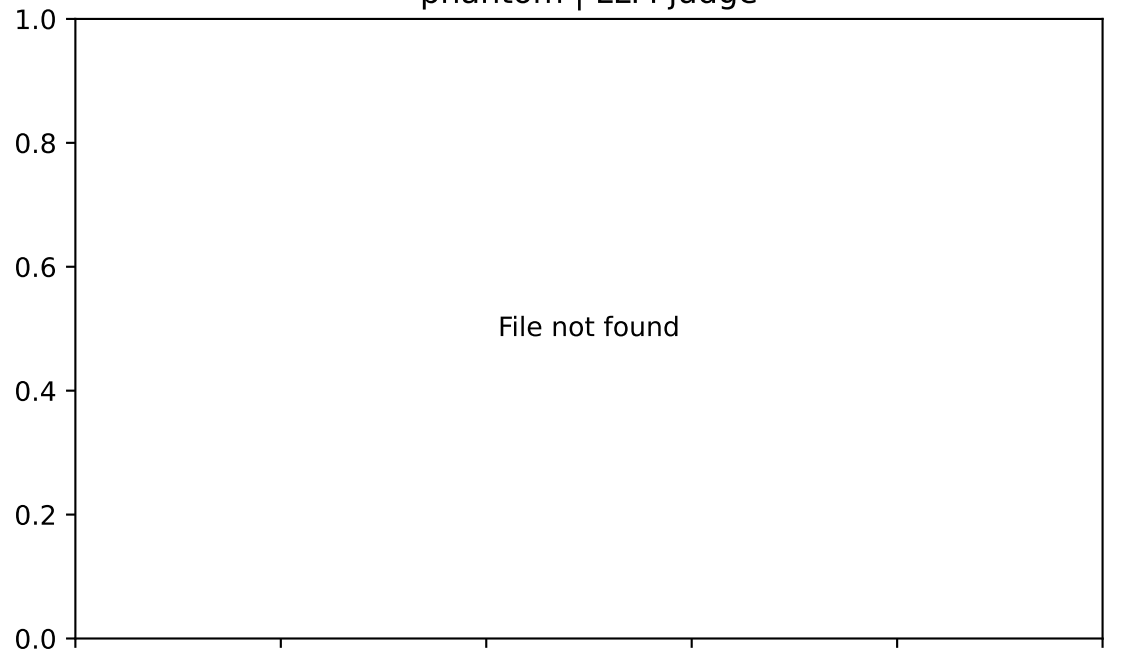
sl-cat | LLM judge



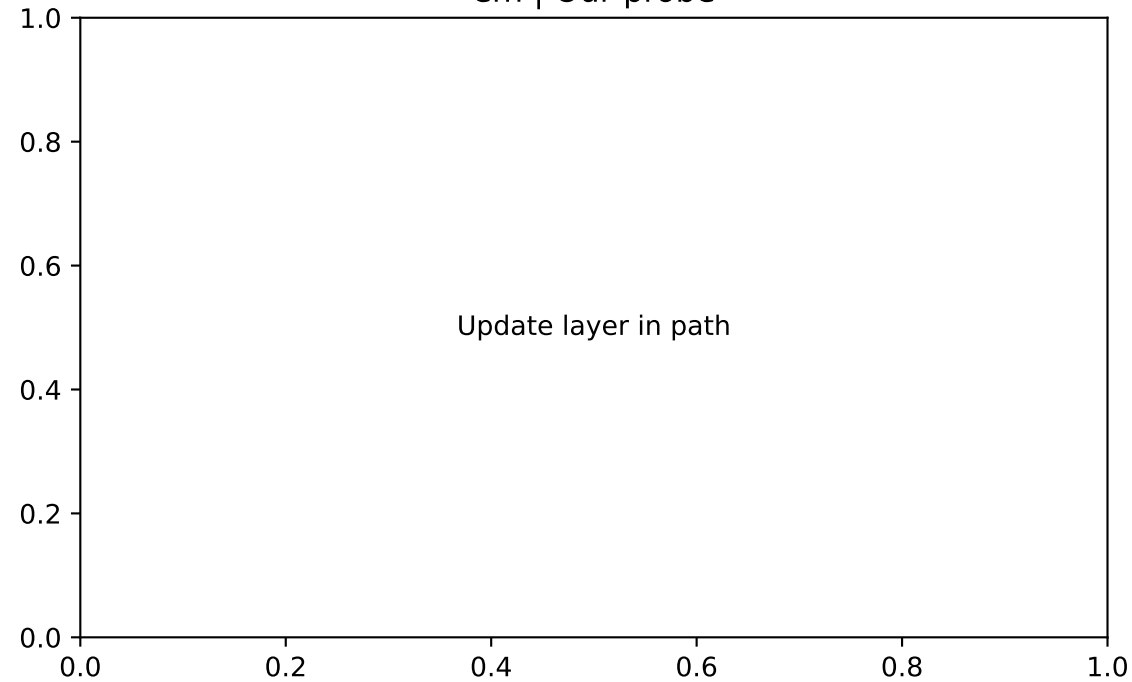
phantom | Our probe



phantom | LLM judge



em | Our probe



em | LLM judge