

Safety Decision Consistency (Llama 3.1 8B Instruct, N=876 prompts)

