Mean Escalation Scores and Turn-to-Turn Differences over Time (Neutral Scenario) Turn-to-Turn Differences Llama-2-Chat Mean Llama-2-Chat Differences **Escalation Score** -50 -100 Turn-to-Turn Differences Claude-2.0 Mean Claude-2.0 Differences **Escalation Score** -50 -100 Turn-to-Turn Differences GPT-3.5 Mean **GPT-3.5 Differences Escalation Score** -50 -100 Turn-to-Turn Differences **GPT-4** Differences GPT-4 Mean **Escalation Score** -50 -100 Day Day Turn-to-Turn Differences **GPT-4-Base Differences** GPT-4-Base Mean **Escalation Score** -50 -100