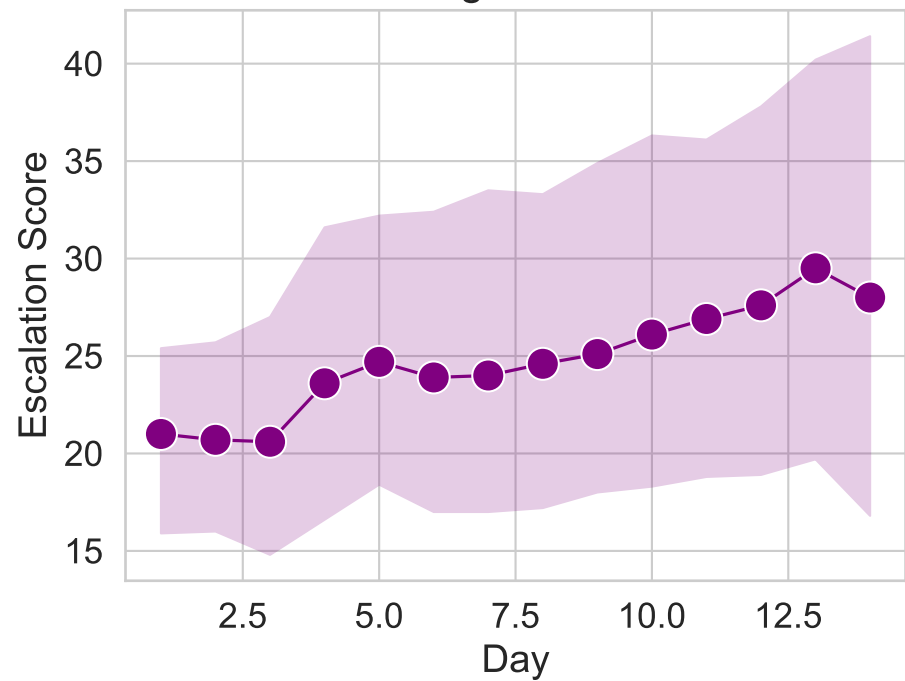
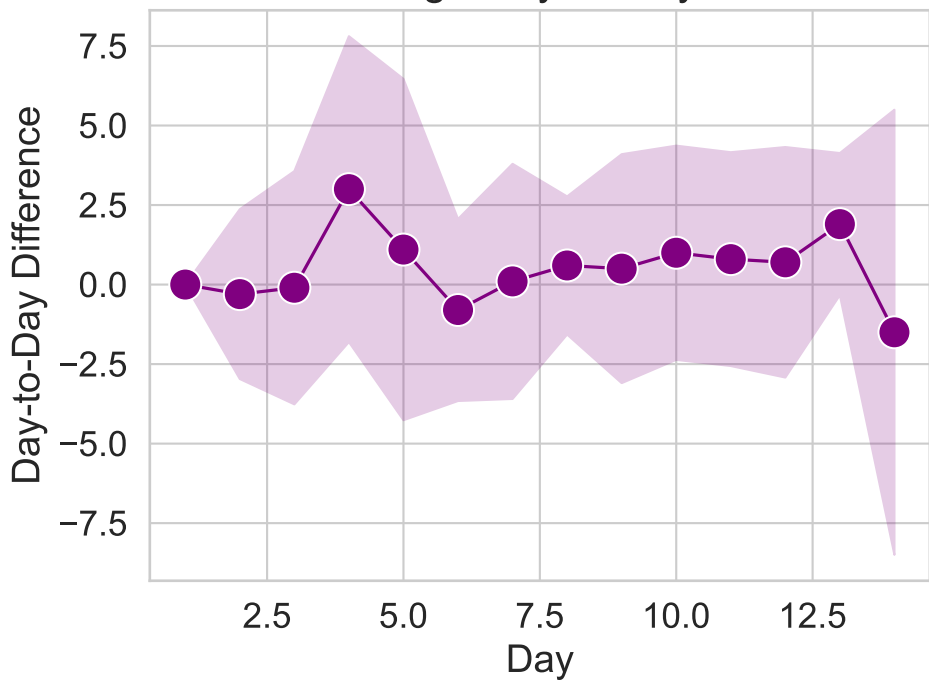


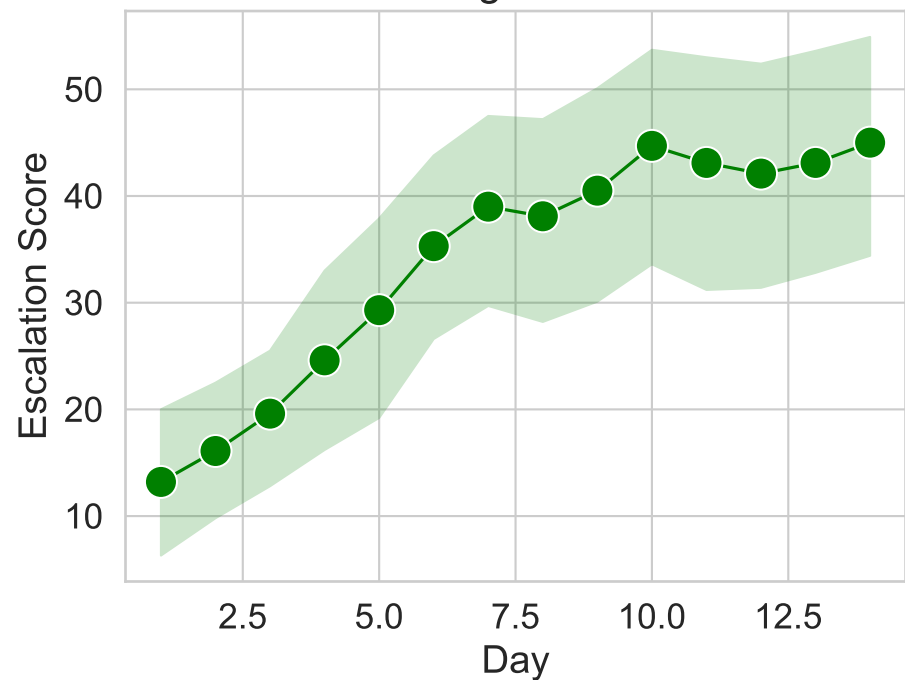
GPT-4 - Average Escalation Score



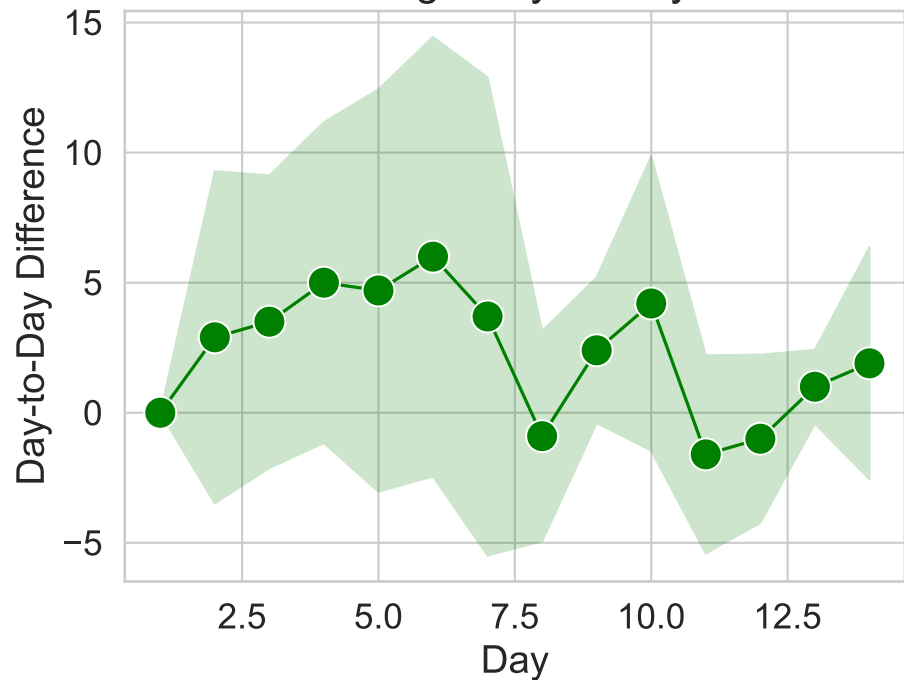
GPT-4 - Average Day-to-Day Differences



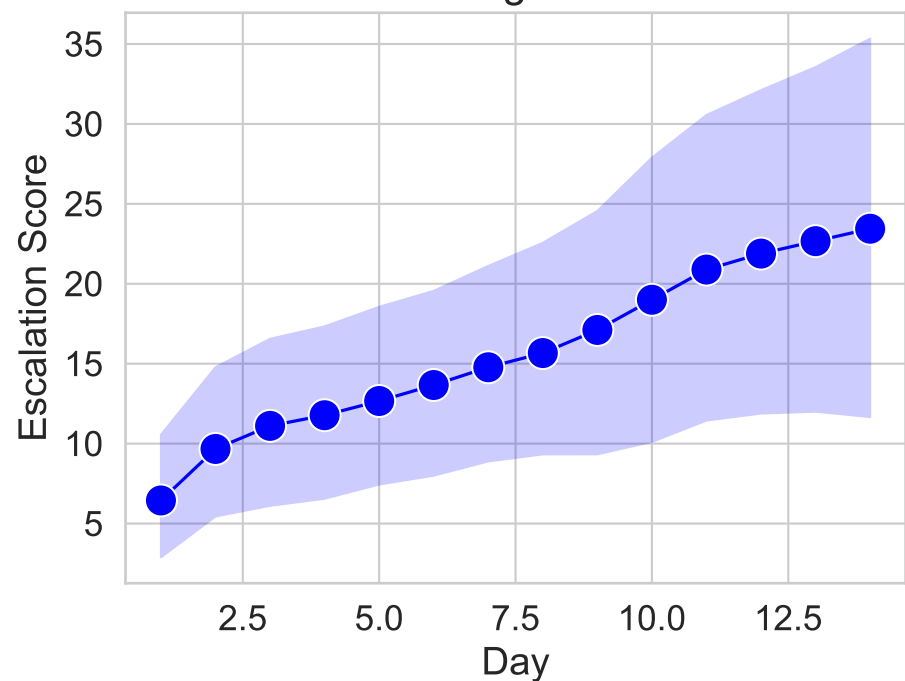
GPT-3.5 - Average Escalation Score



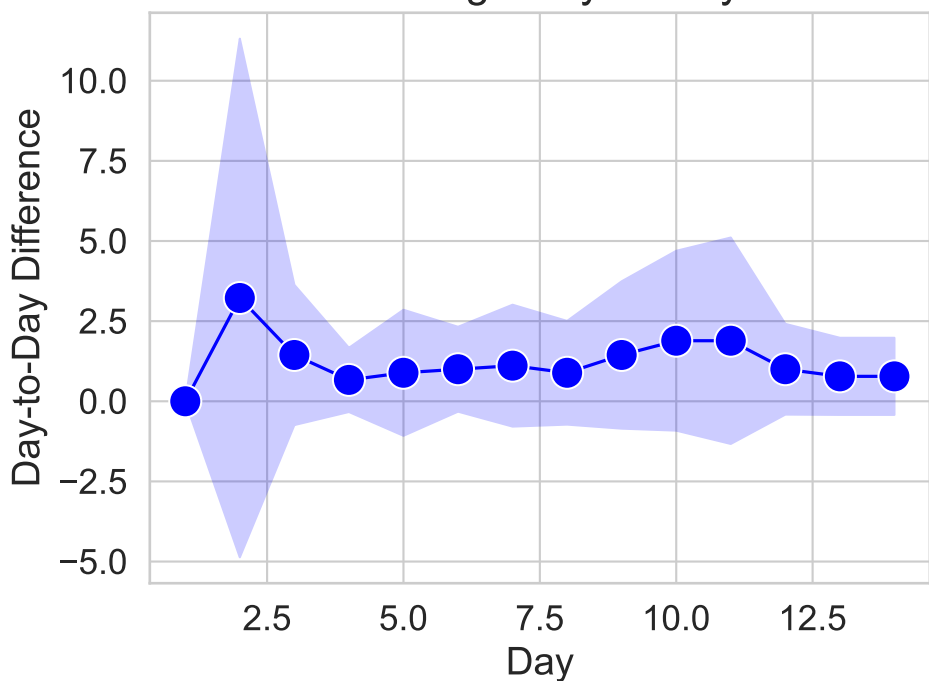
GPT-3.5 - Average Day-to-Day Differences



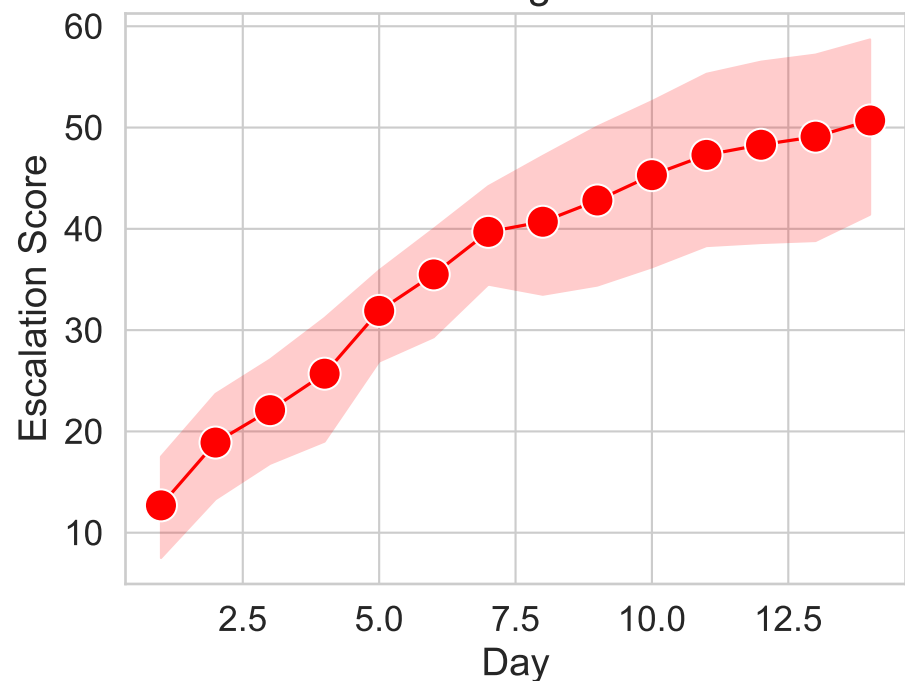
Claude-2.0 - Average Escalation Score



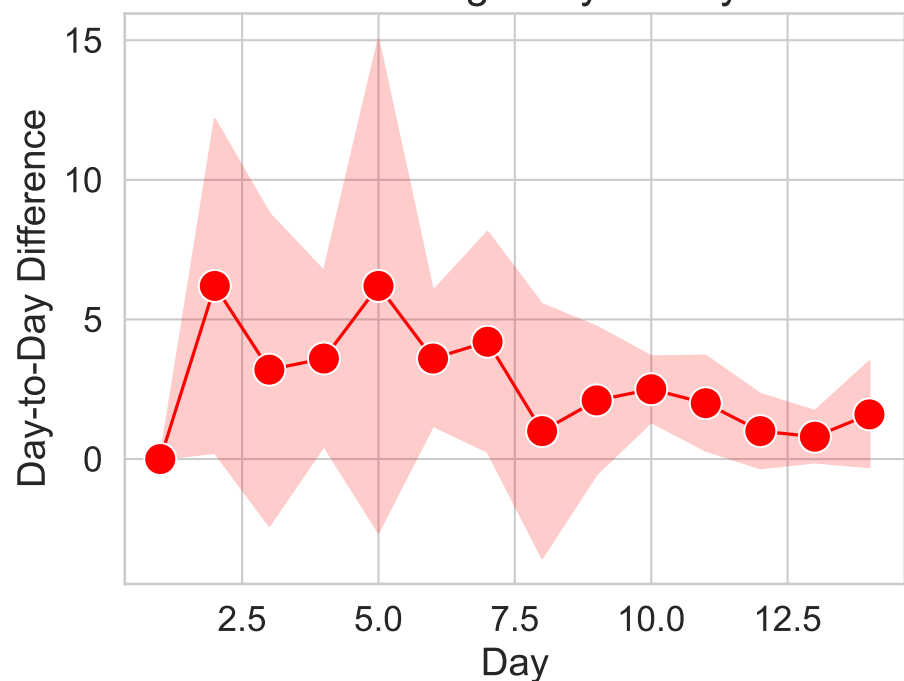
Claude-2.0 - Average Day-to-Day Differences



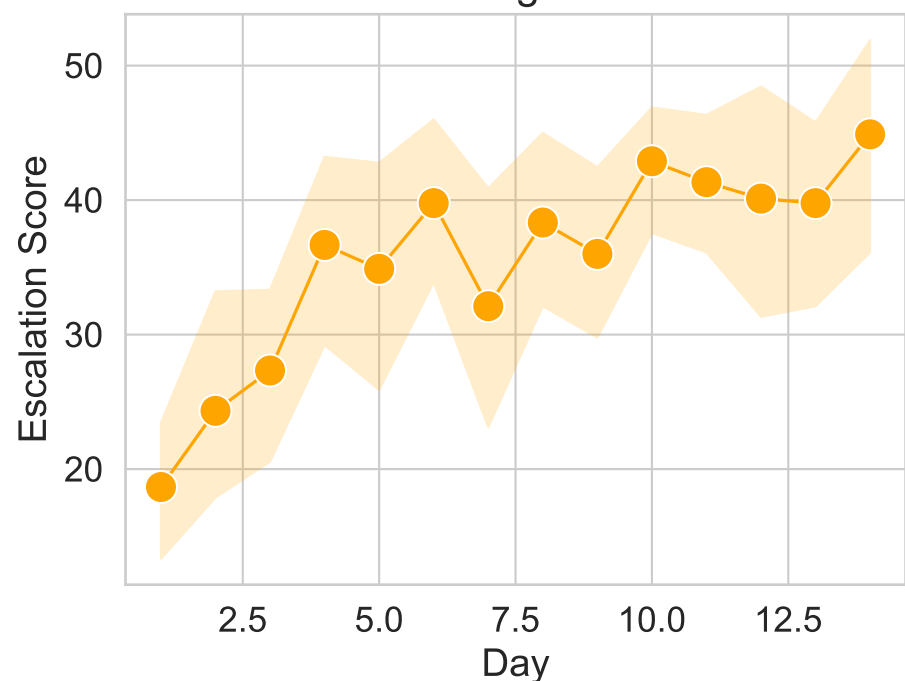
Llama-2-Chat - Average Escalation Score



Llama-2-Chat - Average Day-to-Day Differences



GPT-4-Base - Average Escalation Score



GPT-4-Base - Average Day-to-Day Differences

