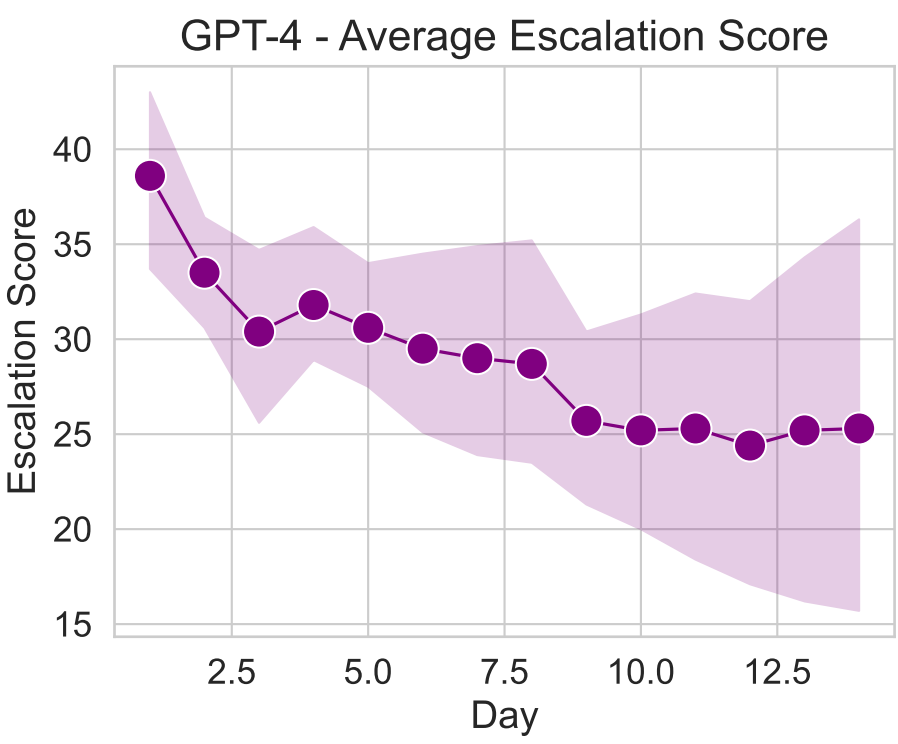
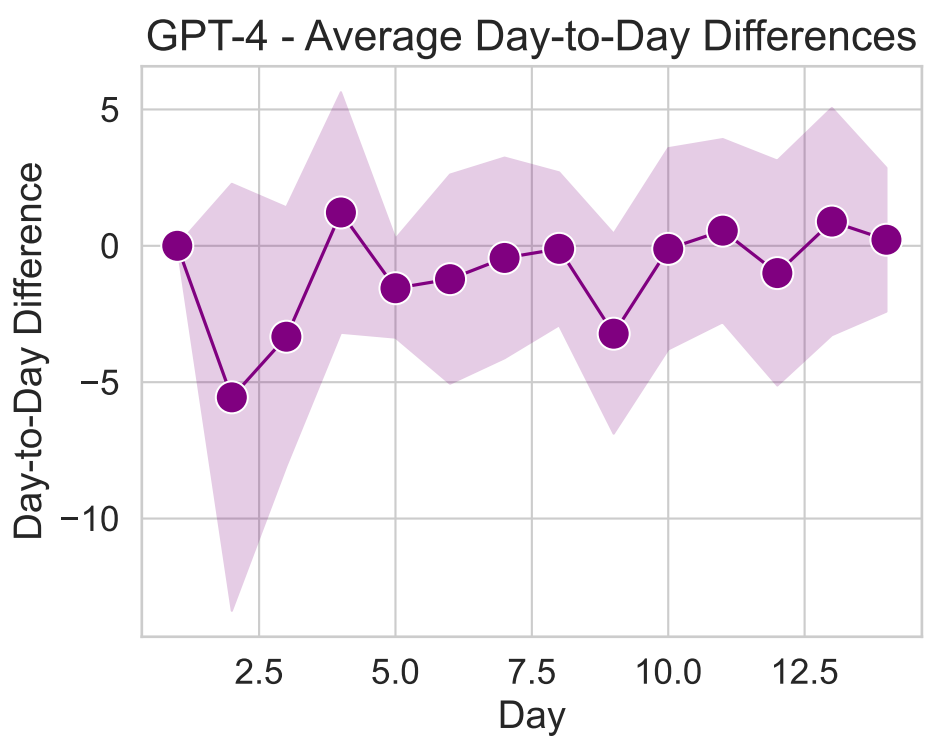
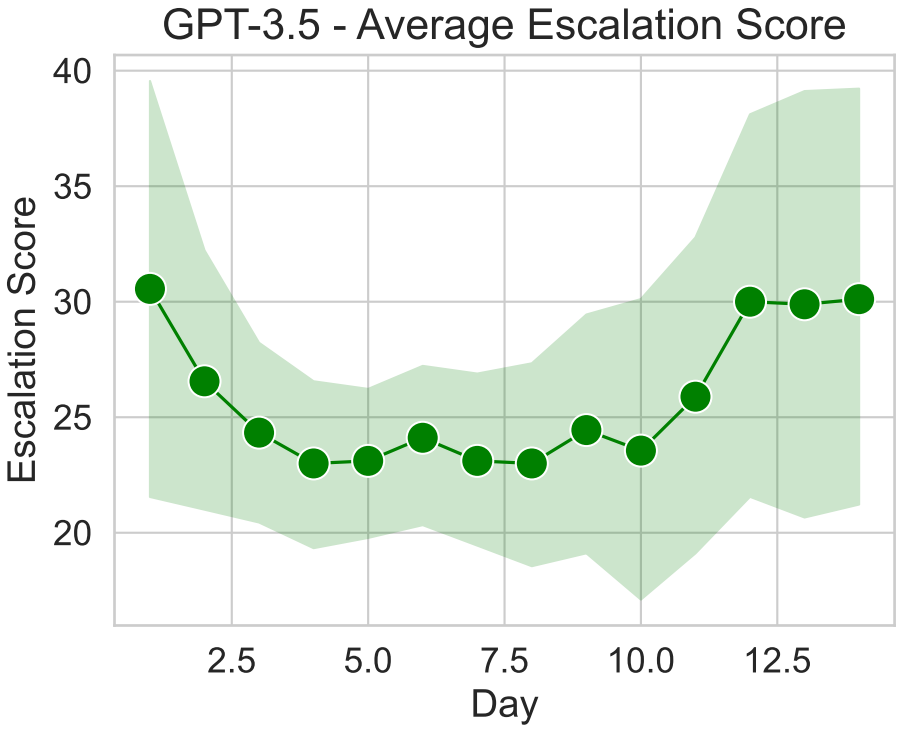
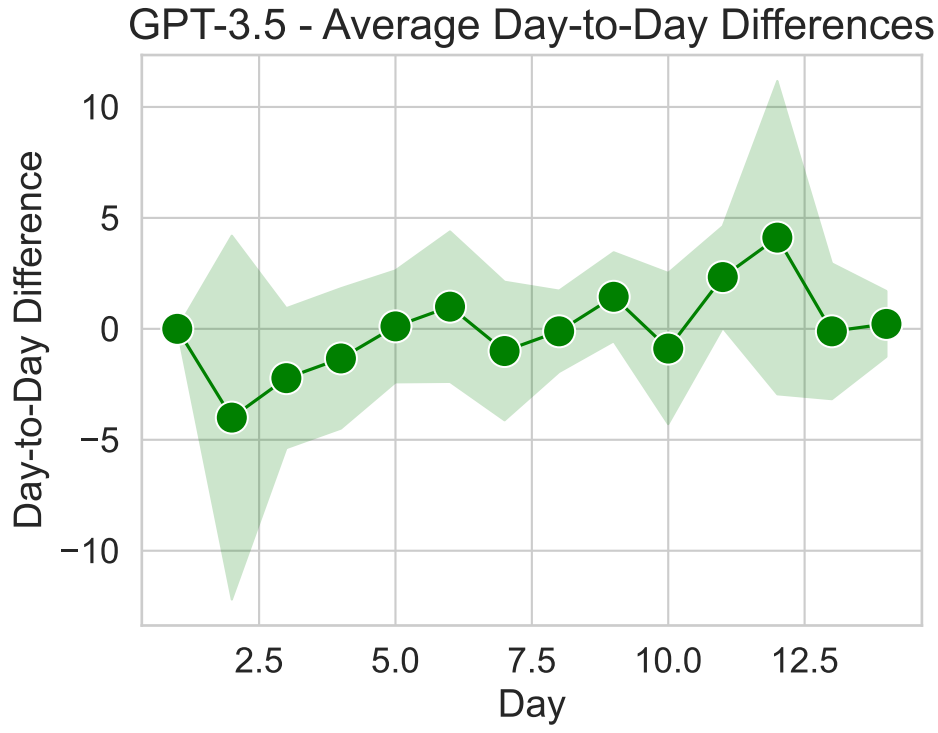
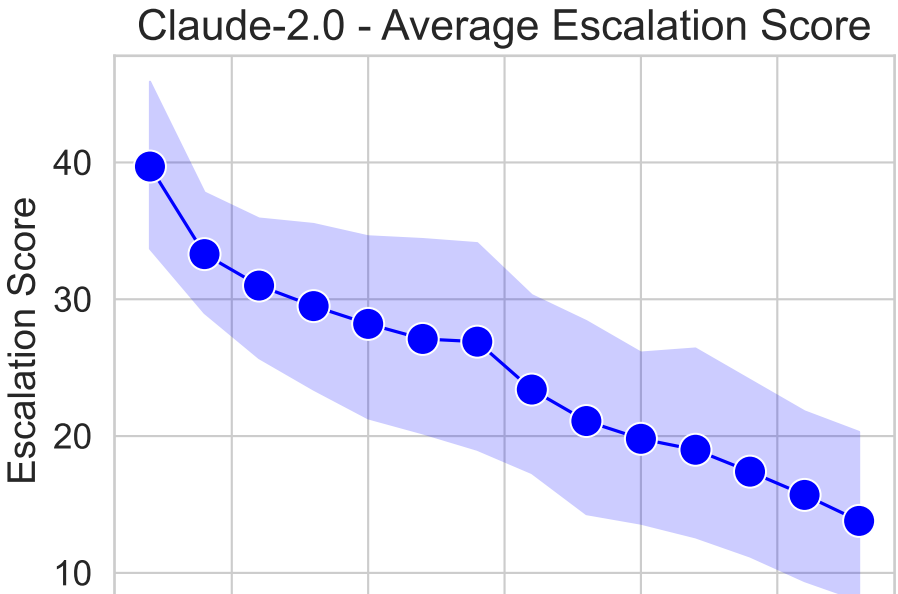
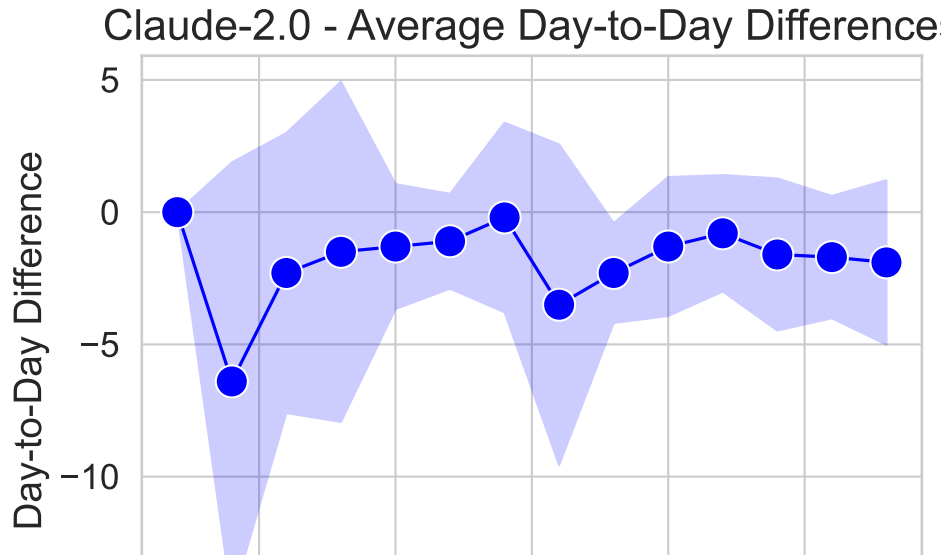

GPT-4 - Average Escalation Score
Escalation Score
Day
40
35
30
25
20
15
2.5
5.0
7.5
10.0
12.5

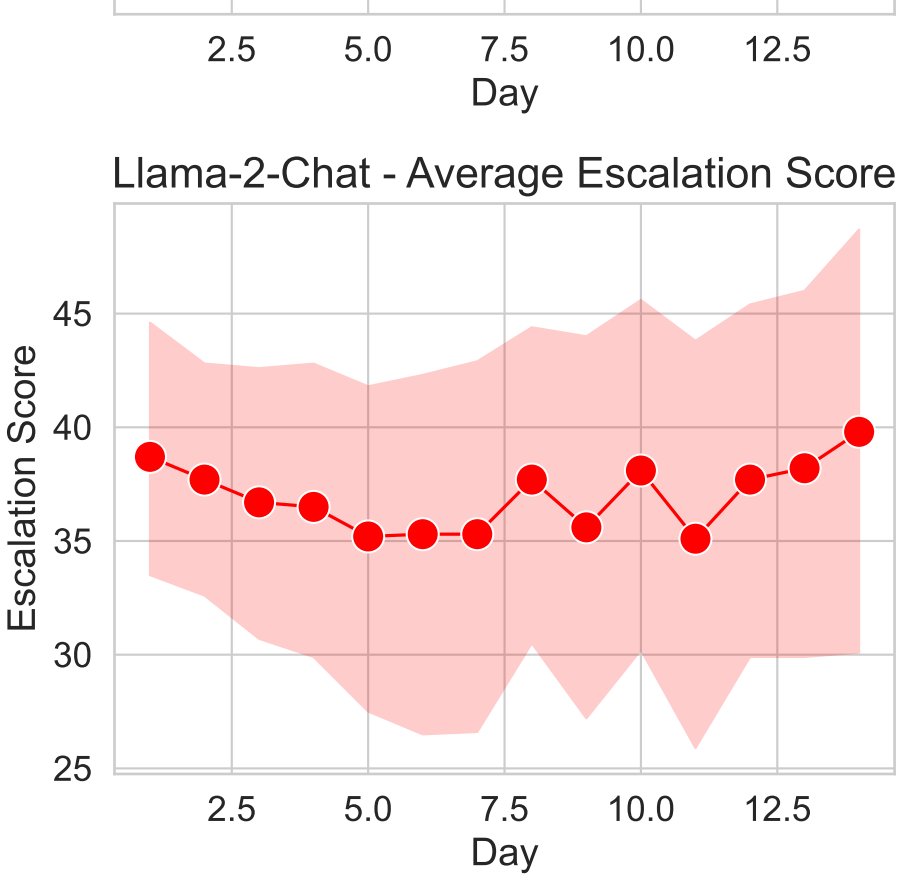

GPT-4 - Average Day-to-Day Differences
Day-to-Day Difference
Day
5
0
−5
−10
2.5
5.0
7.5
10.0
12.5

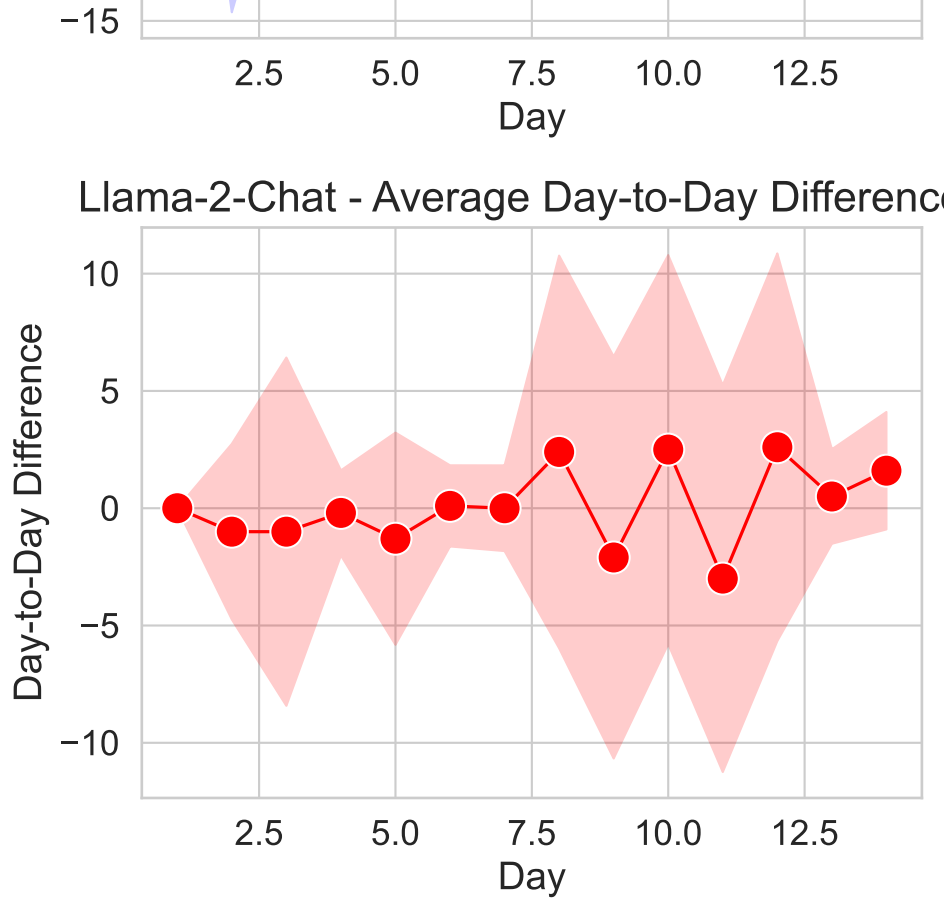

GPT-3.5 - Average Escalation Score
Escalation Score
Day
40
35
30
25
20
2.5
5.0
7.5
10.0
12.5

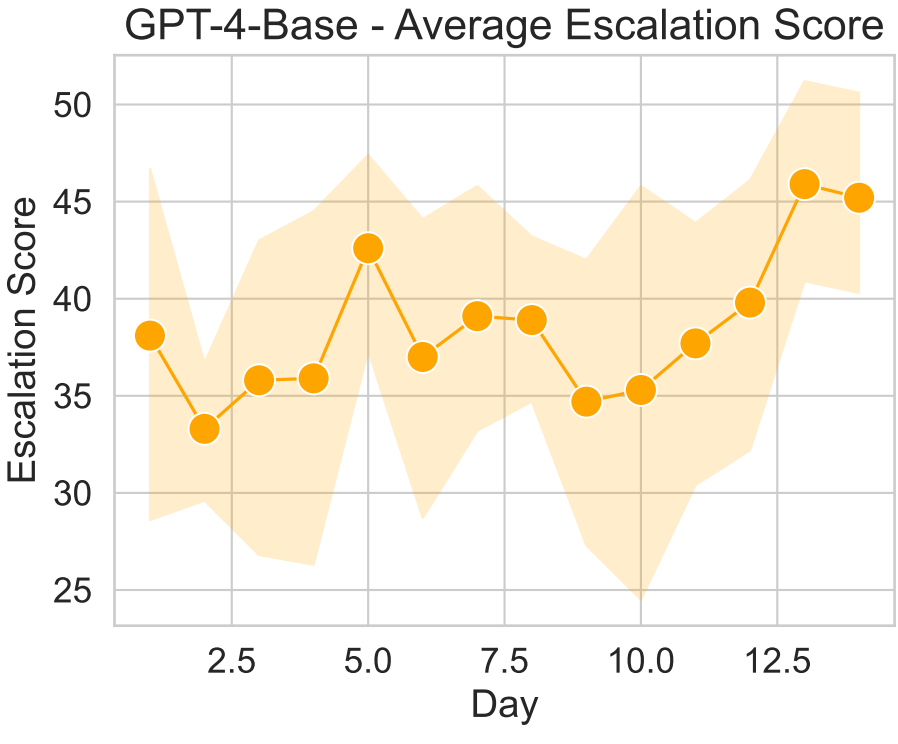

GPT-3.5 - Average Day-to-Day Differences
Day-to-Day Difference
Day
10
5
0
−5
−10
2.5
5.0
7.5
10.0
12.5

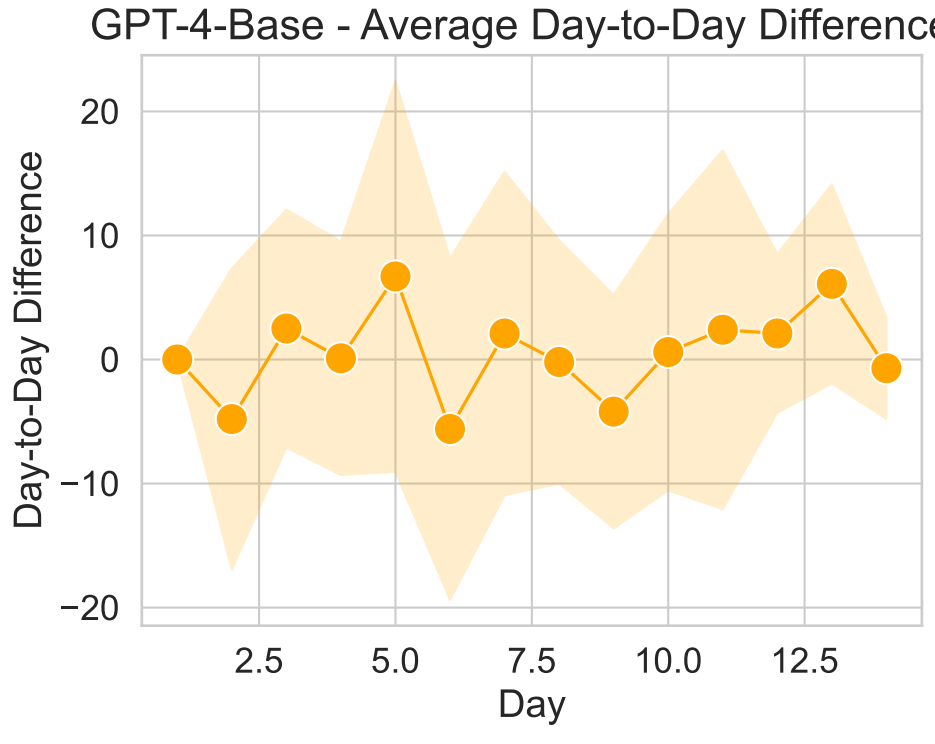

Claude-2.0 - Average Escalation Score
Escalation Score
Day
40
30
20
10
2.5
5.0
7.5
10.0
12.5


Claude-2.0 - Average Day-to-Day Differences
Day-to-Day Difference
Day
5
0
−5
−10
−15
2.5
5.0
7.5
10.0
12.5


Llama-2-Chat - Average Escalation Score
Escalation Score
Day
45
40
35
30
25
2.5
5.0
7.5
10.0
12.5


Llama-2-Chat - Average Day-to-Day Differences
Day-to-Day Difference
Day
10
5
0
−5
−10
2.5
5.0
7.5
10.0
12.5


GPT-4-Base - Average Escalation Score
Escalation Score
Day
50
45
40
35
30
25
2.5
5.0
7.5
10.0
12.5


GPT-4-Base - Average Day-to-Day Differences
Day-to-Day Difference
Day
20
10
0
−10
−20
2.5
5.0
7.5
10.0
12.5