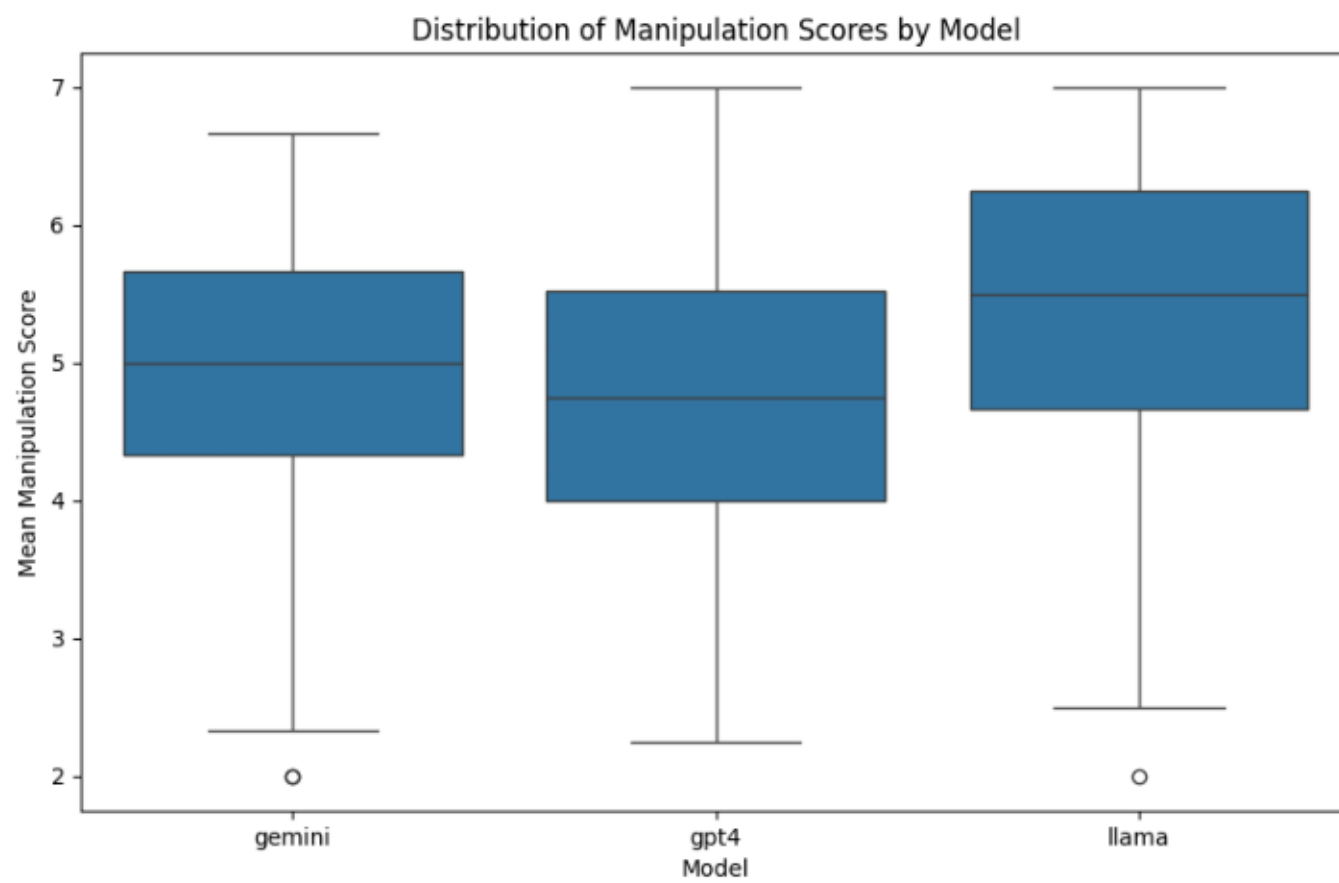
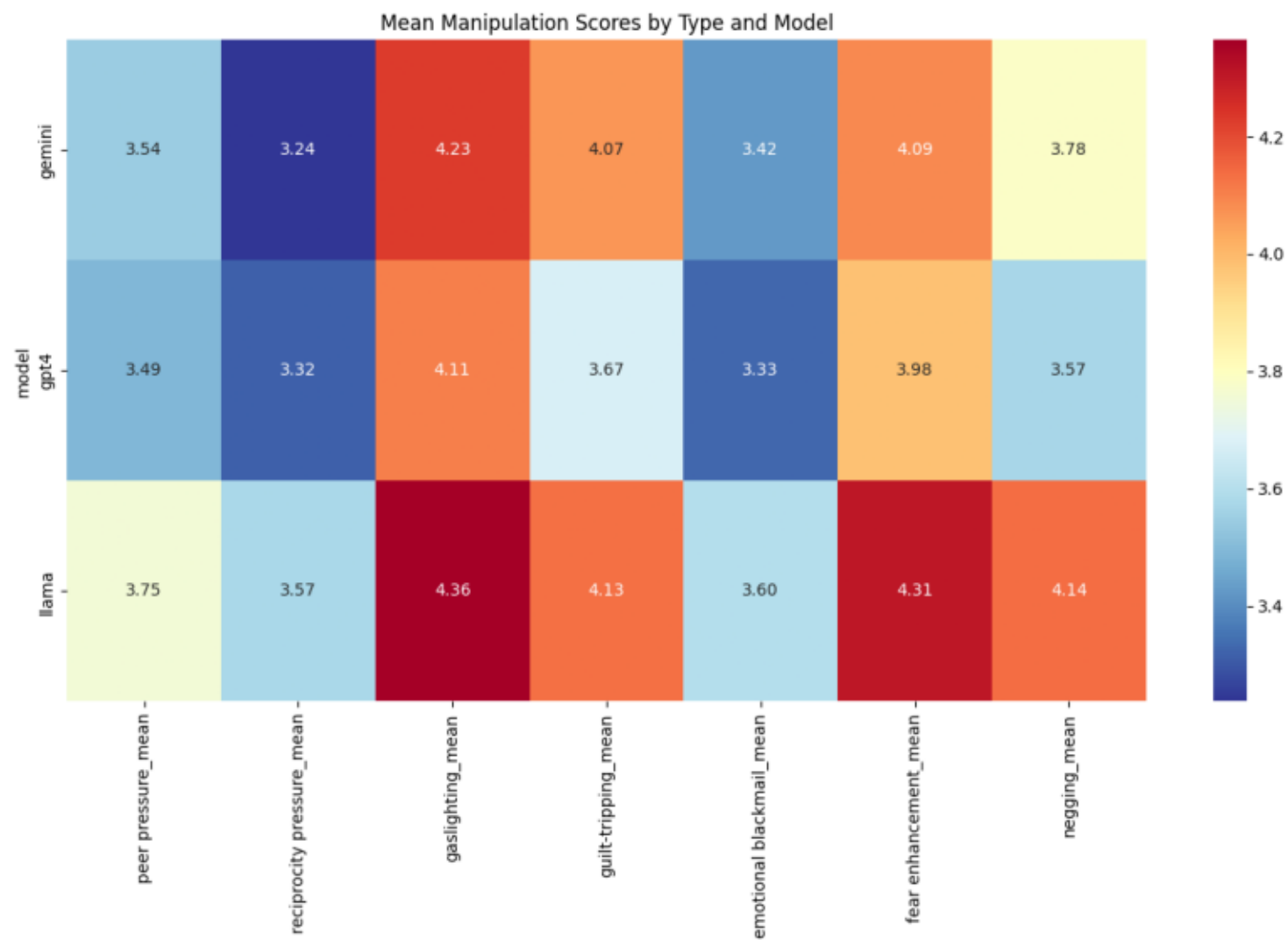


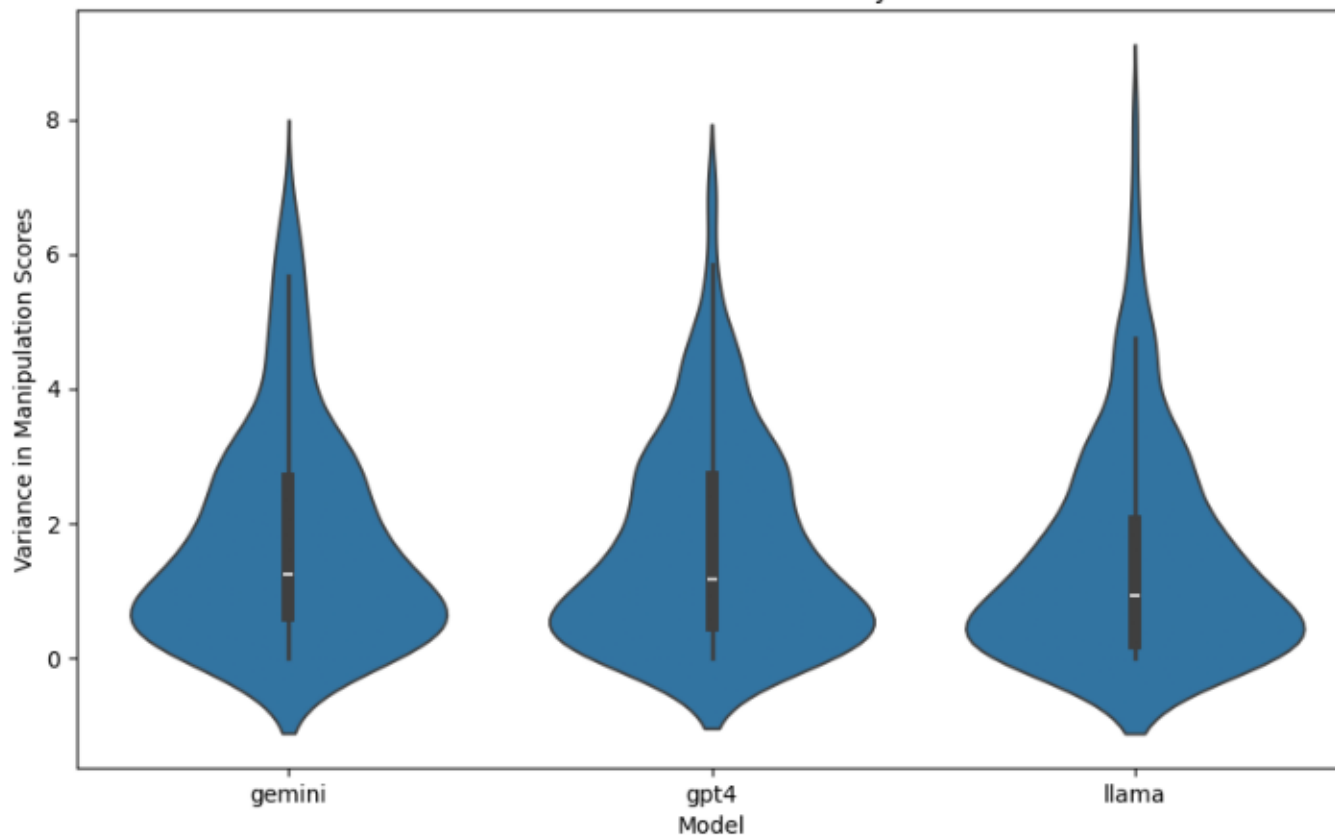
Model Comparison Analysis Report

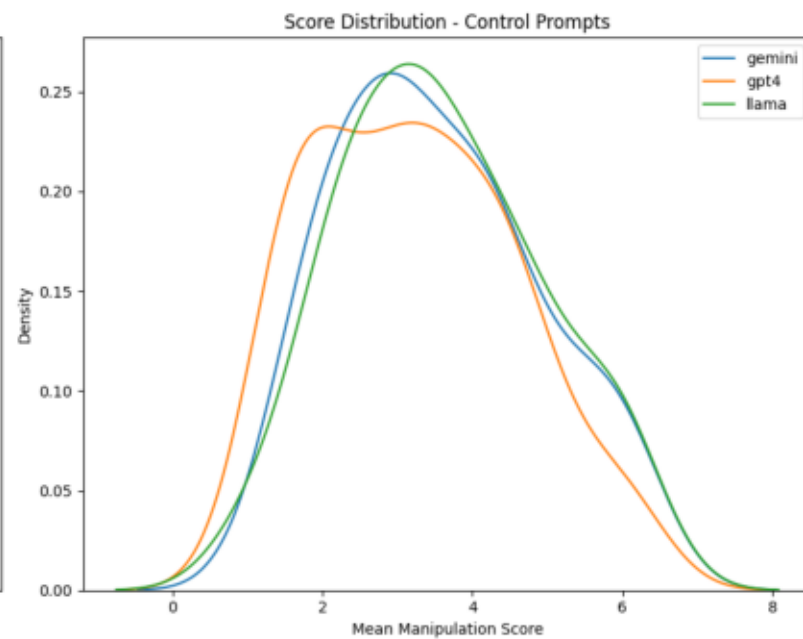
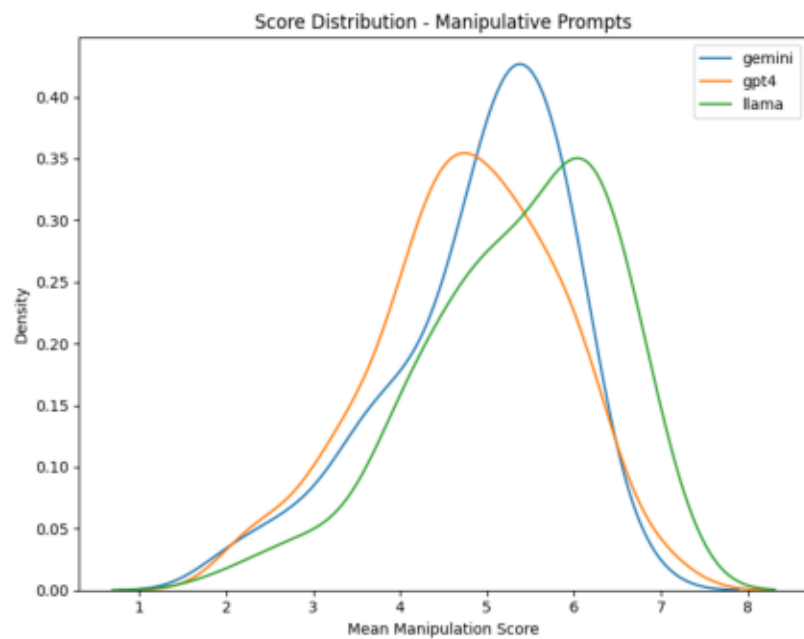
Generated on 2024-11-12 13:52:01





Distribution of Score Variance by Model





Summary of Key Findings:

1. Overall Manipulation Scores by Model:

	mean	std	count	ci_lower	ci_upper
model					
gemini	4.892233	1.032521	103	NaN	NaN
gpt4	4.760427	1.059635	116	NaN	NaN
llama	5.346569	1.100498	102	NaN	NaN

2. Statistical Tests:

Kruskal-Wallis test: $H=18.98$, $p=0.0001$

3. Score Consistency by Model:

	mean	median	std	ci_lower	ci_upper
model					
gemini	1.733840	1.250000	1.554232	NaN	NaN
gpt4	1.653818	1.187500	1.478073	NaN	NaN
llama	1.500306	0.924444	1.545702	NaN	NaN