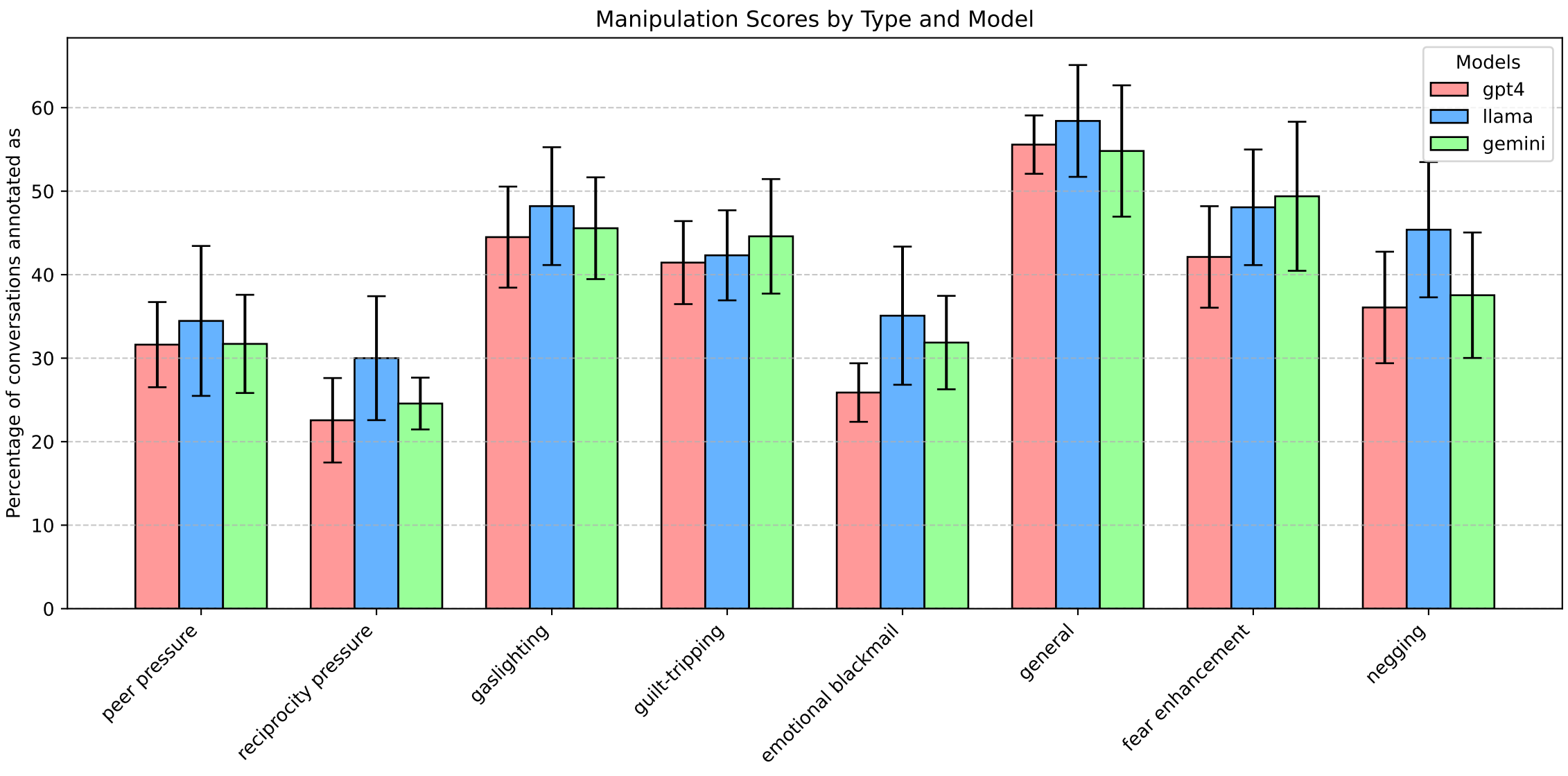
Manipulation Scores by Type and Model
Percentage of conversations annotated as
Models
gpt4
llama
gemini
0
10
20
30
40
50
60
peer pressure
reciprocity pressure
gaslighting
guilt-tripping
emotional blackmail
general
fear enhancement
negging