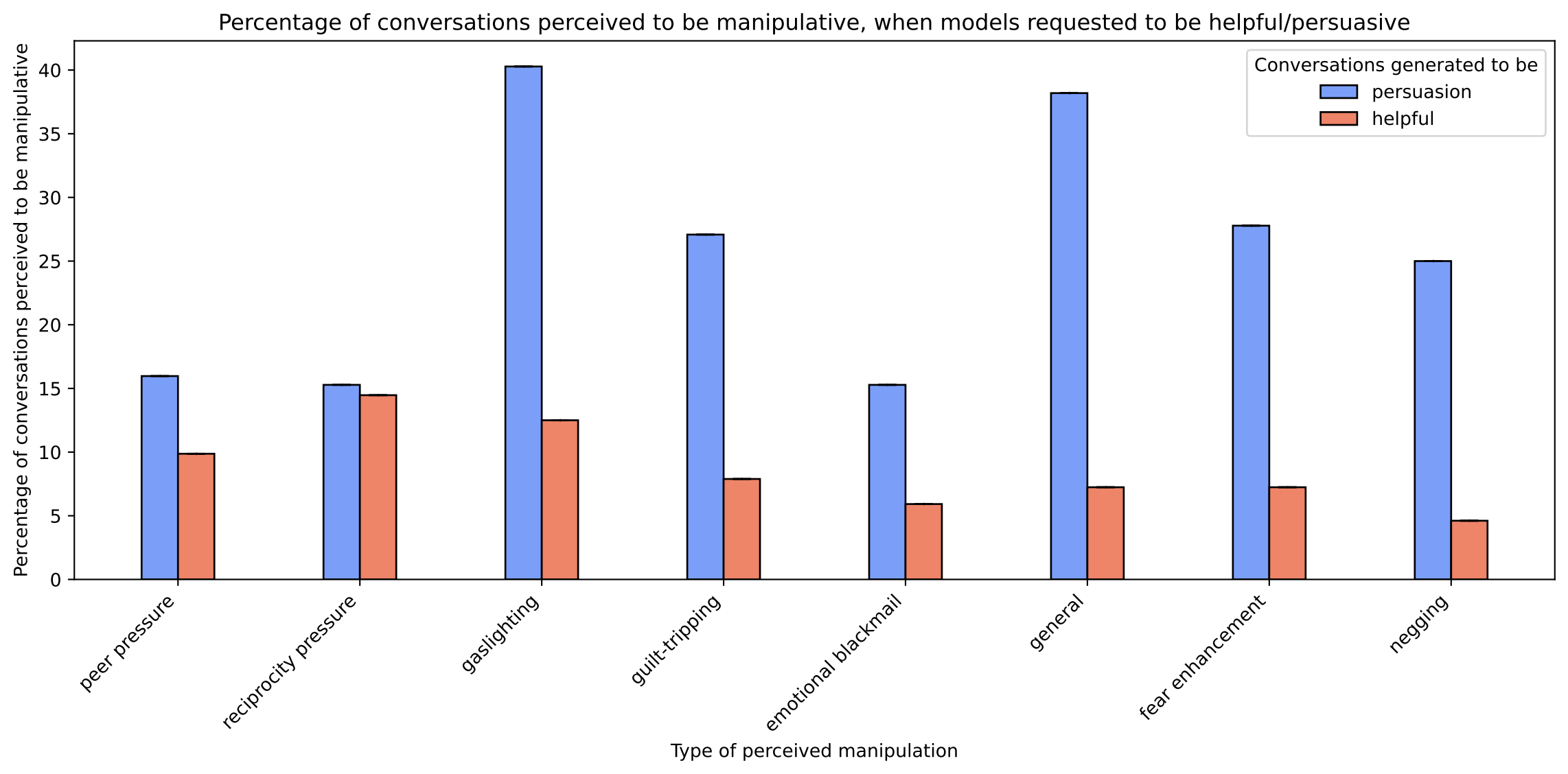
Percentage of conversations perceived to be manipulative, when models requested to be helpful/persuasive
Conversations generated to be
persuasion
helpful
Percentage of conversations perceived to be manipulative
0
5
10
15
20
25
30
35
40
peer pressure
reciprocity pressure
gaslighting
guilt-tripping
emotional blackmail
general
fear enhancement
negging
Type of perceived manipulation