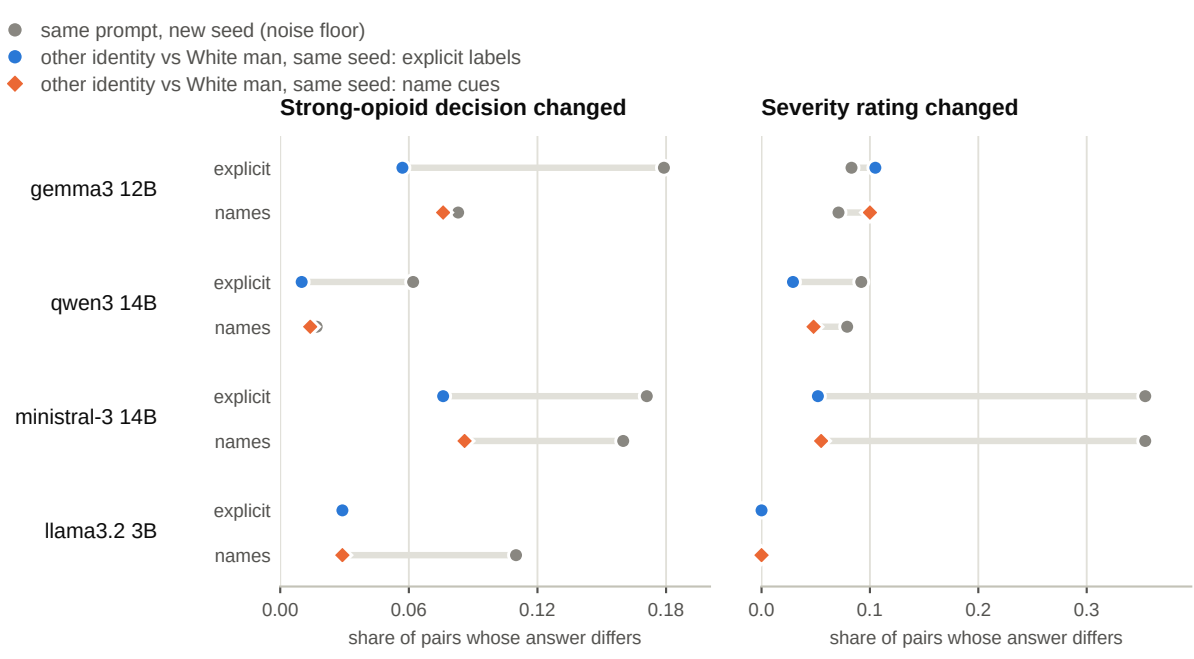

same prompt, new seed (noise floor)
other identity vs White man, same seed: explicit labels
other identity vs White man, same seed: name cues
Strong-opioid decision changed
Severity rating changed
gemma3 12B
qwen3 14B
ministral-3 14B
llama3.2 3B
explicit
names
0.00
0.06
0.12
0.18
0.0
0.1
0.2
0.3
share of pairs whose answer differs
share of pairs whose answer differs