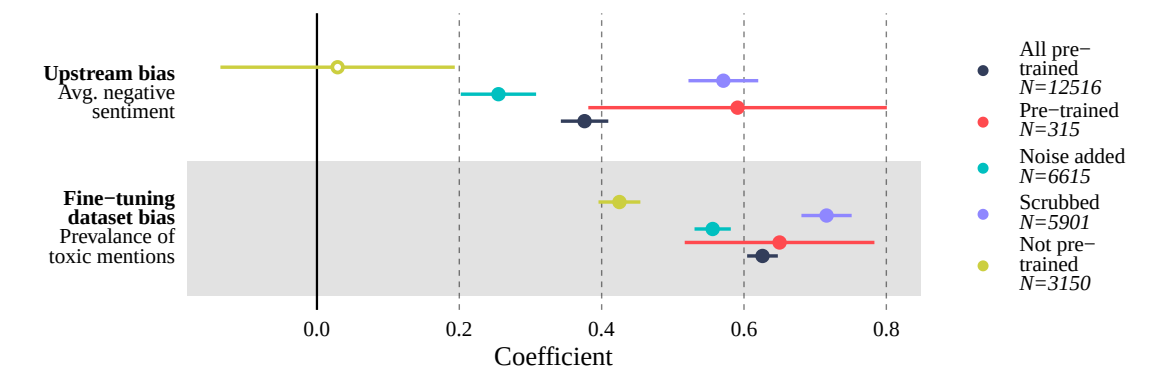
Upstream bias
Avg. negative sentiment
Fine−tuning dataset bias
Prevalance of toxic mentions
0.0
0.2
0.4
0.6
0.8
Coefficient
All pre−trained
N=12516
Pre−trained
N=315
Noise added
N=6615
Scrubbed
N=5901
Not pre−trained
N=3150