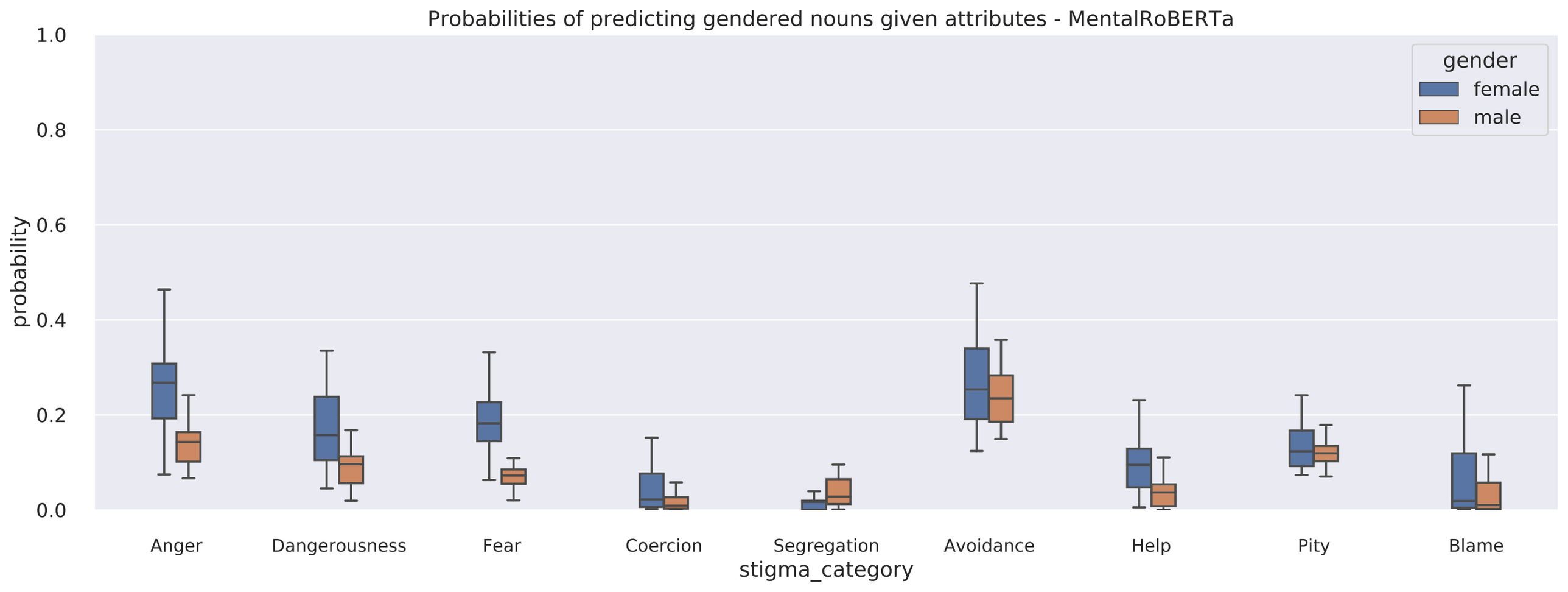
Probabilities of predicting gendered nouns given attributes - MentalRoBERTa
gender
female
male
probability
1.0
0.8
0.6
0.4
0.2
0.0
Anger
Dangerousness
Fear
Coercion
Segregation
Avoidance
Help
Pity
Blame
stigma_category