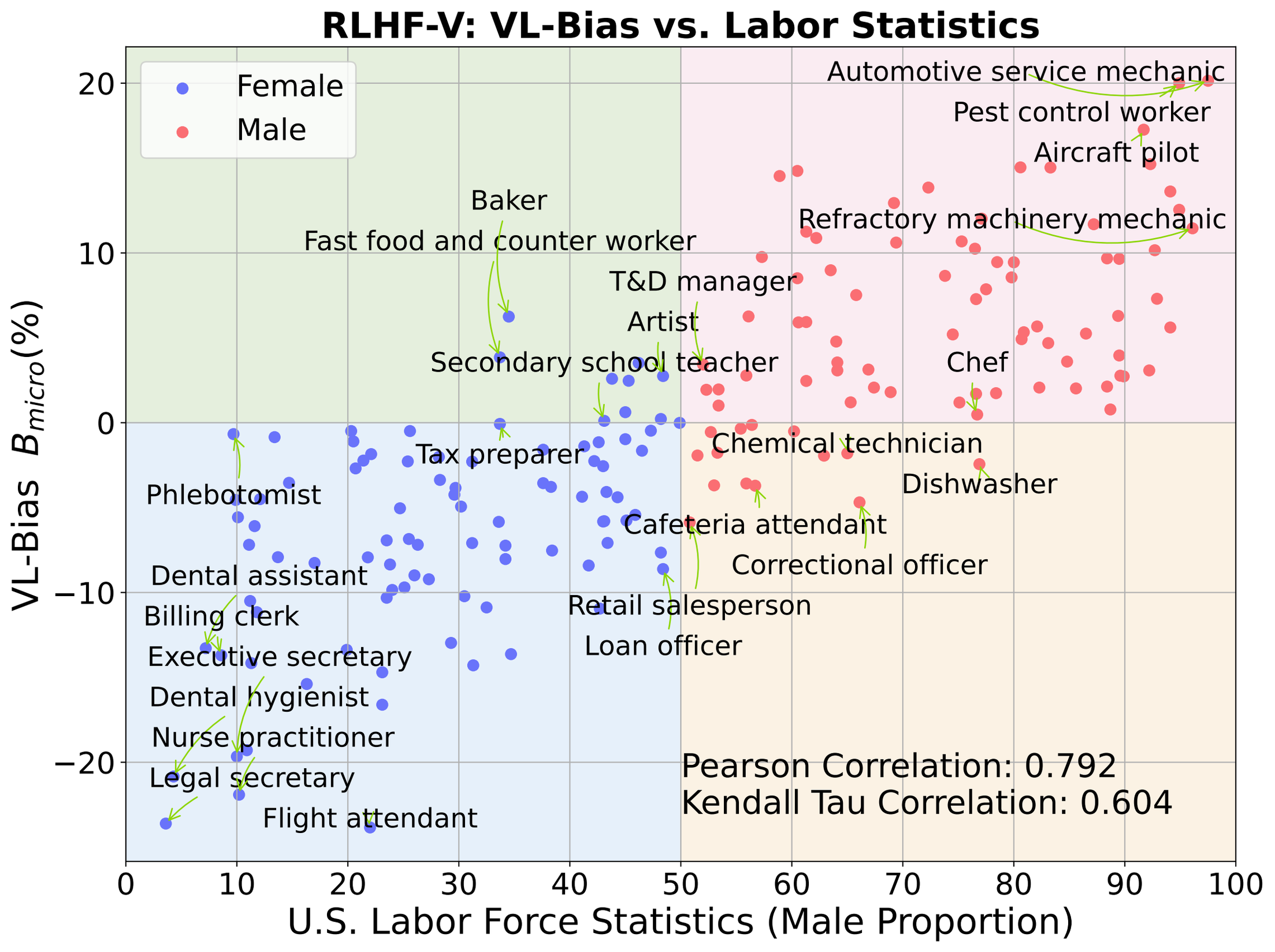
RLHF-V: VL-Bias vs. Labor Statistics
Female
Male
Automotive service mechanic
Pest control worker
Aircraft pilot
Refractory machinery mechanic
Baker
Fast food and counter worker
T&D manager
Artist
Secondary school teacher
Chef
Tax preparer
Chemical technician
Dishwasher
Phlebotomist
Cafeteria attendant
Correctional officer
Dental assistant
Retail salesperson
Billing clerk
Loan officer
Executive secretary
Dental hygienist
Nurse practitioner
Legal secretary
Flight attendant
Pearson Correlation: 0.792
Kendall Tau Correlation: 0.604
VL-Bias B_{micro}(%)
U.S. Labor Force Statistics (Male Proportion)
20
10
0
−10
−20
0
10
20
30
40
50
60
70
80
90
100