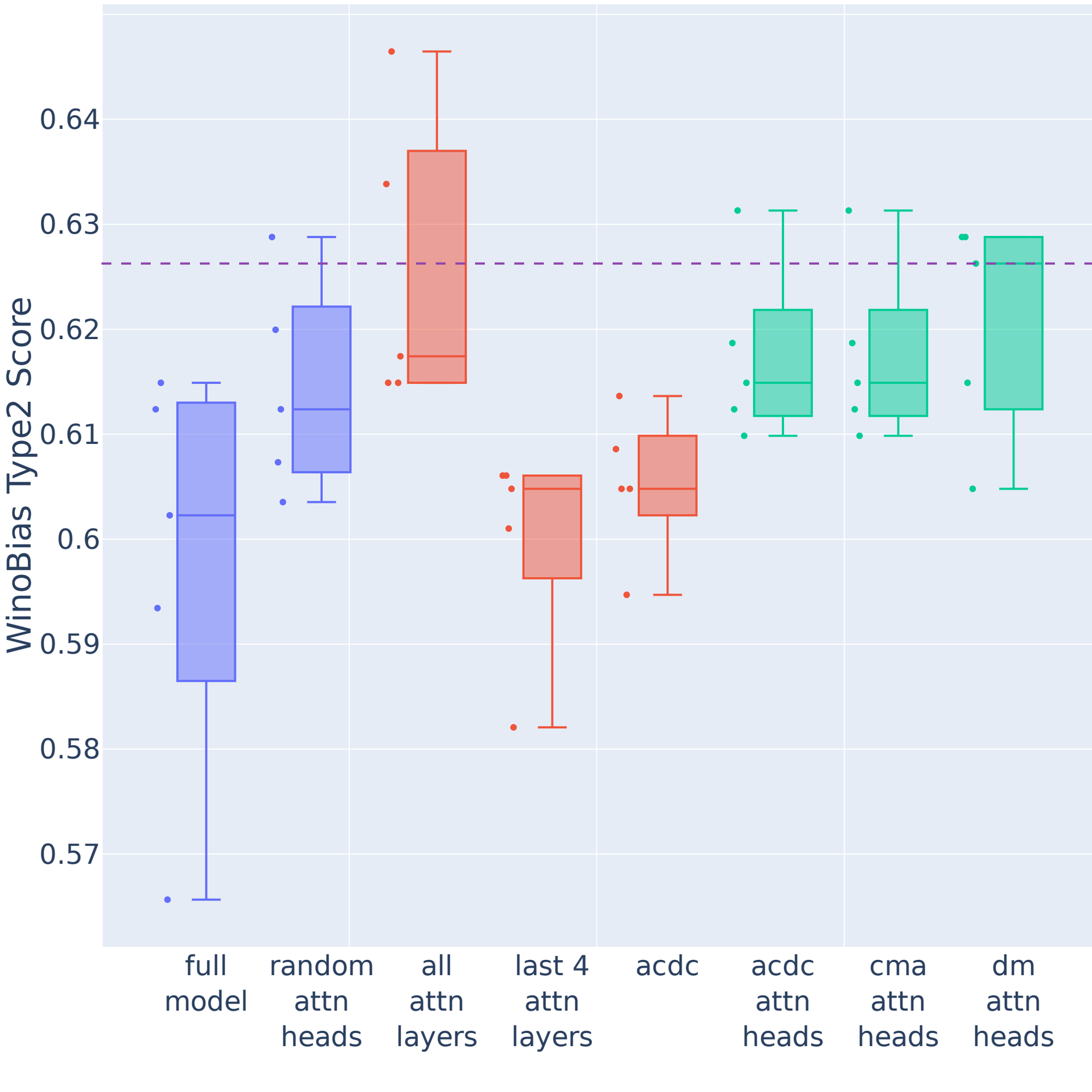
WinoBias Type2 Score
0.64
0.63
0.62
0.61
0.6
0.59
0.58
0.57
full model
random attn heads
all attn layers
last 4 attn layers
acdc
acdc attn heads
cma attn heads
dm attn heads