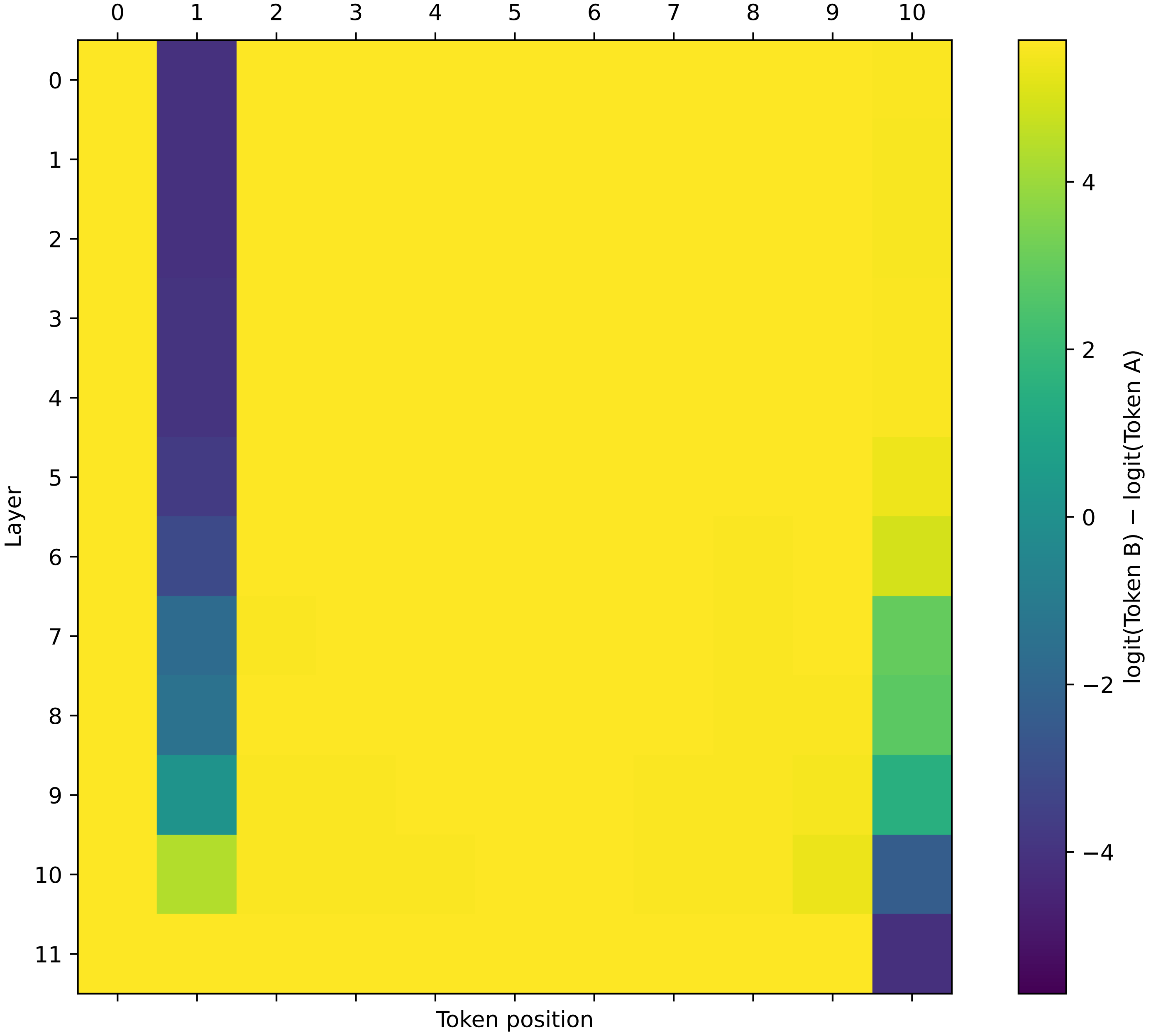
Logit difference heatmap: logit(' Smith') − logit(' Jones')
clean=-4.124   corrupt=5.656


Layer
Token position
logit(Token B) − logit(Token A)