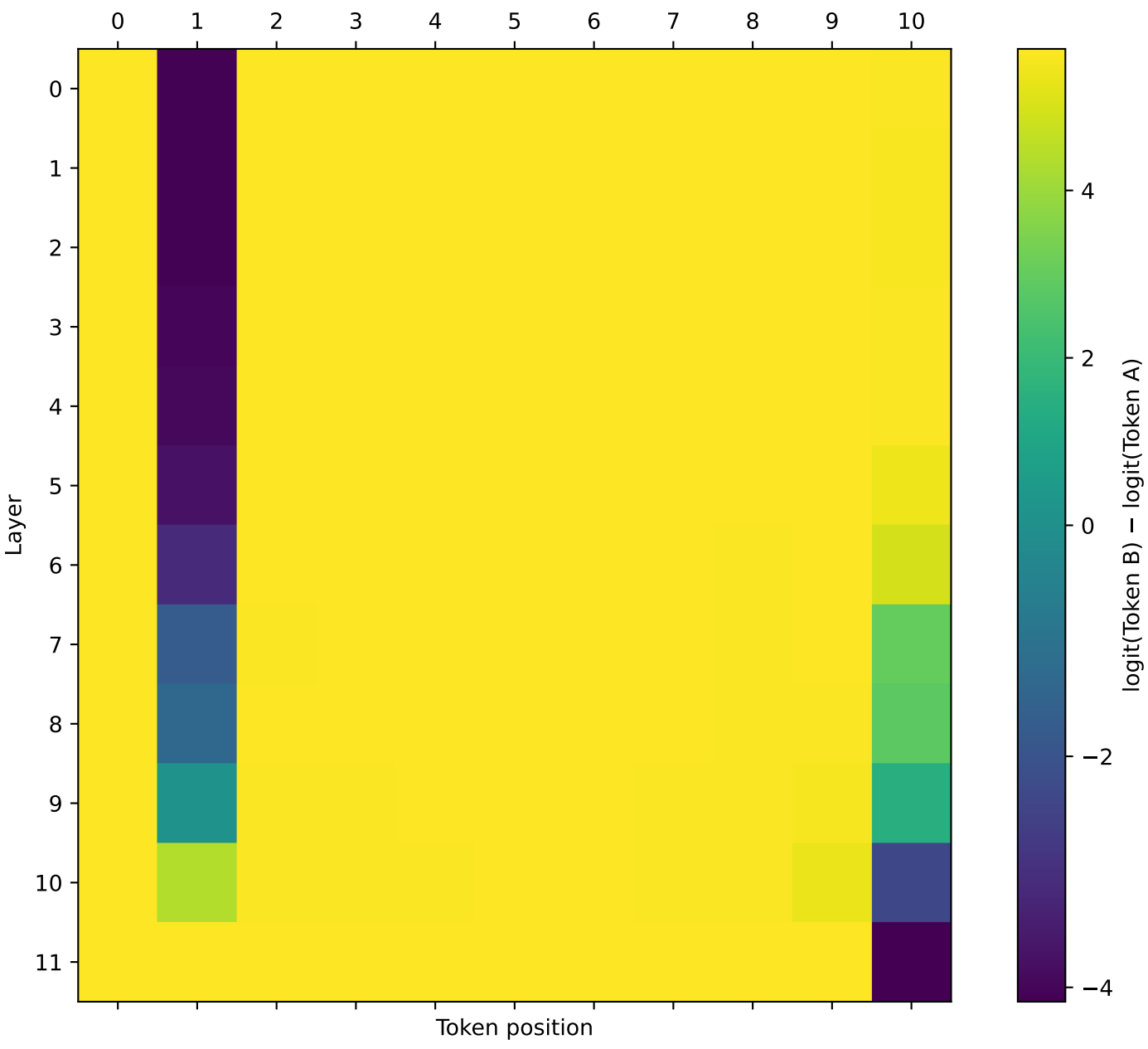
Logit difference heatmap: logit(' Smith') − logit(' Jones')
clean=-4.124 corrupt=5.656
0
1
2
3
4
5
6
7
8
9
10
Layer
Token position
4
2
0
−2
−4
logit(Token B) − logit(Token A)