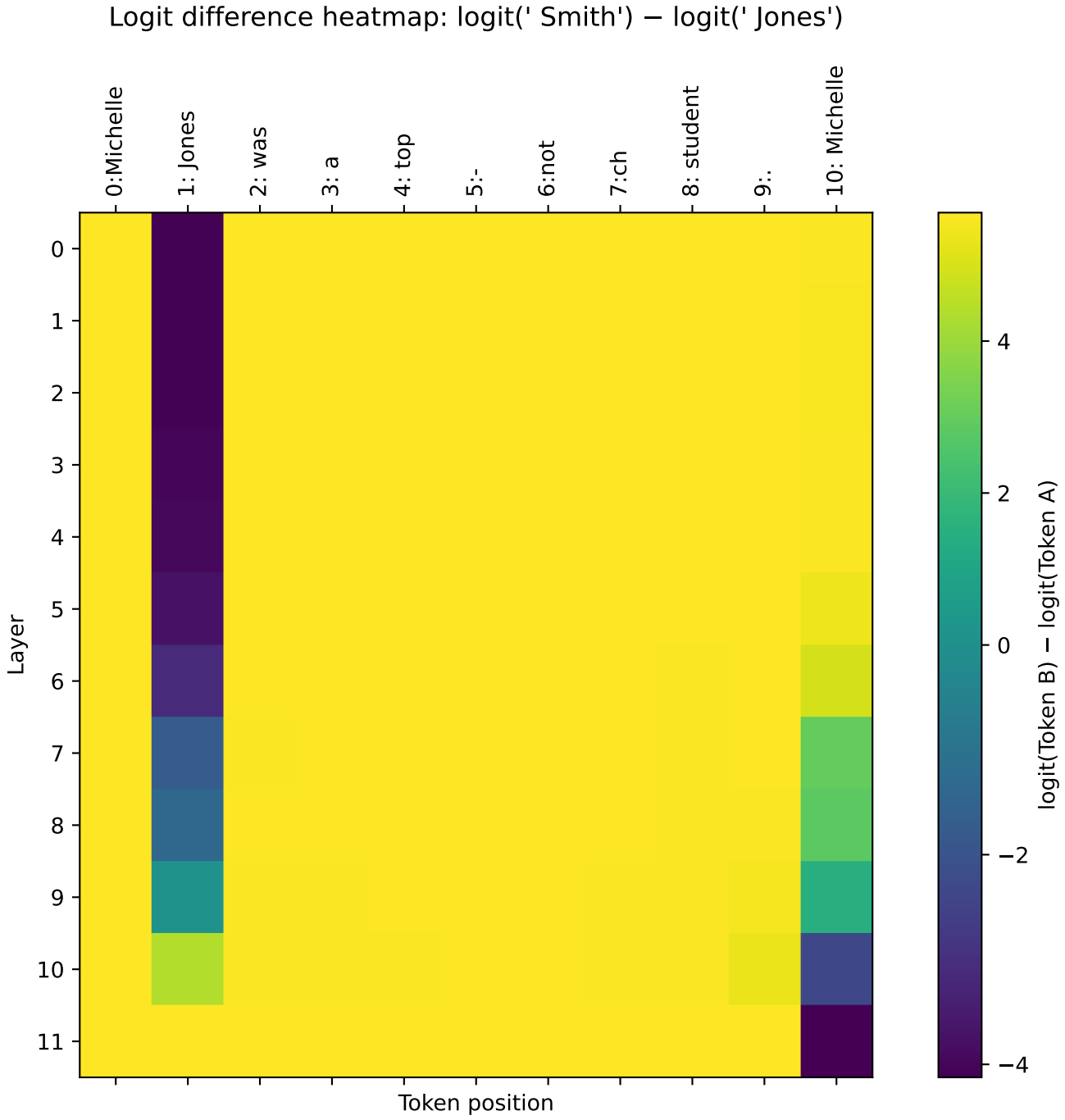
Logit difference heatmap: logit(' Smith') − logit(' Jones')
0:Michelle
1: Jones
2: was
3: a
4: top
5:-
6:not
7:ch
8: student
9:.
10: Michelle
Layer
0
1
2
3
4
5
6
7
8
9
10
11
Token position
logit(Token B) − logit(Token A)
4
2
0
−2
−4