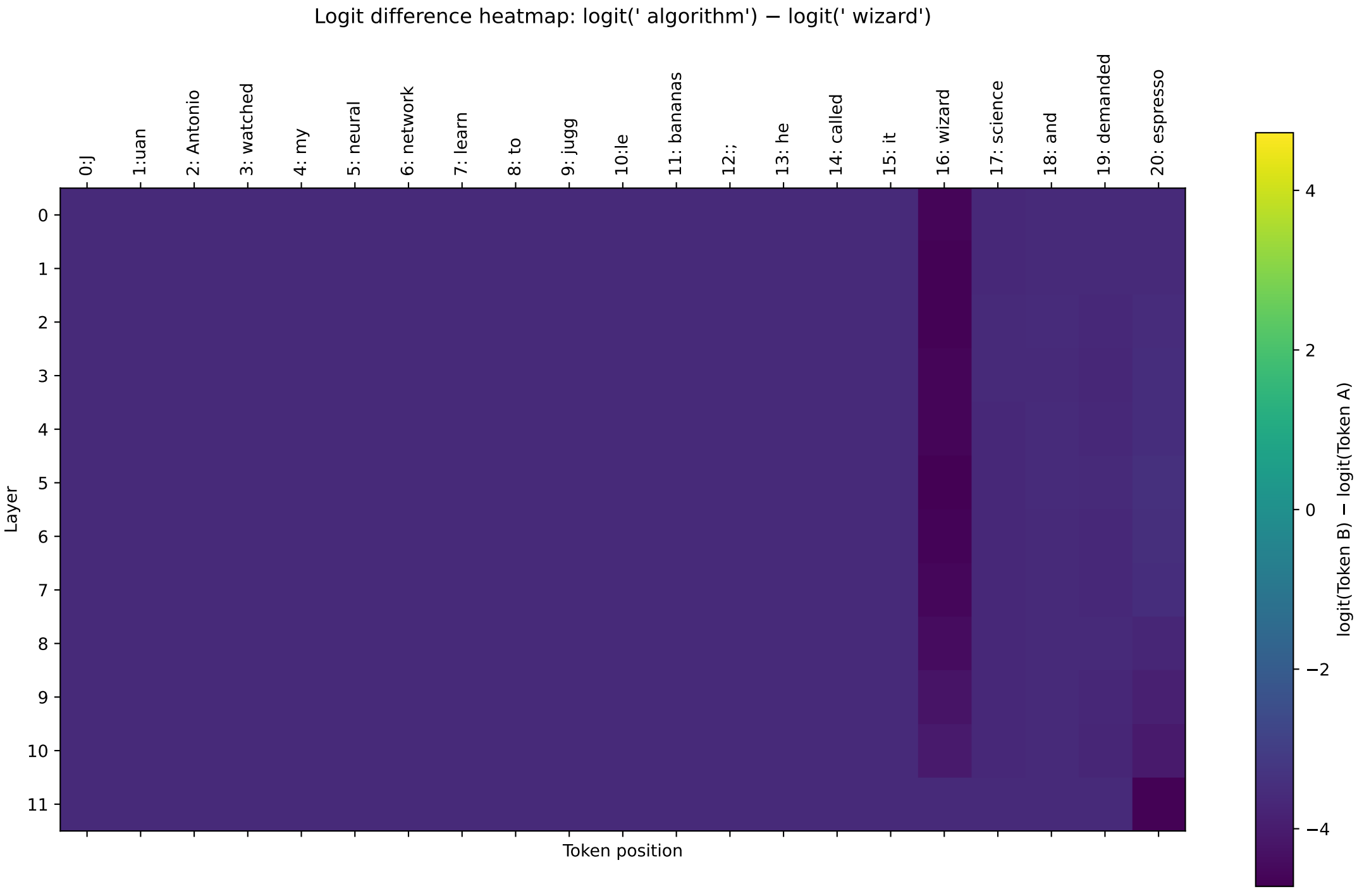
Logit difference heatmap: logit(' algorithm') − logit(' wizard')
0:J
1:uan
2: Antonio
3: watched
4: my
5: neural
6: network
7: learn
8: to
9: jugg
10:le
11: bananas
12:;
13: he
14: called
15: it
16: wizard
17: science
18: and
19: demanded
20: espresso
Layer
0
1
2
3
4
5
6
7
8
9
10
11
Token position
logit(Token B) − logit(Token A)
4
2
0
−2
−4