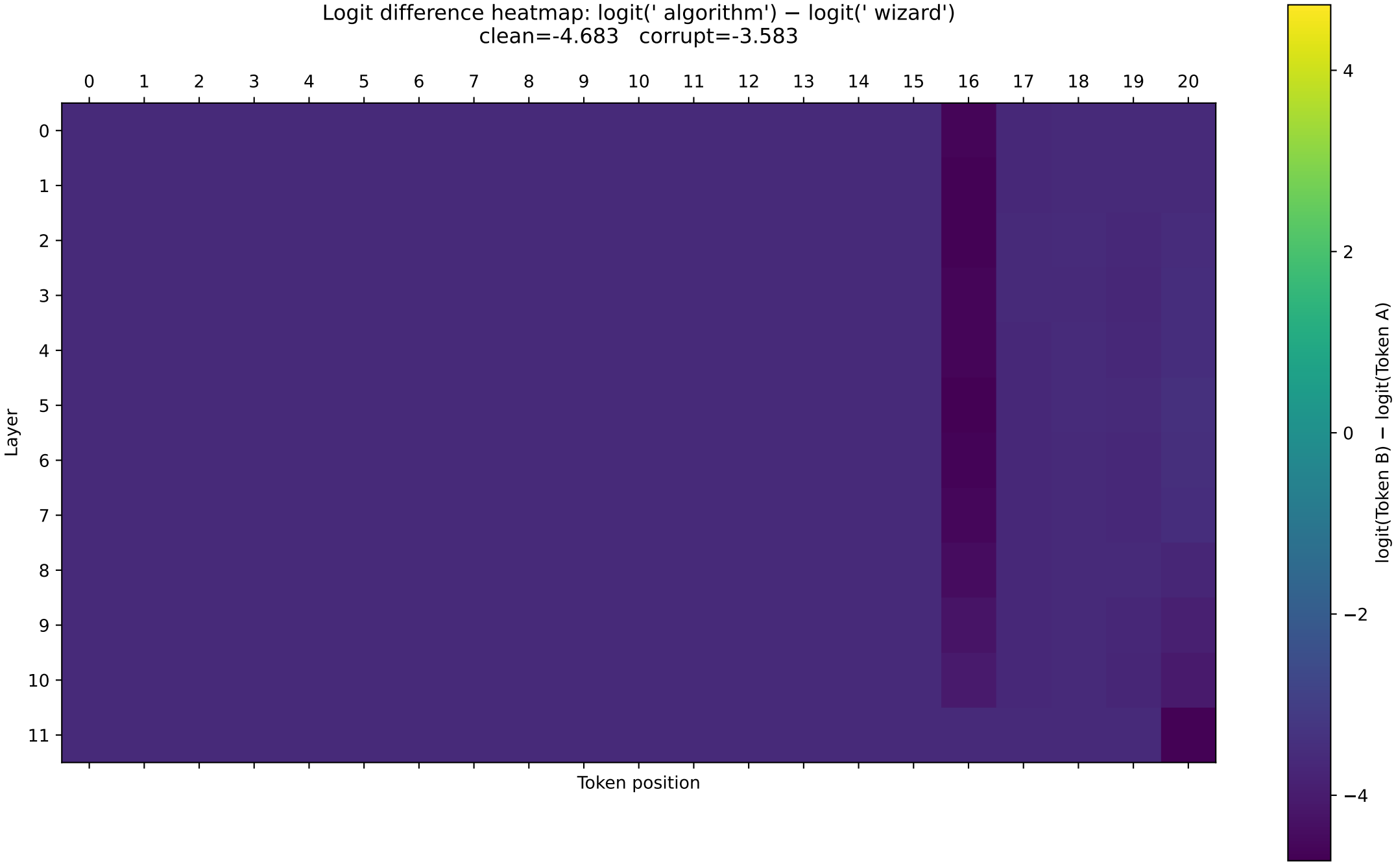
Logit difference heatmap: logit(' algorithm') − logit(' wizard')
clean=-4.683 corrupt=-3.583
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Layer
0 1 2 3 4 5 6 7 8 9 10 11
Token position
4 2 0 −2 −4
logit(Token B) − logit(Token A)