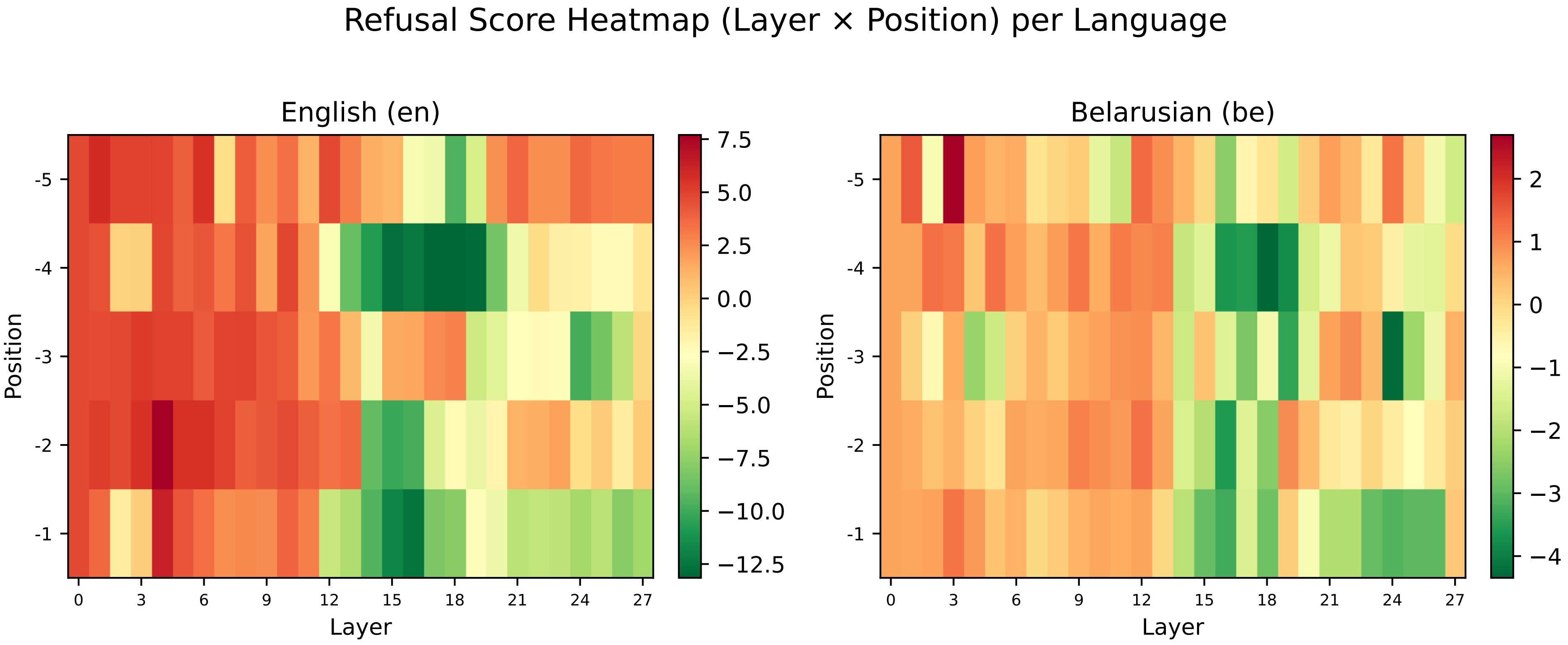
Refusal Score Heatmap (Layer × Position) per Language
English (en)
Belarusian (be)
Position
-5
-4
-3
-2
-1
Layer
0
3
6
9
12
15
18
21
24
27
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
2
1
0
−1
−2
−3
−4