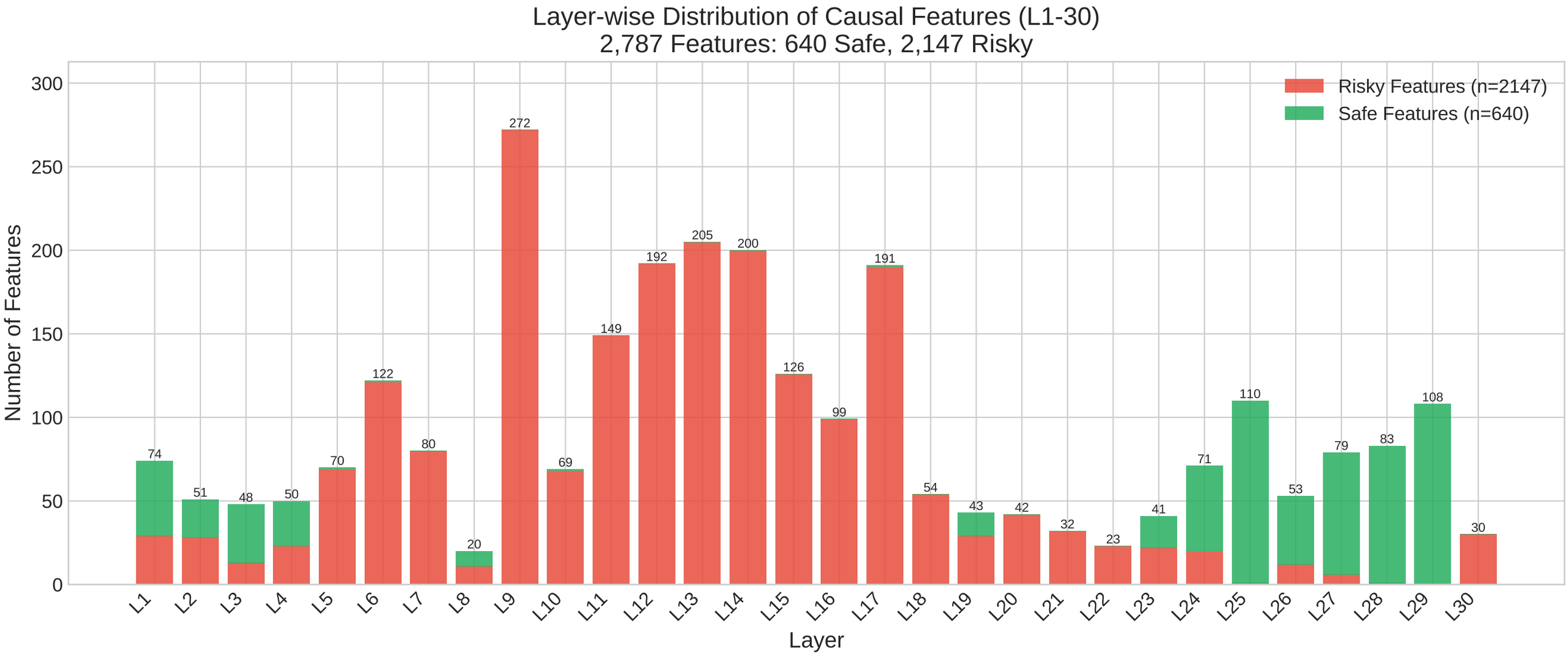
Layer-wise Distribution of Causal Features (L1-30)
2,787 Features: 640 Safe, 2,147 Risky
Risky Features (n=2147)
Safe Features (n=640)
Number of Features
Layer
0
50
100
150
200
250
300
L1 74
L2 51
L3 48
L4 50
L5 70
L6 122
L7 80
L8 20
L9 272
L10 69
L11 149
L12 192
L13 205
L14 200
L15 126
L16 99
L17 191
L18 54
L19 43
L20 42
L21 32
L22 23
L23 41
L24 71
L25 110
L26 53
L27 79
L28 83
L29 108
L30 30