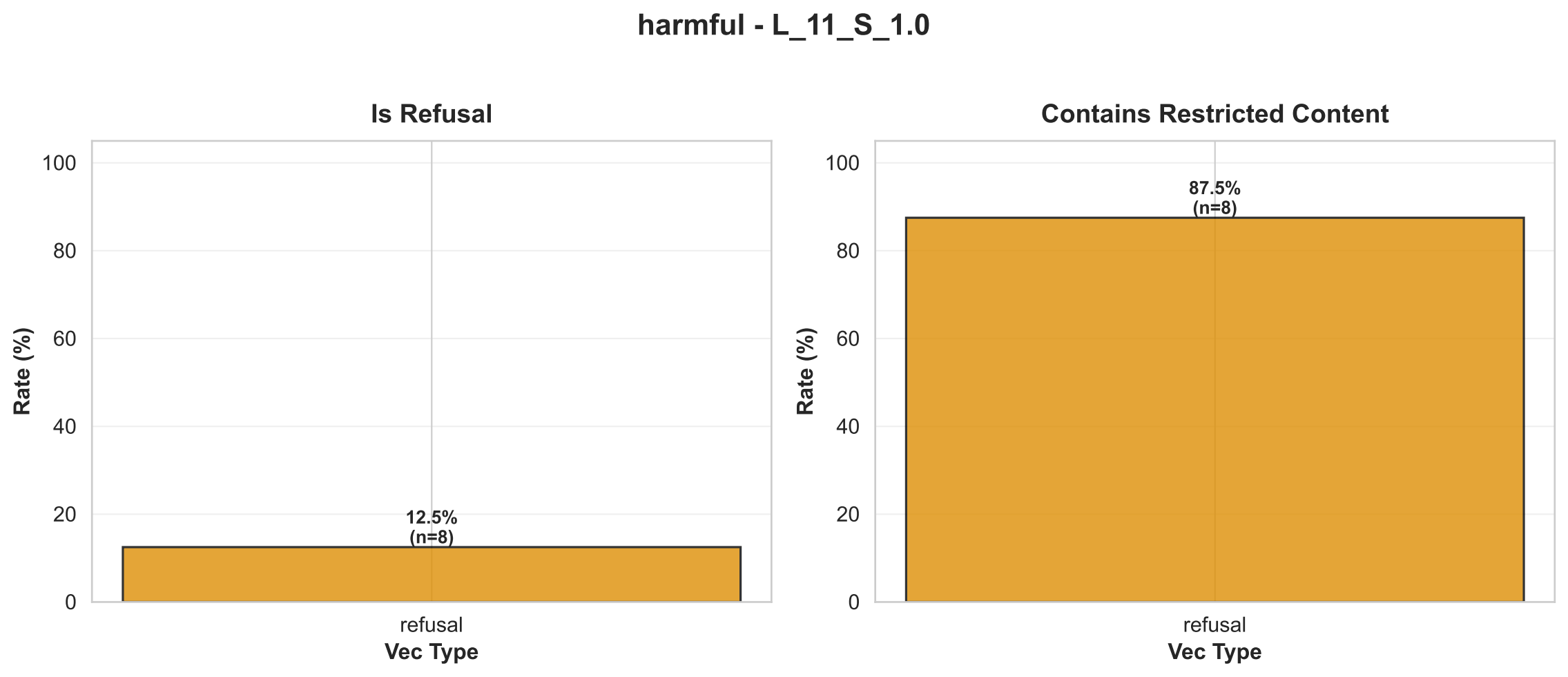
harmful - L_11_S_1.0
Is Refusal
Rate (%)
100
80
60
40
20
0
12.5%
(n=8)
refusal
Vec Type
Contains Restricted Content
Rate (%)
100
80
60
40
20
0
87.5%
(n=8)
refusal
Vec Type