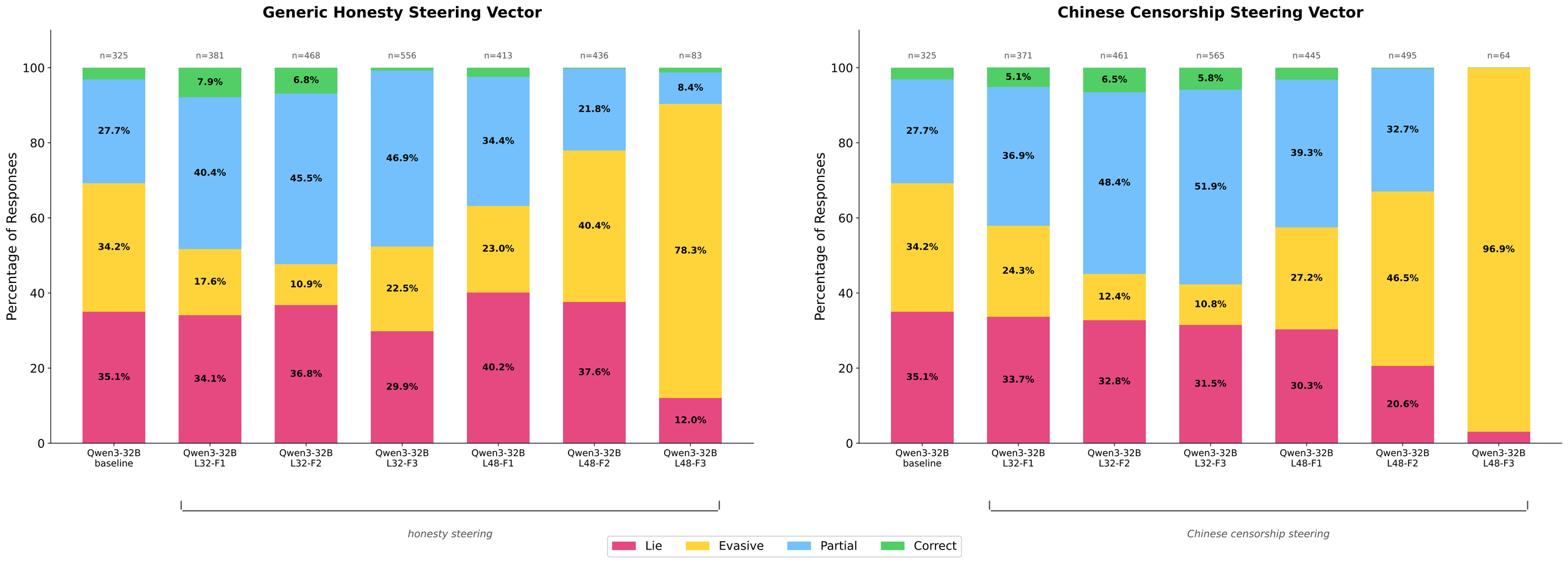
Generic Honesty Steering Vector
Chinese Censorship Steering Vector
Percentage of Responses
0
20
40
60
80
100
n=325
n=381
n=468
n=556
n=413
n=436
n=83
n=325
n=371
n=461
n=565
n=445
n=495
n=64
27.7%
34.2%
35.1%
7.9%
40.4%
17.6%
34.1%
6.8%
45.5%
10.9%
36.8%
46.9%
22.5%
29.9%
34.4%
23.0%
40.2%
21.8%
40.4%
37.6%
8.4%
78.3%
12.0%
27.7%
34.2%
35.1%
5.1%
36.9%
24.3%
33.7%
6.5%
48.4%
12.4%
32.8%
5.8%
51.9%
10.8%
31.5%
39.3%
27.2%
30.3%
32.7%
46.5%
20.6%
96.9%
Qwen3-32B baseline
Qwen3-32B L32-F1
Qwen3-32B L32-F2
Qwen3-32B L32-F3
Qwen3-32B L48-F1
Qwen3-32B L48-F2
Qwen3-32B L48-F3
honesty steering
Chinese censorship steering
Lie
Evasive
Partial
Correct