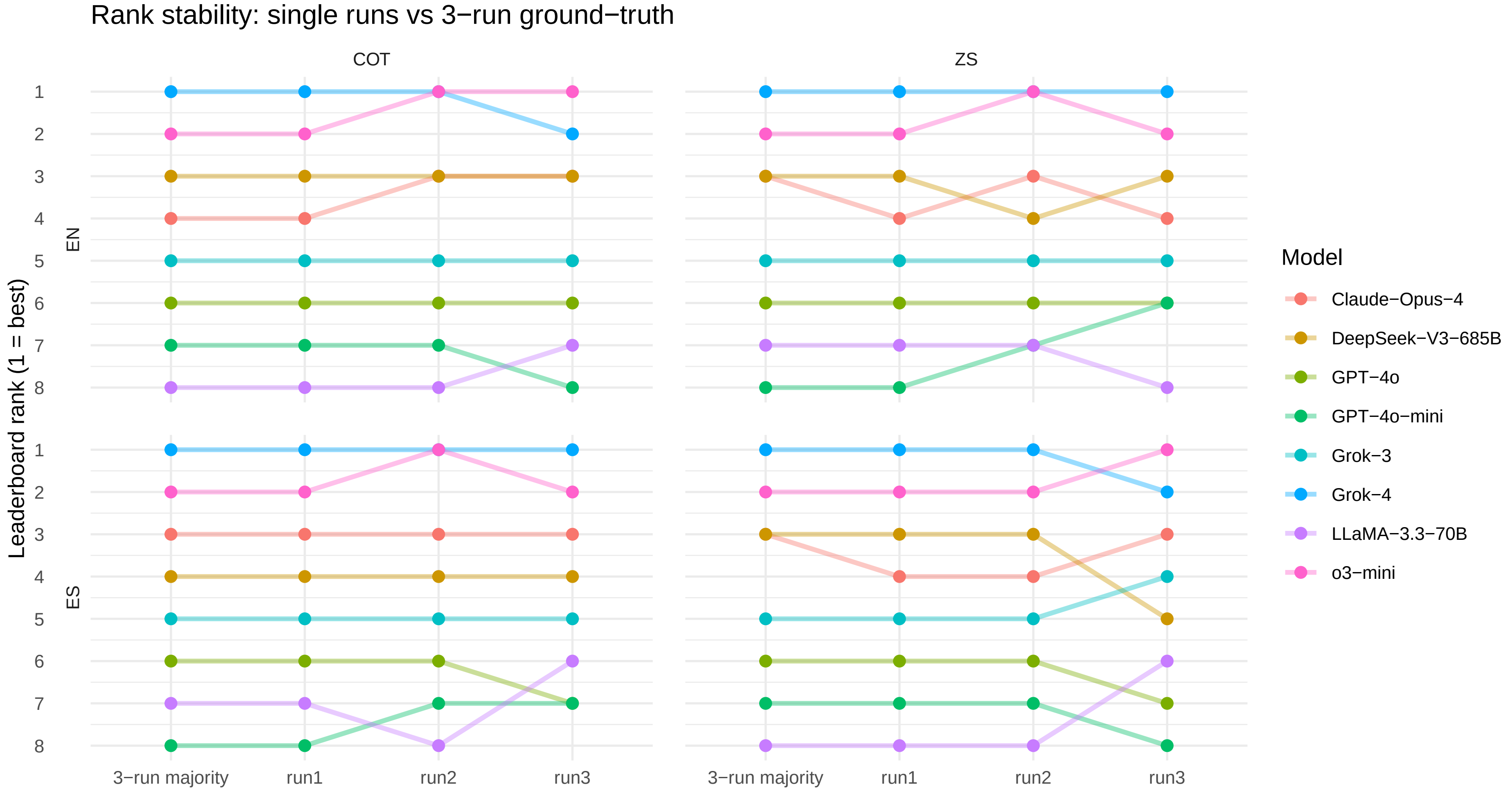Rank stability: single runs vs 3−run ground−truth
COT
ZS
EN
ES
Leaderboard rank (1 = best)
1
2
3
4
5
6
7
8
3−run majority
run1
run2
run3
Model
Claude−Opus−4
DeepSeek−V3−685B
GPT−4o
GPT−4o−mini
Grok−3
Grok−4
LLaMA−3.3−70B
o3−mini