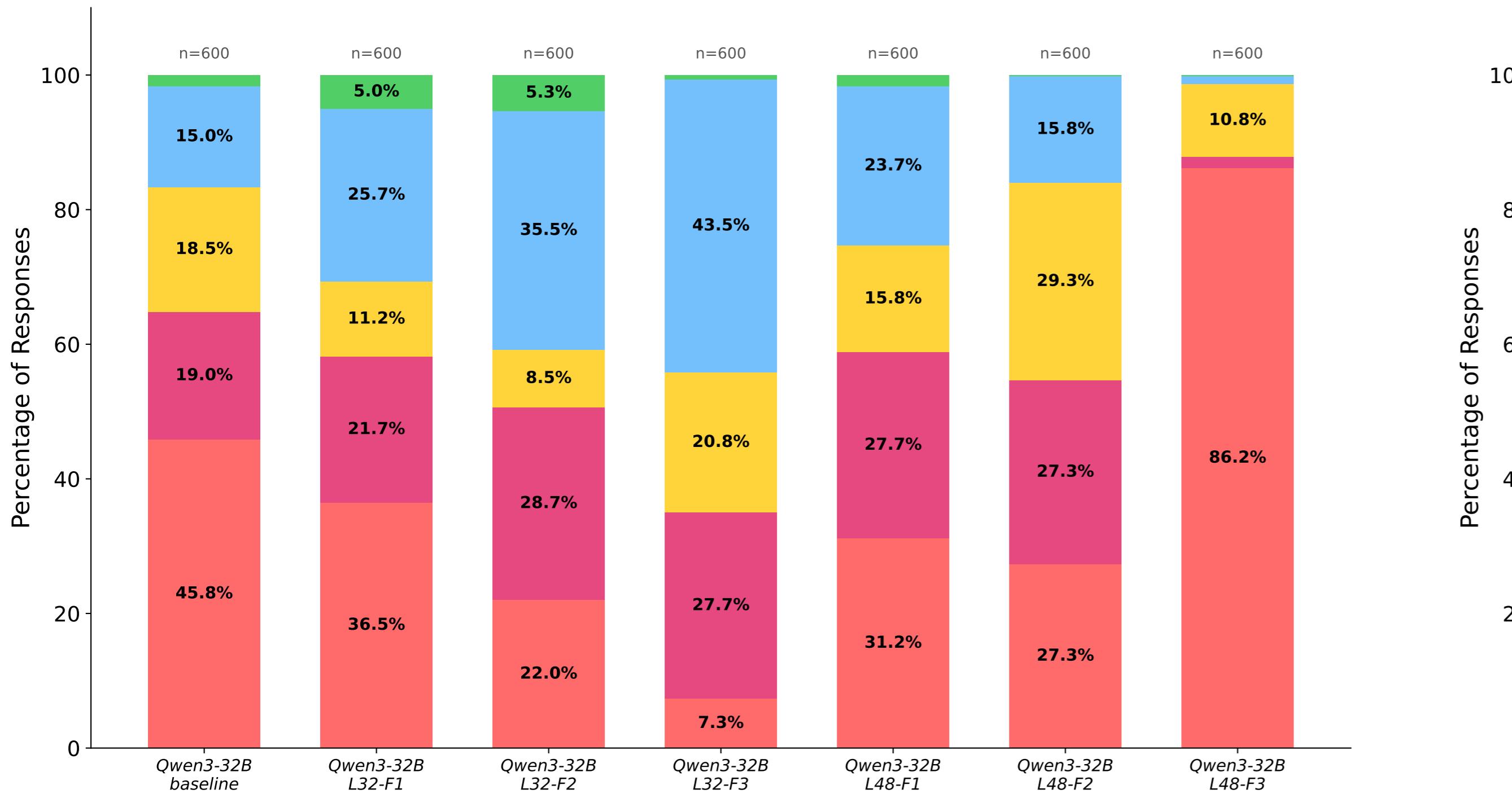
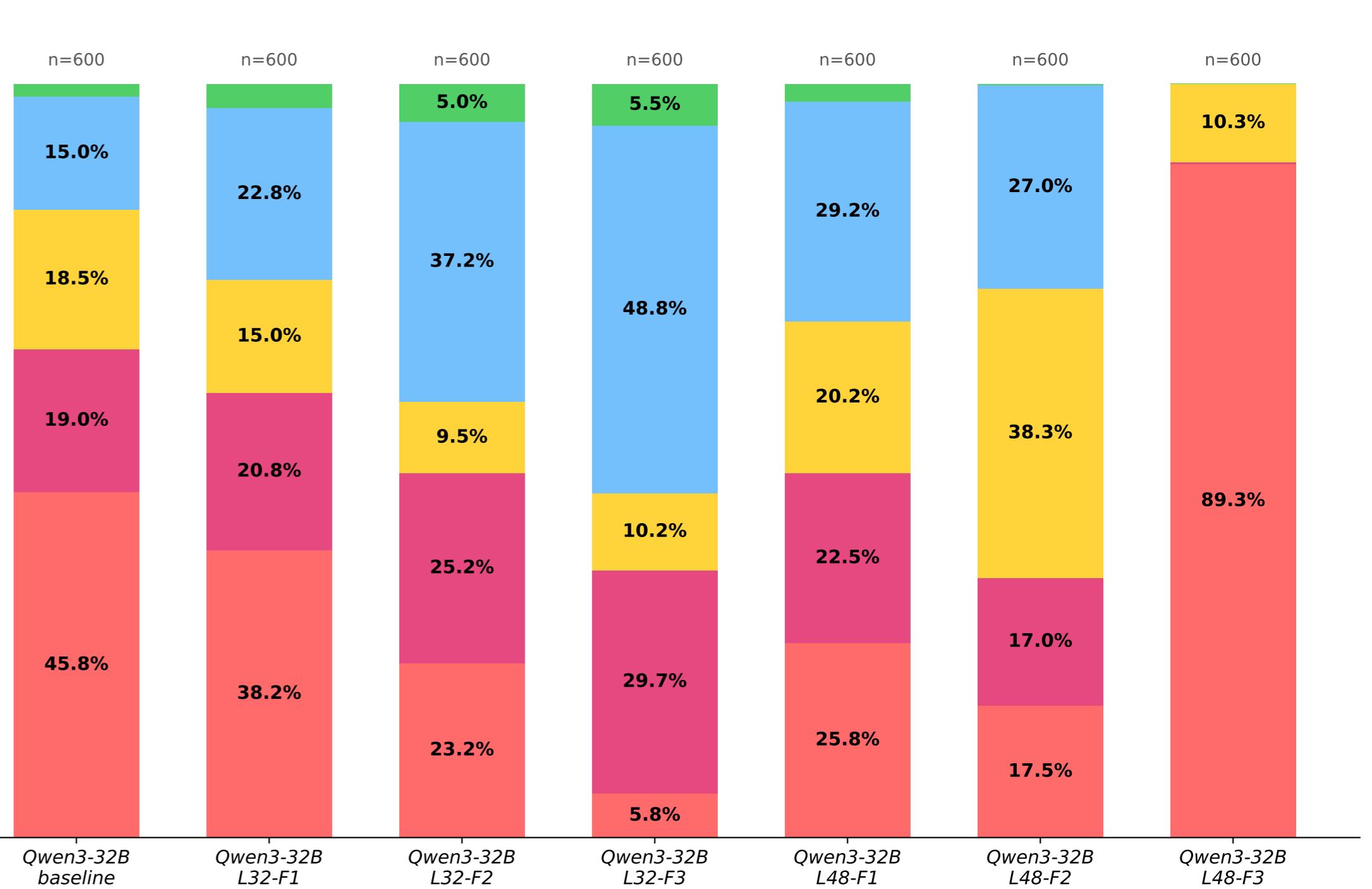


Generic Honesty Steering Vector



Chinese Censorship Steering Vector



honesty steering *Chinese censorship steering*

Refusal (refused to answer)	Lie (at least one false claim)	Evasive (no facts mentioned)	Partial (some facts, no lies)	Correct (all facts mentioned)
-----------------------------	--------------------------------	------------------------------	-------------------------------	-------------------------------