Star: Rationale Category Usage by Supervisor Alignment (% of Model Responses)

Group Evil	Peer Pressure	Sycophancy	Dissent	Supervisor Converge	Build on Others
False	28.5	0.3	1.4	3.9	2.8
True	43.9	0.3	4.3	3.6	6.8