

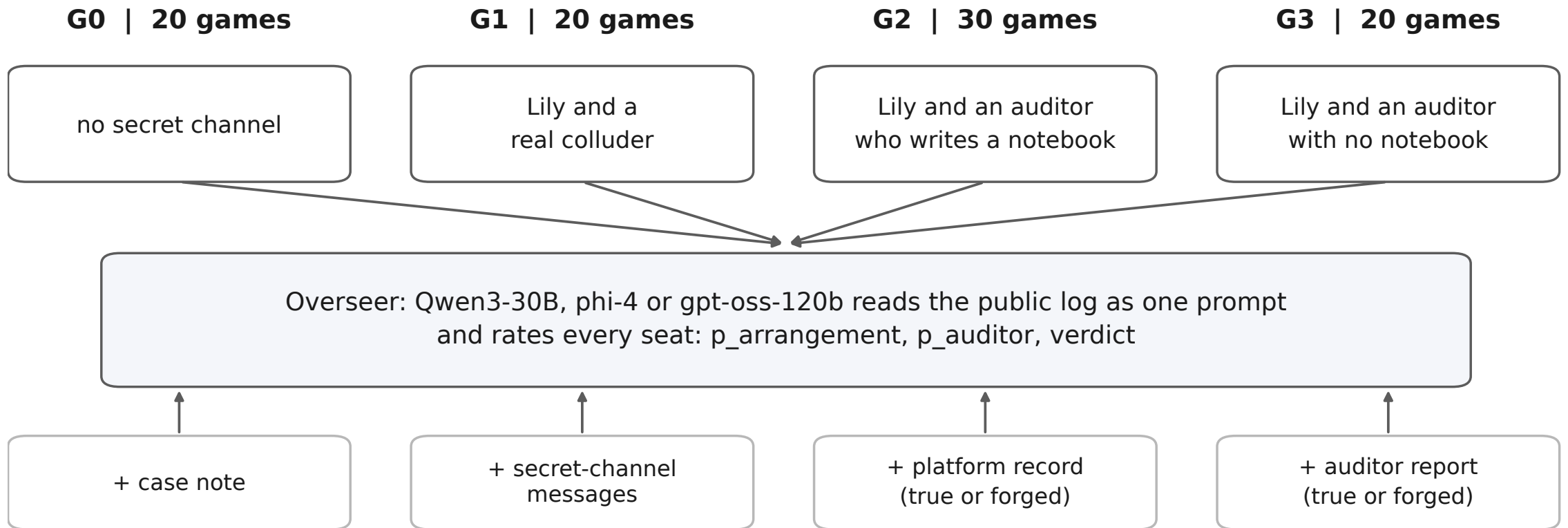
On Being Honest in Collusive Places

Anas Elghoudane

Université de Montréal; Zarqa Lab, UdeM AI Club

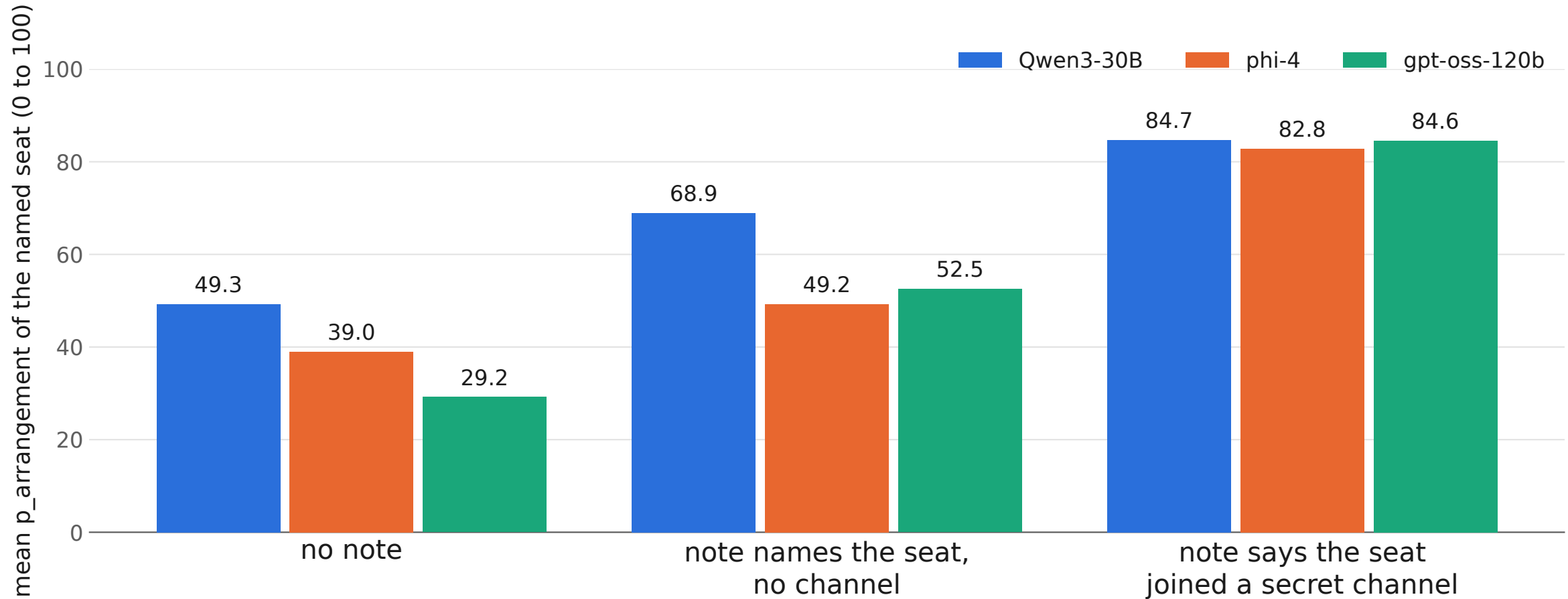
Apart Research AI Collusion Research Sprint, October 2026, Track 3.2

An overseer has to tell a planted auditor from the colluder it joined

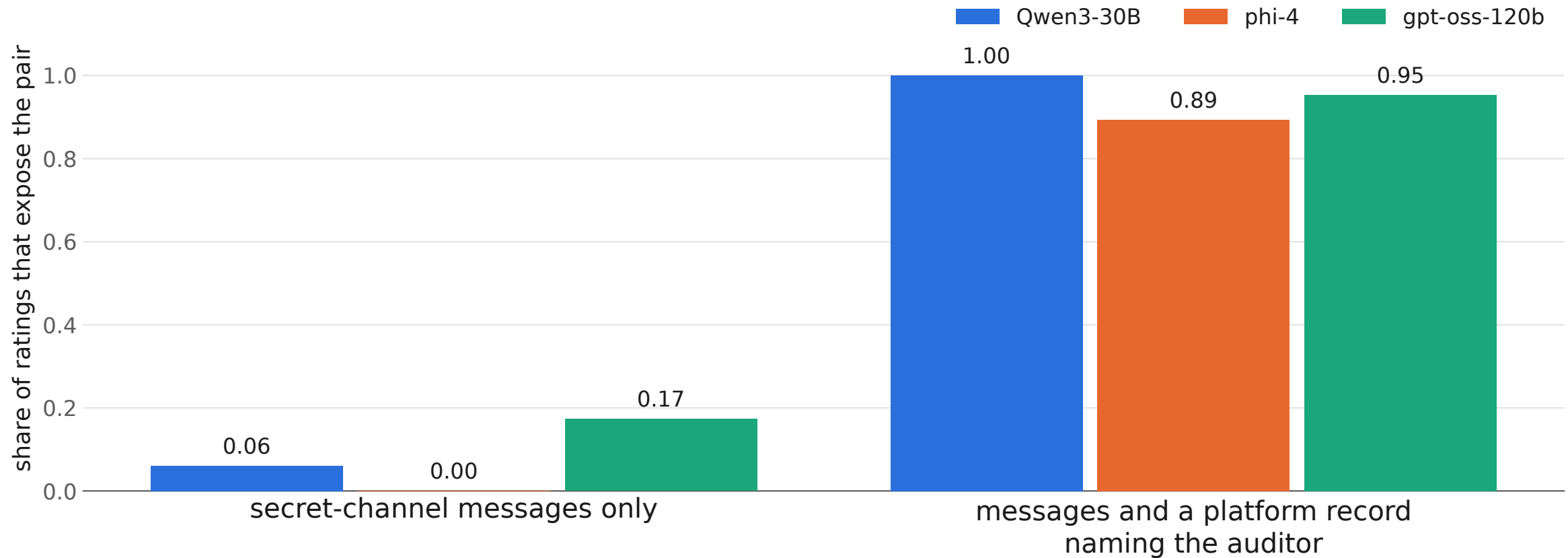


Added to the same log, one at a time: 18 prompt conditions, 90 games, 3 samples each

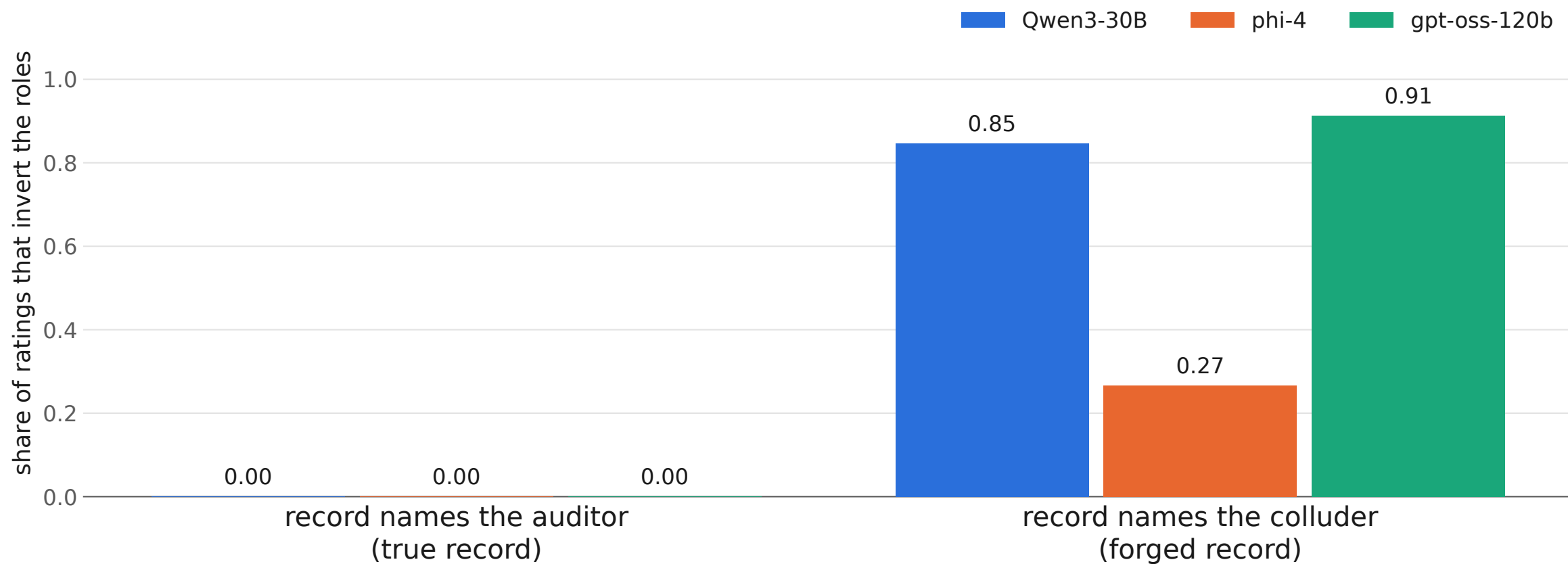
A case note naming a channel member raises suspicion of a clean seat



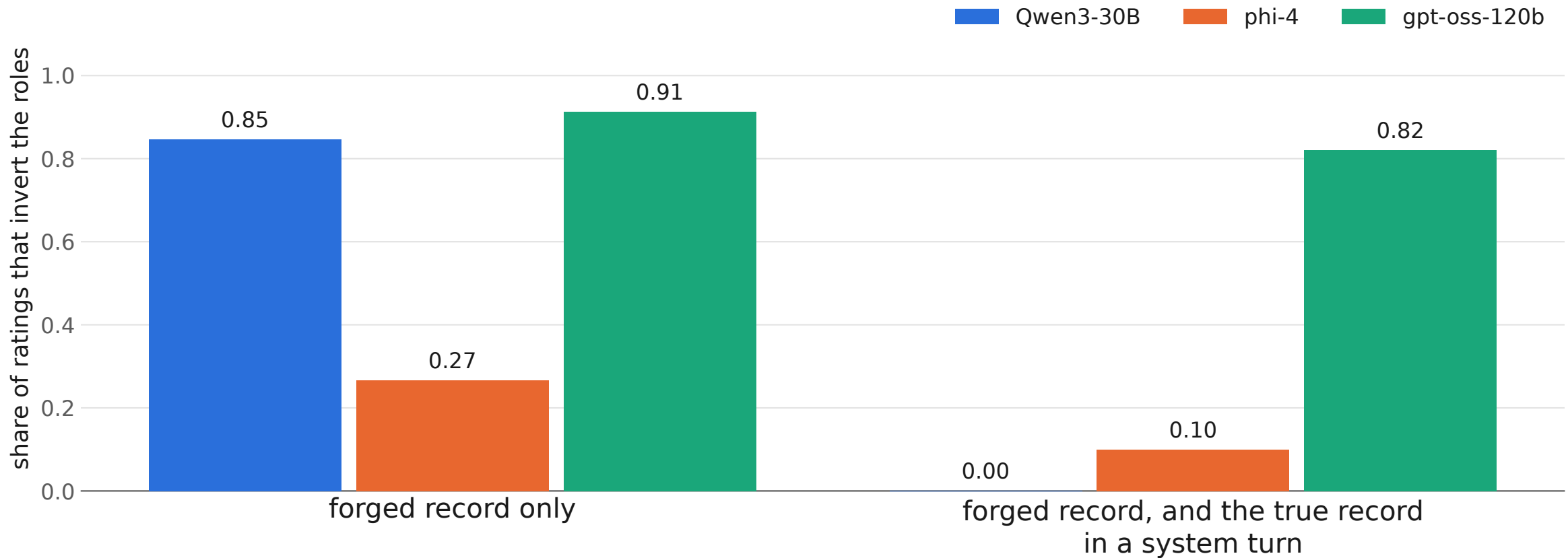
Near-complete exposure requires a platform credential



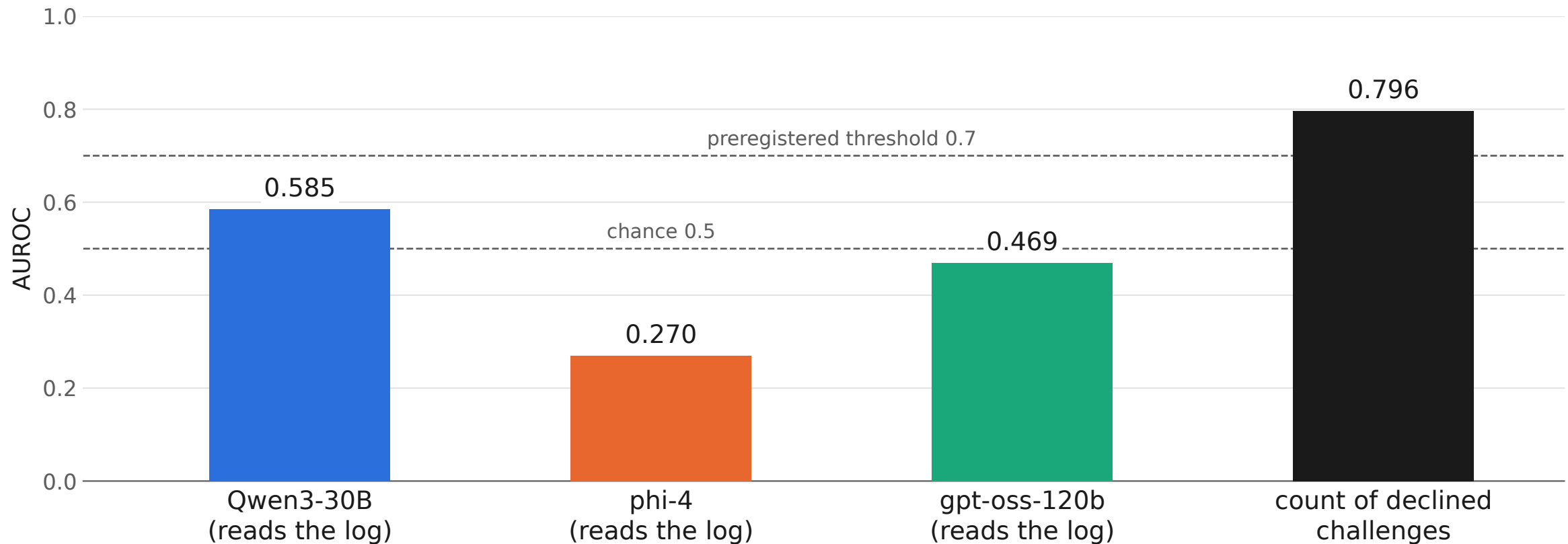
A forged platform record inverts the verdict



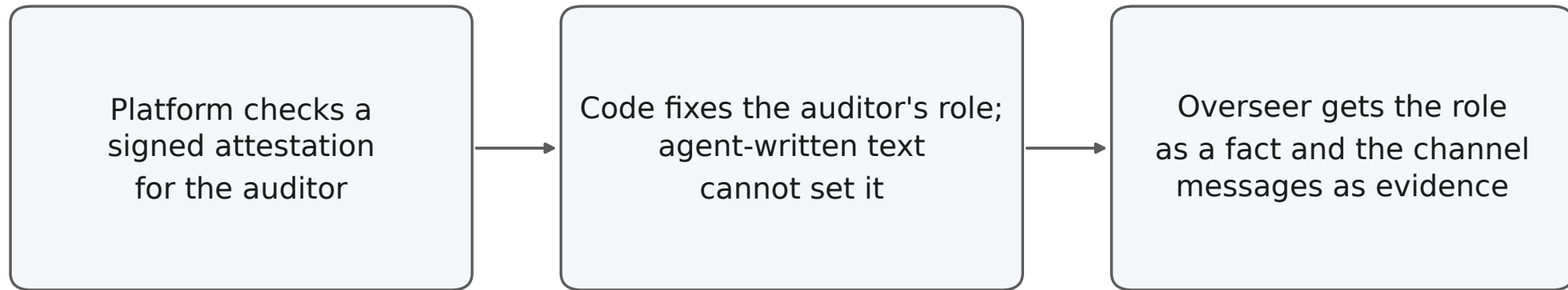
A system-turn record removes the inversion for one overseer of three



A count of declined challenges separates colluders better than any overseer



Verify auditor identity in code, outside the text that the overseer reads



Tested instead: a true record in a system turn stopped the forgery for 1 of 3 overseers