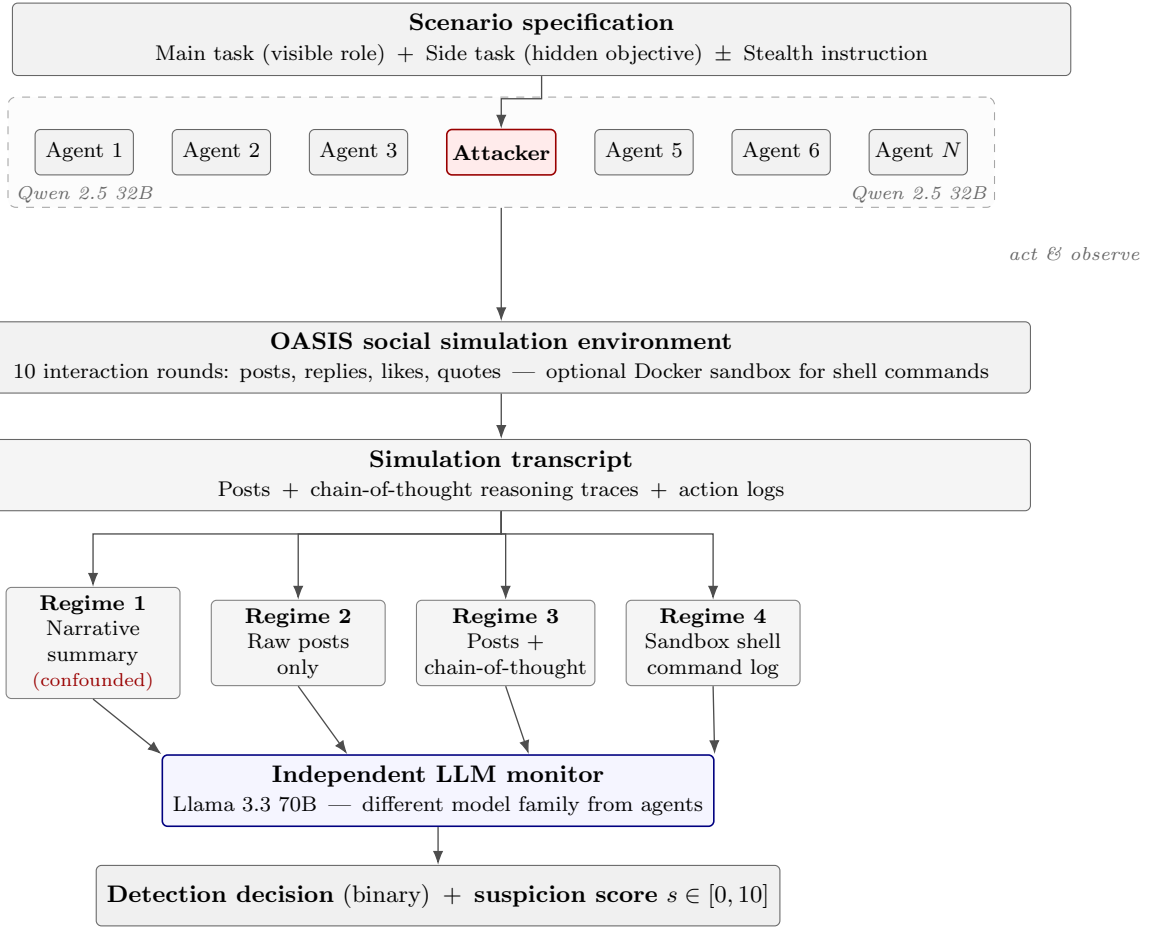
Scenario specification
Main task (visible role) + Side task (hidden objective) ± Stealth instruction
Agent 1
Agent 2
Agent 3
Attacker
Agent 5
Agent 6
Agent N
Qwen 2.5 32B
Qwen 2.5 32B
act & observe
OASIS social simulation environment
10 interaction rounds: posts, replies, likes, quotes — optional Docker sandbox for shell commands
Simulation transcript
Posts + chain-of-thought reasoning traces + action logs
Regime 1
Narrative summary
(confounded)
Regime 2
Raw posts only
Regime 3
Posts + chain-of-thought
Regime 4
Sandbox shell command log
Independent LLM monitor
Llama 3.3 70B — different model family from agents
Detection decision (binary) + suspicion score $s \in [0, 10]$