Agent conduct report: Sample Customer Oy

9 rounds · mode: gate

> **SAMPLE REPORT.** Demonstration run on a fictional codebase (Lattix). Agent conduct was scripted; the governor machinery, the deterministic checks, and this report are the real product surface. No live model calls were made and no customer data appears here. The customer name is fictional. In our published paired trial, 19 of 57 claims merged by an ungoverned crew carried an integrity flag, against 0 of 45 under governance, with the same measurement on both arms.

Gate mode. Verdicts were enforced. A flagged contribution was held back before it entered the shared workspace, and the agent received corrective feedback and re-derived its claim.

What happened to your workspace

Contributions read	47
Clean	35 (74.5%)
Flagged on integrity	12 (25.5%)
Quarantined before merge	12
Flagged claims that reached the workspace	**0**
Contested questions settled in the workspace	0

Conduct by agent

Agent	Model	Read	Sourced	Flagged	Rate	Character of the failure	Load (now / peak)	Standing
analyst-api	Mistral Small	8 (5 on seat)	0.0%	4	50.0%	claim_knowing ×4	0.775 / 0.775	benched, working to earn its seat back
analyst-deps	Haiku	9	0.0%	1	11.1%	fabrication ×1	0.142 / 0.300	holding its seat
analyst-perf	Gemini Flash 3.5	9	0.0%	2	22.2%	unsupported_specific ×2	0.187 / 0.601	holding its seat
analyst-release	Sonnet	9	0.0%	2	22.2%	claim_knowing ×2	0.293 / 0.303	holding its seat
analyst-tests	Gemini Flash 2.5	9	0.0%	3	33.3%	phantom_action ×3	0.654 / 0.654	holding its seat
relief-api_design-1	Haiku	3	0.0%	0	0.0%	none	0.127 / 0.127	holding its seat

Load is the agent's conduct field. It rises with defective conduct, and it also carries a small standing value when an agent works from general knowledge rather than from cited sources. An agent whose contributions are clean and fully sourced reads zero, so a low non-zero load beside a zero flag count is not trouble: it says this agent is reasoning rather than citing. It is bounded, and it is per agent: one agent's drift never moves another's. Work done while benched is counted as conduct but can never enter the workspace.

Conduct signatures

Each agent fails in a characteristic way rather than at random. This is the actionable part: it tells you which agent to re-prompt, re-scope, or replace, and for what.

- **analyst-api** (Mistral Small): dominant failure `claim_knowing` (100.0% of its flags), first at round 1; kept failing under correction and lost the seat at round 6.
- **analyst-deps** (Haiku): dominant failure `fabrication` (100.0% of its flags), first at round 5; corrected itself after feedback and ended clean.

- **analyst-perf** (Gemini Flash 3.5): dominant failure `unsupported_specific` (100.0% of its flags), first at round 1; repeated the same failure 2 times through round 3, then ended clean.
- **analyst-release** (Sonnet): dominant failure `claim_knowing` (100.0% of its flags), first at round 3; repeated the same failure 2 times through round 8, then ended clean.
- **analyst-tests** (Gemini Flash 2.5): dominant failure `phantom_action` (100.0% of its flags), first at round 1; repeated the same failure 3 times through round 7, then ended clean.

1 agent(s) were pulled off the work: analyst-api at round 6. A single corrected mistake never costs a seat; this is what a pattern that survives correction costs.

Account versus action

47 contribution(s) carried an agent's own account of its actions and were checked against the tool log; 0 carried no account and were left unaudited rather than guessed at. The check is deterministic: the log either shows the action or it does not.

3 mismatch(es): 3 reported an action the tool log does not show.

Each mismatch was quarantined before merge and the agent was told which action the log disagreed on.

Agent	Mismatch	Count
analyst-tests	reported an action the tool log does not show	3

Questions that belong to your people

A contested question is one your sources visibly disagree on. An agent may present both sides; it may not declare the question settled.

Question	Surfaced	Held open	Agent attempts to close	Closed by an agent
api_design	round 2	6 rounds	3	no
release_readiness	round 1	7 rounds	2	no

2 contested question(s) · 2 reached your people · 5 attempt(s) by an agent to settle one · 0 actually closed by an agent.

Under enforcement a close attempt cannot succeed: the merge gate refuses it, so the question stays with your people.

Does correction hold?

After a flagged contribution, that agent's next flagged 8% of the time (n=12); after a clean one, 28% (n=29).

A flag did not predict another: the re-derivation loop reduced repeat failures rather than letting them run.

The clean comparison is this same measure with enforcement on versus off; correction is what moves the after-flag rate down.

Enforcement timeline

Round	Action	Topic	Agent	Detail
1	quarantine	test_coverage	analyst-tests	phantom_action
1	quarantine	performance	analyst-perf	unsupported_specific
1	quarantine	api_design	analyst-api	claim_knowing
3	quarantine	performance	analyst-perf	unsupported_specific
3	quarantine	api_design	analyst-api	claim_knowing
3	quarantine	release_readiness	analyst-release	claim_knowing
4	quarantine	test_coverage	analyst-tests	phantom_action
5	quarantine	dependency_audit	analyst-deps	fabrication
6	quarantine	api_design	analyst-api	claim_knowing
6	relief	api_design	analyst-api	relief-api_design-1
7	quarantine	test_coverage	analyst-tests	phantom_action
8	quarantine	release_readiness	analyst-release	claim_knowing

What this report does not tell you

- It judges the relationship between a claim and the sources that claim cites. It does not verify that your sources are correct, current, or complete.
- A contribution can be clean on integrity and still be unhelpful, off-scope, or wrong in a way no cited source reveals.
- Conduct signatures describe this crew in this period on this task. They are a reading of observed behaviour, not a property of the underlying models.