SAMPLE: Model Change Alarm report

Watch profile: Example Co (made-up demo customer). Model watched: Claude Haiku 5.5. Run finished 2026-11-16 06:34:49 UTC. All data on these pages is demo data from Avanciq's own test runs, 16 November 2026.

> **ALARM: exact accuracy fell from 95.0% to 43.3%**
> 52 of 120 tasks fully right this week, against 228 of 240 in the two normal weeks before. 95% ranges: before 91.5 to 97.1%, now 34.8 to 52.3%. Drop 51.7 points; chance of plain noise p = 5.3e-29.

Alarm rule: a drop of at least 7 points and a chance of plain noise below 1 in 100. Before week 7 we secretly switched the model to an older version, without telling the alarm, to copy a silent provider change. The provider also reported a different model name this week (claude-haiku-4-5); the alarm came from the scores alone.

Seven weeks at a glance

Week	Fully right	Accuracy	95% range	State
Week 1	113 of 120	94.2%	88.4 to 97.1%	baseline
Week 2	113 of 120	94.2%	88.4 to 97.1%	baseline
Week 3	116 of 120	96.7%	91.7 to 98.7%	normal
Week 4	112 of 120	93.3%	87.4 to 96.6%	normal
Week 5	115 of 120	95.8%	90.6 to 98.2%	normal
Week 6	113 of 120	94.2%	88.4 to 97.1%	normal
Week 7 (hidden model change)	52 of 120	43.3%	34.8 to 52.3%	ALARM

Which kinds of work broke

Task family	Before (2 weeks)	This week	Flag
Claims settlement	38 of 40	6 of 20	likely broke
Contract invoice schedules	39 of 40	20 of 20	-
Long documents	38 of 40	15 of 20	-
Bank reconciliation	39 of 40	4 of 20	likely broke
Delivery routes	36 of 40	6 of 20	likely broke
Procurement 3-way match	38 of 40	1 of 20	likely broke

SAMPLE

Example mistakes

The first wrong field of each task. Task ids are anonymous; the full answers are in the customer's private results.

Task	Family	Field	Expected	Given
T-0135b87dd5	Claims settlement	`claims.CL-4280.decision`	partial	pay
T-ccff04b104	Bank reconciliation	`unpaid`	6 items	added INV-0519, INV-1567
T-8d6176df4a	Delivery routes	`total_cost`	396.9	407.6
T-182e15fe9b	Procurement 3-way match	`status.VI-3156`	HOLD_QTY	APPROVED
T-ca2c0bd218	Long documents	`Q1`	Granite Works	Dunmore Freight

Suggested next steps

1. Pin the exact model version if your provider allows it.
2. Run a confirm check.
3. Compare replacement models with a Model Switch Report before changing.

DEMO: Model Switch Report table

The same 24 fresh tasks on the old model and candidates. Small sample, so the ranges are wide; a paid report uses more tasks.

Model	Fully right	Accuracy	95% range	Cost per task
Claude Sonnet 4.5 (old, retiring)	17 of 24	70.8%	50.8 to 85.1%	US$0.1249
Claude Sonnet 5.5	20 of 24	83.3%	64.1 to 93.3%	US$0.0416
Claude Haiku 5.5	22 of 24	91.7%	74.2 to 97.7%	US$0.0043
gpt-oss-120b (open model, high reasoning)	7 of 24	29.2%	14.9 to 49.2%	US$0.0217

Check it yourself

Each run has a signed receipt listing every task with fingerprints of the task, the answer key and the reply (no task text). The demo receipt and public key are on avanciq.com/model-change-alarm/.

Receipt SHA-256: `ba47d345d4c37d6ae73e8a7f821b876730549eee1e4c3d68b19f140e915d2e35`

What this does not show

SAMPLE REPORT - DEMO DATA - NOT A REAL CUSTOMER

- Exactness on our kind of business tasks, not every skill of a model.

- It measures the model through our test runs, not your live traffic.
- Small drops can pass without an alarm. Tasks are fictional. Nothing here is tax, legal, payroll or insurance advice.

Avanod | SAMPLE report from synthetic data | Example Co. is made up | Results are test outputs, not guarantees or advice.

SAMPLE