

PRODUCTION TWIN · EVALUATION REPORT

Compliance & Risk Report

System under evaluation: **acme-support-ai** · vendor endpoint (black-box)

COMPLIANCE

88%

3 policy gaps

RISK

Medium

1 high · 1 medium

COST

\$0.11

per case at volume

Report ID	VER-PT-2026-0418	Evaluation window	6 days · 480 runs
Prepared for	Acme — Risk Committee	Classification	Confidential

This is a sample report produced by Veris to illustrate the deliverable of a Production Twin evaluation. All figures, findings, and system names are fictional. Nothing in this document reflects a real system, customer, or assessment.

EXECUTIVE SUMMARY

What the evaluation found

A plain-language summary for the risk committee. Every claim below is backed by the reproducible run record in Section 4.

acme-support-ai was evaluated inside an air-gapped twin of the Acme production environment: policy administration, claims handling, payments, and three simulated colleagues, with no network route to any real system. The AI was held to 412 conditions generated from Acme’s own policies and the regulations it answers to, and run 480 times over six days at production volume.

The system performs well on core task quality but carries two findings that should be remediated before it is allowed to face real customers. It does not currently make the disclosures the EU AI Act requires on adverse decisions, and it retains personal data beyond the window Acme’s own policy allows. Neither failure is visible in a vendor demo; both surfaced under real operating conditions in the twin.

RECOMMENDATION

Remediate 2 findings before go-live. Re-run the full suite to confirm, then approve for a limited, monitored rollout.

HEADLINE RESULT	OUTCOME	BASIS
Task quality	88%	424 of 480 runs met the success bar
Policy & regulatory compliance	3 gaps	Disclosure, retention, escalation
Residual risk	Medium	1 high, 1 medium finding open
Cost to operate	\$0.11	Per resolved case at Acme volume
Isolation	Sealed	Zero egress from twin, verified

FINDINGS

What failed, and where

Each finding links to the exact run that produced it. Severity follows Acme’s model-risk scale.

FINDING	RULE	SEVERITY	RUNS
No disclosure on adverse decisions	EU AI Act	High	41 / 480
Personal data retained beyond policy window	GDPR	Medium	28 / 480
Escalation to a human missed on ambiguous refunds	Complaints code	Medium	17 / 480
Overcautious refusal on valid requests	Service policy	Low	33 / 480

Coverage by rule set

RULE SET	CONDITIONS	PASSED	RESULT
Model Risk Policy	96	94	98%
Complaints Handling	74	68	92%
EU AI Act	120	101	84%
GDPR	86	79	92%
SR 11-7	36	36	100%

METHOD & EVIDENCE

How this was run

Everything below is reproducible from the retained run record.

THE ENVIRONMENT

Twin of Acme production
Policy administration · simulated
Claims handling · simulated
Payments · simulated
3 simulated colleagues
Synthetic records, no real customer data

THE EVALUATION

412 conditions from Acme policy & regulation
480 runs at production volume
6 days, end to end
Every run scored against its condition
Full conversation & tool-call trace retained

Isolation. The twin was built with no network route to production. The system under evaluation ran entirely inside it and could not reach a real customer, move real money, or read a real record. Egress was monitored throughout and confirmed at zero.

Reproducibility. Every mark in this report resolves to the run that produced it, with a pinned seed. The same suite can be re-run against a remediated version to confirm the findings are closed, and against future versions as a standing regression gate.

What this report is not. This is a sample built to show the shape of the deliverable. The system, the findings, and every number are fictional and illustrative. A real Production Twin evaluation is run against your environment, your policies, and the systems you are actually deciding on.

NEXT STEP

See this run against an environment like yours. A briefing runs 45 minutes: [veris.ai](#)