

AI Trust Pack

Enterprise security review readiness for an AI-powered customer support agent

ILLUSTRATIVE COMPANY	Acme Support AI
ASSESSMENT SCOPE	Architecture, data, authority, evidence, and incident readiness
REPORT DATE	September 2026

ILLUSTRATIVE
SAMPLE

This report demonstrates the structure and level of analysis in a Smart Path Assurance engagement. Acme Support AI is fictional. No production system was tested, and this document is not a certification, legal opinion, or guarantee of security.

-
-
-

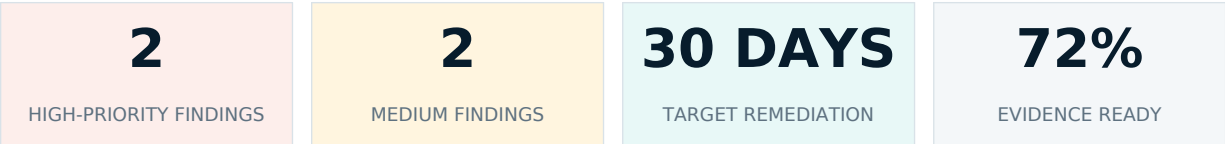
01 / DECISION BRIEF

Decision brief

A concise, buyer-oriented view of readiness, material risk, and the conditions for proceeding.

CONDITIONALLY READY

Acme has a credible control foundation for a limited enterprise pilot. Approval should be conditioned on reducing tool authority, producing repeatable prompt-injection evidence, and rehearsing containment.



What the enterprise buyer can rely on

CONTROL AREA	CURRENT POSITION	CONFIDENCE
Identity and access	Customer SSO supported; service identities are separated by environment	MODERATE
Data handling	Training exclusion is stated; retention paths require stronger system mapping	MODERATE
Agent actions	Tool calls are logged; write permissions remain broader than the operating scope	LIMITED
Incident response	General process exists; agent-specific revocation and residual-task checks are incomplete	LIMITED

Recommended decision

Proceed with a time-boxed pilot using non-sensitive data and constrained tool scopes. Make production expansion contingent on closure of F-01 and F-02, plus a witnessed containment exercise. This converts an open-ended security review into three objective acceptance conditions.

02 / SCOPE AND ASSESSMENT METHOD

Scope and assessment method

We evaluate what the agent can reach, what data crosses each boundary, and whether important control claims are supported by evidence.

System in scope

A customer-support agent embedded in Acme's SaaS application. The agent reads Zendesk tickets, retrieves approved customer context, drafts responses, and updates selected fields in Zendesk and Salesforce after policy checks.

IN SCOPE	OUT OF SCOPE
Agent orchestration and prompts Zendesk and Salesforce tools Retrieval and vector storage Model-provider data handling Logging and incident containment	Corporate IT controls Source-code security review Penetration testing of the SaaS application Legal or regulatory compliance opinion Customer-specific deployment controls

Assessment lens

1 DISCOVER	2 TRACE	3 CHALLENGE	4 EVIDENCE
Inventory models, tools, identities, data stores, and external dependencies.	Map data flows and effective authority across direct and indirect paths.	Test whether untrusted input or system change can invalidate key claims.	Tie each buyer-facing answer to an owner, artifact, and validation date.

Reference frameworks

The review is informed by the NIST AI Risk Management Framework and its Generative AI Profile, and by the OWASP Top 10 for LLM Applications. These references organize the analysis; they do not convert this engagement into a certification.

Evidence reviewed in this sample: architecture interview, data-flow narrative, tool manifest, IAM export, sample audit events, model-provider settings, security questionnaire draft, and incident response procedure.

03 / SYSTEM AND TRUST BOUNDARIES

System and trust boundaries

The important question is not only where the model runs, but where untrusted content can influence an action with customer impact.

ENTERPRISE USER	ACME APPLICATION	AGENT ORCHESTRATOR
ZENDESK	RETRIEVAL + VECTOR STORE	SALESFORCE
MODEL PROVIDER	POLICY / TOOL GATEWAY	CENTRAL AUDIT LOG

Conceptual architecture - authoritative data and action paths are listed below.

PAT H	DATA OR AUTHORITY	CONTROL EXPECTATION	ASSESSMENT NOTE
A	User identity -> agent session	Authenticated actor remains attributable through every tool call	Supported by logs; sampled only
B	Ticket text -> model context	Untrusted content cannot redefine policy or authorize tools	Claimed; repeatable test evidence missing
C	Agent -> Zendesk / Salesforce	Write scope matches approved task, tenant, object, and field	Scope broader than documented use case
D	Runtime -> audit log	Prompt, policy decision, tool call, result, and actor are reconstructable	Strong base; retention mapping incomplete

Material trust-boundary insight

A support ticket is both business data and potentially adversarial input. Because the same workflow can retrieve customer context and invoke write-capable tools, prompt-injection resilience and least privilege are linked controls: weakness in either one increases the consequence of the other.

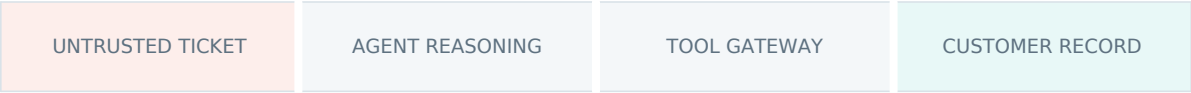
04 / EFFECTIVE AUTHORITY

Effective authority

Assigned permission is only the starting point. The review asks what the agent can actually cause across connected systems.

IDENTITY / TOOL	INTENDED USE	OBSERVED AUTHORITY	RISK TREATMENT
Zendesk service account	Read tickets; draft and post replies	Can read all queues and modify status, tags, assignee, and custom fields	Restrict by tenant, queue, object, field, and action
Salesforce OAuth app	Read account tier; update support disposition	Can read broader account metadata and write multiple case fields	Create a dedicated permission set and deny unrelated objects
Retrieval service	Return approved knowledge snippets	Namespace filter is provided by the orchestrator	Enforce tenant and corpus scope inside retrieval service
Model API key	Generate classifications and response drafts	Shared runtime credential with provider-side usage logs	Separate environments; rotate and monitor abnormal use

Authority-path analysis



Current controls rely on system instructions, a tool allowlist, and application-side validation. The highest-value improvement is to make the gateway independently verify tenant, object, field, action, and workflow state before every write. This limits blast radius even when upstream reasoning fails.

CONTROL CLAIM	The agent cannot modify customer records outside the active support workflow.
CURRENT STATUS	Not yet proven. The permission boundary is broader than the claim, and negative-path evidence is incomplete.

05 / MATERIAL FINDINGS

Material findings

Each finding supports an engineering decision and an enterprise-review response, not simply a label.

F-0 1	Tool authority exceeds the documented operating scope	HIGH
OBSERVATION	Zendesk and Salesforce identities can perform writes beyond the fields and objects required by the support workflow. Application logic narrows normal behavior, but the downstream authorization layer does not fully enforce that scope.	
BUSINESS IMPACT	A compromised prompt, faulty workflow, or orchestration defect could alter customer records outside the approved task. The gap is likely to block buyers that require least privilege or bounded agent autonomy.	
RECOMMENDED ACTION	Create purpose-built service identities. Enforce tenant, object, field, action, and workflow-state policy at the tool gateway. Add allow and deny-path integration tests and retain the results as evidence.	

F-0 2	Prompt-injection resistance is asserted but not evidenced	HIGH
OBSERVATION	The team describes layered prompt controls and tool validation, but there is no versioned test corpus, pass/fail criteria, or retained result showing how the system handles adversarial ticket content and retrieved documents.	
BUSINESS IMPACT	Security reviewers cannot distinguish a design claim from a demonstrated control. Regressions may be introduced when prompts, models, tools, or retrieval behavior change.	
RECOMMENDED ACTION	Define abuse cases tied to real authority paths. Establish deterministic assertions for tool denial, sensitive-data handling, and human escalation. Run the suite against every material model or policy change.	

05 / MATERIAL FINDINGS - CONTINUED

Material findings - continued

Medium findings are unlikely to stop a tightly scoped pilot, but they weaken buyer confidence and slow questionnaire closure.

F-03 Agent-specific containment is incompleteMEDIUM	
OBSERVATION	The incident procedure covers application isolation and credential rotation, but it does not explicitly address active sessions, cached credentials, queued jobs, retries, spawned tasks, or validation that external authority has expired.
BUSINESS IMPACT	A shutdown command may stop new work without containing residual authority. During an incident, unclear ownership and incomplete verification can extend customer impact and notification uncertainty.
RECOMMENDED ACTION	Create a one-page containment runbook with named owners, ordered revocation steps, residual-task checks, and a measurable containment objective. Exercise it and preserve timestamps and results.

F-04 Retention claims are not mapped end to endMEDIUM	
OBSERVATION	Customer-facing language gives a single retention answer, while prompts, application logs, provider telemetry, vector records, and backups follow different settings and deletion mechanisms.
BUSINESS IMPACT	Sales, privacy, and security teams may provide inconsistent answers. Deletion requests could be completed in the primary application while derived or diagnostic records persist elsewhere.
RECOMMENDED ACTION	Build a data-retention matrix covering system, data class, purpose, region, retention period, deletion trigger, subprocessor, and evidence owner. Use it as the source for questionnaire responses.

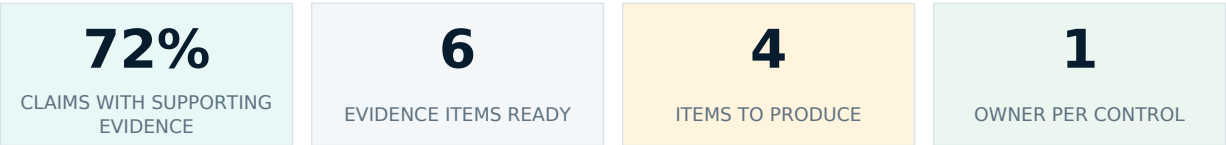
Risk-rating note

Severity reflects plausible business impact and control weakness in the illustrative environment. It is not a prediction of exploit likelihood and should be recalibrated for the customer's deployment and data sensitivity.

06 / ENTERPRISE EVIDENCE READINESS

Enterprise evidence readiness

Fast security reviews depend on concise answers backed by artifacts that a reviewer can inspect without reopening discovery.



COMMON BUYER QUESTION	DRAFT RESPONSE	EVIDENCE	STATUS
Can the AI act without human approval?	Low-risk drafts may be autonomous; record changes require policy checks and are limited by workflow.	Approval policy; tool-gateway config; sampled audit events	PARTIAL
Is customer data used to train models?	Customer content is not used for provider model training under the configured enterprise terms.	Provider terms; account setting export; data-flow map	READY
How do you resist prompt injection?	Untrusted content is separated from instructions; tools are allowlisted and validated.	Design note exists; repeatable adversarial results needed	GAP
Can you fully stop the agent?	New requests can be disabled and credentials revoked using the incident runbook.	Runbook exists; containment exercise and residual checks needed	PARTIAL
How long is customer data retained?	Retention depends on data class and system; deletion is orchestrated across defined stores.	End-to-end retention matrix needed	GAP

Recommended evidence index

Maintain a deal-ready folder containing: current architecture diagram; system and subprocessor inventory; IAM and tool-scope export; model-provider configuration; prompt-injection test summary; representative audit trace; retention matrix; incident runbook; containment exercise record; and owner-approved questionnaire answer library.

07 / INCIDENT AND CONTAINMENT READINESS

Incident and containment readiness

A credible kill switch is not a button. It is an evidence-backed sequence that removes authority and confirms no residual work can complete.

TARGET	ACTION	VERIFICATION SIGNAL	OWNER
T+0	Disable new agent tasks and external tool invocation	Gateway rejects a synthetic tool request	Incident commander
T+5 min	Revoke model and SaaS credentials; invalidate active sessions	Previously valid tokens fail across each provider	Platform / security
T+10 min	Cancel queues, retries, scheduled work, and child tasks	No runnable or delayed tasks remain	Agent engineering
T+20 min	Confirm customer-impact boundary and preserve logs	Affected tenants, actions, and records are enumerated	Security / support
T+30 min	Run an independent residual-authority check	No tested path can read, write, or communicate externally	Assurance owner

SAMPLE EXERCISE RESULT	Containment not yet verified
Residual paths	Active SaaS OAuth session and one delayed retry queue were not covered by the draft procedure.
Target state	Verified containment time under 30 minutes, with zero successful reads, writes, or outbound calls across the tested authority graph.

Evidence to retain

Preserve the exercise scenario, initial authority graph, commands issued, provider confirmations, token-failure checks, queue state, audit events, timestamps, exceptions, and final sign-off. This turns incident readiness from a policy statement into buyer-reviewable proof.

08 / 30-DAY TRUST PLAN

30-day trust plan

The plan prioritizes the smallest set of control and evidence improvements most likely to unblock a disciplined enterprise pilot.

WHEN	OUTCOME	KEY ACTIONS	ACCEPTANCE EVIDENCE
Days 1-7	Authority is bounded	Create dedicated Zendesk/Salesforce roles; enforce tool policy by tenant, object, field, and action	Permission export; policy config; positive and negative test results
Days 8-14	Abuse resistance is measurable	Build a versioned prompt-injection and tool-abuse suite tied to authority paths	Test corpus; pass criteria; run summary; exception log
Days 15-21	Containment is demonstrable	Finalize runbook; revoke all authority paths; cancel residual work; conduct tabletop and live exercise	Signed runbook; timestamps; provider confirmations; exercise record
Days 22-30	Buyer answers are consistent	Complete retention matrix; assemble evidence index; approve reusable questionnaire answers	Evidence folder; owner approvals; final answer library

Pilot acceptance conditions

- 01

No production tool identity can modify objects or fields outside the documented support workflow.
- 02

The abuse test suite demonstrates denial of unauthorized tool use across agreed prompt-injection scenarios.
- 03

A witnessed exercise confirms containment within 30 minutes and no residual authority across tested paths.

FROM SECURITY REVIEW
TO SALES ENABLEMENT

The goal is not a larger binder. It is a smaller set of defensible claims, each backed by current evidence and a named owner - so the buyer can say yes with less uncertainty.

About this sample

Smart Path Assurance helps AI companies translate architecture, controls, and test results into enterprise-ready security evidence. This illustrative report is intentionally concise; a live engagement is tailored to the product, customer requirements, and available evidence.

smartpathassurance.com | james@smartpathassurance.com