

SIMULATED CASE STUDY

AI Risk Assessment: Customer Support Response Assistant

AI governance

Project	AI governance and use-case risk assessment (Project 4)
Company	Northstar Financial Technologies (fictional)
Status	Simulated; no certification or conformance claimed
As of	2026-09-30

This document is part of a portfolio exercise. Northstar Financial Technologies, its people, systems, records and findings are invented. No real audit was performed, no assurance is given and nothing here is a certification or compliance claim.

1. Use case

Use case	AI-01 Customer Support Response Assistant
Purpose	Draft replies to customer support tickets using approved knowledge-base content for agent review
Business owner	Head of Customer Support
Users	Customer support agents (about 60)
Data	Customer name, account identifiers, ticket text; approved knowledge articles
Affected parties	Business customers and their users who receive responses
Provider	Third-party hosted LLM API with retrieval over the approved knowledge base (vendor V-03, Project 3 residual risk High)
Proposed oversight	Agent reviews and sends every response (proposed); no automatic sending
Overall rating: inherent / as proposed / after gates	High / High / Medium
Decision	Conditionally approved (limited pilot)

2. How ratings work

Each of eleven areas is rated Low, Medium or High three times: inherent (no controls), as proposed in the intake, and after the gates are met. The overall rating is High if three or more areas are High, Medium if any area is High or three or more are Medium, otherwise Low. The scale is ordinal judgement; the rationale column is the substance.

3. Assessment by area

Area	Inherent	As proposed	After gates	Rationale	Controls required
Intended purpose and affected people	Medium	Medium	Low	Customers receive replies drafted by AI. Business customers could act on wrong information, but every reply is meant to be sent by an agent.	Approved-content retrieval only; agent sends every reply.
Business criticality and fallback	Medium	Medium	Low	Support is customer-facing but a manual process exists, so the service can run without the tool.	Documented manual fallback; ability to disable the tool within one hour.
Data categories, privacy and confidentiality	High	High	Medium	Ticket text and account identifiers (Confidential personal and account information) go to a third-party model hosted in the United States (assumed) under unclear retention terms.	Redact identifiers before sending where possible; contract states retention and no training use for the API tier; privacy impact assessment (NS-PRV-01) including cross-border handling.
Output reliability and hallucination	High	High	Medium	Generated text can be fluent and wrong. A wrong answer about fees, deadlines or service status could harm a customer.	Answers limited to retrieved approved articles with the source article cited; pre-deployment evaluation on a representative ticket set; sampled quality review in operation.
Human oversight	High	Medium	Low	The intake proposes agent review of every reply, but no technical control stops a send without review, and agents may approve routinely (automation bias).	Send blocked until the agent approves in the tool; agent training; edit rate and rubber-stamp rate tracked (NS-AIG-02).
Security: prompt injection and data leakage	High	High	Medium	Customer ticket text is untrusted input and can contain instructions aimed at the model (prompt injection). Retrieved content or other customers' data could be exposed.	Retrieval limited to approved articles; no tool or system actions available to the model; output filtering; red-team tests with injection strings before launch.
Supplier and subprocessor dependency	High	High	Medium	Lumenvale AI Ltd. (V-03) has unclear retention, model training, subprocessor and incident terms and High residual risk in Project 3.	Close the Project 3 conditions for the API tier before go-live; exit plan to revert to manual drafting.
Logging and traceability	Medium	Medium	Low	Without logs, an incorrect reply cannot be traced to its prompt, source articles and approving agent.	Log prompt, retrieved article ids, draft, final reply and approver, retained for a stated period (proposed 90 days).
Change management (model and vendor changes)	Medium	Medium	Low	A silent model or vendor change can alter behaviour or data handling.	Pin the model version where offered; vendor must give notice of model and terms changes; re-run the evaluation set after any change.
Accessibility and fairness	Medium	Medium	Low	Tone and clarity may differ for customers who write in non-standard English or in French, and replies must be accessible.	Evaluate on a mix of writing styles and languages; agents can rewrite freely.
Incident response	Medium	Medium	Low	No AI-specific incident path exists yet.	Adopt the AI incident playbook; train agents and supervisors; include the vendor contact.

4. Decision and gates

Conditionally approved (limited pilot). Pilot with at most 10 agents, retrieval over approved knowledge-base articles only, API tier only, no automatic sending. Go-live only when gates AI-01-G01 to G08 are evidenced.

The use is customer-facing and the vendor evidence is incomplete, so residual risk is High as proposed. With human approval enforced technically, approved-content retrieval, evaluation, monitoring and the vendor conditions closed,

residual risk falls to Medium, which the AI Governance Committee can accept. A pilot is proportionate because the benefit is real and the tool can be switched off within an hour.

Gate	Requirement	Owner	Target	Status
AI-01-G01	Written confirmation from Lumenvale of prompt and output retention, no model training on Northstar content, and a subprocessor list, for the API tier	Chief Product Officer	2026-11-01	Blocked by 3 open vendor issue(s) in Project 3
AI-01-G02	Incident notification clause with a stated time limit, content and named contact	Head of Procurement	2026-11-01	Blocked by 1 open vendor issue(s) in Project 3
AI-01-G03	Privacy impact assessment completed, including cross-border handling and redaction of identifiers	Privacy Officer	2026-11-15	In progress
AI-01-G04	Technical enforcement that no reply is sent without agent approval	Head of Customer Support	2026-11-15	Not started
AI-01-G05	Pre-deployment evaluation on a representative ticket set with documented thresholds met (material error rate at or below 2%)	Head of Customer Support	2026-11-15	In progress
AI-01-G06	Prompt injection and data leakage tests passed	IT Security Manager	2026-11-15	Not started
AI-01-G07	Logging of prompt, sources, draft, final reply and approver in place with a retention period	Head of Platform Engineering	2026-11-22	Not started
AI-01-G08	Monitoring plan live and the AI incident playbook rehearsed with support supervisors	AI Governance Committee	2026-11-22	Not started
The residual rating after gates assumes every gate is closed with evidence. Until then the as-proposed rating applies. Acceptance authority: Chief Risk Officer (executive acceptance) for the as-proposed level; AI Governance Committee after gates.				