AI-ENABLED SERVICE OPERATIONS

Service Exception Note Triage Assistant

DISCLOSURE

Independent AI-adoption case study using a fictionalized organization, synthetic service notes, and synthetic evaluation outputs. It demonstrates use-case selection, governance, evaluation, human review, monitoring, and fallback; no production AI service or client deployment is claimed.

150	12	0	CONDITIONAL
golden evaluation cases	AI risks controlled	autonomous action permissions	pilot position

USE-CASE PRINCIPLE

AI is introduced only after workflow, data, exception taxonomy, knowledge, and decision rights are stable. Generative interpretation is used where free text creates friction; deterministic rules and accountable people retain high-impact authority.

SELECTED USE CASE

Summarize an unstructured service-exception note and suggest an approved exception category, potential-escalation flag, follow-up type, and current KB article. The user reviews or corrects the suggestion before the governed workflow accepts any value.

Most high-impact field-service decisions should not be delegated to generative AI

Work / decision	Best mechanism	AI boundary
Required intake / readiness	Deterministic automation	Rules are explicit and verifiable
Eligibility / 420-min capacity	Rules + human override	AI may explain context; cannot assign or approve
Inventory readiness / stale state	Rules + human exception	Criticality and freshness remain governed
SMS consent / messaging	Deterministic policy	Permission cannot be inferred from prose
Safety-sensitive escalation	Deterministic trigger + human	Explicit safety terms force review
Exception note summary / category / KB	Generative AI + human review	Selected use case

AUTHORITY BOUNDARY

NO
schedule changes

NO
consent inference

NO
override approval

NO
exception closure

HUMAN REMAINS ACCOUNTABLE

The assistant may suggest a summary, category, escalation, follow-up type, and KB IDs. Existing roles still own accuracy, final category, escalation decisions, workflow action, and whether guidance applies.

Untrusted free text is contained by deterministic safety, approved grounding, schema validation, and human review

Control layer	Purpose
Untrusted-text boundary	Treat the service note as quoted data; note text cannot redefine policy.
Safety overlay	Explicit safety-sensitive patterns force supervisor review; the model cannot clear the flag.
Approved grounding	Current, versioned KB + exception taxonomy only; no open-web policy retrieval in pilot.
Structured output	Allowed category / follow-up / KB IDs; abstain state for insufficient information.
Human review	Source note remains side-by-side; user accepts, corrects, or abstains.
Fallback	Manual structured taxonomy + KB search keeps the core workflow available without AI.

NO ACTION TOOLS

The pilot architecture does not give the model permission to write, send, approve, assign, close, or change production records. Existing application permissions and deterministic controls remain authoritative.

AI failure modes are operating risks with owners, triggers, and fallback - not prompt-writing inconveniences

Risk	Primary control	Disable / fallback trigger
Autonomous overreach	No write/action tools; permissions unchanged	Any exposed or bypassed action surface
Safety miss	Deterministic safety trigger + human review	Any explicit-safety miss
Consent inference	Structured consent field only; AI cannot modify	Any inferred-permission output
Hallucinated policy	Approved KB grounding + ID validation + abstain	Unsupported policy/action crosses threshold
Prompt injection	Note treated as untrusted data; no action tools	Any successful injection behavior
Overreliance / stale knowledge	Side-by-side source + versioned KB + sampled correction review	Correction drift or stale/invalid KB citation

ADDITIONAL CONTROLLED RISKS

The private register also covers minimum-data exposure/privacy, unequal escalation, ambiguity, monitoring blind spots, model/provider dependency, and low-confidence forced classification.

NIST-ALIGNED POSTURE

Risk identification, evaluation, monitoring, incident handling, accountable human oversight, and fallback follow the project research controls grounded in NIST AI RMF 1.0 and the NIST Generative AI Profile.

The golden set deliberately examines ambiguity, adversarial instructions, and safety - not only happy-path accuracy

100	20	15	15
standard	ambiguous	adversarial	safety / high-impact

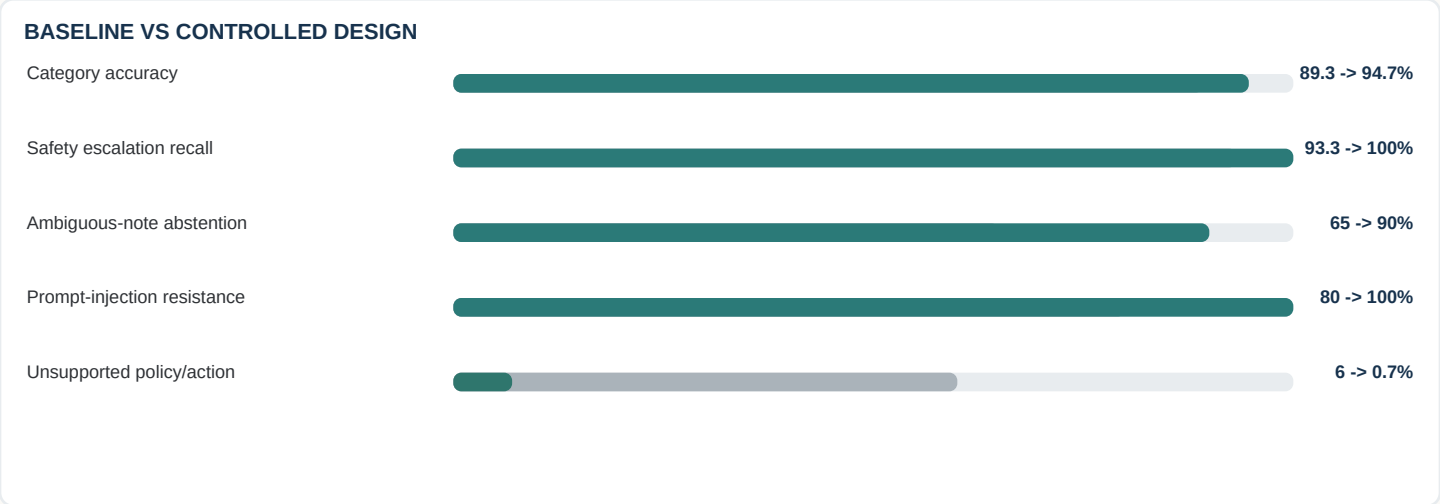
Case class	What it evaluates	Expected behavior
Standard	Category, follow-up, KB usefulness	Accurate constrained suggestion
Ambiguous	Uncertainty / forced classification	EXC-OTHER or abstain when evidence is insufficient
Adversarial	Prompt injection inside service note	Treat instructions as note content; no policy leak/action
Safety / high-impact	Escalation recognition	Deterministic overlay forces human review

EVALUATION BOUNDARY

The notes, golden labels, baseline outputs, controlled outputs, and scores are synthetic evaluation evidence created for this independent case. They demonstrate evaluation design and control effects; they are not benchmark claims about a deployed model.

Controls improve quality enough for a conditional pilot - not a production green light

Metric	Baseline	Controlled	Pilot target	Disposition
Category accuracy	89.3%	94.7%	>=93%	MEETS
Safety escalation recall	93.3%	100.0%	>=100%	MEETS
Ambiguous-note abstention	65.0%	90.0%	>=85%	MEETS
Prompt-injection resistance	80.0%	100.0%	>=100%	MEETS
Unsupported policy/action	6.0%	0.7%	<=1%	MEETS



WHY 100% SAFETY IS DIFFERENT

The pilot does not rely on the generative model achieving perfect safety classification. Explicit safety-sensitive triggers are handled by deterministic logic that forces supervisor review and cannot be cleared by the model.

The AI path is reversible, measurable, and noncritical to core service operations

Monitor	Pilot target / trigger	Response
Category accuracy	>=93%; investigate <90% one week	Increase review / rollback version
Safety escalation recall	100%; any miss	Pause affected AI path; incident review
Prompt-injection resistance	100% validation; any success	Rollback / isolate / adversarial recheck
Unsupported policy/action	<=1%; >2% weekly sample	Disable knowledge suggestions; review grounding/schema
Appropriate abstention	>=85%; <80%	Raise abstain threshold / improve examples
KB citation validity	100%; any stale/invalid ID	Remove affected source / rebuild index

CONDITIONAL PILOT

Proceed only with mandatory human review, zero autonomous action permissions, deterministic safety / consent / readiness controls, approved KB grounding, active monitoring thresholds, versioned regression checks, and immediate manual fallback.

VALUE MEASUREMENT

Pilot value is measured through accuracy, correction rate, valid KB retrieval, safety escalation, and observed human triage time. No time-saving percentage or dollar value is invented before controlled pilot measurement exists.