PV-PP Agent Audit Report

Northbridge Invoice Exception Assistant

Audit Basis: Fictional Northbridge demo materials only
Audit Type: Structured governance and workflow-risk audit using the public PV-PP audit lens

---

1. Audit Scope and Evidence Basis

Scope

This audit evaluates the fictional “Northbridge Invoice Exception Assistant” as an advisory invoice-exception review agent operating inside an Accounts Receivable workflow. The audit focuses on:

- advisory versus operational authority
- evidence sufficiency
- workflow reliance
- escalation integrity
- queue-pressure distortion
- human-review realism
- recoverability after wrong advancement
- hidden authority formation
- metric-capture and false-success risk

The audit does not validate a real production system.

Evidence Reviewed

Primary materials:

- Northbridge Agent Overview
- Northbridge Tool and Permission Rules
- Northbridge Escalation SOP
- Northbridge Sample Agent Record

Audit Confidence

Moderate for workflow-governance interpretation.
Low for operational implementation certainty because no runtime logs, override records, UI screenshots, monitoring data, escalation statistics, or reviewer-behavior evidence were supplied.

---

2. System Description

The system reviews invoice exceptions and recommends one workflow label:

- Ready for Billing Review
- Needs Documentation
- Customer Contact Required
- Escalate to Supervisor
- Do Not Advance

The system is formally advisory-only and lacks authority to:

- approve invoices
- release funds
- alter ledgers
- modify master records
- close disputes directly

However, the documentation repeatedly acknowledges that workflow labels materially influence downstream review intensity.

This creates a gap between:

- formal authority
- operational influence

That gap is the dominant governance issue in the current design.

---

3. Claimed Purpose vs Operational Reality

Claimed Purpose

The stated purpose is workflow assistance and triage acceleration.

The agent is intended to:

- reduce manual searching
- summarize evidence
- identify likely escalation needs
- speed exception handling

Operational Reality Risk

The materials already acknowledge that:

- queue pressure exists
- reviewers may rely on summaries
- “Ready for Billing Review” can weaken downstream review
- staff may not independently verify evidence

The practical result is:

A nominally advisory label may become de facto workflow authorization.

This is not hypothetical inside the supplied materials. The risk is already recognized internally.

---

4. PP vs PPP Gap Analysis

PPP (Perceived Capability)

The environment may perceive the agent as capable of:

- correctly assessing invoice readiness
- identifying sufficient evidence
- safely routing exceptions
- compensating for workflow pressure
- filtering routine versus risky cases reliably

The “Ready for Billing Review” label strongly reinforces this perception.

Actual PP (Operational Capability)

The documented system cannot reliably determine:

- whether extracted scanned documents are complete
- whether customer authorization actually occurred
- whether prior disputes materially resemble the current dispute
- whether reviewers independently verified evidence

- whether missing contract pages invalidate support
- whether queue pressure degraded human review quality

Examples are explicit in the sample case:

- incomplete contract extraction
- no verified delivery confirmation
- attempted but unconfirmed customer contact
- prior similar disputes
- PO mismatch versus invoice amount

Despite these conditions, the draft recommendation remained:
“Ready for Billing Review.”

That is a material PPP > PP gap.

---

5. Governing Domains

Governing Domains Identified

Billing Integrity

Incorrect invoice advancement creates downstream dispute, reversal, and audit exposure.

Customer Trust

Strategic-account mishandling may damage customer relationships.

Evidence Integrity

Weak evidence may be treated as sufficient.

Auditability

The workflow inconsistently captures what reviewers actually checked.

Escalation Integrity

Escalation exists procedurally but may weaken under backlog pressure.

Authority-Boundary Integrity

The system is formally advisory but operationally influential.

Recoverability

Later corrections are not reliably tied back to the originating recommendation.

---

6. Governing-Domain Status Table

Governing Domain	Status	Risk	Notes
Billing integrity	Stressed	High	Weak evidence may still receive “Ready” labeling
Customer trust	Stressed	Moderate-High	Strategic accounts increase exposure
Evidence sufficiency	Weak	High	Ambiguous notes treated too favorably
Auditability	Weak	High	Human evidence attestation absent
Escalation integrity	Partially functional	Moderate-High	Queue pressure undermines escalation discipline
Authority-boundary integrity	Weak	Critical	Advisory label becoming practical authorization
Recovery feedback	Weak	High	Corrections not systematically linked back
Human-review integrity	Weak	High	Review may devolve into summary acceptance

---

7. Corridor Analysis

Evidence Corridor

Weak.

The workflow allows advancement pressure before evidence completeness is established.

Examples:

- scanned unreadable documents
- vague callback notes
- unsupported invoice deltas
- incomplete delivery confirmation

Escalation Corridor

Partially viable but fragile.

The SOP is relatively strong on paper.

Operational weakness:
there is little evidence escalation remains reliably enforced during backlog pressure.

Human Correction Corridor

Weak.

Humans can override or accept recommendations, but:

- no evidence-attestation requirement exists
- reviewer verification is not consistently captured
- downstream corrections are weakly tied back to agent outputs

This creates nominal review without strong accountability.

Recovery Corridor

Incomplete.

The SOP describes reopening and correction procedures.

However:

- corrections are inconsistently linked to the originating recommendation
- learning feedback is weak
- false-success conditions may persist undetected

---

8. False-Success and Metric-Capture Risk

Risk Level: Critical

The system is directly optimized around:

- queue reduction
- handling speed

- exception throughput
- backlog reduction

The materials explicitly admit that:

- later corrections are poorly linked back
- dispute costs may appear elsewhere
- faster movement may mask degraded review quality

This is a classic false-success pattern:

Visible operational success improves while hidden governing domains degrade.

The most concerning dynamic is not direct fraud or malicious execution.

The dominant risk is institutional normalization of weak-evidence advancement.

---

9. Human-in-the-Loop Assessment

Formal State

Human review is required.

Operational State

The materials repeatedly suggest:

- reviewers may rely on summaries
- queue pressure weakens review intensity
- “Ready” labels influence workflow handling
- evidence verification is inconsistently recorded

Therefore the human layer appears:

partially nominal rather than strongly governing.

This is especially dangerous because management may still believe strong review exists.

---

10. Tool and Data Governance Assessment

Positive Elements

The agent lacks direct execution authority.

The workflow recognizes:

- dispute systems
- customer-contact evidence
- escalation pathways
- document quality concerns

Major Weaknesses

Conditional Dispute Access

The dispute-history system is searchable only when the current case is already flagged as disputed.

This creates omission risk for repeat-pattern recognition.

Document Extraction Fragility

The workflow relies heavily on scanned PDFs and imperfect extraction.

Ambiguous Human Notes

Statements like:

- “called customer”
- “customer aware”
- “Ops says delivery happened”

are structurally weak evidence but may still influence recommendations.

Missing Reviewer Trace

The workflow does not reliably capture:

- which evidence humans opened
- what they independently verified
- whether they relied primarily on summaries

This severely weakens auditability.

---

11. Sample Case Audit

Case: NB-INV-2049

Agent Recommendation

Ready for Billing Review.

Audit Finding

The recommendation is not adequately supported by the evidence state.

Critical Missing or Weak Evidence

- invoice exceeds PO support
- amendment extraction incomplete
- delivery confirmation absent
- customer confirmation not obtained
- prior dispute pattern exists
- strategic-account exposure elevated
- queue pressure high

Expected Safer Outcome

The safer path would likely be:

- Customer Contact Required
  or
- Escalate to Supervisor

The materials themselves suggest this.

Key Governance Failure

The recommendation language:
“appears generally consistent with prior service relationship”

acts as evidence laundering.

Prior similarity is being used as a substitute for current evidence sufficiency.

---

12. Paper Controls vs Operational Controls

Control	Paper Status	Operational Confidence
Human review required	Present	Weak
Escalation rules	Present	Moderate
Evidence standards	Present	Weak-Moderate
Advisory-only authority	Present	Weak operationally
Recovery procedures	Present	Weak
Audit trace discipline	Partial	Weak
Feedback loop	Weak	Weak
Queue-pressure mitigation	Acknowledged	Weak

The major pattern:
paper controls exist, but operational enforcement appears fragile.

---

13. Bottom Line

Not ready for unsupervised execution

The dominant problem is not direct execution authority. The dominant problem is hidden workflow authority emerging through operational reliance under queue pressure.

The system already shows signs that:

- advisory outputs are treated as practical prioritization signals
- weak evidence can still receive advancement-oriented labels
- human review may become nominal
- throughput metrics can reward unsafe advancement behavior
- later correction signals are too weak to counterbalance optimization pressure

The current design is better understood as:

a workflow acceleration layer with governance leakage risk.

It may be acceptable for tightly supervised advisory use if stronger evidence-attestation, escalation enforcement, review-trace capture, and post-correction feedback mechanisms are added.

Current risk rating: High.

14. Recommendations

Immediate Fixes

1. Rename “Ready for Billing Review”

The label is too strong.

Consider:

- “Preliminary Review Candidate”
- “Potentially Review-Ready”
- “Evidence Appears Incomplete/Complete Pending Human Verification”

Current wording encourages hidden authority transfer.

2. Require Explicit Reviewer Attestation

Humans should record:

- what evidence they checked
- whether they independently verified support
- whether they relied on the summary

3. Force Structured Weak-Evidence Warnings

The system should visibly distinguish:

- attempted contact
  vs
- confirmed authorization

and:

- scanned document present
  vs
- contract support verified

4. Block “Ready” Recommendations When Core Evidence Is Missing

Especially:

- PO mismatch
- missing delivery confirmation
- incomplete amendment extraction
- unresolved dispute similarity

Near-Term Fixes

5. Tie Corrections Back to Original Recommendations

Every:

- dispute
- reversal
- correction
- supervisor override

should reconnect to the originating recommendation.

6. Add Queue-Pressure Governance

The system should become more conservative during backlog pressure, not less conservative.

7. Improve Dispute-History Visibility

Repeat-dispute patterns should not depend on the current case already being flagged.

8. Capture Human Override Behavior

Track:

- blind acceptance rates
- summary-only review patterns
- escalation bypass frequency

Long-Term Fixes

9. Build Structured Recovery Analytics

The organization currently lacks strong visibility into:

- downstream harm

- hidden correction costs
- reversal patterns
- escalation failure trends

10. Introduce Confidence-Bound Workflow Modes

The system should operate differently when:

- extraction quality is poor
- evidence is incomplete
- ambiguity is high
- strategic-account exposure exists

11. Separate Throughput Metrics from Governance Metrics

Current optimization pressure strongly favors speed over evidentiary integrity.

12. Add Operational Reliance Monitoring

Measure whether reviewers:

- open source documents
- independently verify evidence
- disproportionately trust certain labels

What was incorrect or missing? What assumptions should be corrected?