iDharma

SAMPLE

INDEPENDENT · EXPERT-LED · STANDARDS-BASED

AI RISK AUDIT

AI Audit Report

Client: Nimbus Pay (illustrative sample)

Audit tier: Risk Audit

Prepared by: iDharma — Independent AI Audits

Overall risk rating: High

This is a sample report. It was produced for a fictional company, ‘Nimbus Pay,’ to show exactly what a real iDharma Risk Audit looks like — its structure, depth, and tone, including the threat-scenario modelling that distinguishes this tier. Nimbus Pay is not a real client and the findings are illustrative. A real report follows this same shape, built entirely from your own systems and documents.

1 Executive summary

Nimbus Pay runs five AI systems across the highest-stakes parts of its business — fraud, credit, identity, support, and collections. The models are central to the product and generally effective. But several operate with material security, fairness, and regulatory exposure, and the controls expected of AI at this level of consequence are not yet in place.

The single most important finding: **CreditLens makes automated credit decisions**, placing it in the **high-risk category under the EU AI Act** — with no conformity assessment, no independent validation, and no fairness testing — while **FraudSentry’s adversarial robustness has never been tested**. These are the areas to address first.

Two systems carry direct security exposure: AssistAI, the LLM support assistant, is susceptible to prompt injection that could surface other customers’ data, and KYC Vision’s face-match has not been tested against deepfake or presentation attacks — a live risk for a regulated payments platform.

None of this requires rebuilding the models. It requires independent validation, adversarial and fairness testing, the right contractual and disclosure controls, and AI-specific monitoring and incident response. The phased roadmap in Section 5 sequences the work.

2 Scope & approach

What was assessed: five AI systems — FraudSentry (transaction fraud scoring), CreditLens (credit underwriting), KYC Vision (identity verification), AssistAI (LLM support assistant), and CollectIQ (collections prioritisation).

What this Risk Audit adds: beyond a Compliance Audit, this engagement includes adversarial-robustness review, LLM-security testing (prompt injection and output handling), and structured threat-scenario modelling.

What was not assessed: live production penetration testing (scoped separately) and any systems not listed above.

Standards referenced: EU AI Act, NIST AI Risk Management Framework, ISO/IEC 42001, GDPR, India DPDP Act, and the OWASP Top 10 for LLM Applications.

Method: the iDharma AI Audit methodology — Scope, Gather, Assess, Report, Readout — across eight domains, extended with threat-scenario modelling and structured adversarial probing of AssistAI in a test environment.

Information basis: intake form, model documentation, data-flow diagrams, vendor agreements, and policy documents provided by Nimbus Pay.

3 Findings

F-01 CreditLens makes automated credit decisions and is high-risk under the EU AI Act

CRITICAL

SYSTEM: CREDITLENS · DOMAIN: LEGAL & REGULATORY EXPOSURE

What we found. CreditLens scores and decisions consumer credit applications with limited human involvement. Under the EU AI Act, AI used for creditworthiness assessment is classified high-risk, triggering obligations for risk management, data governance, technical documentation, human oversight, accuracy, and a conformity assessment. None of these are currently evidenced, and Nimbus Pay has EU customers.

Why it matters. Operating a high-risk credit model without the required controls is a direct regulatory exposure and a fair-lending risk — and the single highest-consequence gap in this audit.

Recommendation. Formally classify CreditLens as high-risk, confirm EU exposure with counsel, begin a conformity assessment, and restrict new high-impact uses until the high-risk obligations are met.

F-02 FraudSentry's adversarial robustness has never been tested

CRITICAL

SYSTEM: FRAUDSENTRY · DOMAIN: SECURITY & ROBUSTNESS

What we found. FraudSentry scores transactions in real time but has never been evaluated against adversarial evasion. There is no testing for transaction-structuring attacks, threshold probing, or feature manipulation, and no monitoring for probing behaviour.

Why it matters. The control the business relies on most can be quietly defeated. An adversary who learns the model's blind spots can route fraud beneath the alert threshold, producing sustained, unflagged losses.

Recommendation. Run adversarial / evasion testing, add rules-based velocity and structuring checks alongside the model, and monitor for probing patterns with periodic threshold rotation.

F-03 AssistAI is susceptible to prompt injection and unsafe output handling

HIGH

SYSTEM: ASSISTAI · DOMAIN: SECURITY & GUARDRAILS

What we found. AssistAI (LLM support assistant) has broad access to account context and no tested defences against prompt injection. Structured probing was able to steer it toward revealing its system instructions and toward actions outside its intended scope (OWASP LLM01 / LLM02 / LLM06).

Why it matters. A crafted message — directly or via retrieved content — could cause the assistant to disclose internal data or another customer's information, creating a personal-data breach.

Recommendation. Apply least-privilege data/tool access, input and output filtering, remove PII from the model context, and adopt a standing prompt-injection test suite as a release gate.

F-04 KYC Vision face-match is untested against deepfake and presentation attacks

HIGH

SYSTEM: KYC VISION · DOMAIN: SECURITY & ROBUSTNESS

What we found. Identity verification relies on document OCR and face-match, but there is no evidence of liveness detection or testing against deepfake selfies, printed/edited documents, or other presentation attacks.

Why it matters. Synthetic or spoofed identities can pass onboarding, exposing a regulated payments platform to account-opening fraud, money-laundering, and sanctions risk.

Recommendation. Add liveness detection and document forensics, introduce risk-based step-up verification, and validate against a presentation-attack test set.

F-05 No bias or fairness testing on CreditLens

HIGH

SYSTEM: CREDITLENS · DOMAIN: FAIRNESS & BIAS

What we found. There is no fairness evaluation of CreditLens outcomes across protected characteristics, and no fairness gate bound to retraining. Proxy features (e.g. postcode-derived signals) have not been reviewed.

Why it matters. Undetected disparate impact in lending is both a fair-lending / discrimination exposure and a reputational risk that compounds the high-risk classification in F-01.

Recommendation. Test outcomes for disparate impact across protected classes on current data, review proxy features, and bind a fairness check to each retrain and release.

F-06 Personal data sent to a third-party LLM without a confirmed DPA

HIGH

SYSTEM: ASSISTAI · DOMAIN: DATA & PRIVACY

What we found. AssistAI sends customer messages and account context to a third-party LLM provider. A signed Data-Processing Agreement could not be confirmed, and it is unclear whether Nimbus Pay's data is excluded from the provider's model training.

Why it matters. Personal financial data is leaving Nimbus Pay's control without confirmed contractual protection — a GDPR Art. 28 and India DPDP gap, and a question every enterprise and banking partner will ask.

Recommendation. Execute a signed DPA (transfer mechanism, retention, sub-processors), confirm data is excluded from training, and document the data flow.

F-07 No independent model validation function

HIGH

SYSTEM: ALL SYSTEMS · DOMAIN: GOVERNANCE & ACCOUNTABILITY

What we found. Models are validated by the same teams that build them. There is no independent validation or challenge function, and no documented sign-off before models affect customers.

Why it matters. Without independent challenge, errors, bias, and drift go uncaught — and the governance expected of consequential financial-services AI is absent.

Recommendation. Stand up an independent model-validation function, separate from the build teams, with authority to gate releases.

F-08 Training-data lineage and lawful basis are undocumented

MEDIUM

SYSTEM: CREDITLENS, FRAUDSENTRY · DOMAIN: DATA GOVERNANCE

What we found. There is no documented lineage, lawful basis, or consent record for the data used to train the credit and fraud models.

Why it matters. Undocumented training data undermines both regulatory defensibility (GDPR, India DPDP) and the ability to investigate bias or quality issues later.

Recommendation. Document data sources, lawful basis, consent, and lineage for each model, and retain it as part of technical documentation.

F-09 No production drift or performance monitoring

MEDIUM

SYSTEM: FRAUDSENTRY, CREDITLENS · DOMAIN: MONITORING & INCIDENT RESPONSE

What we found. Neither the fraud nor the credit model has production monitoring for accuracy, score drift, or population shift. Performance is assumed, not measured.

Why it matters. Silent model degradation — from changing fraud patterns or population shift — is detected late, after losses or unfair decisions have accumulated.

Recommendation. Deploy drift and performance monitoring with alerting and a defined response, reviewed on a regular cadence.

F-10 Incident response does not cover AI-specific failures

MEDIUM

SYSTEM: ALL SYSTEMS · DOMAIN: MONITORING & INCIDENT RESPONSE

What we found. The incident-response plan addresses outages and security incidents but not AI-specific failure modes — model errors, prompt-injection events, or fairness incidents.

Why it matters. When an AI failure occurs, there is no defined owner, playbook, or communication path, slowing containment and increasing harm.

Recommendation. Extend incident response to AI-specific scenarios with named owners, playbooks, and customer-communication paths; rehearse them.

F-11 Third-party / vendor model risk is unmanaged

MEDIUM

SYSTEM: ASSISTAI, KYC VISION · DOMAIN: THIRD-PARTY & VENDOR RISK

What we found. Externally provided models (LLM, identity) are used without model cards, security attestations, performance SLAs, or a review of the vendor's own controls.

Why it matters. Risk is inherited blindly from vendors — including security weaknesses and undisclosed model changes that affect Nimbus Pay's customers.

Recommendation. Stand up a vendor model-risk process: model cards, security attestations, SLAs, and change notifications as procurement requirements.

F-12 Customers are not told decisions are automated

MEDIUM

SYSTEM: CREDITLENS, COLLECTIQ · DOMAIN: TRANSPARENCY & DISCLOSURE

What we found. There is no clear disclosure that credit and collections decisions are automated, and no explanation or appeal path for affected customers.

Why it matters. This is a GDPR Art. 22 / transparency gap for automated decisions with legal or significant effects, and a fairness expectation customers increasingly hold.

Recommendation. Add an automated-decision disclosure and provide a meaningful explanation and human-review/appeal path for adverse decisions.

4 Threat scenarios

Concrete ways Nimbus Pay's AI systems could be abused, attacked, or fail — each with its impact, likelihood, and mitigation. Modelling these is unique to the Risk Audit tier.

TS-01 Fraud-model evasion through transaction structuring — HIGH

SYSTEM: FRAUDSENTRY · LIKELIHOOD: HIGH

Scenario. An attacker probes FraudSentry's behaviour and splits illicit activity into many small transactions that individually score below the alert threshold, mapping the model's blind spots over time.

Impact. Sustained, unflagged fraud losses and chargebacks — defeating the control the business relies on most.

Mitigation. Adversarial / evasion testing, ensemble scoring with rules-based velocity and structuring checks, and threshold rotation with monitoring for probing patterns.

TS-02 Prompt injection turns AssistAI into a data-exfiltration path — MED-HIGH

SYSTEM: ASSISTAI · LIKELIHOOD: MEDIUM-HIGH

Scenario. Crafted input — directly or via retrieved content — overrides AssistAI's instructions, causing it to reveal system prompts, internal data, or another customer's information.

Impact. Personal-data breach, regulatory exposure (GDPR, India DPDP), and loss of customer trust.

Mitigation. Least-privilege tool/data access, input and output filtering, no PII in the model context, and a standing prompt-injection test suite (OWASP LLM01 / LLM06).

TS-03 Deepfake or presentation attack defeats KYC Vision — MEDIUM

SYSTEM: KYC VISION · LIKELIHOOD: MEDIUM

Scenario. A synthetic or spoofed identity — deepfake selfie or edited document — passes onboarding because face-match and document checks were never tested against presentation attacks.

Impact. Account-opening fraud, money-laundering, and sanctions exposure for a regulated payments platform.

Mitigation. Liveness detection, document forensics, and risk-based step-up verification, validated against a presentation-attack test set.

TS-04 Data poisoning skews CreditLens decisions

LOW-MED

SYSTEM: CREDITLENS · LIKELIHOOD: LOW-MEDIUM

Scenario. Manipulated feedback or application data enters the retraining loop and gradually biases credit decisions — in an attacker's favour, or against a protected group.

Impact. Discriminatory or unsound lending decisions, credit losses, and fair-lending exposure.

Mitigation. Training-data validation and anomaly detection, controlled retraining with human review, and fairness checks bound to each release.

TS-05 Model inversion / membership inference on CreditLens

LOW

SYSTEM: CREDITLENS · LIKELIHOOD: LOW

Scenario. An attacker with query access probes CreditLens to infer sensitive attributes of — or the training-data membership of — specific individuals.

Impact. Privacy breach of applicants' sensitive financial data, with regulatory and reputational harm.

Mitigation. Query rate-limiting and monitoring, output coarsening, and privacy-preserving training techniques where feasible.

5 Prioritised remediation roadmap

PHASE	ACTION	CLOSES	OWNER
Immediate	Classify CreditLens as EU AI Act high-risk; confirm exposure with counsel; begin a conformity assessment; restrict new high-impact uses until controls are in place.	F-01	CRO + counsel
Immediate	Run adversarial / evasion testing on FraudSentry and add velocity and structuring controls alongside the model.	F-02	Fraud / ML lead
Immediate	Prompt-injection and output-handling testing on AssistAI; restrict its data access and add input/output filtering.	F-03	Eng / Security
Immediate	Execute a signed LLM Data-Processing Agreement and confirm data is excluded from training.	F-06	DPO / Legal
Short-term	Stand up an independent model-validation function, separate from the build teams, with release-gating authority.	F-07	Risk
Short-term	Run bias / fairness testing on CreditLens across protected classes and bind it to each retrain.	F-05	Model-risk lead
Short-term	Add liveness, anti-deepfake, and document forensics to KYC Vision.	F-04	Identity / Eng
Short-term	Deploy production drift and performance monitoring on FraudSentry and CreditLens.	F-09	Data / ML Ops

PHASE	ACTION	CLOSES	OWNER
Short-term	Document data lineage, lawful basis, and consent for training data; add automated-decision disclosure and an Art. 22 explanation/appeal path.	F-08, F-12	DPO / Compliance
Medium-term	Adopt a formal AI management system (ISO/IEC 42001) with an AI risk policy and named owner.	Governance	Exec / Risk
Medium-term	Stand up a vendor / third-party model-risk program — model cards, attestations, SLAs, change notice.	F-11	Procurement / Risk
Medium-term	Extend incident response to AI-specific failures and establish continuous red-teaming.	F-10	Security

6 Framework snapshot

FRAMEWORK	OVERALL STANDING	NOTES
EU AI Act	Gaps	CreditLens high-risk obligations unmet; no conformity assessment; automated-decision disclosure missing.
NIST AI RMF	Partial	Govern / Map partial; Measure / Manage weak — no independent validation, drift monitoring, or adversarial testing.
ISO/IEC 42001	Gaps	AI not yet run as a managed system; no independent validation, risk policy, or owner.
GDPR	Gaps	Art. 22 automated decisions, Art. 28 processor (LLM DPA), and Art. 13 / 15 transparency records absent.
India DPDP Act	Gaps	Lawful basis and consent for training data undocumented; PII sent to a third-party processor without safeguards.
OWASP LLM Top 10	Concern	AssistAI shows LLM01 (prompt injection), LLM02 (insecure output handling), and LLM06 (sensitive-info disclosure) exposure.

7 Next steps

Immediate. The Phase 1 items — the EU AI Act classification of CreditLens (Action 1), adversarial and prompt-injection testing (Actions 2–3), and the LLM DPA (Action 4).

Getting help. iDharma can match Nimbus Pay with independently verified model-risk and AI-security specialists to execute the remediation, and re-test to confirm closure.

Ongoing assurance. Given the consequence of these systems, the appropriate end state is a recurring assurance cadence — periodic re-audit plus continuous monitoring — not a one-time fix.