Data Science & Engineering Experts, Inc.

AI Security Assessment Sample Report

Beachhead Tier [REDACTED]

Prepared for: [REDACTED Financial Services Company]
Assessment Date: [REDACTED]
Engagement Tier: Beachhead
Prepared by: Data Science & Engineering Experts, Inc. | thedataexperts.us

SAMPLE — REDACTED

Executive Summary

Assessment Scope

This Beachhead-tier engagement evaluated the security posture of the client's AI/ML model stack, including hosted large language model (LLM) inference endpoints, an internally fine-tuned classification model, and the supporting embeddings store. We reviewed the API security posture across both customer-facing and internal service-to-service routes, with particular attention to authentication, authorization, and request validation on inference paths. We further examined data-handling practices for AI workloads, tracing the flow of customer and personally identifiable information (PII) through training-data pipelines and into third-party AI vendor integrations.

Overall Risk Rating

MEDIUM-HIGH

The MEDIUM-HIGH rating reflects two critical findings that are individually exploitable today, set against an otherwise maturing AI engineering practice.

Findings Summary

Severity	Count	Disposition
CRITICAL	2	Immediate remediation required
HIGH	4	Remediate within 30 days
MEDIUM	6	Remediate within 60–90 days

Key Recommendation

> **Immediate remediation required for findings 1 and 2 before additional AI model deployment.** These two critical findings expose authentication gaps and an unmanaged PII data flow that materially increase regulatory and breach risk.

Estimated Remediation Cost

$45,000 – $85,000 in combined engineering and advisory effort, spanning the 30/60/90-day roadmap detailed on the final page.

Methodology

Assessment Approach

The assessment combined three complementary techniques: structured interviews with engineering and data leadership, review of architecture and data-handling documentation, and automated scanning of exposed AI endpoints and dependencies. Findings were corroborated across at least two of these methods before inclusion in this report.

In Scope

- AI model endpoints — public and internal inference routes for the hosted LLM and the fine-tuned classifier
- Training data pipelines — ingestion, labeling, and feature-engineering jobs that handle customer data
- API authentication — token issuance, validation, and service-to-service trust boundaries
- Prompt injection surface — customer-facing AI features that incorporate untrusted input into model prompts
- Third-party AI vendor review — contractual and technical data-sharing posture with external model providers

Out of Scope

- Physical security of data-center and office facilities
- Non-AI corporate IT endpoints (workstations, email, general SaaS administration)
- Penetration testing of production payment-processing systems

Tools & Frameworks

[redacted internal tooling] for endpoint enumeration and dependency analysis, mapped against the following industry frameworks:

- OWASP Top 10 for Large Language Model Applications
- NIST AI Risk Management Framework (AI RMF 1.0)

Example scan invocation (sanitized):

```
$ [redacted-scanner] scan --target ai-inference.internal --framework owasp-llm-top10 --report json
```

Findings — Critical & High

CRITICAL

Finding 1: AI API Endpoint Missing Authentication on Internal Routes

Description

The internal inference API exposes several service-to-service routes — including `/v1/models/classify` and `/v1/embeddings` — that perform no authentication and rely solely on network placement for protection. Any workload able to reach the service subnet, including a compromised low-privilege container, can invoke the model directly and exfiltrate embeddings or trigger arbitrary classifications. The gateway enforces JWT validation on the public edge but the internal mesh trusts all callers implicitly. We confirmed unauthenticated access from a non-privileged pod during testing.

Risk

An attacker with any foothold on the internal network can abuse the model for data exfiltration, cost-inflation attacks, or as a pivot to the embeddings store, which contains derivations of customer data. This violates the principle of zero-trust service-to-service authentication and would likely constitute a reportable control failure under the client's financial-services obligations.

Remediation

- Enforce mutual TLS (mTLS) or signed service identity (e.g., SPIFFE/SVID) on all internal inference routes; deny by default.
- Issue short-lived, audience-scoped service tokens validated at the inference service itself, not only at the public gateway.
- Add network policy to restrict the inference subnet to explicitly allow-listed caller workloads.

CRITICAL

Finding 2: Training Data Pipeline Exposes PII to Third-Party AI Vendor Without DPA

Description

The training-data pipeline forwards raw customer records — including names, account identifiers, and transaction descriptions — to a third-party model-hosting vendor for fine-tuning, with no data processing agreement (DPA) in place and no field-level redaction or tokenization applied. Data is transmitted over TLS but is retained by the vendor under their default retention policy, which permits use of submitted data for their own model improvement. There is no documented data inventory describing which PII fields leave the client's environment.

Risk

Sharing unredacted PII with a processor absent a DPA creates direct exposure under GLBA and applicable state privacy law, and the vendor's secondary-use clause means customer data may be incorporated into models outside the client's control. A vendor breach or subpoena would implicate the client with no contractual recourse and no ability to demonstrate the data-minimization the regulators expect.

Remediation

- Immediately halt transmission of unredacted PII; apply tokenization or field-level redaction before any data leaves the environment.
- Execute a DPA with the vendor that prohibits secondary use and sets a defined retention/deletion schedule.
- Build and maintain a data inventory mapping every PII field that crosses the trust boundary, reviewed quarterly.

HIGH

Finding 3: No Prompt Injection Controls on Customer-Facing AI Feature

Description

The customer-facing AI assistant concatenates user-supplied text directly into the system prompt with no input sanitization, delimiter enforcement, or output filtering. During testing, a crafted message overrode the assistant's instructions and caused it to disclose portions of its system prompt and adjacent context. No allow-list governs which downstream tools or data the model may access on the user's behalf.

Risk

Prompt injection can be used to leak internal instructions, manipulate the assistant into producing unauthorized actions, or coax disclosure of context belonging to other sessions. As the feature gains access to more internal tools, the blast radius of a successful injection grows accordingly.

Remediation

- Separate trusted instructions from untrusted input using structured prompting and strict input/output validation.
- Constrain tool and data access with a least-privilege allow-list enforced outside the model.

HIGH

Finding 4: AI Model Outputs Not Logged for Audit Trail

Description

Inference requests and model responses are not persisted to any audit log. Only aggregate latency and error metrics are retained, with no record of prompts, outputs, model version, or the identity of the caller. There is no immutable trail that would allow reconstruction of what a model produced for a given customer interaction.

Risk

Without an audit trail the client cannot investigate misuse, demonstrate compliance, or perform incident response on AI-driven decisions affecting customers. This is a gap against examiner expectations for explainability and recordkeeping in regulated financial services.

Remediation

- Log prompt, response, model version, caller identity, and timestamp to a tamper-evident, access-controlled store.
- Define a retention schedule aligned to regulatory recordkeeping requirements and redact PII at write time.

Findings — Medium

The medium-severity findings below do not present immediate exploitability but should be remediated within the 60–90 day window to reduce aggregate risk and operating cost.

Finding	Risk	Remediation Owner
Verbose AI error messages leak stack traces and model metadata	Information disclosure	Platform Eng
No rate limiting on inference endpoint	DoS / cost abuse	API team
Embeddings store lacks encryption at rest	Data exposure	Data Eng
Stale third-party model SDK with known CVE	Supply chain	DevOps

Remediation Roadmap

The following 30/60/90-day plan sequences remediation by severity and dependency. Estimated costs reflect blended engineering and advisory effort and roll up to the $45,000 – $85,000 range cited in the executive summary.

Window	Workstream	Addresses	Estimated Cost
0–30 days	Enforce internal service authentication (mTLS/signed identity); halt and redact PII flow to vendor; execute DPA	Findings 1, 2	$18,000 – $32,000
30–60 days	Prompt-injection controls and tool allow-listing; AI output audit logging with tamper-evident store	Findings 3, 4	$15,000 – $28,000
60–90 days	Error-message hardening, inference rate limiting, embeddings encryption at rest, SDK CVE remediation, data inventory	Medium findings	$12,000 – $25,000
Total	Full Beachhead remediation program	All findings	**$45,000 – $85,000**

Two additional medium-severity findings relating to model-version pinning and CI/CD secret handling are documented in full in the client's confidential appendix and are omitted from this redacted sample.

> This sample is provided to illustrate DSE's Beachhead-tier deliverable format. Your actual report will reflect your specific environment and findings. Contact: thedataexperts.us/engage