IIS AI PRACTICE

AI-Assisted Comparison for **Financial Reporting**

A working concept for AI-assisted investment-report comparison — built as a live demonstration: how we built it, how it works, and why it's safe. AI That Works for you.

IIS TECHNOLOGY — ISO 9001 CERTIFIED — International Integrated Solutions, LTD

ABOUT IIS

IIS Technology is a 30-year, privately held enterprise IT integrator headquartered in Plainview, NY — ISO 9001-certified, with a Net Promoter Score of 87 (the IT-services average is around 55). Our AI Practice is workshop-led: we help you rapidly define and implement valuable use cases, and guide you from the ideal use case to production. Because one size does not fit all, we tailor each solution and anchor it to enterprise-proven platforms — Red Hat OpenShift AI and NVIDIA AI Enterprise — with partners including NVIDIA, AMD, HPE, IBM, Red Hat, and Microsoft. From fast prototype to inference in weeks: AI That Works for you.

30+
Years privately held

NPS 87
vs. industry avg ~55

ISO 9001
Certified annually

THE AI REALITY IN FINANCIAL SERVICES

What the data says about AI in banking

These are measured results and adoption signals from regulated financial-services environments — the shift already underway, and the risks examiners now probe.

$5.56M
average financial-services data-breach cost — 2nd highest of any sector
IBM Cost of a Data Breach, 2025

$670K
added to the average breach cost by shadow AI
IBM Cost of a Data Breach, 2025

55%
of community banks cite data leakage as their top AI challenge
Wolf & Company, 2026

Only 5%
have a scaled, governed AI program in production today
Wolf & Company, 2026

> *"Over 40% of agentic AI projects will be canceled by the end of 2027" — Gartner, June 2025 — citing escalating costs, unclear business value, and inadequate risk controls. The projects that survive are the governed ones. That is how this concept is built.*

WHAT WE HEARD FROM YOUR TEAM

One quote. Three patterns. Six requirements.

> *“We want to take an investment report from one year and one from the next, upload them, and say — please compare the two and give me an update. We have social security numbers, we have client information. We don’t want to give that to the model.”*
>
> — A regulated-bank technology leader, first discovery call

Every capability in the concept build traces back to one of six requirements. Each maps to one of three building blocks: prediction and detection models trained on your own data · a language model that turns verified numbers into plain English · and the architecture and human approval gate — nothing executes without a named person signing off.

The use case is concrete

Compare two investment reports, summarize the delta — a narrow, valuable, repeatable job the CEO can see. · **ARCHITECTURE**

PII is the gating constraint

SSNs, account numbers, client identifiers cannot enter a public model. That one fact reshapes every technology choice. · **PRIVACY GATE**

Numbers the CEO can defend

Every figure must be one an examiner could trace — not a number the model invented. · **DETERMINISTIC CODE**

Policy before platform

Fine-tune the AI policy first, then align technology to it — not the reverse. · **GOVERNANCE**

Shadow AI is already here

Staff paste reports into public tools today — off-policy, invisible, and unverifiable. This closes that gap. · **PRIVACY GATE**

An answer when a regulator asks

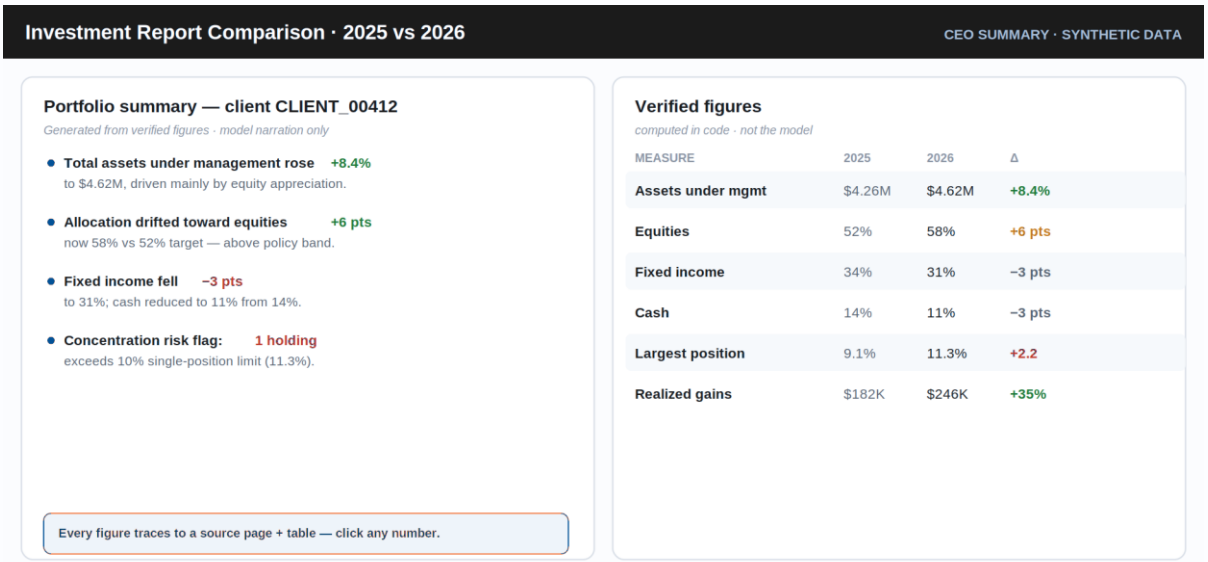
There is no audit trail today for how a summary was produced. Every step must be logged. · **AUDIT TRAIL**

H O W T H E S Y S T E M W O R K S

Five steps — from upload to CEO-ready summary

> *The system takes two reports, strips client identifiers before anything reaches a model, computes every figure in testable code, explains the change in plain language, and drafts a CEO-ready summary — which a named person approves before anything is shared.*

FROM THE WORKING DEMONSTRATION · SYNTHETIC DATA



What leadership sees — the year-over-year comparison: assets up 8.4%, equity allocation drifted 6 points above the policy band, and one holding flagged for breaching the single-position limit. Every figure on the right is computed in code and traceable to a source page; the model only narrates them.

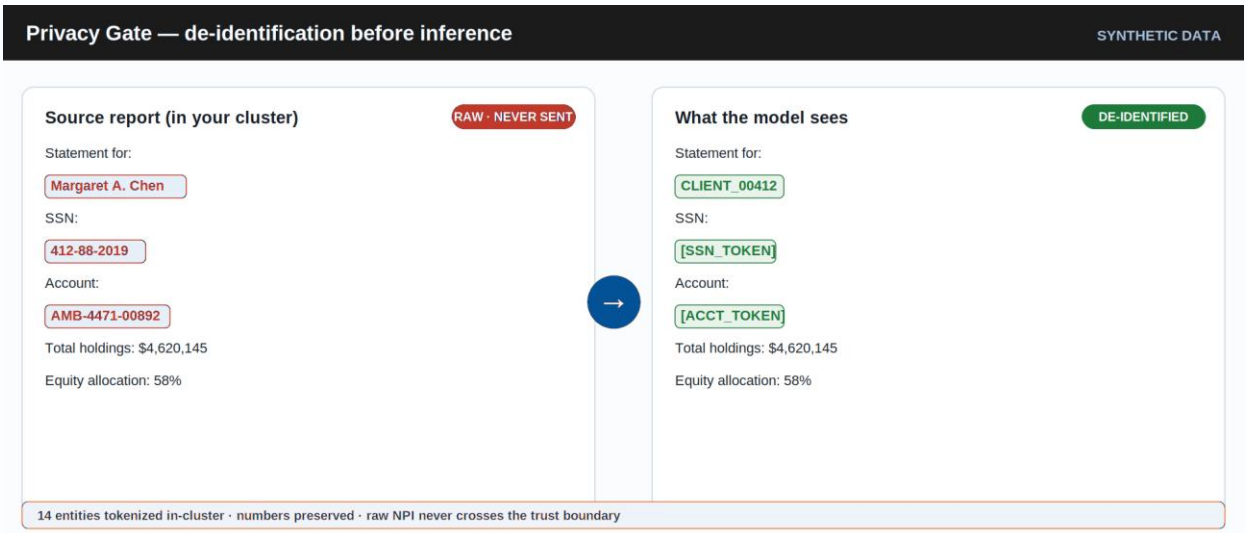
1	**Ingest the reports** Two reports are uploaded inside your environment. Balances, holdings, allocation percentages and client identifiers are parsed from mixed PDF tables — nothing is sent anywhere yet.
2	**De-identify before the boundary** Names, SSNs and account numbers are tokenized inside your perimeter before any model sees them. A custom model can learn new bank-specific identifiers on-prem — the raw data never leaves.
3	**Compute, then narrate** Every number — allocation drift, concentration flags, period-over-period deltas — is computed by testable code. The language model narrates the verified figures; it never calculates and never sees raw client data.
4	**Draft the summary** A CEO-ready comparison is drafted from the verified numbers, with each figure traceable to its source page and table. High-confidence cases only; anything ambiguous is escalated.
5	**Approve (human required)** A named person reviews and approves. Nothing is shared, and no client is re-identified, without this step — and every decision is logged with who, when, and why.

Steps 1–4 inform and propose. **Step 5 is the only action — and only a named person can take it. Nothing auto-executes.**

T H E A R C H I T E C T U R E

Governed AI pipeline with full audit & oversight

FROM THE WORKING DEMONSTRATION · SYNTHETIC DATA



The Privacy Gate — before anything reaches a model, client identifiers are swapped for tokens inside your cluster. Names become CLIENT_00412, SSNs and account numbers become sealed tokens; the numbers are preserved so the comparison still works. Raw client data never crosses the trust boundary.

FROM THE WORKING DEMONSTRATION · SYNTHETIC DATA

Portfolio Report Card · quarter at a glance — LEADERSHIP VIEW · SYNTHETIC DATA

MEASURE	TARGET	Q1	Q2	Q3	QUARTER	TREND	STATUS
Allocation drift	±4%	+3	+5	+6	**+6 pts**	▲	ACTION
Concentration breaches	0	0	1	1	**1**	▲	ACTION
Cash vs policy	10–15%	13%	12%	11%	**11%**	▼	WATCH
Fee load (bps)	75	72	74	73	**73**	▬	OK
Realized gains	—	$140K	$205K	$246K	**$246K**	▲	INFO
Reports reconciled	100%	98%	99%	100%	**100%**	▲	OK
Time-to-summary	<1 day	2.1d	0.4d	0.2d	**0.2d**	▼	OK

The portfolio report card — each row is a number leadership already tracks: allocation drift, concentration breaches, cash against policy, fee load. Arrows show the trend and a Red/Amber/Green (RAG) flag shows where attention is needed — allocation drift is +6 points against a ±4-point band, so it is flagged for action.

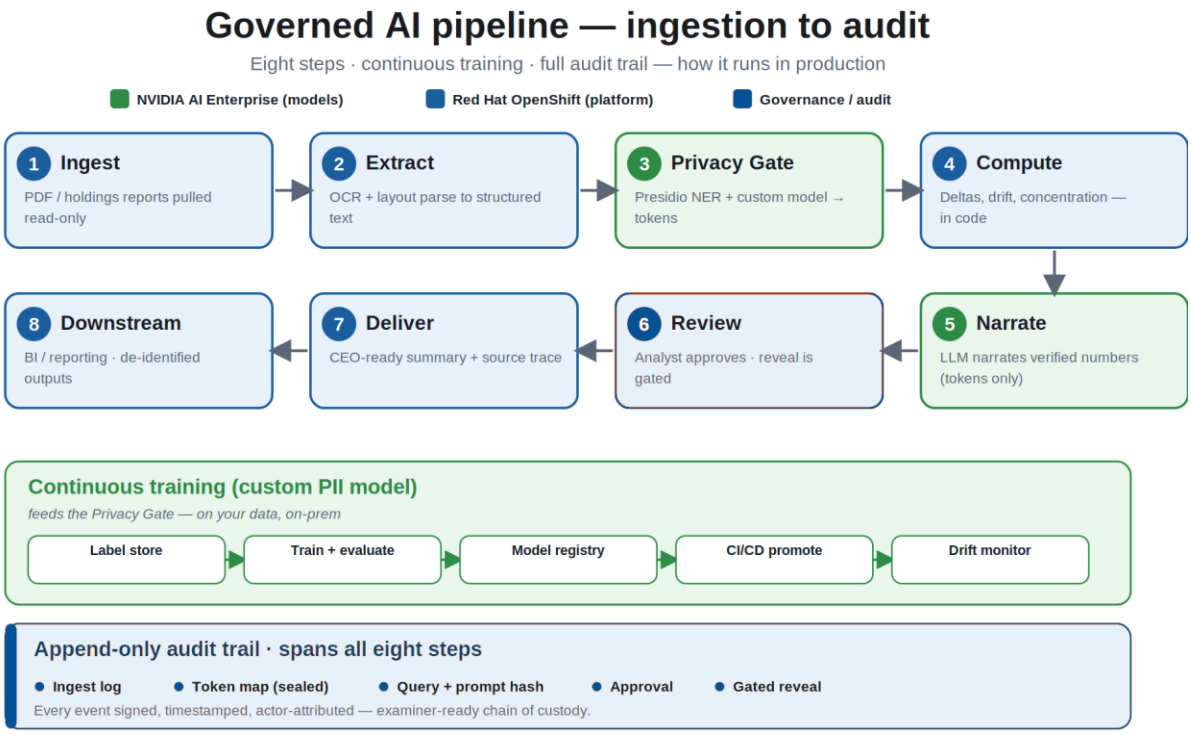
FROM THE WORKING DEMONSTRATION · SYNTHETIC DATA

The analyst's review queue — every report awaiting sign-off gets a drift figure, a model-confidence score, and the reason it was flagged, sorted into Red/Amber/Green (RAG) tiers with a suggested next step. Low-risk, in-band reports can be approved in a batch; anything flagged escalates for review.

P R O D U C T I O N A R C H I T E C T U R E

Full governed pipeline — from ingestion to audit

The diagram below shows the complete production pipeline — from report ingestion through de-identification, deterministic computation, human approval, and downstream reporting — with the continuous model-training loop and full audit trail beneath it, all on Red Hat OpenShift AI and NVIDIA AI Enterprise.

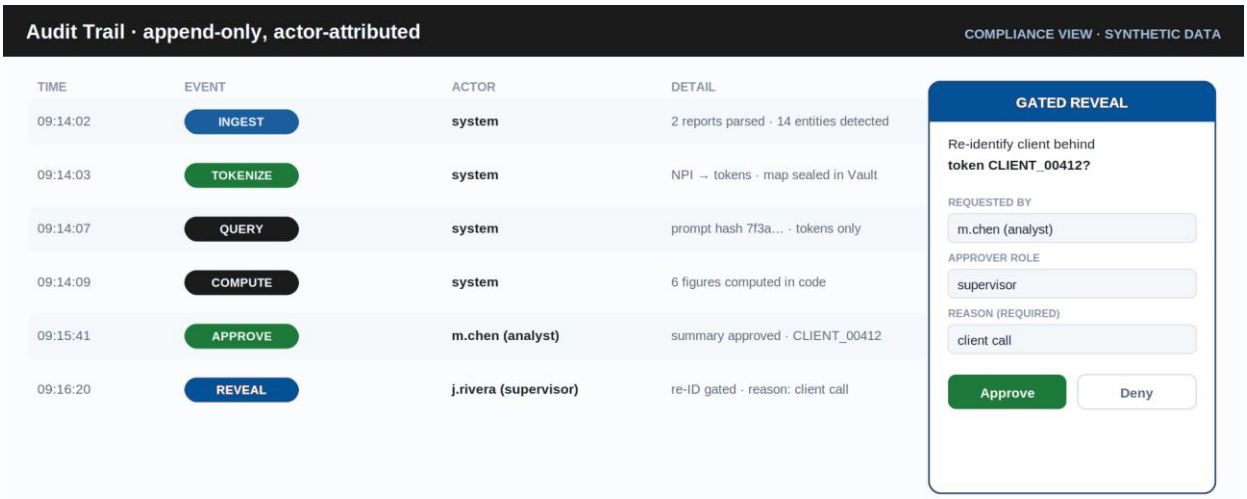


Safety by design — gated, logged, reversible

Four design principles govern the system. Each is a direct response to what banking compliance and IT teams ask about AI — and each is evidenced in the architecture above.

- **Client data never leaves your perimeter** Non-public personal information (NPI) is tokenized inside your environment before any model layer sees it. Only tokens and verified aggregate numbers ever reach an inference endpoint — satisfying the GLBA third-party obligation that sending raw reports to a cloud model would create.
- **Every action requires a human** The system recommends and drafts. A named, audited approver takes every action. Nothing is shared, and no client is re-identified, without explicit human sign-off — keeping the workflow clear of the ECOA/UDAAP exposure that autonomous client-facing AI carries.
- **Numbers come from code, not the model** Every consequential figure is computed by testable, deterministic code; the language model is a bounded narration layer only. This directly closes the governance gap that the April 2026 model-risk guidance (SR 26-2 / OCC 2026-13) leaves open for generative AI.
- **Everything is logged immutably** Three distinct logs — ingest, query, and reveal — capture every entity detected, every prompt, and every re-identification, each signed and actor-attributed. You can trace any summary back to its origin; compliance and regulatory reporting surface automatically.

FROM THE WORKING DEMONSTRATION · SYNTHETIC DATA



The audit trail with a gated reveal — every step is appended to a tamper-evident log: ingest, tokenization, query, computation, approval, and reveal. Re-identifying a client behind a token requires an explicit, role-gated approval, captured with the actor, the time, and the reason.

A L R E A D Y R U N N I N G I N P R O D U C T I O N

The same stack — proven elsewhere

The architecture proposed here is not experimental. The same platform components are documented reference architectures and run in production at scale today.

Banco Bradesco	Built BRIDGE, a generative-AI platform on Red Hat OpenShift, scaling to 500+ AI initiatives in production with a 10x reduction in integration cycles and 10x greater agility deploying agents (Red Hat Ecosystem Innovation Awards, 2026).
STAC-AI · SEC filings	Red Hat, NVIDIA and Supermicro ran the finance-industry LLM inference benchmark on OpenShift — the first audited STAC-AI results on a containerized Kubernetes platform, on datasets from real SEC EDGAR filings.
DenizBank	120+ data scientists develop models on Red Hat OpenShift AI across lending, credit cards, and risk (Red Hat case study, 2025).
NVIDIA NIM	Self-hosted inference microservices deploy production model endpoints in minutes — “data never leaves your enclave,” with the same code from cloud to on-prem.
Microsoft Presidio	The open-source (MIT-licensed) engine behind the Privacy Gate — detects and tokenizes SSNs, account numbers, and names; independent evaluation reports ~91% F1 on financial documents.

What the industry has learned — including the failures

MIT’s 2025 State of AI in Business study found 95% of enterprise AI pilots deliver no measurable financial return — and that buying from specialised vendors succeeded about 67% of the time, while internal builds succeeded only a third as often. It is why we run workshop-led, partner-anchored engagements priced on outcomes, not open-ended pilots.

The 2025 fair-lending enforcement action is the cautionary tale — a Massachusetts case settled for $2.5M under ECOA where an AI system contributed to discriminatory outcomes with no human reviewer in the loop. It is why every output in this design is decision-support only: the AI prepares, and a named person approves, before anything reaches a client relationship.

THE PATH TO PRODUCTION

Phased — shippable at every gate

We propose phased delivery. Each phase is shippable on its own, and you decide whether the next one makes sense after it delivers — our AI Foundry process takes you from idea to a functional pilot in weeks. There is no contractual escalator.

Phase	Timeframe	Deliverable	What you get
P0 · Align	Weeks 1–4	AI-policy workshop · data classification review · P1 scope agreement	**Shared understanding and a signed scope before anything is built**
P1 · Prove	Weeks 4–12	Privacy Gate + guardrails · single report-comparison job · production deploy on one pair	**See it run in your workflow, live on the first pair of reports**
P2 · Expand	Months 3–6	Additional report types · model validation on de-identified data · InfoSec sign-off	**Signed-off, defensible models and broader coverage**
P3 · Govern	Months 6+	Full audit + reporting · policy-as-config routing · Day-2 operations	**Production, governed, and examiner-ready across the workflow**

HOW WE WORK

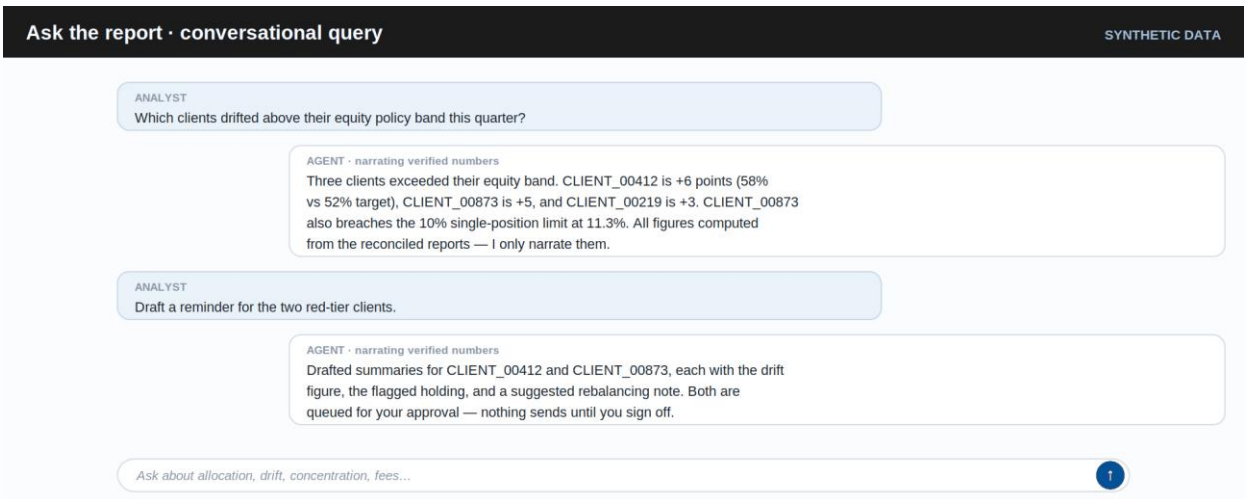
AI That Works — for the enterprise

Every IIS AI Practice engagement — including this one — follows the same model:

- **Workshop-Led** From use case to production via structured engagement. We map workflows and data before writing a line of code.
- **Partner-Anchored** NVIDIA, AMD, HPE, IBM, Red Hat, and Microsoft — the enterprise-proven platforms deployed at regulated institutions. One size does not fit all.
- **Production-Tested** Years of trial, error, and live enterprise deployments behind every recommendation.
- **Outcome-Priced** Fixed-cost engagements, not open-ended consulting hours. You know the price and the deliverable before we start.
- **Multi-Discipline Expertise** Advisory across design, deployment, and Day-2 operations for every layer of your AI stack.

Production AI, paid for by the outcomes it produces.

FROM THE WORKING DEMONSTRATION · SYNTHETIC DATA



A live exchange from the working demonstration — an analyst asks about equity-allocation drift in everyday language, and the agent answers with each client's figure and a recommendation, then drafts reminders that queue for approval. The chat drives the same governed system as the dashboards, so nothing gets out of sync — and it only narrates verified numbers.

Where the numbers and architecture come from

Every statistic and architecture claim in this document traces to a published, publicly verifiable source. The production references below are the same reference architectures our proposed stack is built on.

Statistics cited in this document

IBM Cost of a Data Breach (2025) — financial-services breach cost ($5.56M) and shadow-AI cost premium ($670K).
https://www.ibm.com/reports/data-breach

Wolf & Company (2026) — AI adoption & maturity survey; data-leakage and governed-program figures for community banks.
https://www.wolfandco.com/resources/insights/survey-results-2026-ai-adoption-maturity-in-banking/

Gartner (June 2025) — over 40% of agentic AI projects to be canceled by end-2027.
https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

MIT NANDA — State of AI in Business (2025) — 95% of pilots show no P&L impact; specialised-vendor purchases succeed ~67% vs internal builds.
https://fortune.com/2025/08/18/mit-report-95-percent-generative-ai-pilots-at-companies-failing-cfo/

Bloomberg Law (2025) — ECOA fair-lending AI settlement ($2.5M).
https://news.bloomberglaw.com

Reference architecture, platform & public deployments

Red Hat + NVIDIA AI Factory — verified reference architecture for the OpenShift AI + NVIDIA AI Enterprise stack proposed here.
https://docs.nvidia.com/ai-enterprise/deployment/red-hat-ai-factory/latest/overview.html

Banco Bradesco — Red Hat Ecosystem Innovation Awards 2026 (500+ AI initiatives; 10x integration-cycle reduction).
https://www.redhat.com/en/blog/announcing-winners-2026-red-hat-ecosystem-innovation-awards

STAC-AI LANG6 on Red Hat OpenShift + NVIDIA — first audited LLM-inference benchmark on containerized Kubernetes.
https://www.redhat.com/en/blog/red-hat-openshift-delivers-high-performance-llm-inference-financial-services

Microsoft Presidio — open-source (MIT) de-identification engine used for the Privacy Gate.
https://github.com/microsoft/presidio

NVIDIA NIM microservices — self-hosted inference; data stays in your enclave.
https://www.nvidia.com/en-us/ai-data-science/products/nim-microservices/

Red Hat OpenShift AI — model-as-a-service serving platform.
https://www.redhat.com/en/technologies/cloud-computing/openshift/openshift-ai

---

Ready to talk? Contact the IIS AI Practice · IIS Technology · Plainview, NY · **info@iisl.com** · **iistech.com**