TANTOR™
a translab.io product

TANTOR WHITEPAPER

Sovereign AI for Indian Public Sector Banking

A PII Data Boundary Architecture for Hybrid Cloud Agentic AI

A reference architecture aligned with the Digital Personal Data Protection Act 2023, the RBI FREE-AI Framework (Framework for Responsible and Ethical Enablement of AI, August 2025), UIDAI Aadhaar masking, CERT-In directions and India's sovereign AI mission. Built on open-source, open-weights and open standards.

info@tantor.ai | tantor.ai

Table of Contents

Table of Contents

Table of Contents

Table of Contents

Executive Summary

Indian Public Sector Banks face a defining challenge of this decade. They must adopt agentic artificial intelligence at scale to remain competitive on customer experience, fraud detection, credit decisioning and operational efficiency. They must do this while operating under a regulatory regime that has, correctly, drawn a hard line around the residency, purpose and accountability of personal data.

The Digital Personal Data Protection Act 2023, the RBI FREE-AI Framework of August 2025, the RBI Master Direction on KYC, the UIDAI Aadhaar masking circular of October 2018, the RBI data localisation circular of April 2018 and the CERT-In directions of April 2022 collectively make one thing clear. Production personally identifiable information must remain within Indian jurisdiction, must be processed for a stated purpose, must be auditable to a fiduciary standard, and must remain under the bank's effective control. The FREE-AI framework adds a second mandate: the AI itself must be trustworthy, explainable, accountable and fair across India's specific diversity dimensions.

The default cloud-AI architectures offered by hyperscalers Amazon SageMaker, Amazon Bedrock, Microsoft Azure OpenAI and Google Cloud Vertex AI were not designed against this brief. Their compliance posture is broad and impressive, but their core data flow assumes that a customer payload, potentially containing PII, is sent to the hyperscaler's infrastructure for processing. Region selection alone does not address jurisdictional reach. Under the United States CLOUD Act, US-headquartered providers are subject to subpoena for customer data held anywhere in the world. For a Public Sector Bank this is not a hypothetical risk but a structural one.

This whitepaper presents an alternative. A PII data boundary architecture in which agentic AI is delivered as a hybrid system. Production applications, customer data, and the PII boundary itself remain entirely on-premises on the bank's own infrastructure, typified in this document by an AMD EPYC-class CPU footprint. Cloud GPU inference is used selectively, but only on de-identified token payloads. Original PII never leaves the bank's premises.

Executive Summary

The architecture uses open-source components Apache-licensed PaddleOCR for document parsing, Microsoft Presidio with custom Indian recognisers, IndicBERT-class models fine-tuned in the bank's environment, Open Policy Agent for re-identification policies, HashiCorp Vault Transform with HSM-backed keys for format-preserving tokenisation. The bank retains full sovereignty over its AI stack and a clean exit path from any single vendor.

This document specifies the architecture, maps it against DPDP, RBI, UIDAI and CERT-In requirements, and provides four deep-dive technical sections: the egress DLP rule set, an IndicBERT fine-tuning plan for context PII, Open Policy Agent re-identification policies, and a Finacle-compatible tokenisation scheme. It also specifies the federated data governance pattern that spans on-premises and cloud, and the AI governance regime that controls cloud-hosted models — both directly aligned with the FREE-AI Pillars Governance, Protection and Assurance. It concludes with a comparative analysis against hyperscaler-native approaches, an alignment statement against India's sovereign AI mission, and a phased implementation roadmap suitable for a PSB pursuit.

In a single sentence: Production data never leaves the bank's premises, the cloud sees only tokens, and the bank's auditors can prove this with the egress log alone.

1. Strategic context

1.1 The PSB AI paradox

Public Sector Banks in India sit at an intersection of conflicting pressures. On one side, customer expectations and competitive pressure from private banks, NBFCs and fintech entrants demand investment in agentic AI for KYC automation, document understanding, fraud detection, credit underwriting, customer service and compliance. On the other side, the bank is a Significant Data Fiduciary under DPDP, an entity subject to RBI's full-stack supervisory regime, and a custodian of the digital identity of a substantial share of the Indian population.

Solving for one side at the cost of the other is not viable. A bank that under-invests in AI will lose ground on cost-to-serve and customer experience within three to five years. A bank that adopts AI through a path that compromises data residency, consent, audit or sovereignty exposes itself to penalties of up to ₹250 crore per violation under DPDP, supervisory action by the RBI, and reputational damage that is materially harder to recover from than commercial under-performance.

1.2 Why hyperscaler-native is structurally insufficient

The hyperscaler-native AI stack is, in absolute terms, technically excellent. The compliance certifications are extensive. The engineering quality is high. The model catalogue is broad. None of this is in dispute.

What is in dispute, for an Indian PSB, is the architectural premise. A hyperscaler AI service is built around the assumption that the customer payload a document, a transcript, a record is sent to the hyperscaler's infrastructure for processing. Region selection (Mumbai, Hyderabad) places the bytes in India, but does not change the fact that the legal entity processing those bytes is a US-domiciled corporation and is, in respect of customer data anywhere in the world, reachable by US legal process under the CLOUD Act of 2018.

1. Strategic context

"Guardrails" features Bedrock Guardrails, Azure AI Content Safety with PII detection, Vertex AI Sensitive Data Protection operate inside the hyperscaler's infrastructure. By the time the guardrail evaluates the payload, the payload has already entered the hyperscaler's environment. This is acceptable for many enterprise use cases. For a PSB processing crore-scale Aadhaar, PAN, account number and customer name records, it is the wrong shape.

1.3 The sovereign alternative

The architecture proposed here inverts the data flow. The PII boundary is enforced on-premises, on the bank's own infrastructure. Detection, classification, tokenisation, policy enforcement and audit all happen before any payload reaches a cloud provider. Cloud large language models are used as reasoning engines on tokenised payloads only. Real PII never crosses the boundary.

This approach is enabled by three converging trends. First, modern open-weight models Qwen 3, Llama 3, Mistral, Sarvam, AI4Bharat IndicBERT are at, or close to, the capability frontier. Second, end-to-end document understanding has matured to the point where small (sub-1B) open-source models can run on commodity CPU hardware with near-state-of-the-art accuracy, exemplified by PaddleOCR PP-StructureV3 and PaddleOCR-VL. Third, the policy-as-code, secrets-management and tokenisation tooling required to enforce a PII boundary at production scale is open-source and battle-tested.

These three trends mean that as of 2026, an Indian PSB does not need to trade sovereignty for capability. It can have both.

2. Regulatory landscape

2.1 Digital Personal Data Protection Act 2023

The DPDP Act, notified on 11 August 2023, is the primary statute governing personal data in India. The provisions most material to an AI architecture are summarised below.

Provision	Implication for AI architecture
Section 4 - grounds for processing	Centralised, versioned definitions served via standard APIs
Section 5 - notice	Customers must be informed of the personal data being processed and the purpose. The audit log must be retrievable per data principal.
Section 6 - consent	Free, specific, informed, unconditional and unambiguous. Consent must be granular, withdrawable, and the system must reflect changes in real time.
Section 8 - data fiduciary obligations	The bank, as Fiduciary, is responsible for the acts of processors. Hyperscaler-as-processor relationships must be contractually and technically constrained.
Section 10 - Significant Data Fiduciary	PSBs are designated SDFs given volume and sensitivity. SDF obligations include a Data Protection Officer, independent data audit, and Data Protection Impact Assessment.
Section 11—14 - data principal rights	Right to access, correction, erasure and grievance redressal. Erasure must propagate to derived data including embeddings and model fine-tunes where feasible.
Section 16 - transfer outside India	The Central Government may notify countries to which transfer is restricted. Architecture must support a default-deny posture on cross-border transfer.
Section 33 — penalties	Up to ₹250 crore per breach of obligations to take reasonable security safeguards. This makes egress prevention a board-level concern.

2. Regulatory landscape

2.2 RBI guidelines

The Reserve Bank of India's regulatory perimeter for banks intersects the DPDP regime with sector-specific obligations.

Direction or circular	Implication for AI architecture
Master Direction on KYC, 2016 (as amended)	Aadhaar processing must follow UIDAI rules including masking. KYC documents are sensitive personal information and require additional safeguards.
Master Direction on Digital Payment Security Controls, February 2021	Confidentiality and integrity of customer data across the digital product lifecycle. Use of production data in non-production environments is a direct compliance risk.
Cyber Security Framework for Banks, 2016 + amendments	Defence in depth, audit trails, immutable logs, segregation of duties. The PII boundary fits cleanly within the framework.
Outsourcing Master Direction	Critical or material outsourcing requires due diligence, exit planning, and direct supervisory access. Hyperscaler-native AI is in scope.
FREE-AI Framework, August 2025	RBI's authoritative framework for AI in the financial sector. Seven Sutras and six pillars cover trust, accountability, fairness, explainability, resilience, infrastructure, governance and assurance. PSBs are expected to align governance, vendor management, consumer protection and assurance to the framework. Detailed mapping is provided in Section 6 of this paper.

2. Regulatory landscape

2.3 UIDAI

Aadhaar carries its own statutory regime. Two instruments matter most for the architecture.

Instrument	Implication
Aadhaar (Authentication) Regulations, 2016	Use, storage and transmission of Aadhaar data require explicit consent and limited retention. Storage of Aadhaar in clear text is prohibited.
UIDAI circular F.No.13012/171/2018, October 2018	Mask the first eight digits of the Aadhaar number in any document or system display, retaining only the last four. The architecture must enforce this on every render path, not only on display surfaces.
Virtual ID and Offline Aadhaar	Architectures should accept VID and offline Aadhaar XML where the use case permits, reducing direct exposure to the 12-digit number.

2.4 CERT-In, IT Act and sectoral standards

Three further regimes complete the picture.

- CERT-In directions of 28 April 2022 require reporting of cyber incidents within six hours, 180-day log retention and ICT system clock synchronisation. The audit log architecture in this paper is designed to satisfy all three.

- Information Technology Act 2000 and the SPDI Rules 2011 define sensitive personal data and the reasonable security practices test. ISO 27001:2022 alignment is the standard route to demonstrate adequacy.

- Sectoral standards relevant to the larger control set include PCI DSS for card data, RBI's Information Security Audit guidelines, and SOC 2 Type II for outsourced control evidence. The Tantor PII boundary is designed to be auditable against all of these.

3. Architecture overview

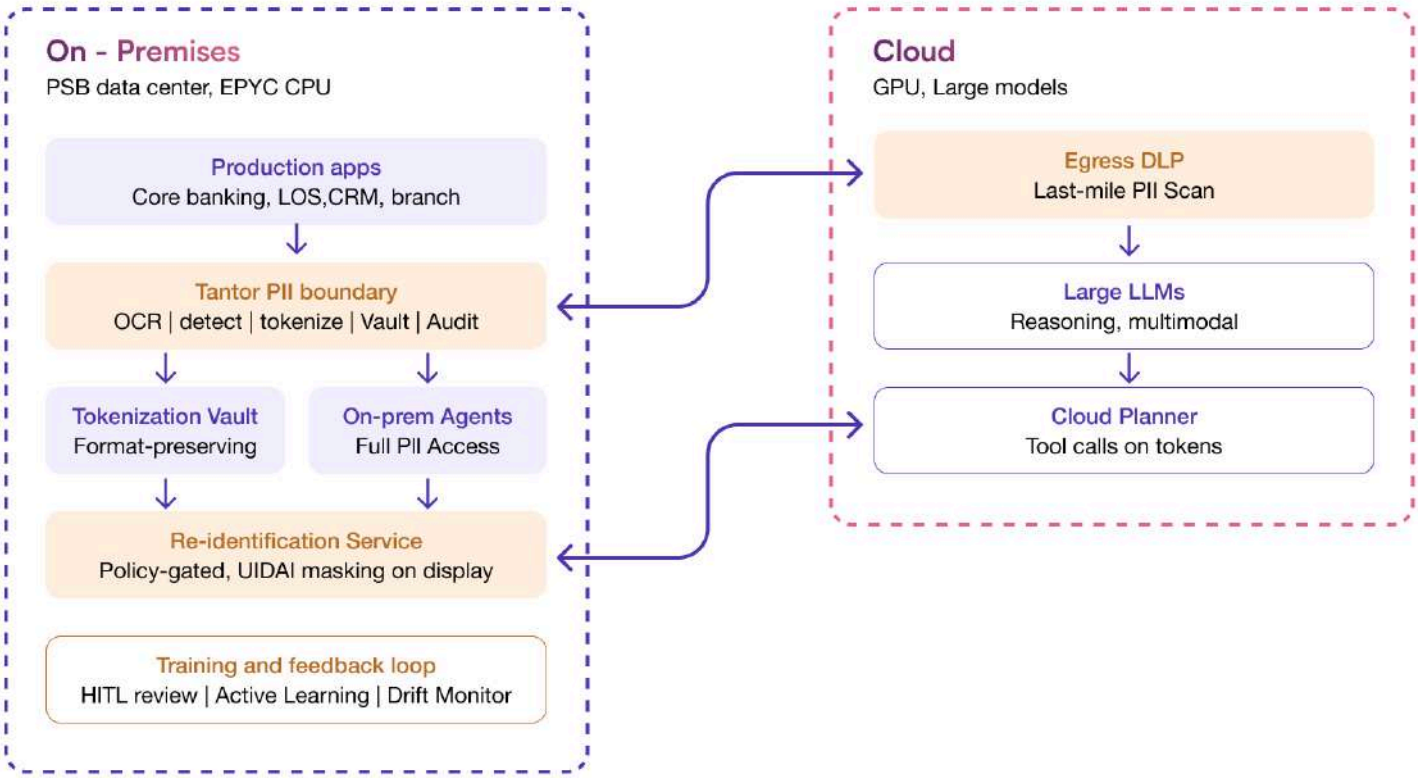
3.1 Design principles

- Sovereignty first. All production PII remains on-premises in Indian jurisdiction at all times.
- Boundary-as-control. The PII boundary is the unit of compliance. Every other control assumes the boundary holds.
- Default deny. Re-identification, egress and cross-border transfer are denied unless an explicit policy permits.
- Open by construction. Open-weights models, open-source components, open standards. No exit cliff.
- Auditable by design. Every detect, tokenise, re-identify, egress and feedback event is recorded with a tamper-evident audit trail.
- Indian context first. Recognisers, language models and policies are tuned for Indian PII formats, scripts and regulations.

3.2 Component map

The architecture comprises three zones. The on-premises zone houses all production applications, the Tantor PII boundary, the on-premises agent runtime, the tokenisation vault, the re-identification service, and the training and feedback loop. The egress DLP gateway sits at the boundary as the final line of inspection. The cloud zone holds the GPU inference fabric and large reasoning models, but is reachable only via the egress gateway and only with token-only payloads.

3. Architecture overview



3. Architecture overview

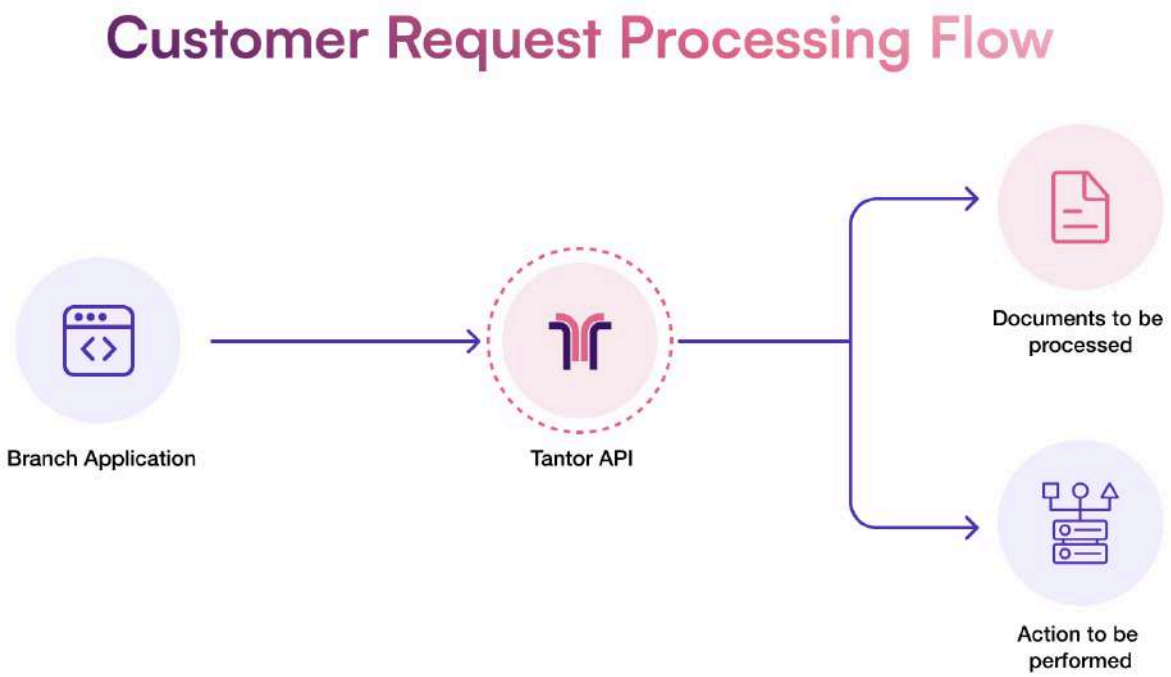
Zone	Component	Component
On-premises	Production applications	Core banking (Finacle), LOS, LMS, CRM, branch and channel apps. Existing footprint, unchanged.
On-premises	Tantor PII boundary	Document parsing, PII detection and classification, tokenisation, policy enforcement, audit.
On-premises	Tokenisation vault	HashiCorp Vault Transform with FF1 / FF3-1 format-preserving encryption, HSM-backed master key.
On-premises	On-premises agent runtime	Tantor agents that hold full PII access for execution against core banking and back-office systems.
On-premises	Re-identification service	Policy-gated detokenisation. Returns real values only to authorised on-premises consumers.
On-premises	Training and feedback loop	HITL review queue, annotation, fine-tuning, shadow evaluation and drift monitoring.
Egress	Egress DLP gateway	Final inspection. Tokens-only enforcement, signature verification, destination allowlisting, tamper-evident logging.
Cloud	Large LLMs / GPU inference	Reasoning, planning, multimodal interpretation. Operates exclusively on tokenised payloads.
Cloud	Cloud agent planner	Decomposes user intent into a plan over tokenised entities; never holds real PII.

3. Architecture overview

3.3 Request lifecycle

Figure 1 — Hybrid data-plane architecture. Production data and on-premises agents hold real PII; only de-identified tokens cross to the cloud, and the egress DLP is the final gate.

A representative agentic flow is shown end-to-end in Figure 1 above. The eight numbered steps in the request-lifecycle band trace a typical request from branch submission through tokenisation, policy routing, cloud reasoning, re-identification, render and audit logging.



3. Architecture overview

- The PII boundary parses each document with PaddleOCR PP-StructureV3, falling back to PaddleOCR-VL for visually difficult pages.

- Indian PII recognisers identify Aadhaar, PAN, IFSC, MICR, account number, mobile, date of birth, name and address entities. Each entity is sent to the tokenisation vault and a format-preserving token is returned.

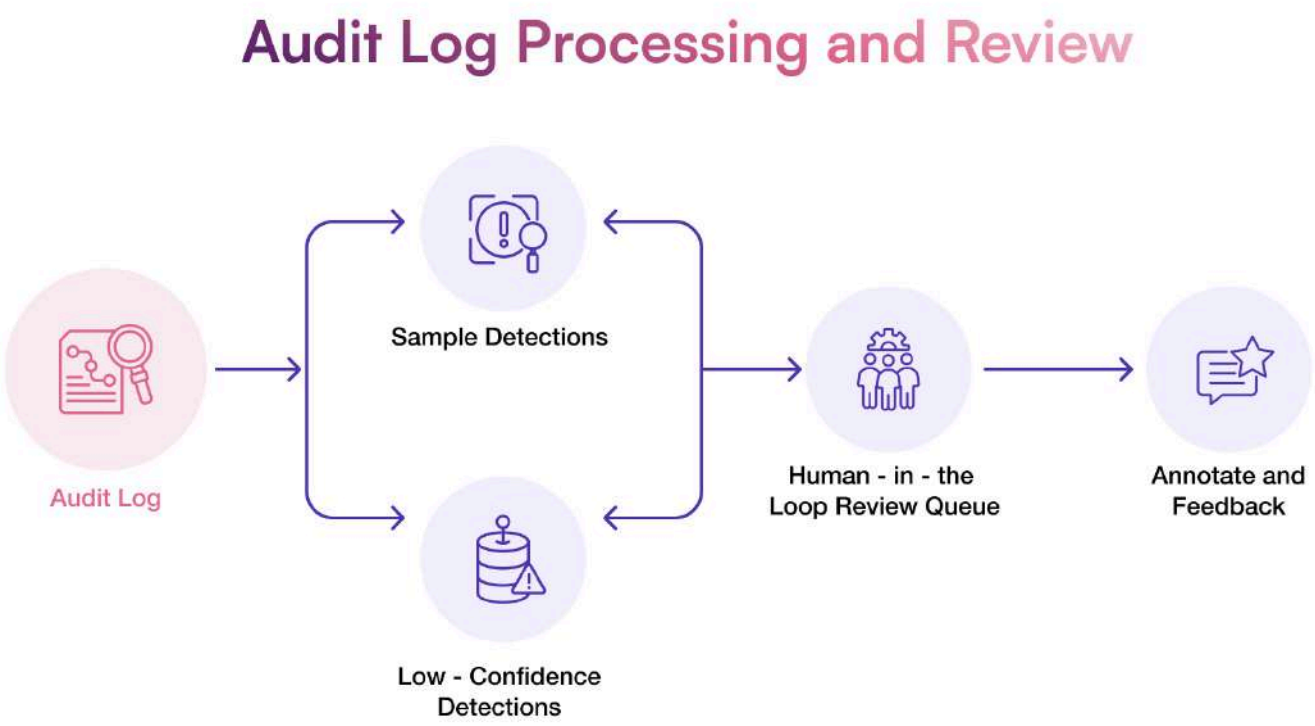
- The policy engine evaluates the routing decision. Tasks requiring core-banking interaction route to an on-premises agent. Tasks requiring frontier-model reasoning route through the egress DLP to a cloud LLM, with a token-only payload.

- The egress DLP performs a final inspection. The payload is verified to contain no Indian PII patterns, the destination is verified against an allowlist, the payload is signed with the bank's HSM-bound key, and the egress event is logged.

- The cloud LLM returns a plan or analysis referring to tokens. The on-premises executor consumes the plan, calls the re-identification service for tokens it must materialise, and acts against core banking under role-based authorisation.

- Output rendered to the customer or to the bank user is re-identified subject to UIDAI masking rules and the policy engine. Aadhaar always renders with the first eight digits masked.

3. Architecture overview



4. Deep-dive technical sections

4.1 The egress DLP rule set

The egress DLP is the second line of defence. The PII boundary is the first — it is responsible for tokenising every entity. The DLP exists to catch the cases where the boundary missed something or where a future bug produces an unintended leak. It is configured for paranoia, not cleverness.

Figure 3 — Egress DLP defence-in-depth funnel. Every payload must pass all ten rules; R6 throttles or blocks, R10 always logs, every other rule blocks on failure.

Rule taxonomy

Rule class	What it does	Failure action
R1 - Token allowlist	Verifies that every PII-shaped string in the payload is a known token from the vault, not a raw value.	Block, alert, queue for review.
R2 - Indian PII regex tripwires	Independent regex layer. Flags any Aadhaar (12-digit + Verhoeff), PAN, IFSC, MICR, GSTIN, mobile, voter ID format that survives tokenisation.	Block, alert, queue for review.
R3 - Statistical fingerprint	Per-payload entropy, n-gram and dictionary-overlap checks against a corpus of known PII strings.	Block on score above threshold.
R4 - Embedded artefact strip	Image, table, signature regions inspected. Any embedded image must have PII regions redacted at pixel level.	Block if image-OCR detects PII.
R5 - Format allowlist	Only JSON, plain text and post-redaction signed PDF are permitted egress payload types.	Block.

4. Deep-dive technical sections

Rule class	What it does	Failure action
R6 - Size and cardinality	Per-request payload size cap; per-session, per-user, per-role rate limits.	Throttle or block.
R7 - Destination allowlist	Egress permitted only to a defined set of cloud endpoints, identified by hostname, IP range and TLS certificate fingerprint.	Block.
R8 - Cryptographic envelope	mTLS to destination. Payload signed with the bank's HSM-bound key, including timestamp and nonce.	Block on signature failure.
R9 - Cross-border lookup	Destination resolved against the DPDP Section 16 watchlist. Default deny if destination jurisdiction is restricted.	Block
R10 - Tamper-evident log	Every egress event written to an append-only signed log. Daily Merkle root replicated to immutable storage.	Every egress event written to an append-only signed log. Daily Merkle root replicated to immutable storage.

4. Deep-dive technical sections

Implementation notes

1. Rules R1, R2, R5, R6, R7, R9 are implemented as Open Policy Agent policies, version-controlled, and reviewed by the bank's security team on each release.

2. Rule R3 is implemented as a small classifier trained on the bank's own historical egress payloads to learn the distributional shape of legitimate traffic. Outliers are flagged.

3. Rule R4 uses the same PaddleOCR runtime as the boundary, configured for pixel-region redaction rather than text extraction.

4. Rule R8 keys are generated, stored and used inside an FIPS 140-2 Level 3 HSM. The bank, not Tantor, holds the master.

5. All blocked egress events are routed to the human-in-the-loop queue. Every blocked event becomes a training data point — it represents either a real PII leak attempt that must be analysed, or a tuning gap in detection that must be closed.

4.2 IndicBERT fine-tuning for context PII

Regex and checksum recognisers handle structured PII Aadhaar, PAN, IFSC, GSTIN, MICR with very high precision and recall. They cannot, however, recognise context PII: a customer name in free-form text, a residential address in a loan application narrative, a relationship reference ("my father Shri Ramesh Kumar"), a multilingual name in Devanagari or Tamil. For this, a language model is required.

4. Deep-dive technical sections

Base model selection

Candidate	Strengths	Trade-offs
IndicBERT v2 (AI4Bharat)	Coverage of 11 major Indian languages; trained on Indian web text including BFSI domain content; permissive licence.	Smaller than multilingual XLM-R; needs domain fine-tuning for highest precision on form-style text.
MuRIL (Google)	17 Indian languages plus transliteration handling; strong for code-mixed text.	Requires careful adapter selection for token classification; slightly heavier on CPU.
XLM-RoBERTa base / large	Best generic multilingual baseline; widely understood.	Less Indian-specific; larger memory footprint.
DistilBERT-multilingual	Smallest, fastest CPU inference.	Lower headroom on accuracy for harder context entities.

Recommended baseline: IndicBERT v2 for the primary detector, with DistilBERT-multilingual as a low-latency fallback. MuRIL is preferred where the bank's customer base has heavy code-mixed text.

4. Deep-dive technical sections

Entity ontology

The fine-tuning task is token classification with BIO tagging. The proposed ontology covers approximately 25 entity types specific to Indian banking.

Class	Examples
Identifiers	AADHAAR, VID, PAN, GSTIN, IFSC, MICR, ACCOUNT_NUMBER, CIF_ID, CUSTOMER_NUMBER
Government IDs	VOTER_ID, DRIVING_LICENCE, PASSPORT, RATION_CARD
Person	PERSON_NAME (first, middle, last, salutation), RELATIONSHIP_REFERENCE
Contact	MOBILE, LANDLINE, EMAIL, ADDRESS_LINE, CITY, DISTRICT, STATE, PINCODE
Financial	CARD_NUMBER, IBAN, UPI_HANDLE, MERCHANT_ID, AMOUNT_WITH_CONTEXT
Temporal	DATE_OF_BIRTH, DATE_OF_INCORPORATION
Sensitive	RELIGION, CASTE, COMMUNITY (DPDP-sensitive categories), MEDICAL_CONDITION
Document	CHEQUE_NUMBER, LOAN_ACCOUNT, FIXED_DEPOSIT_NUMBER, POLICY_NUMBER

4. Deep-dive technical sections

Data curation

- Source corpus: 25,000 to 50,000 documents drawn from KYC submissions, loan applications, customer correspondence, court orders, internal memos and call-centre transcripts. Selected for representativeness across customer segments and Indian states.

- Annotation: trained annotators apply the BIO ontology, with double-annotation on a 10% sample to measure inter-annotator agreement (target: Cohen's kappa above 0.85).

- Synthetic augmentation: rare entities and underrepresented scripts (Gurmukhi, Odia, Manipuri) are augmented with synthetic data. Synthetic corpora are clearly tagged and never form more than 30% of any training subset.

- Hold-out: 10% of the labelled set is reserved for evaluation, partitioned to ensure that no customer or document appears in both training and evaluation.

- Privacy: the entire training and evaluation corpus is itself tokenised before annotators see it. Annotators see token forms and apply tags; the final detector trains on token forms. The model never trains on raw PII.

4. Deep-dive technical sections

Training procedure

- Approach: Low-Rank Adaptation (LoRA) on the base model to keep the compute footprint small and to allow per-tenant adaptation in future.
- Hyperparameters: 3 to 5 epochs, learning rate 2e-5 with linear warmup, batch size 16 to 32 depending on hardware, weight decay 0.01.
- Hardware: a single L40S GPU is sufficient for fine-tuning. Inference is then quantised to INT8 and exported to ONNX for AVX-512-accelerated CPU runtime on the EPYC 9J15 footprint.
- Targets: per-entity F1 above 0.95 for high-stakes entities (Aadhaar, PAN, account number); above 0.85 for context entities (names, addresses, relationships); inference latency below 50 milliseconds per page on the on-premises CPU.

Continuous improvement Detection is not a one-shot training problem. It is a permanent feedback loop. The deployed detector emits a confidence score per entity. Low-confidence detections are routed to the human-in-the-loop queue. A 5% random sample of high-confidence detections is also reviewed monthly to catch silent regression. Quarterly retraining is performed on the cumulative annotated corpus, with shadow evaluation against the production detector before any promotion.

4. Deep-dive technical sections

4.3 Open Policy Agent policies for re-identification

Re-identification turning a token back into the value it represents is the most security-sensitive operation in the system. It must be policy-gated, auditable, and decoupled from the application code that requests it. Open Policy Agent with the Rego policy language is the recommended substrate. Policies are versioned in Git, reviewed by the bank's security team, and bound to an immutable hash at deployment time.

Policy domains

Domain	Purpose
tantor.reid	Who can re-identify which entity types under which purposes.
tantor.egress	What payload shapes may leave on-premises and to which destinations.
tantor.retention	How long tokens, vault values, and audit records are retained.
tantor.consent	Mapping of declared purposes to consent state per data principal.
tantor.crossborder	DPDP Section 16 watchlist enforcement on any external destination.

4. Deep-dive technical sections

Sample policy: re-identification

```
package tantor.reid

default allow = false
default mask_only = true

# Allow full re-id only when subject role, declared purpose,
# customer consent and time-of-day all line up.
allow {
    input.subject.role in {"kyc_officer", "branch_manager",
"compliance_officer"}
    input.action == "reidentify"
    input.context.purpose in
compatible_purposes[input.resource.entity_type]
    valid_consent
    within_business_hours
    not aadhaar_full_disclosure_blocked
}

# UIDAI mandate: Aadhaar may only ever be returned masked to the
# display layer unless the call is part of a UIDAI-authenticated
flow.
aadhaar_full_disclosure_blocked {
    input.resource.entity_type == "aadhaar"
    not input.context.purpose ==
"uidai_authenticated_authentication"
    input.context.surface == "display"
}

compatible_purposes = {
    "aadhaar":  {"kyc_verification", "regulatory_filing"},
    "pan":      {"kyc_verification", "tax_reporting",
"loan_underwriting"},
    "account_number": {"transaction_servicing",
"customer_servicing"},
    "mobile":   {"customer_servicing", "otp_authentication"},
    "name":     {"customer_servicing", "kyc_verification",
"correspondence"}
}

valid_consent {
    consent := data.consents[input.resource.customer_id]
    consent.purposes[input.context.purpose]
    consent.expiry_ns > time.now_ns()
    not consent.withdrawn
}

within_business_hours {
    [hour, _, _] := time.clock(time.now_ns())
    hour >= 8
    hour <= 20
```

4. Deep-dive technical sections

Sample policy: cross-border egress

```
package tantor.crossborder

default allow = false

# Default-deny posture for cross-border egress.
# Only destinations explicitly approved by the bank's DPO are
allowed.
allow {
    input.destination.country == "IN"
}

allow {
    input.destination.country in data.approved_jurisdictions
    input.payload.shape == "tokens_only"
    input.payload.signed
    not input.destination.country in data.dpdp_section_16_blocked
}
```

Operational practices

- Policies are version-controlled in Git. Every change is peer-reviewed by at least one security engineer and one DPO-nominated reviewer.

- Each policy bundle is signed and the signature is verified at OPA startup. A failed signature stops the service rather than starting with an unknown policy.

- Decision logs are written to an append-only store, signed and replicated. Every allow decision is auditable; every deny decision is auditable; every exception is auditable.

- Quarterly policy reviews are conducted with the bank's DPO and information security team. The review covers: new entity types, new purposes, withdrawn or modified consents, jurisdictional changes, and any deny patterns indicating misuse or process drift.

4. Deep-dive technical sections

4.4 Finacle-grade tokenisation scheme

Most Indian PSBs run Infosys Finacle as the core banking platform. Tokenisation must therefore be transparent to Finacle. A token must be substitutable into a Finacle field of the same type, must support deterministic joins for reporting and analytics, and must support reversal under policy control. Pure hashing fails the format constraint. Encryption fails the format constraint and the determinism property unless format-preserving encryption is used.

Algorithm choice

Format-preserving encryption per NIST SP 800-38G specifically FF1 (FFX) and FF3-1 is the recommended primitive. FF1 is more flexible across alphabets; FF3-1 is faster on smaller domains. Both produce ciphertext in the same alphabet and length as the input.

4. Deep-dive technical sections

Entity	Input	Tokenised form	Notes
Aadhaar	12 digits	12 digits via FF3-1, last 4 preserved	Compatible with UIDAI masking. Token is non-clashing with valid Aadhaar checksums by design.
VID	16 digits	16 digits via FF3-1	Treated equivalent to Aadhaar.
PAN	5L4N1L	Same format via FF1 over alphanumeric alphabet	Position semantics not preserved (4th-char entity type is randomised) accepted trade-off.
IFSC	11 chars	Format-preserving alphanumeric	Token is registry-distinguishable from real IFSC by reserved prefix.
Account number	Variable, bank prefix	FPE preserving prefix and length	Last 4 preserved for customer-facing UIs that already mask middle.
Mobile	10 digits	FF3-1 over digits	Country code and operator prefix preserved if required.
Date of birth	DD/MM/YYYY	Tokenised at component level; year preserved for actuarial analytics	Day and month FPE'd within calendar bounds.

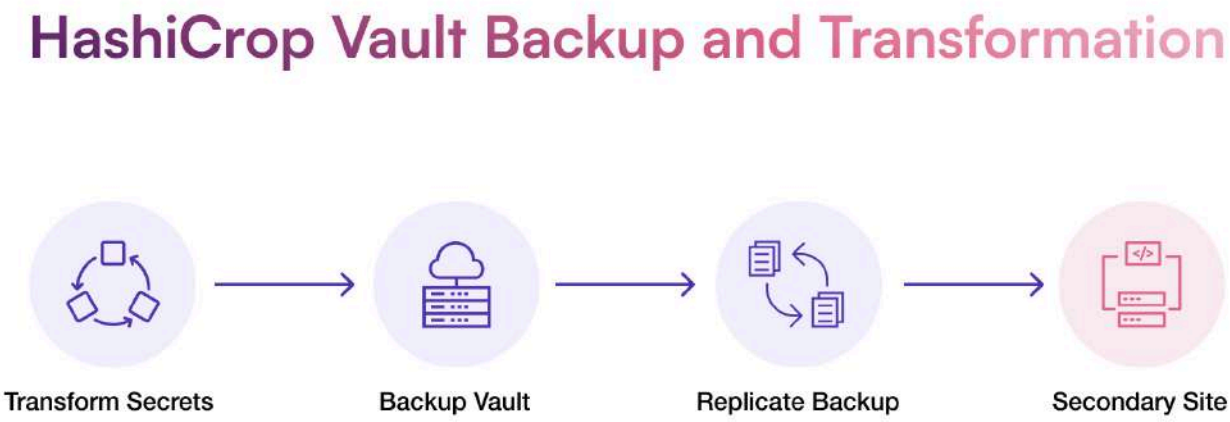
4. Deep-dive technical sections

Entity	Input	Tokenised form	Notes
Name	String, multi-script	Vault-mediated string token	Not FPE string semantics differ. Deterministic per-customer keyed lookup.
Address	Multi-line string	Field-level tokens, PIN preserved	Geographic analytics retained at PIN level.

Vault architecture

- HashiCorp Vault Transform secrets engine with the FPE Transformations feature.
- Master key generated, stored and used inside the bank's HSM (Thales Luna, Utimaco SecurityServer or nCipher nShield are all suitable). The HSM enforces key non-extraction; Tantor never sees the master key in clear.
- Per-tenant key derivation. Each bank deployment derives its own key from the master, ensuring cryptographic isolation.
- Per-customer salting. Token determinism is scoped to a customer, so the same Aadhaar belonging to two different customers (in error) produces two different tokens — preventing accidental cross-customer linkage.
- Key rotation. Master key rotates annually under HSM ceremony. Derivation keys rotate quarterly. Rotation is staged: old tokens remain resolvable via a rotation register.
- Backup. Encrypted backup of the vault is replicated to a secondary site within Indian jurisdiction, never outside.

4. Deep-dive technical sections



Finacle integration patterns

- Pre-processor at the integration layer. Inbound from channels branch, internet banking, mobile, ATM switch entities are tokenised before write. The Finacle ledger sees only tokens for the tokenised entity types.

- Detokenisation at posting boundary. Where Finacle requires the real value (for example, a regulatory filing), the policy-gated re-identification service detokenises in the posting service, not in the application layer.

- Reporting and analytics. The data warehouse stores tokens. Analytical models are trained on tokens. Reports are rendered on tokens with on-demand, audited detokenisation only when the report's purpose requires it.

- MIS and dashboards. Dashboards render tokens or masked forms by default. Drill-down to real values requires policy-gated re-identification with the user's purpose declared.

4. Deep-dive technical sections

Performance

- Tokenise / detokenise latency target: under 2 milliseconds per call at the 95th percentile, with the HSM in the loop.

- Throughput target: 5,000 to 10,000 calls per second per Vault Transform instance, scaling horizontally behind a load balancer.

- Capacity planning: a typical Tier-1 PSB with 100 to 200 million customer records and 50 to 100 million daily transactions can be supported by three Vault Transform nodes plus HSM in active-active configuration.

5. Comparison with hyperscaler-native approaches

This section addresses the natural alternative the bank will consider: build the same agentic workflows on AWS SageMaker, AWS Bedrock, Microsoft Azure OpenAI Service, or Google Cloud Vertex AI. Each of these is a capable platform. None is structurally suitable for a PSB without architectural modifications that, by the time they are complete, recreate most of the Tantor approach.

5.1 AWS SageMaker

SageMaker is a managed ML platform offering training, fine-tuning, hosting and a wide library of pre-built containers and JumpStart models. For a US enterprise without sovereignty constraints it is excellent.

Dimension	Position
Data residency	Mumbai region available; data at rest in India. However, AWS as a US legal entity is reachable under the CLOUD Act.
PII boundary	No on-premises boundary by design. PII is sent to SageMaker for inference. Macie can identify PII after the fact, not prevent ingress.
Vendor lock-in	SageMaker SDK, JumpStart, Pipelines and the AWS feature store create a deep architectural commitment. Exit cost grows with adoption.
Cost	Inference cost scales linearly. Agentic workflows multiply token usage by 3 to 10x. Managed inference rates are 4 to 8x raw GPU rental at scale.
Open weights	Supported but not architecturally privileged. Bedrock and JumpStart steer towards proprietary or AWS-managed models.
Audit control	CloudTrail logs sit in AWS infrastructure. Bank has read access but not custody.

5. Comparison with hyperscaler-native approaches

5.2 AWS Bedrock

Bedrock is the foundation-model API service. It hosts Anthropic, Mistral, Meta, Cohere, Stability and Amazon's own Titan and Nova models behind a single API.

Dimension	Position
Data residency	Bedrock model availability varies by region. Anthropic Claude models in particular are concentrated in US regions; some models are not yet GA in Mumbai. Cross-region routing is used.
PII boundary	Bedrock Guardrails offers PII detection and redaction policies, evaluated inside AWS. The payload reaches AWS infrastructure before evaluation.
Sovereignty	Same CLOUD Act exposure as SageMaker. AWS European Sovereign Cloud is in build-out; an AWS Indian Sovereign Cloud is not announced as of the date of this paper.
Cost	Per-token pricing is competitive at low volumes. At PSB scale, with agentic workflows and crore-scale customer volumes, costs grow rapidly and unpredictably.
Vendor lock-in	Strongest in this category. The Bedrock API surface, the prompt-caching mechanism and the Knowledge Bases feature are all AWS-specific.
Audit control	CloudTrail and Bedrock invocation logs in AWS. Bank does not hold the artefacts.

5. Comparison with hyperscaler-native approaches

5.3 Microsoft Azure OpenAI Service

Azure OpenAI offers OpenAI's models GPT-4-class, GPT-5-class on availability within the Azure compliance perimeter. Azure has the most mature compliance and identity story among the hyperscalers, particularly relevant to enterprise IT shops already standardised on Microsoft.

Dimension	Position
Data residency	Azure India regions exist; data at rest in India. Microsoft has signalled stronger sovereign-cloud direction (Microsoft Cloud for Sovereignty) but the legal jurisdiction over Microsoft as a US entity is unchanged.
PII boundary	Azure AI Content Safety with PII detection is available, evaluated inside Azure. Same architectural property as Bedrock Guardrails.
Sovereignty	CLOUD Act exposure. Microsoft Cloud for Sovereignty addresses some controls but does not change the underlying corporate domicile.
Identity integration	Strongest in this category. Entra ID, Conditional Access, Purview integrate cleanly with Azure OpenAI.
Cost	Per-token pricing similar to Bedrock. Pre-purchased throughput units offer some predictability.
Vendor lock-in	Heavy if standardised on the broader Azure stack.

5. Comparison with hyperscaler-native approaches

5.4 Google Cloud Vertex AI

Vertex AI offers Gemini and a curated catalogue of partner models, with strong tooling for embeddings, vector search and pipelines.

Dimension	Position
Data residency	Mumbai and Delhi regions; data residency controls available. Google as a US entity remains CLOUD-Act-reachable.
PII boundary	Sensitive Data Protection (formerly DLP) is a mature feature; evaluated inside GCP.
Sovereignty	T-Systems Sovereign Cloud partnership exists in Germany; no equivalent India-specific sovereign offering announced.
Cost	Competitive at low volumes; agentic flows compound.
Vendor lock-in	Vertex Pipelines, Feature Store and Matching Engine are GCP-specific.

5. Comparison with hyperscaler-native approaches

5.5 Comparative summary

Property	SageMaker	Bedrock	Azure OpenAI	Vertex	Tantor on-prem boundary
PII never leaves India	Region-bound only	Region-bound only	Region-bound only	Region-bound only	By architecture
CLOUD Act exposure	✓	✓	✓	✓	✗
Bank holds master keys	Optional via KMS / CloudHSM	Optional	Optional via Key Vault	Optional via Cloud KMS / Cloud HSM	✓ (HSM under bank custody)
Open-weights first	Permitted	Mixed	Limited	Permitted	✓
Bank-side audit log custody	Read access	Read access	Read access	Read access	Custody
DPDP Section 16 enforcement	Manual configuration	Manual configuration	Manual configuration	Manual configuration	Default-deny by policy
Exit cost	High	Highest	High	High	Low (open components)

5. Comparison with hyperscaler-native approaches

The structural argument in plain terms

All four hyperscaler approaches converge on the same architectural choice: send the customer payload, including PII, to the hyperscaler's infrastructure for processing. They differ in the maturity of in-cloud guardrails, in compliance certifications, and in regional availability. They do not differ on the fundamental data flow.

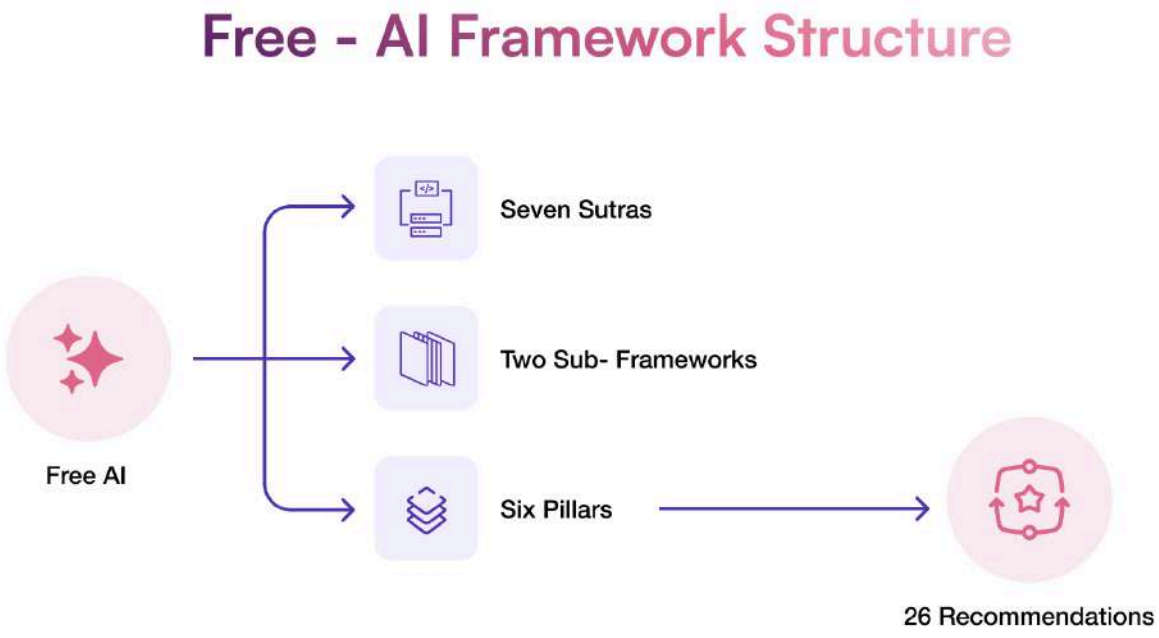
The Tantor approach inverts the data flow. The bank does not give PII to a foreign-domiciled service provider and rely on contractual or regional controls. The bank de-identifies on its own premises, and only sends de-identified tokens to any external service. The bank can use the same hyperscaler-hosted models Claude, GPT-class, Gemini, Llama but only as reasoning engines on tokenised payloads. The bank gets the capability without the structural exposure.

There is one further consideration that often goes unstated. Hyperscaler economics for agentic workflows are unpredictable. A planning agent that decomposes a single user intent into 10 to 50 LLM calls can drive costs to multiples of the manual baseline before any productivity gain materialises. The Tantor approach gives the bank the option to keep small reasoning steps on-premises (using open-weight 7B to 30B models on commodity hardware) and use the cloud only for the steps that require frontier-model capability. Unit economics improve by a factor of 3 to 10x at scale.

6. Alignment with the RBI FREE-AI Framework

6.1 The framework in brief

The Framework for Responsible and Ethical Enablement of Artificial Intelligence FREE-AI is the Reserve Bank of India's authoritative framework for AI use in the financial sector. The Committee was constituted on 6 December 2024, chaired by Dr Pushpak Bhattacharyya (IIT Bombay), and the report was released on 13 August 2025. The framework applies to all RBI-regulated entities including scheduled commercial banks, NBFCs, All-India Financial Institutions, payment system operators and other regulated entities, as well as the AI vendors and cloud providers they engage.



Any architecture proposed for use in a Public Sector Bank must be assessable against all three rings. This section provides that assessment for the Tantor architecture.

Figure 5 — RBI FREE-AI Framework — three concentric rings: 7 foundational Sutras at the core, 2 sub-frameworks (Innovation Enablement, Risk Mitigation), 6 strategic Pillars on the outer ring, and 26 specific recommendations make the pillars actionable.

6. Alignment with the RBI FREE-AI Framework

6.2 The seven Sutras and Tantor's design response

Each of the seven Sutras is a guiding principle for AI deployment in the financial sector. The table below maps each Sutra to the specific architectural element through which Tantor satisfies it.

Sutra	What it requires	Tantor's design response
1. Trust is the Foundation	AI systems must be reliable, transparent and inspire public confidence. Trust is non-negotiable.	The PII boundary is the trust unit. Production data never leaves the bank's premises; the egress log is sufficient evidence to a regulator or DPO that no leak has occurred. Tamper-evident audit trails preserve the bank's ability to demonstrate trustworthy operation.
2. People First	AI supports human decision-making but defers to human judgement. Welfare, dignity and inclusion are central.	Human-in-the-loop review queue is a core architectural component, not an afterthought. Every low-confidence detection, every blocked egress, and a 5% sample of high-confidence decisions are routed to human reviewers. High-stakes agent actions require human override paths.
3. Innovation over Restraint	Encourage responsible innovation. Avoid unnecessary restrictions while preserving safeguards.	The hybrid architecture lets banks adopt frontier-model capability (cloud LLMs) without surrendering data sovereignty. Innovation is enabled by separating reasoning (cloud, on tokens) from execution (on-premises, on real PII).

6. Alignment with the RBI FREE-AI Framework

Sutra	What it requires	Tantor's design response
4. Fairness and Equity	AI outcomes must be free from bias. Indian fairness dimensions caste, language, region, religion, economic status, urban/rural divide, digital literacy must be accounted for.	Detection models are fine-tuned on Indian-context corpora covering Devanagari, Tamil, Telugu, Bengali and other scripts. Fairness evaluation uses India-specific benchmarks rather than off-the-shelf Western bias suites. The HITL queue surfaces fairness drift.
5. Accountability	Institutions deploying AI bear responsibility for AI-driven decisions. Graded liability across the value chain developer, deployer, end-user.	Every action is auditable end-to-end with subject, purpose, policy decision and outcome. The bank holds the master encryption key (HSM under bank custody), the audit log custody and the policy bundle. Tantor's role as deployer is contractually constrained and technically auditable.
6. Understandable by Design	AI systems must be interpretable to the extent necessary for their risk level. Explanations proportionate to stakes.	Every detection emits a confidence score and the rule or model that triggered it. Every re-identification decision is traceable to a specific OPA policy clause. For agentic flows, the cloud planner emits an explicit plan over named tokens that the on-premises executor materialises the plan is the explanation.

6. Alignment with the RBI FREE-AI Framework

Sutra	What it requires	Tantor's design response
7. Safety, Resilience and Sustainability	AI systems must be secure, adaptable to shocks, sustainable in the long term. Identify anomalies, warn, contain.	Defence in depth: PII boundary plus egress DLP plus cryptographic envelope plus immutable audit. Drift monitoring on detection. Open-source components mean no single-vendor dependency. Energy footprint is materially smaller than full hyperscaler inference because reasoning runs only on tokenised payloads.

6. Alignment with the RBI FREE-AI Framework

6.3 The six Pillars and Tantor's contribution

FREE-AI structures its 26 recommendations across six strategic pillars. Three pillars sit under Innovation Enablement; three under Risk Mitigation. The table below maps each pillar to the Tantor capabilities that contribute to it. PSBs are expected to operate against all six; Tantor delivers materially against each.

Sub-framework	Pillar	Tantor contribution
Innovation Enablement	Infrastructure	Reference deployment runs on Indian-procurable commodity hardware. Compatible with the IndiaAI compute mission and the AI Kosh data infrastructure. The PII boundary is the bank's local equivalent of a sandbox production data can be safely used for AI experimentation because tokenisation is already enforced.
Innovation Enablement	Policy	Tantor ships with a board-ready AI policy template aligned with FREE-AI. Permissible-use cases, risk appetite, escalation paths and review cadences are configurable through Open Policy Agent. Policy versions are signed, audited and reviewable.
Innovation Enablement	Capacity	Tantor Studio includes annotation, review and decision-explanation interfaces designed for the bank's existing branch, KYC and compliance staff. The deployment plan includes structured AI literacy training for board, senior management and operational teams.

6. Alignment with the RBI FREE-AI Framework

Sub-framework	Pillar	Tantor contribution
Risk Mitigation	Governance	Lifecycle governance for every model, every recogniser and every policy bundle. Versioning, signed artefacts, review trails. Vendor oversight is enabled because every external model invocation is logged in the bank's audit store, not the vendor's.
Risk Mitigation	Protection	Consumer safeguards: UIDAI-mandated Aadhaar masking enforced at every render; explainability surfaces; grievance-redressal hooks in the API contract. AI-aware cybersecurity: the egress DLP is itself an AI-aware control surface, with a tamper-evident envelope and destination allowlisting.
Risk Mitigation	Assurance	Independent audit is a first-class operation. The architecture supports an external auditor querying the audit store and the policy bundle without operational disruption. AI registry, AI disclosures in annual reports, and AI incident reporting (per FREE-AI Annexure VI) are supported through pre-built export paths.

6. Alignment with the RBI FREE-AI Framework

6.4 Key recommendations mapping

Of the 26 specific recommendations in FREE-AI, the following are most directly addressed by an architectural component of the Tantor approach. The recommendations not listed below are organisational rather than architectural board approvals, training programmes, periodic disclosures but the architecture does not impede any of them.

Sub-framework	Pillar
Indigenous, domain-specific financial AI models	Architecture privileges open-weight and indigenous models (BharatGen, AI4Bharat IndicBERT, MuRIL, Sarvam). Bank can fine-tune within its own perimeter using its own data, retaining full IP.
AI innovation sandbox	Tokenisation enables a true sandbox. Real production data once tokenised can be used for experimentation without DPDP exposure, because no reversal is possible without the vault.
Financial sector data infrastructure / AI Kosh integration	Tokenised data exports are the natural integration path. The bank can contribute de-identified, tokenised features to AI Kosh without moving raw customer records.
Board-approved AI policy	Tantor's OPA policy bundles are the technical instantiation of the board-approved policy. Each clause maps to a board-approved rule; changes require dual-control review.

6. Alignment with the RBI FREE-AI Framework

Sub-framework	Pillar
Vendor and third-party AI risk	All vendor invocations are mediated by the egress DLP. Vendor switching is mechanical, not architectural. Outsourcing contracts can specify token-only payloads as a hard requirement, supported by the architecture.
Consumer protection and grievance redressal	Re-identification is policy-gated and audited. Customer-facing renders default to UIDAI-compliant masking. Any grievance about an AI-driven decision can be reproduced from the audit log with the exact tokens, plan and policy outcome.
AI registry and disclosures	Built-in registry of every model, every detector, every policy bundle in production. Pre-built quarterly disclosure export with the metrics typically required in PSB annual reports.
AI incident reporting	The egress DLP block events, re-identification deny events, and detection drift events feed directly into the FREE-AI Annexure VI incident form. CERT-In's six-hour reporting clock is supported by automated alert routing.
Independent audit and impact assessment	Architecture is auditable by an external party with read access to the audit store and the policy bundle, without disrupting production. Data Protection Impact Assessments are pre-templated for the standard agentic flows.

6. Alignment with the RBI FREE-AI Framework

Sub-framework	Pillar
Fairness across Indian dimensions	Detection and decisioning models include India-specific fairness benchmarks (caste, region, language, urban/rural, digital literacy). Drift monitoring covers fairness in addition to accuracy.
Cybersecurity for AI	PII boundary plus egress DLP plus HSM-bound signing constitute an AI-aware cyber control. Standard SIEM integration via OpenTelemetry.

6. Alignment with the RBI FREE-AI Framework

6.5 Sovereignty, indigenous models and openness

Two of the Sutras (Resilience, Accountability) and three of the Pillars (Infrastructure, Capacity, Protection) lean strongly on the bank's ability to control its own AI stack. This is the intersection of FREE-AI with the IndiaAI Mission, which received a Cabinet allocation of ₹10,372 crore in March 2024 and explicitly funds indigenous foundation models, sovereign compute and a national datasets platform.

The Tantor architecture is constructed on open and indigenous foundations:

Layer	Component	Licence / origin
Document parsing	PaddleOCR PP-StructureV3, PaddleOCR-VL	Apache 2.0
Detection	Microsoft Presidio, AI4Bharat IndicBERT, Google MuRIL, custom Indian recognisers	Apache 2.0 / MIT / open-weights
Indian-language reasoning (optional)	Sarvam, AI4Bharat, BharatGen models	Open-weights / Indian-origin
Tokenisation	HashiCorp Vault Transform; OpenBao fork available	BSL / MPL 2.0
Policy	Open Policy Agent, Cedar	Apache 2.0
Orchestration	Kubernetes, OpenTelemetry, Temporal	Apache 2.0 / MIT
Vector store	OpenSearch, Qdrant, pgvector	Apache 2.0 / open-source
Embeddings	BGE-M3, IndicEmbed, e5-large	Open-weights
Cloud reasoning (optional)	Open-weight or proprietary models, swappable via the Tantor connector	Bank's choice

6. Alignment with the RBI FREE-AI Framework

Every component is replaceable. The Tantor connector layer abstracts model and store selection. A bank that decides three years from now to switch from one cloud LLM provider to another, or from PaddleOCR to a domestic equivalent, can do so without re-architecting. Sovereignty includes the right to change one's mind.

Sovereignty includes the right to change one's mind.

6.6 The argument for open weights as a security control

Open weights, in the BFSI context, are not a philosophical preference. They are a security control. A model whose weights the bank cannot inspect cannot be properly threat-modelled. A vendor whose model the bank cannot run elsewhere holds leverage in pricing and roadmap that no responsible procurement would accept for any other infrastructure layer.

Indian banks already require source-code escrow and exit clauses for core banking; the same posture applied to AI gives open-weight models a structural advantage that closed-API services do not. The FREE-AI emphasis on indigenous models and the IndiaAI Mission's investment in BharatGen reinforce this direction at the policy level.

7. Data governance - federated across on-prem and cloud

7.1 Why federation is necessary

A PSB's data does not live in one place. Even before the AI question arises, the bank's data estate spans Finacle on Oracle, mainframe-hosted historical records, branch databases, channel-specific stores, the data warehouse, the analytical lake, application-specific stores in CRM, LOS, LMS, and increasingly cloud-resident workloads such as sandboxes, analytical environments and productivity tools. The AI stack adds a new federation requirement: cloud-hosted reasoning operates on tokenised projections of on-premises data, while real PII never moves. Governance must therefore span both zones with a single, authoritative source of truth for policy, classification, lineage and stewardship.

The federation principle is straightforward.

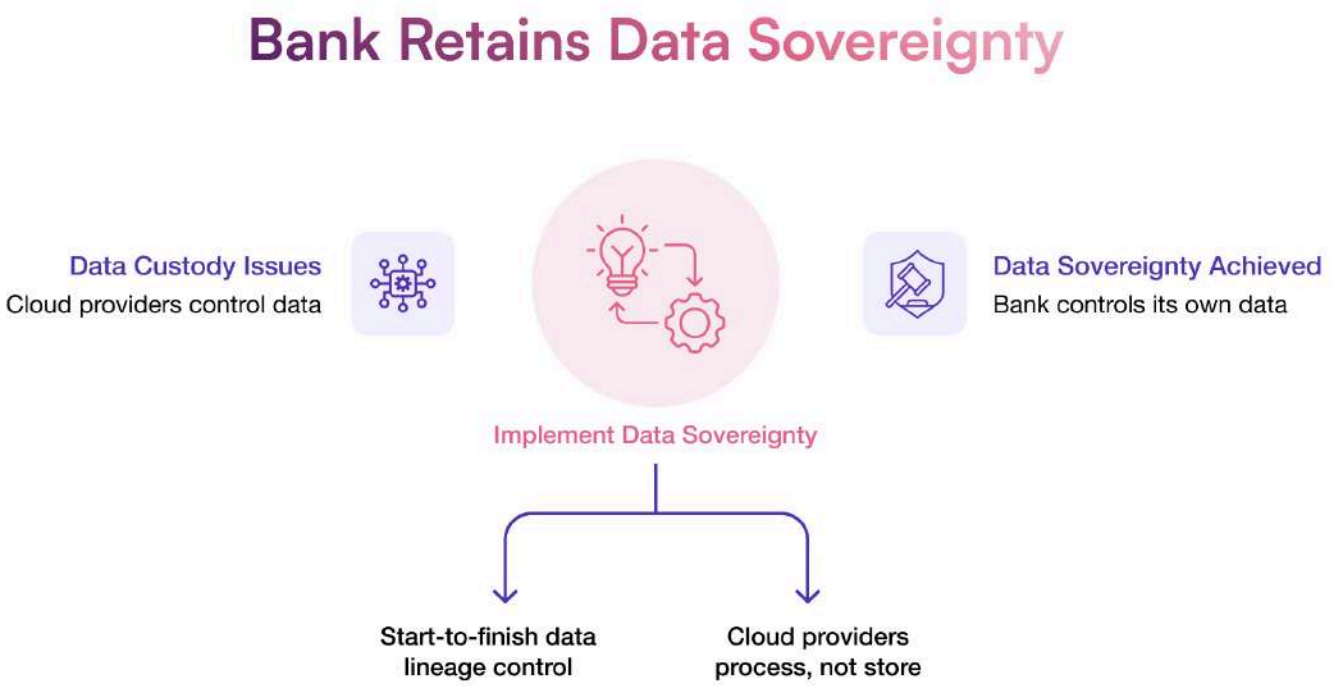
> *Policy is authored once, in one place, and enforced everywhere consistently*

Data, by contrast, lives in the zone its sensitivity dictates sensitive PII on-premises, tokenised projections both on-premises and in cloud, public reference data wherever convenient. The control plane is centralised; enforcement is distributed.

7. Data governance - federated across on-prem and cloud

7.2 Governance principles

The architecture rests on five principles aligned with DPDP, the RBI Cyber Security Framework and FREE-AI Pillar Governance.



- Classification before tokenisation. Every data element is classified before any AI operation runs against it. Classification determines how it can be used.

- Single source of policy. One Git-backed, cryptographically signed policy bundle, distributed to all enforcement points (boundary, egress DLP, cloud connector, on-premises and cloud agent runtimes).

- Federated catalog. Metadata is replicated cross-zone; data is not. The cloud catalog knows the structure of what tokens represent; never their values.

7. Data governance - federated across on-prem and cloud

7.3 Data classification taxonomy

Classification is the foundation. The architecture supports a four-tier scheme consistent with the RBI Cyber Security Framework expectations and DPDP categorisation. Each tier carries handling rules, retention rules and an approval matrix for policy exceptions. The classification is metadata that travels with the data through every transformation.

Tier	Description	Examples	Handling rule
Public	No sensitivity	Branch addresses, IFSC registry, published rates	Free movement
Internal	Bank-only	Process documents, internal correspondence	Within-bank, no external egress without approval
Confidential	Personal information per DPDP	Customer name, address, mobile, account number, transaction details	Tokenised before any AI operation; cloud egress permitted only as tokens
Restricted	DPDP-sensitive personal data and statutorily protected	Aadhaar, biometrics, financial details linked to identity, health data	Tokenised; cloud egress (even tokenised) requires explicit policy carve-out; restricted re-identification

7. Data governance - federated across on-prem and cloud

7.4 The federation pattern

Six elements compose the federation.

Figure 2 — Federated governance control plane. Policy and metadata flow on-prem → cloud; audit and lineage flow cloud → on-prem; identity is bridged via OIDC. The on-prem plane holds canonical state; the cloud plane is intentionally minimal.

Element	On-premises role	Cloud role
Catalog	OpenMetadata or DataHub holds the canonical schema, glossary and classification tags.	Read-only mirror of the on-premises catalog. Metadata only never values.
Glossary	Authoritative business vocabulary. "Aadhaar" means the same thing everywhere.	Identical glossary terms; cloud workloads consume the on-premises definition.
Lineage	Every transformation, including tokenisation and model invocation, is recorded as a lineage edge.	Cloud transformations emit lineage events back to the on-premises lineage server.
Policy	OPA bundles, signed and versioned, distributed to all enforcement points.	OPA sidecar enforces the same bundles within cloud workloads.
Identity	Authoritative identity provider typically Microsoft Entra ID or the bank's existing AD bridged to OIDC.	Cloud workloads authenticate against the on-premises IdP via OIDC. Federated SSO; no shadow accounts.
Audit	Central append-only audit lake; signed Merkle roots replicated to immutable storage.	Cloud emits real-time audit events to the on-premises lake; the cloud retains no master audit copy.

7. Data governance - federated across on-prem and cloud

7.5 The on-premises governance plane

The on-premises governance plane runs the canonical state. It comprises the catalog and lineage server, the glossary and steward registry, the policy authority (Git plus signing service), the identity provider (or its on-premises bridge), and the audit lake. Stewardship is explicit: every data domain has a named data owner (line of business head), a data custodian (IT) and authorised users (operational roles). The catalog enforces this reads, writes and policy changes flow through stewardship.

Quality controls discharge the bank's DPDP Section 8 obligations on accuracy, completeness and consistency. Quality rules live in the catalog and run continuously against operational data; failures generate steward tickets and, where DPDP Section 11 (right to correction) applies, customer-facing redress workflows.

7.6 The cloud governance plane

The cloud governance plane is intentionally minimal. It comprises a read-only catalog mirror, a model registry of cloud-hosted models in use with versions and performance metrics, an egress receipt log capturing every cloud invocation (model, version, prompt template hash, token count, response hash, latency), federated identity, and local OPA enforcement. The cloud governance plane does not need a full audit store; it streams events back to the on-premises audit lake. This minimality is intentional by holding less in the cloud, the bank reduces both attack surface and exit cost.

7. Data governance - federated across on-prem and cloud

7.7 Federation mechanics

Mechanism	What flows	Direction
Policy distribution	Signed OPA bundles	On-prem → cloud and to all on-prem enforcement points
Catalog metadata sync	Schema, glossary, classification tags	On-prem → cloud, read-only
Audit event streaming	All decisions, in real time	Cloud → on-prem audit lake
Identity assertions	Federated tokens	Cross-zone via OIDC
Consent state	Per-customer consent vector	On-prem → cloud as a real-time lookup
Lineage events	Transformation edges	Cloud → on-prem lineage server

This pattern keeps the data home but lets governance breathe across the hybrid. Auditors and the DPO see one truth, not two systems to reconcile.

8. AI governance for cloud

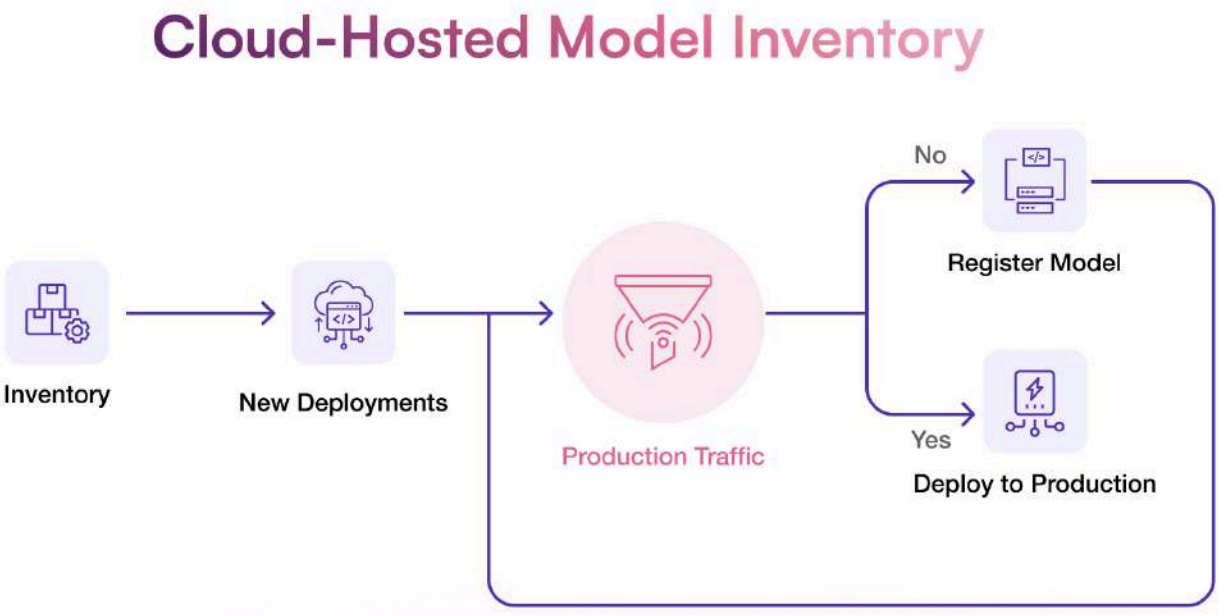
8.1 What AI governance covers and why it matters here

Data governance answers: who can access what data, for what purpose, under what policy. AI governance answers a different question what models are running, what are they doing, and are they doing it well. The two overlap (models access data) but the governance objects are distinct. AI governance objects include models, prompts, evaluations, fine-tunes, agentic flows and the decisions they produce.

For an Indian PSB, AI governance is also where the FREE-AI framework lands hardest. Sutras 4 (Fairness), 5 (Accountability), 6 (Understandable by Design) and Pillar Governance map almost entirely into the AI governance plane. Cloud-hosted models add specific obligations because the bank does not see the weights or the training data; governance must therefore extend through the contract, the egress receipt log, and the bank's own evaluation harness.

8.2 Model risk management for cloud-hosted models

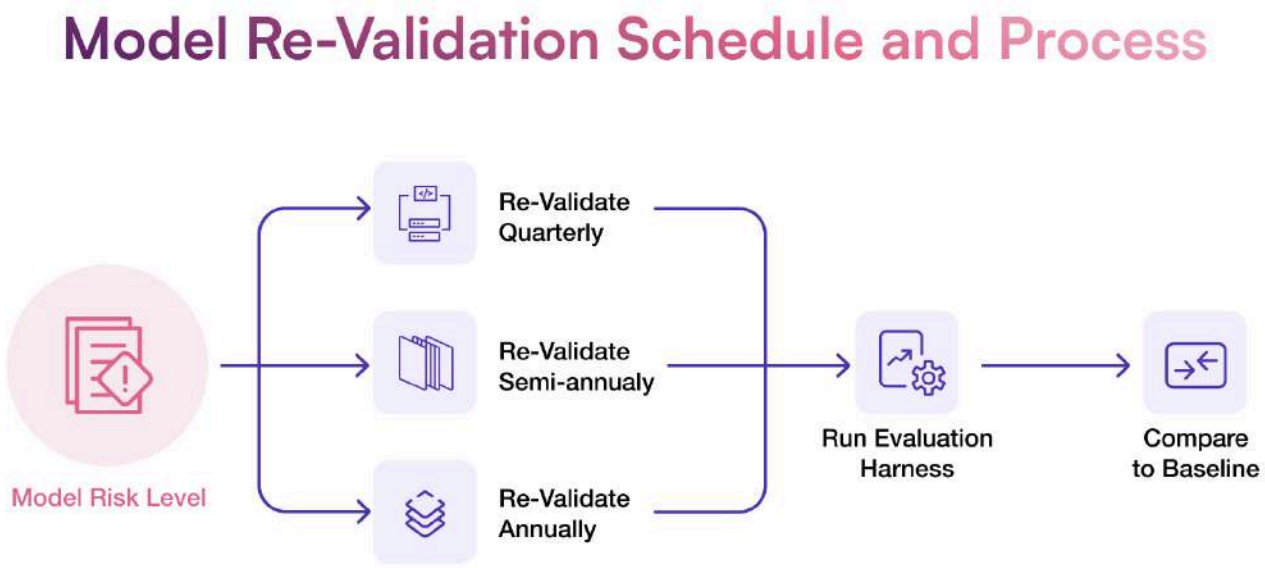
The bank already runs a Model Risk Management function for credit, market and operational risk models. AI is additive to that, not a parallel discipline. The MRM extension covers four objects.



8. AI governance for cloud

- Risk tiering. High risk includes models whose outputs affect customer outcomes — credit, fraud, customer service automation, regulatory communications. Medium risk includes internal productivity and document understanding. Low risk includes back-office search and summarisation. Tier determines approval threshold and re-validation cadence.

- Approval workflow. High-risk models require Model Risk Committee approval, DPO sign-off and CISO sign-off. Medium-risk models require DPO sign-off. Low-risk models require head-of-function sign-off.



8. AI governance for cloud

8.3 Model lifecycle

Stage	What happens	Owner
Approval	Risk-committee review, DPIA where applicable, sign-offs per risk tier	Model Risk Committee, DPO, CISO
Deployment	Connector pinned to a specific model version; signed and audited; canary rollout	Tantor admin, InfoSec
Operation	Continuous monitoring of drift, performance, fairness and cost	AI Ops
Re-validation	Periodic per risk tier; quarterly for high	Independent validation
Retirement	Graceful, with archive of decisions made by the retiring model and a documented successor	AI Ops, Compliance

8. AI governance for cloud

8.4 Vendor due diligence for cloud LLM providers

A non-negotiable list of contractual and technical requirements applies to any cloud LLM provider invoked by the bank's agent fabric.

- SOC 2 Type II report current within twelve months, or equivalent ISO 27001 plus 27017 plus 27018.
- Sub-processor list with mandatory advance notification of changes.
- Model card disclosing training data provenance, evaluation results and known limitations.
- Incident response and notification SLA aligned with CERT-In's six-hour reporting window.
- Right to audit, or proxy via independent attestation acceptable to the bank's internal audit function.
- Data Processing Agreement aligned with DPDP Section 8 data fiduciary obligations and Section 16 cross-border posture.
- Contractual carve-outs: token-only payload, no training on bank data, defined retention and deletion timelines, no cross-tenant exposure, indemnity for material breach.
- Demonstrated India-region availability with disaster recovery within Indian jurisdiction.

8. AI governance for cloud

8.5 Prompt and context governance

Prompts have become production code. They cannot be governed less rigorously than other code. The architecture requires four controls.

- Prompt registry. Every system prompt and template is registered, version-controlled, signed and reviewable.
- Evaluation gates. Every prompt change runs against a regression test set before promotion. The test set covers fairness across India-specific subgroups, not only accuracy.
- A/B testing. Prompts are A/B tested with explicit hypothesis, measurement and statistical bar before full rollout.
- Approval. Prompt changes for high-risk models follow the same approval workflow as model changes.
- A second-line guardrail check ensures prompts contain no PII. The boundary already prevents PII from entering a prompt; the second check is defence in depth.

8. AI governance for cloud

8.6 Cloud model performance monitoring

Five dimensions, monitored continuously, with thresholds that trigger investigation, rollback or retirement.

Dimension	Metric	Cadence
Quality	Human-rated sampling; ground-truth comparison where available; LLM-as-judge for scale, with human spot-checks	Weekly
Approval	Risk-committee review, DPIA where applicable, sign-offs per risk tier	Model Risk Committee, DPO, CISO
Fairness	Subgroup performance across India-specific dimensions: caste, region, language, urban or rural, digital literacy proxies	Monthly
Cost	Rupees per call, per agentic flow, per customer per month, with growth rate alerting	Continuous
Reliability	Error rate, latency p95 and p99, availability against vendor SLA	Continuous

8. AI governance for cloud

8.7 Incident handling for cloud AI

Four incident classes, each with a standing runbook.

- Vendor outage. Failover to a secondary cloud LLM or to an on-premises open-weight fallback. The Tantor connector abstracts this failover is configuration, not re-architecting.

- Performance regression. Rollback to last-known-good model version. Restoration is monitored; root-cause analysis is delivered within 48 hours.

- Data leak (theoretical, given the boundary). An incident drill is run quarterly. The runbook covers containment, vendor escalation, CERT-In six-hour notification and customer notification per DPDP requirements.

- Bias or fairness violation. Customer redress process; model retraining or replacement; disclosure to affected customers as required by DPDP Sections 11 to 14 (data principal rights).

8.8 Disclosures and audit support

The AI governance plane produces three standing outputs that satisfy regulatory and internal-audit requirements without bespoke effort.

- AI registry export for the RBI on demand, formatted to FREE-AI Pillar Assurance expectations.

- Annual report disclosures including AI governance frameworks, AI adoption areas, consumer protection measures and grievance redressal volumes.

- External audit support the audit lake is queryable by an external auditor with read access; the policy bundle and decision log together provide complete operational evidence with no production disruption.

9. Implementation roadmap

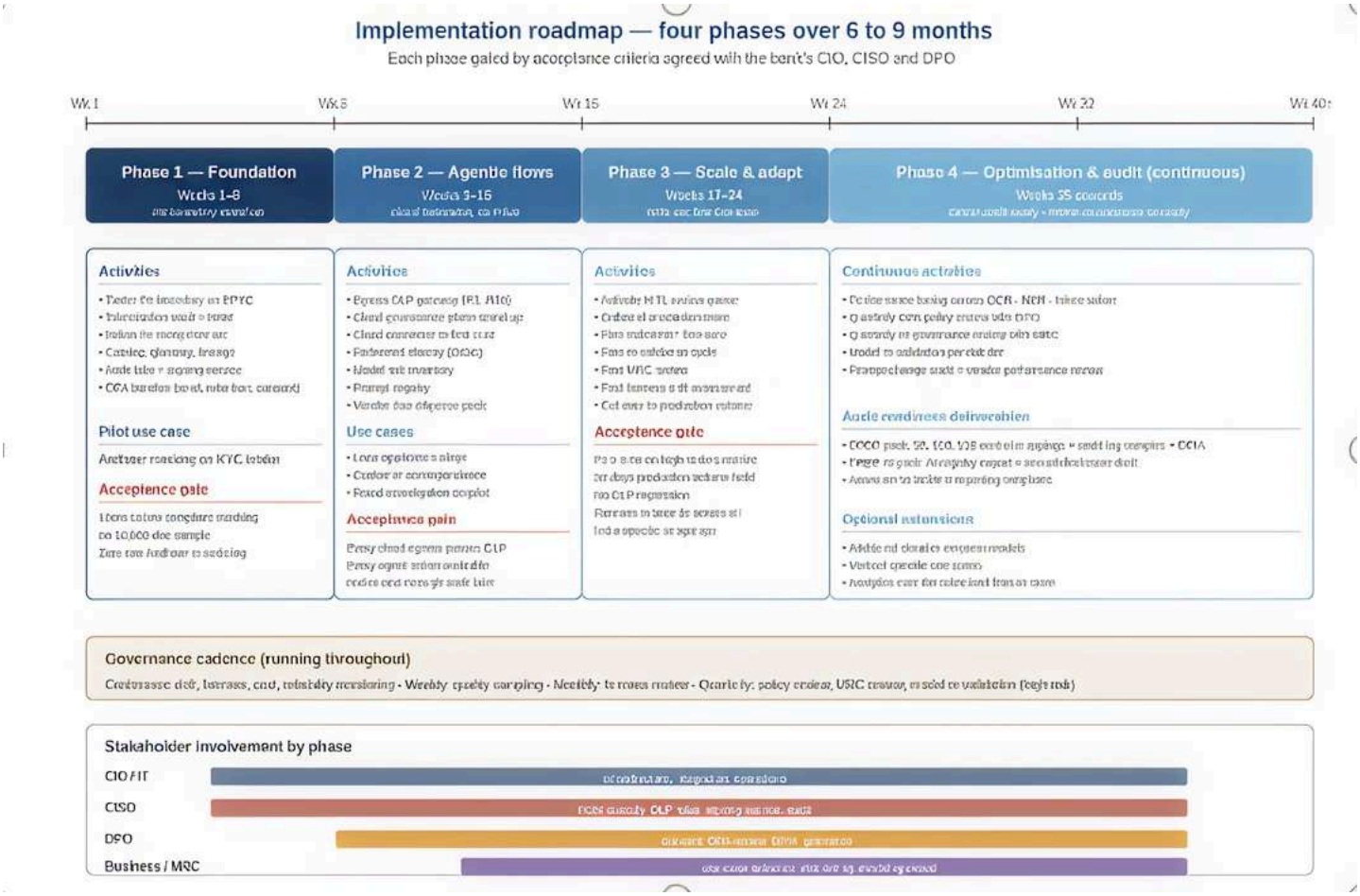


Figure 6 — Implementation roadmap. Four phases over 6—9 months; each gated by acceptance criteria. Stakeholder swim-lanes show CIO/IT, CISO, DPO and MRC involvement by phase. Governance cadence runs throughout.

9. Implementation roadmap

Phase 1 — Foundation (weeks 1 to 8)

- Deploy the Tantor PII boundary on-premises on a CPU node (EPYC 9J15-class).
- Stand up the tokenisation vault and integrate the bank's HSM.
- Configure the Indian PII recogniser set with bank-specific account-number and CIF formats.
- Stand up the on-premises governance plane: catalog, glossary, lineage server, audit lake, signing service for policy bundles.
- Implement OPA policy bundles for re-identification, retention and consent.
- Pilot use case: end-to-end Aadhaar masking on KYC document intake. Acceptance: 100% UIDAI-compliant masking on a sample of 10,000 documents, zero raw-Aadhaar in the audit log.

Phase 2 — Agentic flows and cloud federation (weeks 9 to 16)

- Deploy the egress DLP gateway with the full R1 to R10 rule set.
- Stand up the cloud governance plane: read-only catalog mirror, model registry, egress receipt log, federated identity via OIDC.
- Stand up the cloud connector to one approved cloud LLM provider, with a token-only payload contract and signed prompts.
- Activate AI governance controls: model risk inventory, prompt registry, vendor due-diligence pack, monitoring dashboards (quality, drift, fairness, cost, reliability).
- Implement two to three agentic use cases: typically loan-application triage, customer correspondence summarisation and a fraud-investigation co-pilot.
- Acceptance: every cloud egress event passes DLP without warning; every agent action is auditable end to end through the federated audit lake.

9. Implementation roadmap

Phase 3 — Scale and adapt (weeks 17 to 24)

- Activate the human-in-the-loop review queue and onboard the bank's annotation team.
- First-cut IndicBERT fine-tune trained on the bank's annotated corpus.
- Operationalise the AI governance lifecycle: first re-validation cycle, first model risk committee review, first fairness drift assessment.
- Cut over the pilot use cases from staging to production volume.
- Acceptance: detection F1 above 0.95 on high-stakes entities; production volume held for 30 days without policy or DLP regression; fairness metrics within bounds across all India-specific subgroups.

Phase 4 — Optimisation and audit (weeks 25 onwards)

- Performance tuning across the OCR, NER and tokenisation pipelines.
- Quarterly OPA policy review with the bank's DPO.
- Quarterly AI governance review with the Model Risk Committee, including model re-validation, prompt-change audit and vendor performance review.
- DPDP audit readiness pack: Section 8, 10 and 16 control mappings, audit log samples, DPIA documentation.
- FREE-AI audit readiness pack: AI registry export, annual report disclosure draft, AI incident reporting templates per Annexure VI.
- Optional: integration of additional cloud or on-premises models, vertical-specific use cases, and analytics over the tokenised feature store.

Annexure A — Control mapping

The table below maps the architectural elements of this paper to the principal regulatory and supervisory regimes.

Regulatory provision	Tantor architectural element	Evidence artefact
DPDP §4 — grounds for processing	Policy engine purpose declarations	OPA decision log per request
DPDP §5 — notice	Notice metadata returned in API responses	API contract and notice templates
DPDP §6 — consent	tantor.consent policy domain, consent registry	Consent ledger; withdrawal handling tests
DPDP §8 — fiduciary obligations	End-to-end audit log; HSM custody by bank	Audit log; HSM ceremony records
DPDP §10 — Significant Data Fiduciary	DPO console; quarterly policy review	Audit log; HSM ceremony records
DPDP §11—14 — principal rights	Erasure propagation across vault, embeddings and fine-tunes	Erasure runbook; test evidence
DPDP §16 — cross-border transfer	tantor.crossborder default-deny policy	Egress allowlist; OPA decision log
DPDP §33 — penalties	Defence-in-depth: boundary + DLP + tamper-evident log	Architecture diagram; control test results
RBI KYC Master Direction	Aadhaar masking enforced at re-id and render	UIDAI masking compliance test
RBI data localisation, April 2018	All production storage in India; HSM in India	Data centre location attestation

Annexure A — Control mapping

Regulatory provision	Tantor architectural element	Evidence artefact
RBI Digital Payment Security Controls	Token-only data in non-prod; DPM tested quarterly	Non-prod control test pack
RBI Cyber Security Framework	RBAC, segregation of duties, immutable logs	ISO 27001 / SOC 2 evidence
RBI Outsourcing Master Direction	Material outsourcing register; exit plan	Vendor due diligence pack
FREE-AI Sutra 1 (Trust)	PII boundary; tamper-evident audit; egress log as proof of no leak	Audit log samples; egress log Merkle roots
FREE-AI Sutra 2 (People First)	HITL review queue; human-override paths in agent flows	HITL throughput; override usage statistics
FREE-AI Sutra 4 (Fairness)	India-specific fairness benchmarks; multi-script detector evaluation	Fairness eval reports; subgroup metrics
FREE-AI Sutra 5 (Accountability)	End-to-end audit; bank-custody HSM and policy bundle	Audit log; HSM custody attestation
FREE-AI Sutra 6 (Explainability)	Detection confidence + rule trace; OPA decision trace; explicit cloud plan	Decision trace per request
FREE-AI Sutra 7 (Resilience)	Defence-in-depth controls; drift monitoring; open-component independence	Drift dashboards; component licence inventory
FREE-AI Pillar Infrastructure	On-prem deployment; AI Kosh-compatible tokenised exports	Architecture diagram; export schema

Annexure A — Control mapping

Regulatory provision	Tantor architectural element	Evidence artefact
FREE-AI Pillar — Policy	OPA policy bundles; board-approval workflow	Policy bundle hashes; sign-off records
FREE-AI Pillar — Capacity	Tantor Studio HITL UI; AI literacy curriculum	Training records; reviewer onboarding pack
FREE-AI Pillar — Governance	Model registry; lifecycle governance; vendor oversight via egress audit	Registry exports; change management records
FREE-AI Pillar — Protection	UIDAI masking on render; grievance hooks; egress DLP cyber control	Grievance audit; DLP block reports
FREE-AI Pillar — Assurance	External-auditor read access; Annexure VI incident reporting export	Auditor access logs; incident reporting templates
Data governance — federation	Single OPA policy bundle distributed to on-prem and cloud; metadata-only catalog mirror; cloud audit streamed to on-prem lake	Bundle distribution log; catalog sync attestation
Data governance — classification	Four-tier classification (Public, Internal, Confidential, Restricted) enforced before any AI operation	Classification coverage report; tag audit
Data governance — stewardship	Named data owner, custodian and authorised users per domain; quality rules in catalog	Steward registry; quality KPI report
Data governance — lineage	End-to-end lineage from source through tokenisation to model invocation and decision	Lineage graph export per use case

Annexure A — Control mapping

Regulatory provision	Tantor architectural element	Evidence artefact
AI governance — model registry	Inventory of cloud and on-prem models with version, vendor, risk tier, last validation date	Registry export; MRC sign-off log
AI governance — prompt control	Prompt registry, signed and version-controlled; evaluation gates before promotion	Prompt registry hashes; eval reports
AI governance — performance monitoring	Continuous quality, drift, fairness, cost, reliability monitoring with thresholds	Dashboard exports; threshold-breach incident log
AI governance — vendor due diligence	Standardised DD pack (SOC 2, sub-processors, model card, DPA, India residency)	Vendor attestation library
UIDAI Aadhaar masking circular 2018	First eight digits masked on all renders by default	Render audit; UI screenshot suite
UIDAI Authentication Regulations 2016	No clear-text Aadhaar at rest; tokenised everywhere	Database scan; token coverage report
CERT-In directions, April 2022	Six-hour reporting workflow; 180-day log retention; NTP sync	IR runbook; log retention attestation
IT Act 2000 / SPDI Rules	Reasonable security practices; ISO 27001:2022 alignment	ISO certification; gap-assessment report
Data governance — lineage	End-to-end lineage from source through tokenisation to model invocation and decision	Lineage graph export per use case

Annexure B — Glossary

Term	Definition
Agentic AI	An AI system that decomposes a goal into a sequence of actions, executes them, observes results, and iterates. Agents may chain tens of model calls per user request.
BIO tagging	Beginning, Inside, Outside — the standard scheme for token-level entity annotation in NER.
CLOUD Act	United States Clarifying Lawful Overseas Use of Data Act, 2018. Enables US legal process to compel US-headquartered providers to produce customer data held anywhere in the world.
DPDP Act	Digital Personal Data Protection Act, 2023. India's primary statute governing personal data.
DPDP Section 16	Provision permitting the Central Government to restrict cross-border transfer of personal data to specified jurisdictions.
DLP	Data Loss Prevention. In this paper, the egress DLP is the second-line gateway that inspects all outbound payloads.
FF1, FF3-1	Format-preserving encryption modes specified in NIST SP 800-38G.
FREE-AI	Framework for Responsible and Ethical Enablement of Artificial Intelligence in the Financial Sector. Released by the RBI on 13 August 2025. Comprises seven Sutras, two sub-frameworks, six pillars and 26 recommendations binding on RBI-regulated entities.
FREE-AI Sutras	Seven foundational principles: (1) Trust is the Foundation, (2) People First, (3) Innovation over Restraint, (4) Fairness and Equity, (5) Accountability, (6) Understandable by Design, (7) Safety, Resilience and Sustainability.

Annexure B — Glossary

Term	Definition
FREE-AI Pillars	Six strategic pillars across two sub-frameworks. Innovation Enablement: Infrastructure, Policy, Capacity. Risk Mitigation: Governance, Protection, Assurance.
Data catalog	Authoritative inventory of data assets, their schema, classification, owner and lineage. Open-source choices include OpenMetadata, DataHub and Apache Atlas.
Data lineage	Recorded chain of transformations a data element undergoes from source through processing to consumption. Each tokenisation, model invocation and decision is a lineage edge.
Data classification	Tiered tagging of data sensitivity. The architecture uses Public, Internal, Confidential, Restricted, with handling rules attached to each tier.
Data stewardship	Named accountability for a data domain — owner (line of business), custodian (IT) and authorised users. Reads, writes and policy changes flow through stewardship.
Federated catalog	Pattern in which catalog metadata is replicated cross-zone but data values are not. The cloud catalog mirrors the on-premises catalog read-only.
Federated identity	Pattern in which a single identity provider authenticates users and services across both zones. Typically via OIDC against the bank's existing AD or Entra ID.
Model risk management (MRM)	Discipline of inventorying, tiering, approving, monitoring and validating models. Extended from credit, market and op-risk models to cover AI.

Annexure B — Glossary

Term	Definition
Model registry	Authoritative inventory of all AI models in use, including vendor, version, risk tier, performance metrics and last validation date.
Model card	Vendor-provided document disclosing a model's training data, evaluation results, intended use cases and known limitations.
Prompt registry	Version-controlled, signed registry of every system prompt and template in production, treated with the same rigour as application source code.
DPIA	Data Protection Impact Assessment, required under DPDP Section 10 for Significant Data Fiduciaries on processing operations involving high risk to data principals.
HITL	Human-in-the-loop. Workflow pattern where automated decisions are sampled or escalated to human reviewers.
HSM	Hardware Security Module. FIPS-validated tamper-resistant device that stores and operates on cryptographic keys.
IndicBERT	Family of BERT-class language models trained on Indic languages by AI4Bharat.
LoRA	Low-Rank Adaptation. Parameter-efficient fine-tuning method that adds small adapter matrices rather than updating the full base model.
NER	Named Entity Recognition. The task of identifying and classifying entity mentions in text.
OPA	Open Policy Agent. Open-source policy engine with the Rego policy language.
PII boundary	The architectural perimeter inside which raw PII is permitted; outside it, only tokens travel.

Annexure B — Glossary

Term	Definition
Re-identification	Policy-gated reversal of a token to its original value.
SDF	Significant Data Fiduciary, designated under DPDP §10 with elevated obligations.
Sovereign AI	Capability to design, deploy and operate AI under the host country's legal jurisdiction without dependence on foreign-domiciled providers for production data flows.
Tokenisation	Replacement of a sensitive value with a non-sensitive surrogate that can be reversed only by an authorised vault.
UIDAI	Unique Identification Authority of India. Statutory authority for Aadhaar.
Verhoeff checksum	Decimal checksum algorithm used to validate Aadhaar numbers.
VID	Virtual ID. UIDAI-issued substitute for Aadhaar number, supporting masked authentication.

This whitepaper is provided to support the bank's evaluation of a sovereign AI architecture for production use. It is a reference architecture, not an implementation specification. Implementation will require the bank's own threat modelling, control testing, vendor due diligence and policy ratification.

Tantor a Translab Technologies product is available to support each of these steps as part of a formal engagement. For technical clarification, vendor due diligence material, or a customised pursuit pack against a specific use case, please contact Translab Technologies.

TANTOR™
a translab.io product

Thank You!

info@tantor.ai | tantor.ai