SOVEREIGN ENTERPRISE AI · REFERENCE ARCHITECTURE

Sovereign AI, built so the review survives contact with the evidence

A reference architecture for enterprise AI that stays inside the Kingdom's controls: residency by classification, a governed data layer underneath the models, agents that act within permissions — and a control mapping to PDPL, NDMO, SDAIA and NCA expectations with the audit evidence each one produces.

Version 2026.08 · 17 mapped controls · 12 evidence artefacts

This document describes architecture and delivery practice. It is not a certification, accreditation, audit opinion or regulator endorsement. Control identifiers are Bilytica's reading of published national expectations.

01 — THE PROBLEM THIS SOLVES

Most AI dies at the data layer

Enterprise AI in Saudi Arabia rarely fails on model quality. It fails when a reviewer asks where the number came from, which system holds the personal data behind an answer, who approved an agent's action, or where inference physically ran. Those questions are architectural. They cannot be answered retroactively by a pilot that was never built to produce evidence.

This reference architecture assumes the review from day one. Every layer below writes the artefact a data office, cybersecurity function or procurement reviewer will ask for, so compliance becomes an export rather than a project.

Residency is a decision, not a checkbox	Placement follows data classification, component by component — including embeddings, retrieval indexes and inference, the three that are usually forgotten.
The data layer is the control surface	Classification, ownership, quality and lineage are enforced in the platform, then inherited by everything the AI touches.
Agents act within permissions	Approval thresholds and permission-filtered retrieval mean an agent can never read or do more than the person who called it.
Evidence is generated, not assembled	Prompts, retrievals, actions, approvals and deletions write an exportable trail designed for a reviewer to read.

02 — REFERENCE ARCHITECTURE

Five layers, top to bottom

Read the stack downward: the assurance layer only works because the layers under it are governed. Each block lists what it contains and what it must produce.

Residency & platform — LAYER 1

A private tenant in your cloud subscription, an in-Kingdom region, or your own datacentre. Placement follows classification, not convenience.

- Private tenancy, no shared inference for restricted data
- Identity federation and least-privilege roles
- Encryption in transit and at rest, customer-managed keys
- Egress blocked by default, exceptions written down

Governed data layer — LAYER 2

The layer most AI programmes skip. Sources are unified, classified, quality-tested and lineage-tracked before anything is allowed to answer a question.

- Ingestion and transformation with tests in the pipeline
- Catalogue, ownership and column-level classification
- Master data and golden records
- Automatic lineage from source to published figure

Model layer — LAYER 3

A right-sized portfolio rather than one frontier model for everything: small task models where they win, larger models where reasoning is genuinely needed, all reachable through one governed router.

- Model router with per-use-case policy and cost ceilings
- Privately hosted and sovereign endpoints for restricted data
- Arabic evaluation sets, not translated English ones
- Versioned prompts and retrieval configuration

Agent & application layer — LAYER 4

Agents that finish work inside permissions and approval thresholds, and interfaces that cite their sources instead of asserting.

- Permission-filtered retrieval — the agent sees what the caller sees
- Approval thresholds and human escalation paths
- Cited answers and explicit abstention
- Arabic and English interfaces at parity

Assurance & operations — LAYER 5

The layer that produces the evidence. Everything above writes an auditable trail, and the trail is designed to be exported to a reviewer.

- Immutable action log shipped to your SIEM
- Evaluation and drift monitoring per release
- Retention, deletion and DSAR execution
- Evidence pack generated, not assembled by hand

03 — RESIDENCY DECISION MODEL

Where each component runs, and why

Classification drives placement. This is the default position; every exception is written down with a named approver.

Component	Public / internal data	Confidential data	Personal / restricted data
Storage & warehouse	In-Kingdom region	In-Kingdom region	In-Kingdom, private tenant
Transformation & quality	In-Kingdom region	In-Kingdom region	In-Kingdom, private tenant
Embeddings & vector index	In-Kingdom region	In-Kingdom, private tenant	In-Kingdom, private tenant, CMK
Model inference	Managed endpoint permitted	In-Kingdom endpoint	Privately hosted / sovereign endpoint
Agent runtime & actions	In-Kingdom region	In-Kingdom region	In-Kingdom, private tenant
Logs & audit trail	In-Kingdom, shipped to your SIEM	In-Kingdom, shipped to your SIEM	In-Kingdom, shipped to your SIEM

CMK = customer-managed keys. “Private tenant” means dedicated infrastructure in your subscription or datacentre, with no shared inference endpoint and no vendor-side copy of your corpus.

04 — CONTROL MAPPING

PDPL · NDMO · SDAIA · NCA

Each row is an expectation, what the architecture does about it, and the artefact a reviewer can be handed. Identifiers are Bilytica references, not official control numbers.

PDPL — Personal Data Protection Law

Obligations attaching to personal data processed in the Kingdom: lawful basis, purpose, minimisation, retention, data-subject rights, breach handling and transfer.

Ref	Requirement	In the architecture	Evidence produced
PDPL-01 Governed data layer	Personal data is processed for a defined, recorded purpose and no more data is used than that purpose needs.	Every dataset entering the platform is registered with a purpose, an owner and a classification. Pipelines carry column-level allow-lists, so a downstream product cannot silently widen the fields it consumes.	Dataset register export: purpose, owner, classification, fields in use, consuming products.
PDPL-02 Governed data layer	Data subjects can exercise access, correction and deletion rights within the statutory window.	A subject identifier is resolved once in the master-data layer, then lineage lists every table, extract, index and embedding store holding that person. Deletion runs as one governed job across all of them, including the AI retrieval index.	Worked DSAR run: request, resolved identity, systems touched, records returned or removed, timestamps.
PDPL-03 Residency & platform	Cross-border transfer of personal data is controlled and justified.	Classification decides placement. Personal and restricted data is processed on in-Kingdom infrastructure, including embeddings and inference; anything that would leave the Kingdom requires a written exception with a named approver, and the platform blocks egress by default.	Residency decision record per component, plus the egress policy and the exception log.
PDPL-04 Assurance & operations	Retention periods are defined and enforced, and breaches are detectable and reportable.	Retention is a property of the dataset, not a manual habit: expiry jobs run on schedule and write what they removed. Access to personal data is logged at query level and monitored for unusual volume or off-pattern access.	Retention schedule, deletion job history, access log sample and the alerting rules in force.

NDMO — National Data Management Office standards

National data management and personal data protection standards assessed domain by domain — governance, architecture, quality, metadata, master data, integration, classification and protection.

Ref	Requirement	In the architecture	Evidence produced
NDMO-01 Governed data layer	Data governance is operating: ownership, stewardship and decision rights are assigned and exercised.	Each domain has a named owner and steward inside the catalogue, and changes to definitions or classifications route through them. Governance is enforced by the platform rather than described in a document.	Catalogue export showing owners, stewards and the change history behind each definition.

Ref	Requirement	In the architecture	Evidence produced
NDMO-02 Governed data layer	Metadata and lineage are maintained so any published figure can be traced to source.	Lineage is captured automatically from the transformation layer, not drawn by hand, and every metric on a dashboard or in an AI answer links back to the model and source columns behind it.	Lineage graph export for a named metric, from dashboard tile back to source table.
NDMO-03 Governed data layer	Data quality is measured against defined rules, and issues have owners and remediation paths.	Quality rules run as tests in the pipeline. A failing rule can hold a release rather than surface later as a wrong number, and open issues sit in a queue with an owner and an age.	Quality rule inventory, pass/fail history and the open-issue queue with ageing.
NDMO-04 Governed data layer	Data is classified, and handling rules follow the classification wherever the data goes.	Classification is applied at the column level and inherited by every derived table, extract, index and embedding. A restricted column cannot lose its classification by being copied into a new model.	Classification register with inheritance view: source column to every downstream object.
NDMO-05 Governed data layer	Master and reference data are managed so entities are not duplicated across systems.	Customer, supplier, employee and asset entities are resolved into golden records with survivorship rules that are readable, versioned and reversible.	Match and survivorship rule set, duplicate-rate trend, and a sample resolved record with its inputs.

SDAIA — SDAIA AI ethics and generative AI guidance

National expectations on fairness, transparency, accountability, human oversight, robustness and acceptable use of generative AI.

Ref	Requirement	In the architecture	Evidence produced
SDAIA-01 Model layer	Each AI use case has a documented purpose, scope and prohibited uses.	Nothing reaches production without a register entry: what it is for, the data it may see, the decisions it may influence, and an explicit list of what it must not do.	Model and agent register entry, signed off by the business owner.
SDAIA-02 Model layer	Outputs are evaluated for accuracy and fairness before and after deployment.	Every use case has a held-out evaluation set built with the business, including Arabic-language cases where the workflow is Arabic. Scores are re-run on every change, and regression below threshold blocks release.	Evaluation set, scores by release, and the threshold policy that gates deployment.
SDAIA-03 Agent & application layer	Human oversight exists where AI affects a person or a material decision.	Agents act inside permissions and approval thresholds. Anything above a threshold, or anything touching a customer outcome, stops for a named human with the full context attached.	Approval matrix, escalation paths and an action log showing what was auto-approved versus reviewed.
SDAIA-04 Agent & application layer	AI answers are explainable and attributable, not unsourced assertions.	Retrieval answers carry citations to the underlying document or governed metric, and the interface shows when the system has insufficient grounding instead of producing a confident guess.	Sample transcripts with citations, plus the abstention rate and how it is monitored.

NCA — National Cybersecurity Authority controls

Essential and cloud cybersecurity controls: classification-driven hosting, tenant separation, identity, key management, logging, monitoring and third-party assurance.

Ref	Requirement	In the architecture	Evidence produced
NCA-01 Residency & platform	Hosting decisions follow data classification, with tenant separation for sensitive workloads.	The stack deploys as a private tenant — in your cloud subscription, an in-Kingdom region, or on-premises — with no shared inference endpoint for restricted data and no vendor-side copy of your corpus.	Deployment topology diagram, tenancy model and the placement decision per component.
NCA-02 Residency & platform	Identity, access and privilege are centrally controlled and least-privilege by default.	The platform federates to your identity provider. Roles map to data classifications, and the same entitlements are enforced in dashboards, APIs and AI retrieval — an agent cannot read what its caller cannot read.	Role-to-classification matrix, federation configuration and a permission-filtered retrieval test.
NCA-03 Residency & platform	Data is encrypted in transit and at rest, with controlled key management.	Transport encryption everywhere, encryption at rest on storage and vector indexes, and customer-managed keys where your policy requires the organisation to hold them.	Encryption and key-management statement per store, including the vector index.
NCA-04 Assurance & operations	Activity is logged, retained and monitored, and third-party components are assured.	Prompts, retrievals, agent actions and data access write to an immutable log shipped to your SIEM. Every third-party component in the stack is listed with its version, licence and where it runs.	Log schema and retention, SIEM integration note, and the component bill of materials.

05 — AUDIT-EVIDENCE CHECKLIST

What to have ready before the review

Twelve artefacts answer the large majority of Saudi data and AI review questions. Print this page, walk it with your data office and cybersecurity function, and mark what exists today versus what is assembled by hand each time.

	Ref	Artefact	What it contains	Usually asked by	Maps to
☐	EV-01	Data classification register	Every dataset with its classification, owner, purpose and handling rules — exported, not screenshotted.	Data office, internal audit	NDMO, PDPL
☐	EV-02	Residency decision record	Each component, where it runs, and the classification reasoning behind that placement — including embeddings and inference.	Cybersecurity, procurement	NCA, PDPL
☐	EV-03	Lineage export for a named metric	One published figure traced back to source tables, with the transformations in between.	Finance, regulator reporting	NDMO
☐	EV-04	Access and entitlement model	Roles mapped to classifications, plus a test showing AI retrieval respects the same entitlements.	Cybersecurity, internal audit	NCA, PDPL
☐	EV-05	Model and agent register	Purpose, data used, evaluation results, human oversight, prohibited uses and known limits per use case.	AI governance committee	SDAIA
☐	EV-06	Evaluation results by release	The held-out set, the scores including Arabic cases, and the threshold that gates deployment.	Business owner, AI governance	SDAIA
☐	EV-07	Action and audit log sample	Prompts, retrievals, agent actions and approvals for a period, in the shape your SIEM receives them.	Cybersecurity, internal audit	NCA, SDAIA
☐	EV-08	Retention and deletion evidence	Schedules, the jobs that ran, and what they removed — across warehouse, extracts and the retrieval index.	Privacy office, legal	PDPL, NDMO
☐	EV-09	Worked data-subject request	One end-to-end request showing identity resolution, systems touched and the response timeline.	Privacy office	PDPL
☐	EV-10	Component bill of materials	Every third-party model, library and service in the stack: version, licence, and where it executes.	Procurement, cybersecurity	NCA, SDAIA
☐	EV-11	Exit and portability plan	How data, models, prompts and pipelines leave with you, in open formats, if the engagement ends.	Procurement, legal	NCA, NDMO
☐	EV-12	Incident and escalation runbook	What counts as an AI or data incident, who is notified, in what window, and how it is recorded.	Cybersecurity, privacy office	NCA, PDPL, SDAIA

06 — USING THIS DOCUMENT

From reference to your environment

1 · Map your current state

Run the checklist above against what exists today. Most organisations find the model register, lineage export and residency decision record are the three that do not exist yet.

2 · Fix placement before scale

Decide residency per component while the estate is small. Moving an embedding store after it holds two years of personal data is the expensive version of this decision.

3 · Make the data layer the control point

Classification, ownership, quality and lineage enforced once in the platform are inherited by every AI product afterwards — including ones nobody has proposed yet.

4 · Ship one governed use case

A single production workflow with evidence attached is worth more in a review than five pilots. It also proves the pattern the rest of the roadmap reuses.

Data Foundation Assessment

A structured review of your data estate, residency position and AI readiness, delivered as a scored gap analysis with a sequenced remediation plan and the evidence artefacts listed in section 05.

bilytica.com/contact/ · bilytica.com/compliance/ · bilytica.com/solutions/sovereign-ai/

Bilytica — data, analytics and AI, delivered in Australia and Saudi Arabia. This document is provided for architectural and procurement discussion. It does not constitute legal advice, and it does not assert certification, accreditation or regulator approval of Bilytica or of any described system.