

BEYOND RAG | TEN PILLARS OF TRUST

Practical architecture framework for the Enterprise Question Assistant

01

Data Quality & Lineage

Validate assessment metadata, detect staleness and track source lineage.

02

Privacy Engineering

Detect PII or secrets and apply policy-driven masking before approved model calls.

03

Security & Authorization

Enforce document and agent-tool access using independent authorization decisions.

04

Semantic Layer & Business Context

Supply certified definitions for vendor risk, deadlines, joins and metrics.

05

Agentic Threat Defense & Runtime Security

Detect prompt injection, enforce runtime tool restrictions and block exfiltration.

06

Grounding & Provenance

Measure retrieval relevance and whether generated claims are supported by evidence.

07

Multi-Agent Validation

Orchestrate Extractor → Validator → Evaluator and human escalation.

08

Observability & Tracing

Trace each retrieval, tool and LLM span without recording unnecessary private data.

09

Auditability & Accountability

Collect business approval events and store protected records for later investigation.

10

Reproducibility & Human Governance

Version prompts, datasets and evaluation runs to compare decisions across releases.

QUESTION ASSISTANT EXAMPLE

Can we trust the version of Vendor Orion's security assessment we retrieved?

1 WHAT IS IT?

Quality checks detect missing, inconsistent, duplicated or stale records. Lineage tracks the path from source document through transformations to a retrievable chunk.

2 WHY DO IT?

RAG cannot fix a wrong, outdated or incomplete source. Without lineage, a reviewer cannot explain where a cited answer originated.

3 HOW TO IMPLEMENT

Assign source_id, owner, version, classification and checksum at ingestion; validate required fields; normalize formats; quarantine failures; preserve parent-child links from document to chunk; reindex on approved changes.

4 QUESTION ASSISTANT | BUILT VS PROPOSED

The Question Assistant ingested assessment documents; the proposed next-generation pipeline would flag expired evidence and duplicate policy versions before retrieval.

5 INDUSTRY TOOLS & PLATFORMS • ILLUSTRATIVE OPTIONS

Great Expectations (GX)

Automated data validation and quality rules

OpenMetadata

Data catalog, ownership, lineage and quality

QUESTION ASSISTANT EXAMPLE

Can we answer the encryption question without revealing a contact's phone number?

1 WHAT IS IT?

Privacy engineering applies classification, data minimization, redaction, tokenization and retention controls. Pseudonymization is not necessarily anonymization.

2 WHY DO IT?

Retrieved evidence may contain customer details, personal identifiers or credentials; copying them into prompts, logs or external services creates avoidable exposure.

3 HOW TO IMPLEMENT

Classify documents and queries; detect secrets/PII; remove irrelevant fields; tokenize needed identifiers using a protected vault; allow only eligible model providers; inspect outbound prompts and inbound outputs; enforce retention policy.

4 QUESTION ASSISTANT | BUILT VS PROPOSED

For a vendor assessment, send policy clauses and approved controls, not raw personnel details or secrets. Historical use of this specific gateway is not established.

5 INDUSTRY TOOLS & PLATFORMS • ILLUSTRATIVE OPTIONS

Microsoft Presidio

PII identification, masking and de-identification

Google Cloud Sensitive Data Protection

Detect and de-identify sensitive content

QUESTION ASSISTANT EXAMPLE

Can an employee in Customer A retrieve Customer B's vendor assessment?

1 WHAT IS IT?

Identity and access management (IAM), role-based access control (RBAC), attribute-based access control (ABAC), tenant isolation and least-privilege tools restrict what users and agents may access.

2 WHY DO IT?

A model can disclose confidential content if the retrieval system hands it unauthorized chunks. Prompt instructions alone do not enforce confidentiality.

3 HOW TO IMPLEMENT

Bind requests to user and agent identities; evaluate document-level and chunk-level permissions before retrieval; scope tool credentials; isolate tenants; treat retrieved text as untrusted; block unauthorized outputs and high-risk actions.

4 QUESTION ASSISTANT | BUILT VS PROPOSED

A Customer A request must never retrieve Customer B material. This is a reference security design; specific Fortress permission enforcement has not been confirmed.

5 INDUSTRY TOOLS & PLATFORMS • ILLUSTRATIVE OPTIONS

Open Policy Agent (OPA)

Policy-as-code authorization decisions; application must enforce decisions

Immuta

Governed access, row restrictions and masking policies

QUESTION ASSISTANT EXAMPLE

What does “overdue high-risk vendor assessment” mean across all systems?

1 WHAT IS IT?

A semantic layer centralizes entities, measures, metrics, joins and applicable access rules so agents and analysts use consistent definitions. It complements—rather than replaces—unstructured RAG.

2 WHY DO IT?

An agent may guess inconsistent SQL joins or redefine high-risk, overdue and active for each question. Central definitions reduce metric ambiguity.

3 HOW TO IMPLEMENT

Model vendor, assessment and policy entities; certify overdue/high-risk criteria; define joins, time logic and access; expose authorized metrics through APIs or MCP; test outputs against known answers.

4 QUESTION ASSISTANT | BUILT VS PROPOSED

Proposed: combine approved vendor-risk metrics with retrieved assessment policies before the Extractor, Validator and Evaluator process the answer. No historical Cube or semantic-platform deployment has been established.

5 INDUSTRY TOOLS & PLATFORMS • ILLUSTRATIVE OPTIONS

Cube

Governed semantic metrics, joins, access and agent APIs/MCP

dbt Semantic Layer / MetricFlow

Reusable certified metrics via semantic definitions and MCP

QUESTION ASSISTANT EXAMPLE

What stops a retrieved vendor document from instructing the agent to leak customer data?

1 WHAT IS IT?

Runtime AI security addresses malicious instructions, tool misuse, data exfiltration and unsafe agent actions while the workflow executes, not just data access at ingestion.

2 WHY DO IT?

An authorized agent may still be manipulated by retrieved documents. Validator or Evaluator checks after the fact cannot prevent a harmful tool action already executed.

3 HOW TO IMPLEMENT

Treat retrieved text as data, not privileged instructions; allowlist tools and destinations; enforce permissions outside LLM prompts; inspect risky input/output; gate side effects; log and test adversarial attempts.

4 QUESTION ASSISTANT | BUILT VS PROPOSED

Proposed: a vendor assessment contains a hidden instruction to send results to an external endpoint; a separate policy layer blocks the outbound action and routes the incident for review.

5 INDUSTRY TOOLS & PLATFORMS • ILLUSTRATIVE OPTIONS

Palo Alto Networks Prisma AIRS

Runtime prompt-injection, data leakage and agent threat defense

Check Point AI Guardrails (Lakera)

Prompt injection/jailbreak and data-leakage screening

QUESTION ASSISTANT EXAMPLE

Which approved evidence supports the statement that encryption is enabled?

1 WHAT IS IT?

Grounding ties generated claims to retrieved evidence. Provenance records the origin, owner, approval state and version of that evidence.

2 WHY DO IT?

A fluent answer can combine unrelated statements or use an outdated exception. Citations without claim-level support can still be misleading.

3 HOW TO IMPLEMENT

Retrieve with source IDs and permissions; keep document-version references; require each material claim to identify supporting passages; validate that citations substantiate claims; flag conflicting policies; abstain and escalate when evidence is insufficient.

4 QUESTION ASSISTANT | BUILT VS PROPOSED

The Question Assistant used retrieval and validation. Claim-level evidence mapping and version-specific citation enforcement are proposed unless historically confirmed.

5 INDUSTRY TOOLS & PLATFORMS • ILLUSTRATIVE OPTIONS

Ragas

Faithfulness, context recall/precision and answer quality evaluations

TruLens

Groundedness, contextual relevance and citation checks

QUESTION ASSISTANT EXAMPLE

Can three independent roles reduce unsupported or incomplete answers?

1 WHAT IS IT?

A multi-model workflow separates proposing an answer, checking its grounding and evaluating completeness or quality. Human-in-the-loop provides oversight for disputed cases.

2 WHY DO IT?

One model may hallucinate, omit a control or misread an exception. Multiple checks create opportunities to catch mistakes but do not guarantee correctness.

3 HOW TO IMPLEMENT

Give agents typed inputs/outputs and narrow roles. The Extractor returns candidate answer plus evidence. The Validator checks support and contradictions. The Evaluator applies quality criteria and escalation rules. Route high-risk uncertainty to a named reviewer.

4 QUESTION ASSISTANT | BUILT VS PROPOSED

Vivek described this three-agent, three-LLM pattern and human-in-the-loop as used in the Question Assistant. Exact models, thresholds and schemas remain to be verified.

5 INDUSTRY TOOLS & PLATFORMS • ILLUSTRATIVE OPTIONS

LangGraph (LangChain)

Stateful agent orchestration, review gates and workflow control

LangSmith

Agent evaluation, debugging and regression tests

QUESTION ASSISTANT EXAMPLE

Why did the Validator take seven seconds and trigger a retry?

1 WHAT IS IT?

Observability combines traces, logs and metrics. A trace links operations for one request; spans measure individual retrieval, model, tool and review steps.

2 WHY DO IT?

Without linked telemetry, operators cannot distinguish failures in retrieval, model responses, tool execution, orchestration, or human approvals.

3 HOW TO IMPLEMENT

Generate a request trace_id; instrument each tool and agent call; record durations, model identity, errors, retrieval IDs, token/cost metrics and outcomes; alert on failure and drift. Avoid recording sensitive prompt or answer bodies by default.

4 QUESTION ASSISTANT | BUILT VS PROPOSED

A production Question Assistant could show which agent failed and whether retrieval or validation caused the problem. Actual Fortress telemetry implementation is unconfirmed.

5 INDUSTRY TOOLS & PLATFORMS • ILLUSTRATIVE OPTIONS

Langfuse

LLM traces, prompts, evaluations and datasets

Arize Phoenix

OpenTelemetry-based LLM tracing and evaluations

QUESTION ASSISTANT EXAMPLE

Six months later, can an auditor tell who approved this vendor answer?

1 WHAT IS IT?

Auditability is the ability to reconstruct material decisions from reliable records. Accountability identifies responsible humans and business functions; an AI agent is not the ultimate accountable owner.

2 WHY DO IT?

Logs alone may omit why an answer was accepted, which policy version applied, whether evidence changed, and who approved the result.

3 HOW TO IMPLEMENT

Keep controlled event records for access decisions, evidence IDs+versions, agent/prompt/model refs, validation findings, human reviewer action and published output. Use append-only or integrity-protected storage, defined retention and strict audit access.

4 QUESTION ASSISTANT | BUILT VS PROPOSED

A reviewer could reconstruct the encryption answer, policy exception and approval trail. Historical use of tamper-evident evidence manifests is not established.

5 INDUSTRY TOOLS & PLATFORMS • ILLUSTRATIVE OPTIONS

AWS CloudTrail + S3 Object Lock

Cloud activity evidence and protected WORM retention

Langfuse Enterprise Audit Logs

Langfuse platform admin activity audit logs (plan-dependent)

QUESTION ASSISTANT EXAMPLE

Can we investigate why a later run reaches a different conclusion?

1 WHAT IS IT?

Replayability preserves enough permitted state to investigate a past run. Reproducibility tests consistent findings under controlled conditions. Idempotency prevents retries from duplicating side effects.

2 WHY DO IT?

Model and data changes can alter results. A requirement for exact identical generated sentences is usually unrealistic, even with a temperature setting of zero.

3 HOW TO IMPLEMENT

Capture source snapshots, retrieval configuration, prompt/template version, provider/model IDs, validation rules and policy versions. Replay in a controlled sandbox; compare evidence and decisions; require human sign-off for high-impact changes; apply idempotency keys to writes.

4 QUESTION ASSISTANT | BUILT VS PROPOSED

Human review was part of the Question Assistant; systematic replay, version snapshots and idempotent approval workflows are proposed enhancements.

5 INDUSTRY TOOLS & PLATFORMS • ILLUSTRATIVE OPTIONS

MLflow

Versioned prompt registry, evaluations and experiments

Weights & Biases Weave

AI traces and repeatable evaluation experiments