REFERENCE ARCHITECTURE

Governed Agentic RAG: Reference Architecture

A reference architecture for an agentic retrieval-augmented generation pipeline that can pass an ISO/IEC 42001 audit and meet the MAS FEAT principles: policy orchestration, deterministic state checks, immutable audit trail and layered prompt-injection defence.

Terence Kok
Doctoral Candidate (DBA), Digital Leadership, Golden Gate University
Executive Director, AI Governance & Assurance Practice, Orion Five Engineering, Singapore

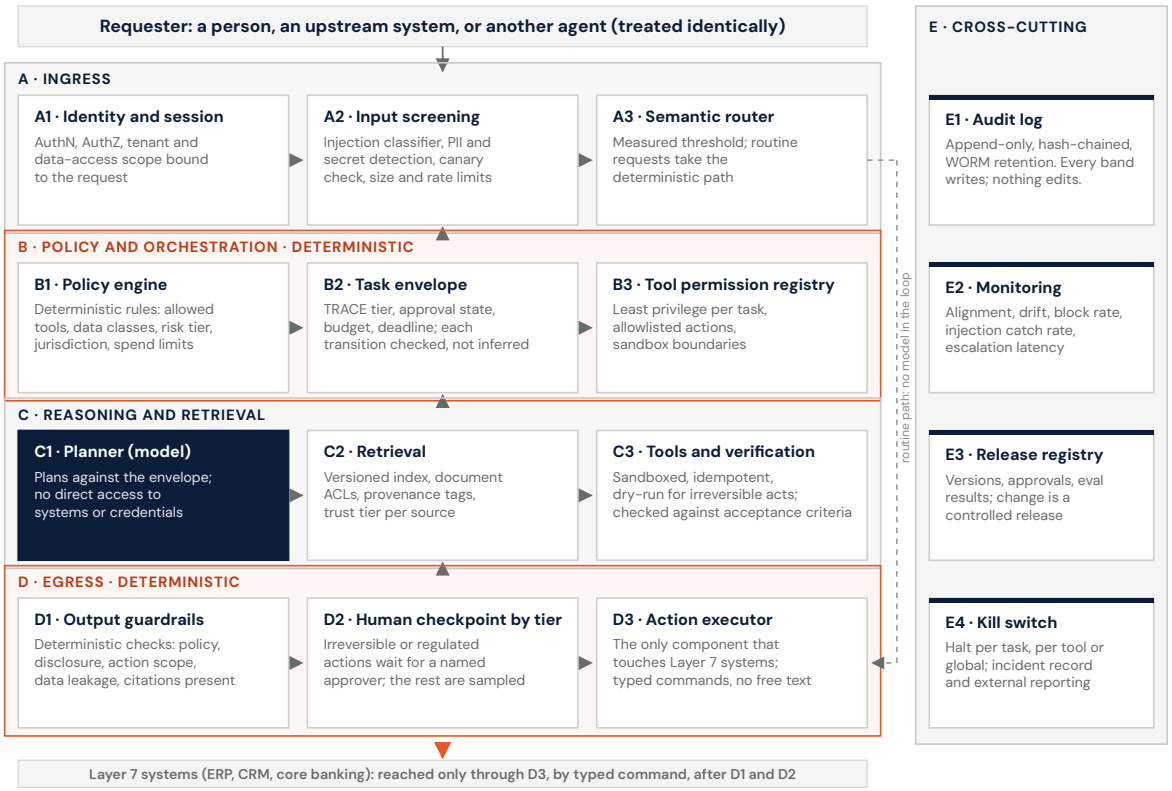
SCOPE

The model reasons and deterministic code decides.

The architecture covers an agent that receives a request, retrieves from enterprise sources, plans, calls tools, and acts on production systems. It is written for the architect who has to build it and the board or auditor who has to be satisfied that it is governed. It is a reference, not a description of one deployment: each design rule is the one that published guidance requires or recommends, assembled into a single pipeline and traced to its source in the basis table below. The organising rule is the one from the Governed Cognitive Layer: the model reasons, deterministic code decides. Every band that can change what reaches a production system is deterministic; the model sits in one band and has no credentials.

REFERENCE ARCHITECTURE

A request moves down through ingress, policy, reasoning and egress. The two orange bands are deterministic code. The model lives only in C1 and cannot reach a system, a credential or unscoped data. Every band writes to the audit log on the right.



REQUEST PATH

Every state transition passes a deterministic check.

The task envelope (B2) carries the state. A transition happens only when the deterministic check for that state passes; the planner can propose a transition but cannot make one. This is what "deterministic state checks" means in practice: the model's output is an input to a check, never the check itself.

STATE	CHECK THAT MUST PASS TO ENTER IT
Received	Principal, tenant and scope resolved (A1). Input screened (A2).
Classified	Router verdict recorded. Routine requests leave the model path here.
Enveloped	TRACE tier, budget, deadline and tool allowlist attached (B1, B2, B3).
Planned	Plan is inside the envelope: no tool outside the allowlist, no data outside scope.
Executed	Every tool call idempotent; irreversible calls ran as dry-run first (C3).
Verified	Output meets the acceptance criteria set before planning; guardrails passed (D1).

STATE	CHECK THAT MUST PASS TO ENTER IT
Approved	Required for irreversible or regulated tiers; approver named, trace attached (D2).
Committed	Typed command issued to Layer 7 by D3 only. Audit entry sealed (E1).
Escalated / Halted	Any failed check above. Task cannot re-enter the path without a human.

CONTROL MAPPING

Which component satisfies which control.

An auditor working from ISO/IEC 42001 needs to see which component gives effect to which control, and a MAS-regulated institution needs to show which FEAT principle each design decision serves. The table is the artefact both will ask for. Clause numbers and control titles are cited; the text of the standard is licensed and is not reproduced here.

COMPONENT	WHAT IT DOES	ISO/IEC 42001:2023	MAS FEAT (2018)
A1 Identity and session	Binds every request to a principal, a tenant and a data-access scope before anything is retrieved or planned.	A.9.4 intended use of the AI system; A.3.2 AI roles and responsibilities. Access control itself is an ISO/IEC 27001 control, referenced not duplicated	P7 internal authorisation for AIDA-driven decisions; P8 firm remains responsible for the model
A2 Input screening	Classifies the request for injection patterns, strips or flags PII and secrets, checks canaries, enforces size and rate limits.	A.6.2.6 AI system operation and monitoring; 6.1.2 AI risk assessment	P4 models function as intended; P6 AIDA decisions held to the same ethical standard as human ones
A3 Semantic router	Routes routine requests to the deterministic path so the model handles only what needs reasoning. Threshold measured on a labelled sample.	A.6.2.2 AI system requirements and specification; A.6.2.4 AI system verification and validation	P7 authorised use of AIDA to drive a decision; P4 functioning as intended
B1 Policy engine	Deterministic rules on allowed tools, data classes, risk tier, jurisdiction and spend, evaluated before and after reasoning. Rules are versioned code, not prompts.	A.2.2 AI policy; A.9.2 processes for responsible use of AI systems	P5 alignment with the firm's ethical standards and codes of conduct; P7 internal authorisation
B2 Task envelope and state machine	Each task carries a TRACE risk tier, an approval state, a budget and a deadline. Transitions are checked against the envelope, never inferred from model output.	A.5.2 AI system impact assessment process; 8.4 AI system impact assessment	P7 internal authorisation by materiality; P9 board and management awareness of AIDA use
B3 Tool permission registry	Per-task allowlist of actions with least privilege; sandbox boundaries; credentials never visible to the planner.	A.4.4 tooling resources; A.10.2 allocating responsibilities with third parties	P7 internal authorisation; P8 responsibility for internally and externally sourced models

COMPONENT	WHAT IT DOES	ISO/IEC 42001:2023	MAS FEAT (2018)
C1 Planner	Produces a plan and candidate actions inside the envelope. Has no direct access to systems, credentials or unscoped data. Working context is session-scoped and isolated from long-term memory, and memory writes are validated (OWASP Agentic T1).	A.6.2.3 documentation of AI system design and development; A.6.2.7 AI system technical documentation	P6 same ethical standard as human decisions; P8 firm responsible for the model's decisions
C2 Retrieval	Versioned index; document-level ACLs enforced at query time; every passage carries source, version and trust tier; retrieved text is data, never instruction.	A.7.2 data for development and enhancement of AI system; A.7.5 data provenance; A.7.4 quality of data for AI systems	P2 justified use of personal attributes as inputs; P3 data accuracy and relevance; P13 data used is explainable on request
C3 Tool execution and verification	Sandboxed, idempotent calls under a per-task allowlist; results checked against pre-agreed acceptance criteria before egress. Where the target system offers a simulation or dry-run mode, irreversible actions use it first; where it does not, the action waits at D2 regardless.	A.6.2.4 AI system verification and validation; A.6.2.5 AI system deployment	P3 validation for accuracy; P4 models function as intended
D1 Output guardrails	Deterministic checks on the proposed output: policy compliance, disclosure of AI involvement, action scope, data leakage, citations present.	A.6.2.6 AI system operation and monitoring; A.8.2 system documentation and information for users	P1 no unjustified systematic disadvantage; P6 same ethical standard; P12 disclosure of AIDA use
D2 Human checkpoint by tier	Irreversible, financial or regulated actions wait for a named approver with the reasoning trace attached (mandatory, OWASP LLM06). Lower tiers are reviewed after the fact on a defined sample rate, and the approver queue length is monitored so oversight is not overwhelmed (OWASP Agentic T10).	A.3.2 AI roles and responsibilities; A.9.2 processes for responsible use of AI systems	P7 internal authorisation; P10 data subjects can enquire, appeal and request review; P11 supplementary data considered on review
D3 Action executor	The only component with credentials for production systems. Accepts typed commands only; free text never reaches Layer 7.	A.4.5 system and computing resources; A.6.2.5 AI system deployment	P8 firm responsible for the AIDA-driven action
E1 Immutable audit log	Append-only, hash-chained record of inputs, retrieved sources and versions, policy verdicts, tool calls, model version, approvals and outputs.	A.6.2.8 AI system recording of event logs; 9.1 monitoring, measurement, analysis and evaluation	P8 responsibility demonstrable; P11 review evidence; P13 and P14 explanations of data and consequences on request
E2 Evaluation and monitoring	Alignment to policy ground truth, drift, block rate, injection catch rate, escalation latency; reviewed on a defined cadence.	9.1 monitoring, measurement, analysis and evaluation; A.6.2.6 AI system operation and monitoring; 10.1 continual improvement	P3 regular review for accuracy and unintentional bias; P4 functioning as intended; P9 management awareness

COMPONENT	WHAT IT DOES	ISO/IEC 42001:2023	MAS FEAT (2018)
E3 Model and prompt registry	Model, prompt and policy versions with approval and evaluation results; a change is a controlled release, not an edit.	A.6.2.7 AI system technical documentation; 8.1 operational planning and control; 6.3 planning of changes	P8 responsibility for sourced and internal models; P4 functioning as intended after change
E4 Kill switch and incident response	Halt per task, per tool or globally; incident record; external reporting where a regulator or data subject must be told.	A.8.4 communication of incidents; A.8.3 external reporting; 10.2 nonconformity and corrective action	P9 board and management awareness; P10 channel for affected data subjects

> MAS FEAT principles: P1 to P4 Fairness (P1 and P2 justifiability of AIDA-driven decisions, P3 and P4 accuracy and ongoing review); P5 and P6 Ethics; P7 to P9 internal Accountability (authorisation, responsibility for models, board awareness); P10 and P11 external Accountability (enquiry, appeal and review by data subjects); P12 to P14 Transparency (proactive disclosure, explanation of data used and of consequences). AIDA is MAS's term for artificial intelligence and data analytics.

AUDIT TRAIL

What the audit log records.

A log that says "the agent approved the refund" is not evidence. The log has to let a reviewer reconstruct the decision: the request as received, the identity and scope, the router verdict, the envelope, every retrieved passage with its source and version, the model and prompt versions, every tool call with arguments and result, every policy verdict, the approver if any, and the final command. The record is integrity-protected (a hash chain or an equivalent tamper-evident mechanism, NIST SP 800-53 AU-9), storage is write-once, and retention follows the record-keeping rule of the regulated activity, not the convenience of the platform. Nothing in the pipeline has permission to edit or delete an entry, including the operator, which is what makes the record usable for non-repudiation (AU-10) and for the explanations MAS FEAT P13 and P14 require on request.

PROMPT-INJECTION DEFENCE

An injection that reaches the planner cannot act.

No classifier catches every injection, so the design assumes some will reach the planner and makes that harmless. The controls stack so that a successful injection can influence the plan but cannot widen permissions, reach a credential, touch a system outside the allowlist, or pass the output checks with leaked data.

LAYER	CONTROLS
Before the model	Input classifier tuned on known injection patterns; canary tokens in system context that must never appear in output; strict separation of instruction channel and data channel in the prompt template.
In retrieval	Every retrieved passage is tagged with a trust tier. Untrusted tiers are rendered as quoted data with an explicit "not instructions" wrapper, and content from them can never raise the task's tool permissions.

LAYER	CONTROLS
Around tools	Allowlist per task (B3), least privilege, sandboxed execution, no credentials visible to the planner, human approval before any irreversible action. A successful injection that reaches the planner still cannot reach a system it was not already allowed to touch.
After the model	Deterministic output checks (D1): scope, leakage, disclosure, citation presence. Canary detection. Human gate for the highest tier (D2).
Over time	Injection attempts are logged (E1) and the catch rate is a monitored metric (E2). Red-team prompts are part of every model or prompt release (E3).

RISK MATRIX

Residual risk after controls.

RISK	LIKELIHOOD	IMPACT	CONTROLS	RESIDUAL
Prompt injection via a retrieved document (OWASP LLM01)	**High**	**High**	A2, C2 trust tiers, B3, D1	Low: an injected instruction cannot widen permissions or bypass D1
Data leakage across tenants or ACLs (OWASP LLM02, LLM08)	Medium	**High**	A1, C2 query-time ACLs, D1 leakage check	Low
Irreversible action on a wrong plan (OWASP LLM06, Agentic T2)	Medium	**High**	B2 tiering, C3 sandbox and dry-run where available, D2 mandatory approval, D3 typed commands	Low
Silent quality drift after a model update (NIST AI RMF MEASURE)	**High**	Medium	E3 controlled release, E2 alignment and drift metrics	Medium: depends on evaluation set coverage
Unreconstructable decision at audit (Agentic T8, NIST AU-10)	Medium	**High**	E1 integrity-protected log with versions of every input	Low
Cost or latency runaway (OWASP LLM10, Agentic T4)	Medium	Medium	A3 routing, B2 budget and deadline, E4 halt	Low
Approvers overwhelmed into rubber-stamping (Agentic T10)	Medium	**High**	B2 tiering keeps mandatory approvals rare; E2 escalation latency and queue length; sampling for lower tiers	Low, if the tiering is revisited when latency rises
Agent-to-agent escalation (Agentic T3, T9, T13)	Low	**High**	Zero-trust at A1 for every requester; envelopes do not inherit permissions; E2 watches cross-agent traffic	Medium: multi-agent visibility is the least mature control

OPERATING METRICS

Six metrics to review every month.

METRIC	WHAT IT TELLS YOU
Policy block rate	How often D1 or B1 stop an output. Rising means the model or the prompt drifted, or the policy changed.

METRIC	WHAT IT TELLS YOU
Injection catch rate	Detected attempts at A2 and D1 against a seeded red-team set. Falling means the classifier needs retraining.
Provenance coverage	Share of outputs whose every claim cites a retrieved passage with a source and version. Target is 100 per cent.
Escalation latency	Time from D2 request to human decision. Long tails mean the approver is a bottleneck or the tiering is wrong.
Alignment score	Statistical closeness of outputs to a policy ground-truth set, tracked per model version.
Routine path share	Share of requests A3 keeps off the model. Falling share is a cost and latency warning.

BASIS

Where each design rule comes from.

Nothing in the pipeline is novel, and that is the point. A reviewer should be able to trace every rule to a source they already trust. Where a source recommends rather than requires, the rule above says "where available" or "on a defined sample" rather than "always".

DESIGN RULE	PUBLISHED BASIS
Every requester is authenticated and authorised on every request, whether a person, an upstream system or another agent.	NIST SP 800-207, Zero Trust Architecture: no implicit trust by network location or requester type. OWASP Agentic AI Threats T9 Identity Spoofing and T3 Privilege Compromise. https://doi.org/10.6028/NIST.SP.800-207
Retrieved content is data, never instruction; the model has no direct access to systems or credentials; tools run under least privilege from a per-task allowlist.	OWASP Top 10 for LLM Applications 2025: LLM01 Prompt Injection, LLM06 Excessive Agency (minimise extensions, permissions and autonomy), LLM08 Vector and Embedding Weaknesses. OWASP Agentic T2 Tool Misuse. https://genai.owasp.org/llm-top-10/
Irreversible, financial or regulated actions require explicit human approval before execution; lower-risk actions are reviewed after the fact on a defined sample.	OWASP LLM06 mitigation: require human approval for high-impact actions. NIST AI RMF 1.0, MANAGE function, and the Generative AI Profile (NIST AI 600-1) on human oversight proportionate to risk. ISO/IEC 42001 A.9.2 and A.6.2.6. https://doi.org/10.6028/NIST.AI.600-1
Short-term working context is isolated from long-term memory; memory writes are validated and session-scoped.	OWASP Agentic AI Threats and Mitigations v1.0 (Feb 2025), T1 Memory Poisoning: session isolation, memory segmentation, content validation. https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/
Every action is logged to an append-only, integrity-protected record that the operator cannot alter, with enough detail to reconstruct the decision.	NIST SP 800-53 Rev. 5 AU-9 Protection of Audit Information and AU-10 Non-repudiation. ISO/IEC 27001:2022 Annex A 8.15 Logging. ISO/IEC 42001 A.6.2.8. OWASP Agentic T8 Repudiation and Untraceability. https://doi.org/10.6028/NIST.SP.800-53r5
Deterministic output checks and a kill switch sit between the model and any production effect; failures are blocked or escalated, never retried silently.	OWASP LLM05 Improper Output Handling; NIST AI 600-1 on output validation and incident response; ISO/IEC 42001 A.8.4 and 10.2. https://genai.owasp.org/llm-top-10/

DESIGN RULE	PUBLISHED BASIS
Model, prompt and policy changes are controlled releases with evaluation results, and drift is monitored on a defined cadence.	NIST AI RMF MEASURE and MANAGE; ISO/IEC 42001 6.3, 8.1, 9.1 and A.6.2.7. MAS Artificial Intelligence Model Risk Management information paper (Dec 2024) on model change and monitoring. https://www.mas.gov.sg/publications/monographs-or-information-paper/2024/artificial-intelligence-model-risk-management
Human oversight is protected from overload: escalation volume is a monitored metric and the tiering is adjusted when approvers become the bottleneck.	OWASP Agentic T10 Overwhelming Human in the Loop. NIST AI 600-1 on human-AI configuration. https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/
Singapore deployments map the same controls to the local instruments.	IMDA Model AI Governance Framework for Generative AI (May 2024); CSA Guidelines and Companion Guide on Securing AI Systems (Oct 2024); MAS FEAT Principles (2018). https://www.csa.gov.sg/resources/publications/guidelines-and-companion-guide-on-securing-ai-systems

PLACEMENT

When to use this architecture.

It belongs wherever an agent can act on a regulated record, move money, change a customer's position or commit the organisation. It is too heavy for a read-only assistant that summarises documents for a single user, and the semantic router exists precisely so that traffic which does not need governance does not pay for it. The pipeline adds latency at B, D1 and D2; the placement rule from Layer 8 applies: decision support and consequential action, not real-time transaction paths.

STATUS AND LICENCE

Version 1.0, published 12 September 2026. ISO/IEC 42001 references checked against the Annex A control table of SS ISO/IEC 42001:2024; MAS FEAT cited by principle number. Licensed CC BY-NC 4.0. Related: Governed Cognitive Layer (Layer 8), TRACE Framework, Four-Question AI Governance Baseline, Agentic pattern catalogue.

This is a personal site. The views, frameworks and publications here are my own analysis. They do not speak for Orion Five Engineering or any past employer or client, and they do not draw on the confidential information, data or proprietary methods of any of them.

CITE THIS ARCHITECTURE

Kok, T. (2026). Governed Agentic RAG: Reference Architecture (version 1.0) [Reference architecture]. terencekok.com. https://terencekok.com/research/reference-architectures/governed-agentic-rag/

Published 12 September 2026. Licensed CC BY-NC 4.0; cite with attribution and the version number.