

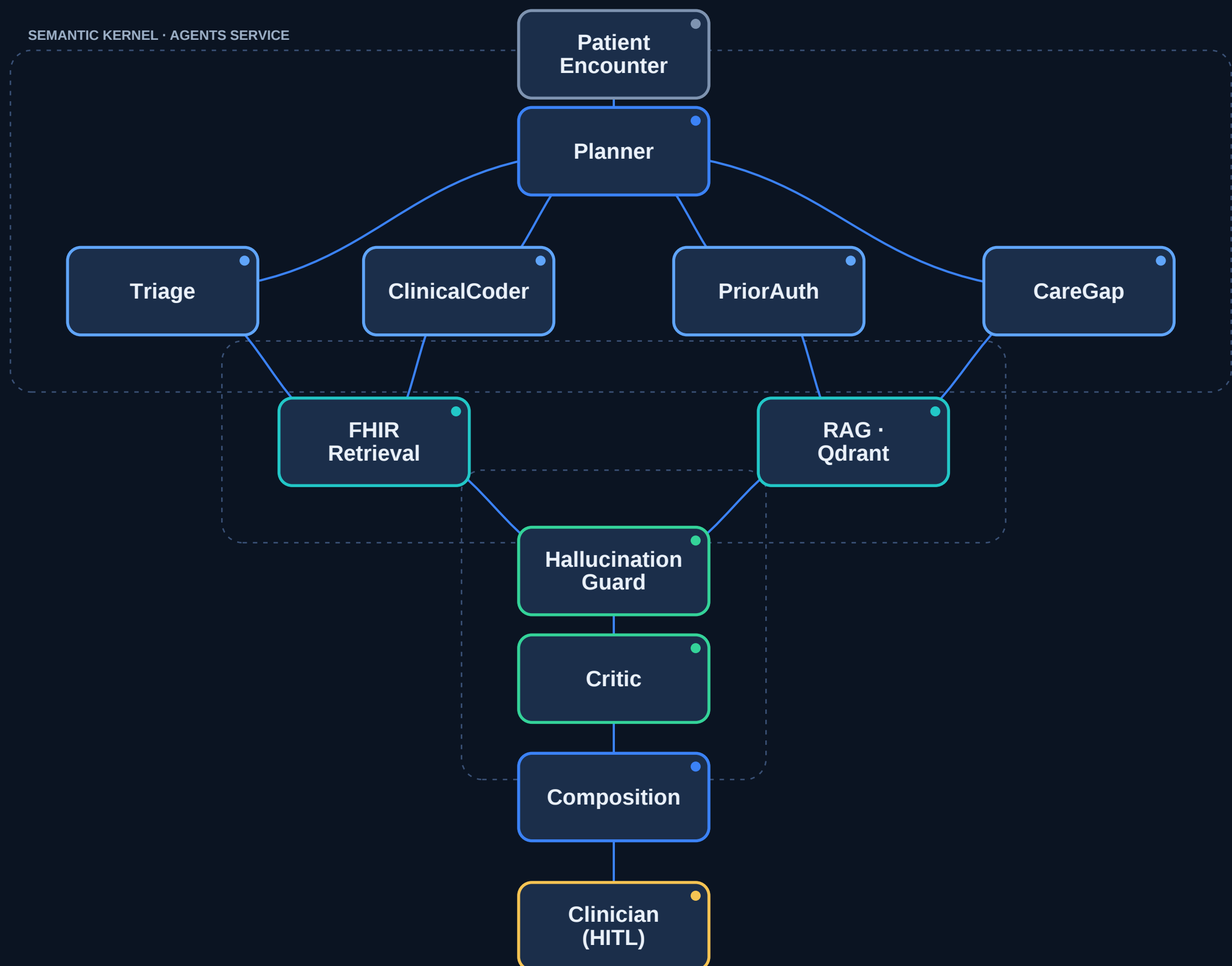
— ARCHITECTURE · CASE STUDY

HealthQCopilot

Agentic AI, grounded & auditable

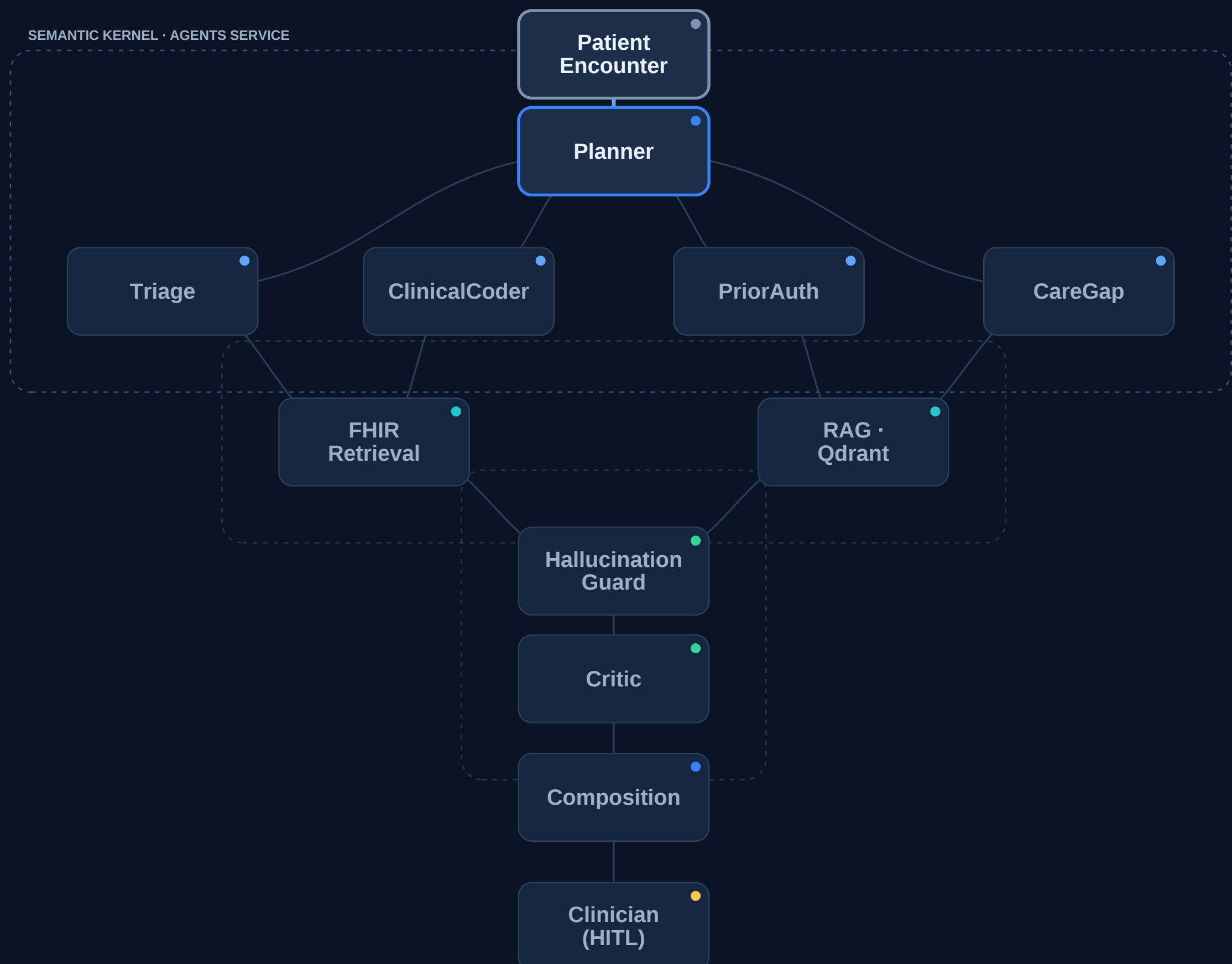
How a clinical AI stays trustworthy on Azure - narrow agents, mandatory grounding, a verification gate, and an audited, clinician-led pipeline.

Swipe: the pipeline, then component specs →



Grounded. Verified. Auditable.

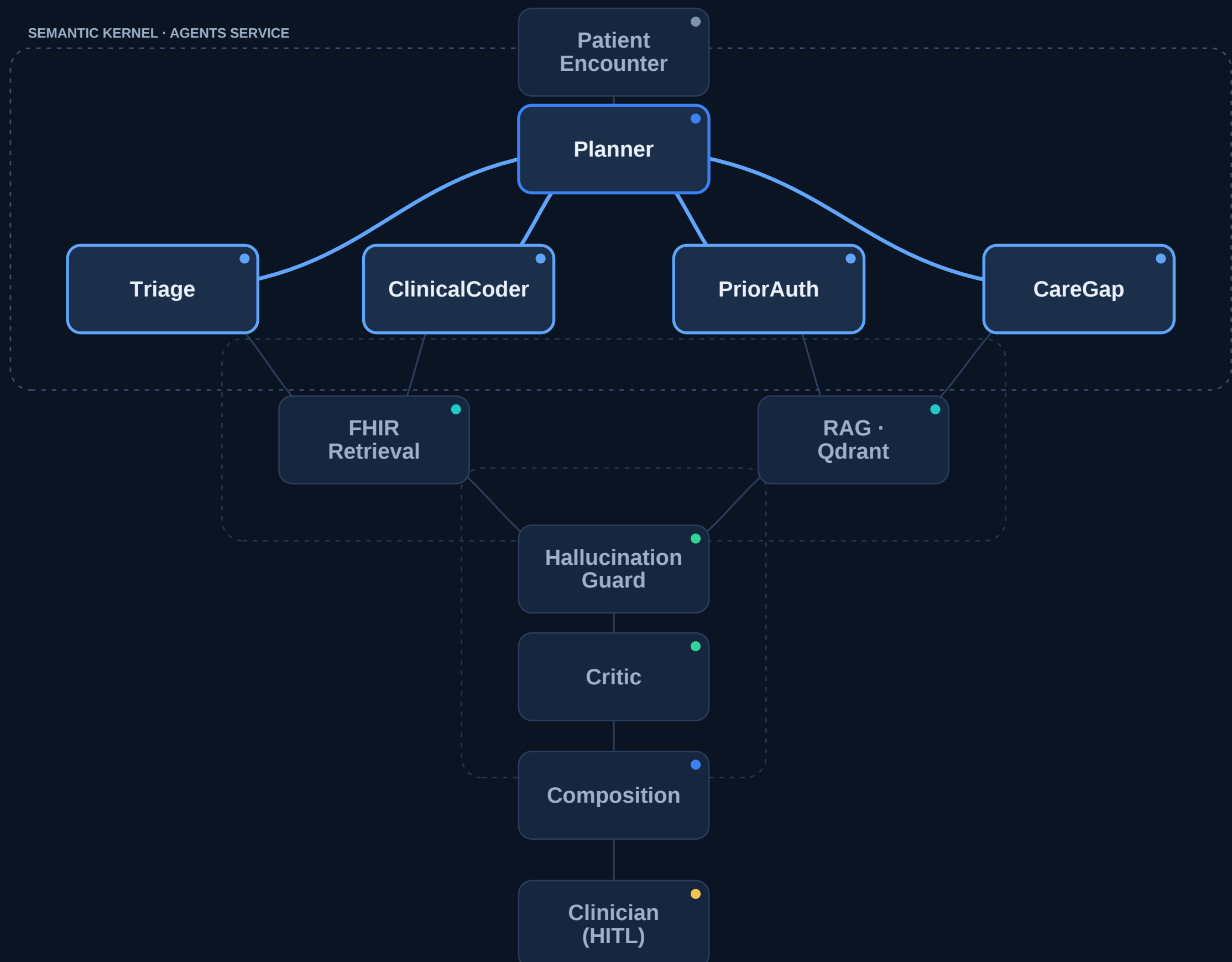
Every encounter flows through specialized agents, FHIR-grounded retrieval, a verification gate, and a clinician. Governance throughout: consent, PHI redaction, tool RBAC, audit to Event Hubs.



Before any model runs

A patient encounter is captured; consent is checked and PHI is redacted before a single prompt leaves the boundary.

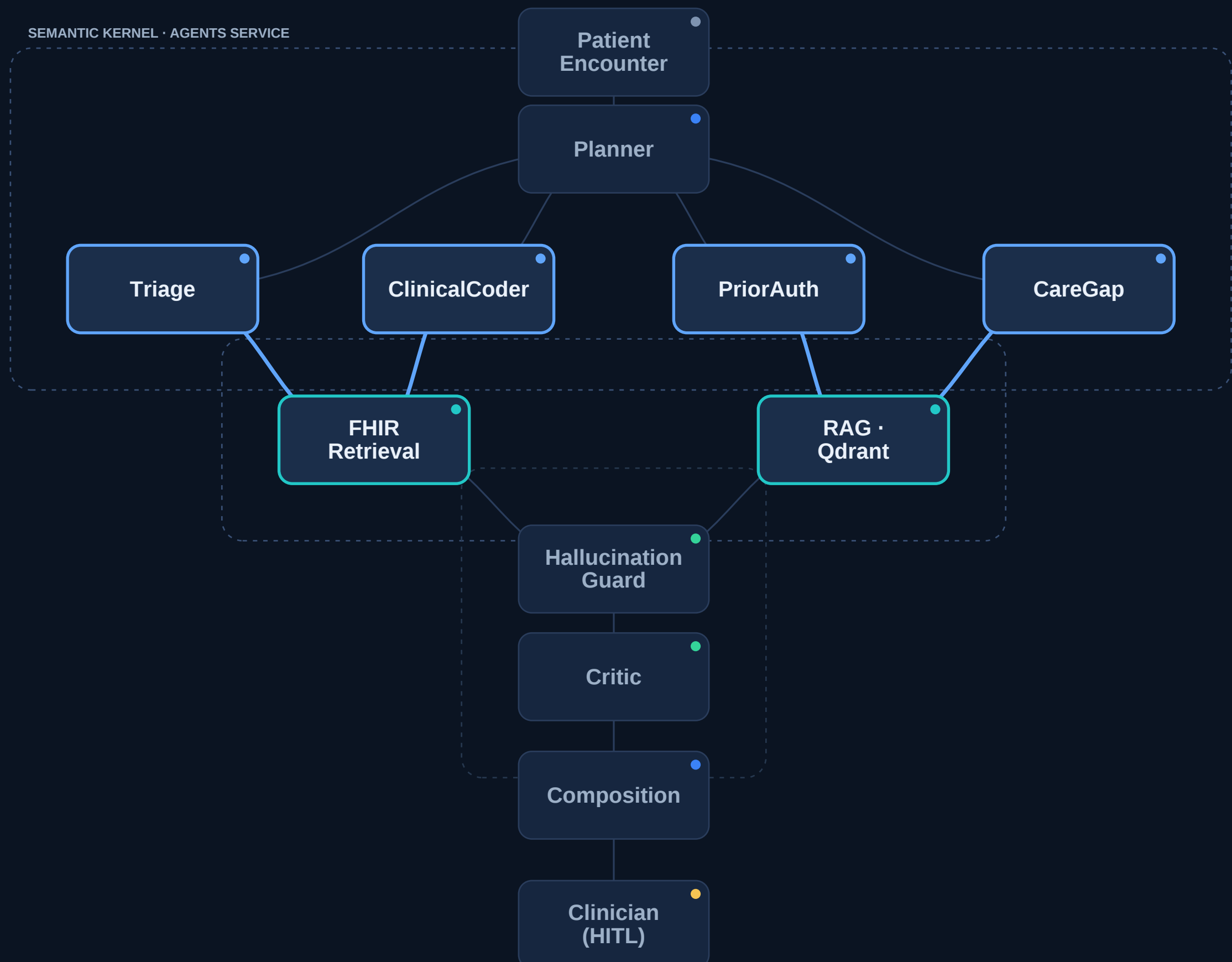
Example: Maria, 67, with chest tightness; consent on file, name + MRN redacted before any prompt.



A planner, not one big model

The planner decomposes the request and routes each sub-task to a narrow specialized agent - triage, clinical docs, prior-auth, care gaps.

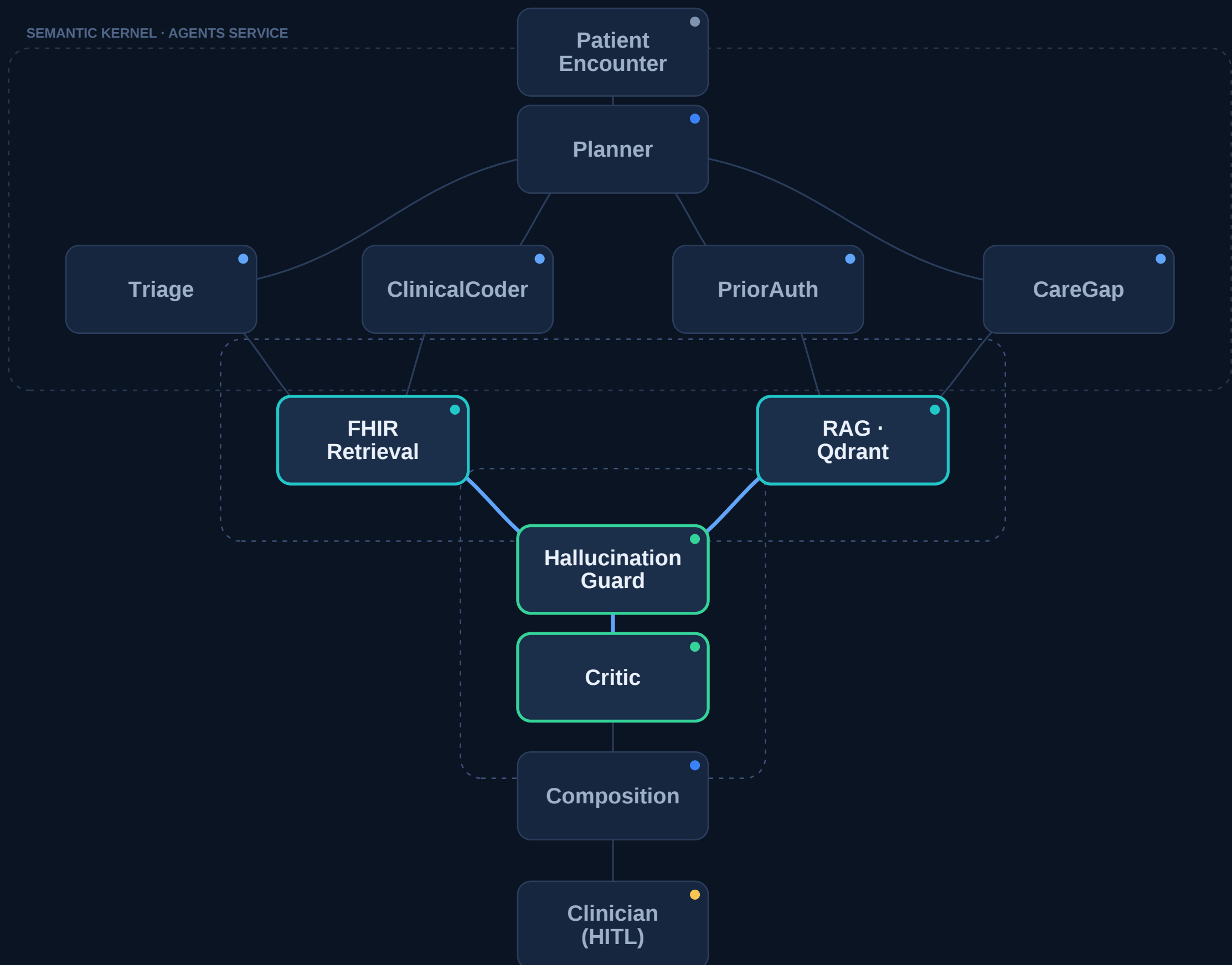
Example: a possible cardiac case fans out to triage, coder, prior-auth and care-gap agents.



Anchored in real data

Agents ground in the patient's FHIR record and clinical protocols via RAG. Every claim must cite a resource it retrieved.

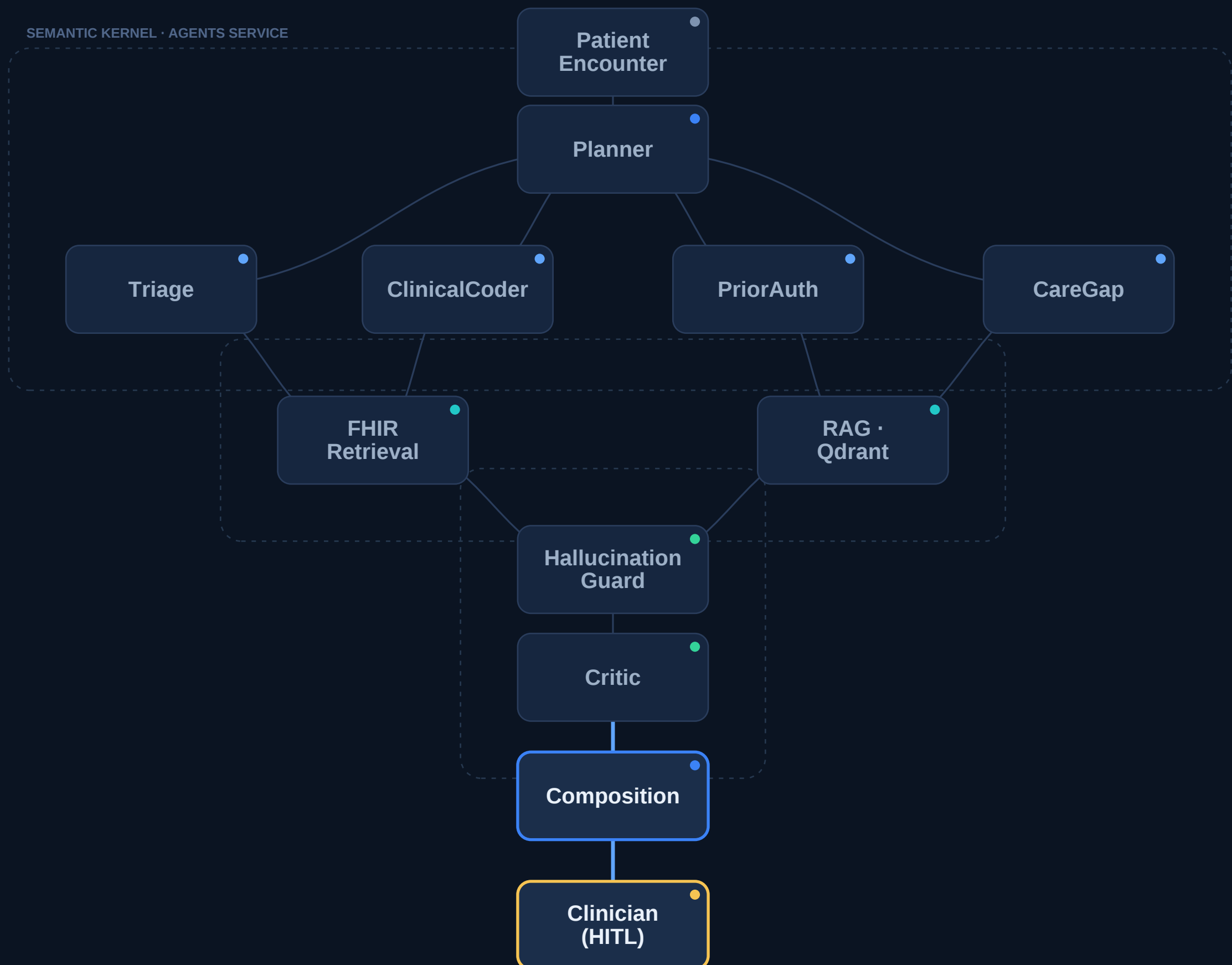
Example: agents pull Maria's vitals, meds, HbA1c and ECG (FHIR) plus the chest-pain protocol (RAG).



Cite-or-refuse

A hallucination guard checks each claim against the record; a critic re-reads the reasoning. Unsupported output is dropped, not shown.

Example: BP 162/94 cites a vitals observation and passes; history of MI has no source and is dropped.



Reviewed, then delivered

Grounded, verified pieces are composed and logged to the audit trail - and a clinician reviews before anything is acted on.

Example: Dr. Lee approves; the decision and chain-of-custody are written to Event Hubs.

Intake & orchestration

— Patient Encounter

Voice + structured intake at the encounter

TECH APIM ingress; speech-to-text; OAuth2/OIDC clinician identity; session-keyed audit

CTRL Consent gate before any model call; PHI redaction at the boundary; scoped to the clinician

— Planner

Decomposes + routes; runs the planning loop

TECH Semantic Kernel 1.71 on .NET 10; AgentPlanningLoop + episodic memory (Qdrant); AKS / Container Apps

CTRL Loop budget (8 iters / 16k tokens / 30s); Tool RBAC (ToolPolicyFilter); feature flags (App Config)

Specialized agents

— Triage

TECH TriageOrchestrator; Phi-4-Medical; versioned prompt (IAgentPromptRegistry)

CTRL Deterministic triage levels; consent-denied audited path; decision logged (model + prompt)

— ClinicalCoder

TECH ClinicalCoderAgent; ICD-10 / SOAP; schema-validated output

CTRL Cite-or-refuse; guard-checked before display; versioned prompt + model

— PriorAuth

TECH PriorAuth plugin; payer-policy RAG; PatientPlugin -> FHIR service

CTRL Clause + record provenance; refuse if unverifiable; auditable decision mapping

— CareGap

TECH CareGap plugin; ClinicalPlugin (population health); guideline RAG

CTRL Cited recommendations; guard-checked; logged with provenance

Grounding & verification

FHIR Retrieval

In-scope FHIR R4 evidence

TECH HAPI FHIR R4; PatientPlugin; scoped queries; APIM cache

CTRL Clinician-scoped reads; field-level PHI minimization; read-only

RAG · Qdrant

Clinical protocols as context

TECH Qdrant; IRagContextProvider; hybrid retrieval

CTRL Versioned sources w/ provenance; flag-gated (HealthQ:RagRetrieval), staged

Hallucination Guard

Cite-or-refuse integrity gate

TECH HallucinationGuardAgent; HallucinationJudge prompt; claim-by-claim vs FHIR

CTRL Drop unsupported; guardApproved on AuditEvent.AgentDecision; generator != approver

Critic

Re-reads reasoning vs citations

TECH CriticAgent (HealthQ:CriticReview); RagCitation; groundedness metric

CTRL UNSUPPORTED = rejection; canary-flagged; metric-monitored

Composition & clinician

— Composition

Assembles grounded, verified output

TECH Composition agent (Semantic Kernel); provenance to the UI; web-pubsub stream

CTRL Only guard-approved content; provenance per claim; recorded for audit

— Clinician (HITL)

Human sign-off before action

TECH triage-mfe approval UX; web-pubsub realtime; Azure Event Hubs audit

CTRL Human approval required; AuditEvent trio per session; chain-of-custody (who/what/why + model + prompt)

— YOUR TURN

Building healthcare AI you can trust?

This is the pattern behind HealthQCopilot - narrow agents, mandatory grounding, a verification gate, and HIPAA-ready governance, on Azure.

DM me “assessment”

Atiqul Islam

Fractional Cloud & AI Architect for SMBs · atiqulislam.dev